



UNIVERSITY OF PIRAEUS
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
DEPARTMENT OF INFORMATICS

PHD THESIS

Thesis Title:	(English) Analysis and Detection of Deviant and Malicious Behaviors in Social Media and Beyond (Greek) Ανάλυση και Εντοπισμός Παραβατικών και Κακόβουλων Συμπεριφορών στα Κοινωνικά Δίκτυα και Πέραν Αυτών
Student's name-surname:	NIKOLAOS LYKOUSAS
Father's name:	VASILEIOS
Student's ID No:	ΠΛΔ/1802
Supervisor:	Constantinos Patsakis, Associate Professor

May 2022

PhD Thesis
was prepared during the Programme of Doctoral Studies
of the Department of Informatics of the School of Information and
Communication Technologies of the University of Piraeus
for the degree of Doctor of Philosophy

3-Member Supervising Committee

Constantinos Patsakis
Associate Professor
University of Piraeus
School of Information and
Communication Technologies
Department of Informatics
(Supervisor)

Emiliano De Cristofaro
Professor
University College London
Department of Computer
Science
(Member)

Vicenç Gómez i Cerdà
Associate Professor
Universitat Pompeu Fabra
Department of Information
and Communications
Technologies (Member)

PhD Thesis
was presented before the 7-Member Examination Committee and approved on
May 3, 2022

7-Member Examination Committee

Constantinos Patsakis
Associate Professor
University of Piraeus
School of Information and
Communication Technologies
Department of Informatics

Emiliano De Cristofaro
Professor
University College London
Department of Computer Science

Vicenç Gómez i Cerdà
Associate Professor
Universitat Pompeu Fabra
Department of Information and
Communications Technologies

Alastair Beresford
Professor
University of Cambridge
Department of Computer
Science

Vasilios Katos
Professor
Bournemouth University
School of Computing

Julio Hernandez-Castro
Professor
University of Kent
School of Computing

Efthimios Alepis
Associate Professor
University of Piraeus
School of Information and
Communication Technologies
Department of Informatics

Ευχαριστίες

Μέσα από τις επόμενες γραμμές θα ήθελα να εκφράσω τις ευχαριστίες μου σε όλους εκείνους που συνέβαλαν ουσιαστικά στην πραγματοποίηση και ολοκλήρωση της παρούσας διδακτορικής διατριβής.

Θα ήθελα, πρώτα απ' όλα, να εκφράσω την ευγνωμοσύνη και τις ευχαριστίες μου στον επιβλέποντα και μέντορά μου, Καθηγητή Κωσταντίνο Πατσάκη, για την αμέριστη ηθική υποστήριξη, το συνεχές, αδιάπτωτο ενδιαφέρον και την ακαταπρόκλητη καθοδήγησή του, που με συνοδεύουν και μου δείχνουν τον δρόμο από τα φοιτητικά μου χρόνια.

Επιπλέον, θα ήθελα να ευχαριστήσω τον Καθηγητή Vicenç Gómez, που στο διάστημα της φοίτησής μου στο Universitat Pompeu Fabra συνέβαλε καταλυτικά στη διαμόρφωσή μου ως ερευνητή και μου έμαθε να αποζητώ την αλήθεια χωρίς συμβιβασμούς.

Επιθυμώ, επίσης, να εκφράσω τις ευχαριστίες μου στον συνάδελφο και φίλο μου Δρ. Fran Casino, για όλες τις ενδιαφέρουσες συζητήσεις που είχαμε, την πολύτιμη βοήθεια και τις κατευθύνσεις που μου προσέφερε για την ολοκλήρωση της διατριβής μου.

Θα ήταν ακόμα ασυγχώρητη παράλειψη εκ μέρους μου, αν δεν ανέφερα με ευγνωμοσύνη την Καθηγήτρια μου Δέσποινα Πολέμη που διέβλεψε σε εμένα ικανότητες και μου ενέπνευσε την αγάπη για την έρευνα.

Τέλος, ευχαριστώ, μέσα από την καρδιά μου, όλους εκείνους που πιστεύουν σε εμένα, τους δικούς μου ανθρώπους, την οικογένειά μου και τους φίλους μου, που με ενθαρρύνουν και με στηρίζουν έμπρακτα.

Πειραιάς, Μαΐος 2022

Επιτελική Σύνοψη

Αναμφίβολα η άναρχη φύση του Διαδικτύου ευνοεί την άνθιση κάθε μορφής παραβατικών και εγκληματικών συμπεριφορών. Αυτές δυνητικά έχουν καταστροφικό αντίκτυπο στους χρήστες, στις υποδομές, και, κατ' επέκταση, στον κοινωνικό ιστό του ολοένα και περισσότερο διασυνδεδεμένου κόσμου του σήμερα. Σκοπός αυτής της διατριβής είναι να αναδείξει και να κατανοήσει διάφορες πτυχές κακόβουλων συμπεριφορών τόσο στον κόσμο των μέσων κοινωνικής δικτύωσης όσο και στον κόσμο του κυβερνοεγκλήματος. Συγκεκριμένα, η παρούσα μελέτη απαρτίζεται από δύο μέρη:

Το πρώτο μέρος εστιάζεται στην μελέτη αποκλινοσών συμπεριφορών συνυφασμένων με την παραγωγή και κατανάλωση πορνογραφικού υλικού σε κοινωνικά δίκτυα ζωντανής μετάδοσης εικονοροών (Social Live Streaming Services - SLSS). Τα εν λόγω δίκτυα επιτρέπουν στους χρήστες να μοιράζονται την καθημερινότητά τους μέσω της κάμερας των κινητών συσκευών τους. Το επίπεδο διεπαφής που προσφέρουν τέτοιες πλατφόρμες, σε συνδυασμό με την αδυναμία επαλήθευσης της ηλικίας των χρηστών τους καθώς και την έλλειψη αποτελεσματικών μέσων επιβολής των κανόνων ασφαλούς χρήσης τους, δίνει σε κακόβουλους χρήστες τη δυνατότητα να προσεγγίζουν εύαλωτα άτομα (π.χ. ανηλίκους) με απώτερο σκοπό την αποπλάνηση και τη σεξουαλική εκμετάλλευσή τους. Στα πλαίσια αυτής της διατριβής έγινε εκτενής μελέτη συναφών συμπεριφορών σε δύο κοινωνικά δίκτυα αυτής της κατηγορίας, από τα οποία συλλέχθηκε μεγάλος όγκος από αλληλεπιδράσεις μεταξύ χρηστών. Τα δεδομένα που προέκυψαν, κατόπιν αναλύθηκαν ώστε να μοντελοποιηθούν τέτοιας φύσης παραβατικές συμπεριφορές και να “αποκρυπτογραφηθούν” τα μοτίβα επικοινωνίας κακόβουλων χρηστών που έχουν διαμορφωθεί με τρόπο τέτοιο ώστε να παρακάμπτουν τις δικλίδες ασφαλείας που εφαρμόζονται από αυτές τις πλατφόρμες. Επίσης, η μελέτη επεκτείνεται και στην ανάλυση της παράνομης εμπορευματοποίησης και διάχυσης πορνογραφικού υλικού σε δημοφιλή μέσα κοινωνικής δικτύωσης.

Το δεύτερο σκέλος της διατριβής επικεντρώνεται στον κόσμο της σύγχρονης ηλεκτρονικής εγκληματικότητας. Συγκεκριμένα, τα προϊόντα και οι επί πληρωμή υπηρεσίες που προσφέρουν κυβερνοεγκληματίες στον ιστό επιφανείας (Surface Web), εκμεταλλευόμενοι την αποκεντροποιημένη και ανώνυμη φύση που χαρακτηρίζει εμπορικές πλατφόρμες όπως το Shoppay. Τέτοια παράνομα “προϊόντα” περιλαμβάνουν κλεμμένα διαπιστευτήρια για ψηφιακές υπηρεσίες, κακόβουλο λογισμικό κ.α. Εστιάζοντας στο κακόβουλο λογισμικό, η παρούσα διατριβή επιχειρεί να συμβάλει στην αντιμετώπιση της εξάπλωσής του, μέσω της συλλογής και ανάλυσης ενός εκτενούς συνόλου διευθυνσεων που έχουν δημιουργηθεί από αλγορίθμους δημιουργίας διευθύνσεων (Domain Generation Algorithms) και χρησιμοποιούνται για την ενορχήστρωση επιθέσεων και τον απομακρυσμένο έλεγχο προσβεβλημένων συστημάτων. Επίσης, παρουσιάζει και αξιολογεί ένα νέο σύνολο χαρακτηριστικών για την αποτελεσματική ανίχνευση τους με μεθόδους μηχανικής μάθησης. Τέλος, η διατριβή σκιαγραφεί τους κινδύνους που ελλοχεύουν στον αναδυόμενο κόσμο των συστημάτων διευθυνσιοδότησης που στηρίζονται στην τεχνολογία blockchain (Blockchain DNS), αναλύοντας τα οικοσυστήματα Namecoin και Emercoin που ως επί το πλείστον γίνονται αντικείμενο κατάχρησης από κυβερνοεγκληματίες για κακόβουλους σκοπούς.

Abstract

Deviant and malicious behaviors are inevitably interwoven into the very fabric of the Internet, due to its ubiquitous nature and its inherent imperviousness to authoritative regulation and control. Such behaviors can potentially have a devastating impact on today's hyper-connected world. The purpose of this thesis is to study several relatively unexplored facets of deviant and malicious online behaviors in the areas of social media and cybercrime. Concretely, this work is organized in two parts:

The first one focuses on the adult-content related deviant behaviors flourishing in the domain of Social Live Streaming Services (SLSS). This kind of social platforms allow a new level of social interaction, by enabling their users to share their daily lives through the cameras of their mobile devices. However, as they lack the mechanisms to effectively enforce their community guidelines, they are rife with adult content. This work examines in depth the mechanics of the adult production and consumption phenomenon in two large SLSS platforms in terms of interactions between their users and characterizing attributes of their behavior. Additionally, the largest-to-date dataset of chats and user interactions in the context of adult content live streams is constructed and analyzed to unveil evidence of sexual exploitation and grooming targeting underage users, as well as to disentangle the strategies adopted by malicious users to evade the moderation mechanisms of such platforms. Furthermore, this thesis sheds light on the semi-illicit adult content market layered on the top of popular social media platforms, its offerings, and the demographics of adult content producers.

The second part concentrates on the world of cybercrime. Specifically, this dissertation studies the modus operandi of cybercrime vendors who use anonymous marketplace platforms on the Surface Web to sell illicit products and services such as leaked credentials, breached accounts and malware, while hiding in plain sight. Particularly for the case of malware, this work delves into the problem of detecting algorithmically generated domains, an approach employed by modern malware and botnets to enhance and scale their persistence and orchestration capabilities over millions of infected devices. To this end, this work presents the largest-to-date dataset of such domains, and proposes a novel set of features useful for the resilient and robust detection of a wide set of domain generation algorithms through machine learning approaches. Finally, this thesis explores the novel threats of the emerging field of blockchain-based DNS alternatives, focusing on the Namecoin and Emercoin ecosystems which are found to be abused for malicious purposes.

Contents

I	Introduction	1
1	Context and Motivation	3
1.1	Deviant Online Behaviors in Social Media	4
1.1.1	Deviant behaviors related to adult content production and consumption in SLSS	5
1.1.2	Grooming in Social Live Streaming Services	7
1.1.3	Inside the realm of “premium” social media accounts	8
1.2	Malicious Behaviors Beyond Social Media	8
1.2.1	Cybercrime Markets on the Surface Web	9
1.2.2	Domain Generation Algorithms	10
1.2.3	Emerging threats in Blockchain-based DNS	11
2	Publications and Contributions	13
2.1	Publications of the compendium and Contributions	13
2.2	Other publications and Contributions	16
II	Deviant Online Behaviors in Social Media	19
3	Detecting and characterizing users involved with adult content production and consumption in SLSS	21
3.1	LiveMe & Loops Live functional overview	21
3.2	Data collection methodology	23
3.2.1	Sampling the social graphs	23
3.2.2	labeling the users	26
3.2.3	Effectiveness of SLSS moderation systems	28

3.3	Profiling deviant users	29
3.3.1	Features	29
3.3.2	Deviant relationships	31
3.4	Modeling deviant behavior	34
3.4.1	Notation	34
3.4.2	Proposed features	34
3.4.3	Evaluation	36
4	Analysis of Grooming Behaviors in Social Live Streaming Services	39
4.1	The dataset	39
4.2	Identifying the characterizing attributes of grooming behavior	41
4.3	Topical analysis	44
4.3.1	Preprocessing	46
4.3.2	LDA models	48
4.3.3	Interpretation and analysis of topics	51
4.3.4	Topic relatedness	53
5	Inside the realm of “premium” social media accounts	59
5.1	The FanCentro Platform	59
5.2	Data Collection	60
5.3	Characterizing Performers	61
5.4	Exploring the supply and demand	64
5.4.1	FanCentro content	65
5.5	Monetization means of premium social media accounts	67
III	Malicious Behaviors Beyond Social Media	69
6	Analysis of a cybercriminal marketplace on the Surface Web	71

6.1	The case of Shippy	72
6.2	Data Collection	72
6.3	Shippy Dataset Exploration	76
6.3.1	Shippy in numbers	76
6.3.2	Characterizing products and services	79
6.4	Products and services related to cybercrime	83
7	Detecting Algorithmically Generated Domains	87
7.1	The HYDRAS Dataset	88
7.2	Proposed Features for AGD detection	90
7.2.1	Feature Extraction Approach	90
7.2.1.1	Gibberish Detection	92
7.3	Classification of AGDs	92
7.3.1	Binary Classification using the HYDRAS Dataset	93
7.3.2	Classification of Adversarially Designed AGDs	95
7.3.3	Detection of Unknown Families	96
8	Emerging threats in Blockchain-based DNS	99
8.1	Blockchain-based DNS	99
8.2	Threats in the context of Blockchain-based DNSs	101
8.2.1	Malware	102
8.2.2	Underlying registrar mechanism	103
8.2.3	Domain registration market	104
8.2.4	Phishing	105
8.2.5	Lack of motivation	106
8.2.6	Immutability	106
8.3	Analysis of real-world data	107
8.3.1	Data collection and labeling	108

8.3.2	Representation of malicious activities in Emercoin and Namecoin ecosystems	109
IV	Closure	115
9	Conclusions and Future Work	117
9.1	Conclusions	117
9.2	Future work	120
	Bibliography	125
V	Publications	139

List of Tables

3.1	Network statistics of the crawled graphs: number of nodes $ N $, number of edges $ E $, number of banned nodes $ B $, average degree $\langle k \rangle$, density D , and reciprocity ρ	25
3.2	Distribution of users according to their class.	28
3.3	Proportion (total in parenthesis) of banned accounts in each class.	28
3.4	Top 5 features for differentiating the three classes.	30
3.5	Comparison in terms of link density D between the crawled networks of producers, consumers normal users and a corresponding random network.	33
3.6	Classification Performance. P: Producers, C: Consumers, N: Normal, MF1: macro-F1.	37
4.1	Top 15 verbs (simple and phrasal) associated with clothing items.	50
4.2	Top 10 nearest neighbors (cosine distance) of the word “pussy”.	50
4.3	Top 10 nearest neighbors (cosine distance) of the word “boobs”.	51
4.4	Top 10 most frequent clothing terms.	51
4.5	Top 5 interaction features relevant for characterizing grooming broadcasts	54
4.6	Topics	56
4.7	Illustrative chat messages from broadcasts where Topic #7 is dominant. Key terms of Topic #7 are in bold.	57
5.1	Sexual identity and orientation	62
5.2	Premium services	67
6.1	Collected usernames and Shoppay links.	76

6.2	Shopyy products per category	76
6.3	Some illustrative false “services”, priced $\geq 500\$$	78
6.4	Different topic classes and their corresponding key terms, sorted according to the number of documents found.	80
6.5	Sample products for each topic.	82
6.6	Indicative products modeled by Topic #5 in descending price order.	84
6.7	Known breaches sold through Shopyy, as identified by Topic #13’s most salient terms.	85
7.1	Distribution of records per DGA in our dataset. DGAs in green denote those which were frequently underrepresented, so they were run to create more samples, while purple indicates adversarial ones.	89
7.2	Features used in the proposed approach and their corresponding description.	91
7.3	Performance measures for binary classification (in percentage).	95
7.4	A detection of unknown DGA families represented by adversarially designed DGAs (leave-one-out experiment).	96
7.5	Binary classification against adversarially designed DGAs. First row of each family denotes the reported results in the original work.	97
8.1	Technical characteristics of the most relevant DNS systems. Although Blockstack is blockchain agnostic, it is mainly used with Bitcoin blockchain.	101
8.2	Main characteristics of blockchain DNSs.	101

List of Figures

1.1	Some illustrative Google Play reviews of LiveMe, highlighting the problem of deviant behaviors in SLSS.	6
3.1	Data collection methodology. A proxy intercepts the messages from the smartphone to the SLSS. To decrypt the traffic and derive the API, a root certificate is installed in the smartphone (steps 1-6). Then, an adult keyword list is used to get an initial set of seed users, which are then queried to collect even more users. Based on their properties (e.g. banned, gifts) the dataset is constructed accordingly (steps 7-9).	24
3.2	Cumulative distribution function of the adult content score values.	27
3.3	Cumulative distribution functions (CDFs) of the different user profile attributes.	30
3.4	User relationship insights, provided by HITS.	32
3.5	Cumulative distribution function (CDF) for the novel features for each dataset.	36
4.1	Cumulative distribution functions (CDFs) of broadcast metadata features and interactions.	41
4.2	Broadcasters count per country	42
4.3	Cumulative distribution functions (CDFs) of the chat messages per broadcast and per user.	42
4.4	Top emoji collocations for clothing related emojis.	45
4.5	Most frequent semantically-similar words to LIWC sexual terms, as learned by FastText.	46
4.6	Top collocates of sexual words in LiveMe dataset.	47
4.7	Collocates of the word “show”.	48

4.8	Verbs extracted from chats containing clothing terms.	49
4.9	C_v metric according to no of topics.	52
4.10	Cumulative distribution functions (CDFs) of the features of Table 4.5.	54
5.1	An example performer’s profile page in FanCentro.	61
5.2	Weekly registrations	62
5.3	Age Distribution	63
5.4	Descriptive characteristics of performers profiles: tags and links to external sites.	63
5.5	Cumulative distribution functions (CDFs) of (non-zero) revenue, number of followers and posts.	65
5.6	Monthly posting activity	65
5.7	Characteristics of posts in terms of text content and reactions (likes, comments).	66
5.8	Number of offerings per service and subscription duration.	68
5.9	Bar plots of monthly subscription price (normal and discounted) per service and subscription duration.	68
6.1	Screenshots from hacking forums indicative of the use of Shoppo for selling illicit products.	73
6.2	A store of a cybercriminal vendor hosted by Shoppo platform.	74
6.3	Workflow of the Shoppo data collection and analysis pipeline implemented.	75
6.4	CDFs of products and prices offered by Shoppo cybercrime vendors.	77
6.5	Fractional price bins of Shoppo products.	78
6.6	C_v metric according to the number of topics.	79
7.1	The <i>modus operandi</i> of a typical DGA-powered botnet [115].	88

7.2	Exemplified overview of the feature extraction process.	93
8.1	Workflow of the browser extensions procedure to enable resolution of EmerDNS, Namecoin, New Nations and OpenNIC domains. The extension analyses the TLD extension of the requested domain and directs the query to the corresponding DNS system.	102
8.2	An overview of main threats of blockchain DNSs.	103
8.3	Outline of the methodology for analyzing blockchain DNS data.	108
8.4	Graph-based representation of the Emercoin ecosystem.	112
8.5	Graph-based representation of the Namecoin ecosystem.	113

Part I

Introduction

Chapter 1

Context and Motivation

The largely unsupervised wilderness of the Internet has allowed a broad spectrum of deviant behaviors to flourish, posing significant risks for its users. Those risks, once realized, can lead to severe consequences in the real world, spanning from the disruption of economic and business processes to the corruption of human conscience, societal norms, and moral standards. Deviant behavior is unarguably a characteristic that can be observed in every human society. The extent of this behavior, in terms of how many people exhibit it, and the harm that is caused by it define the ethics of the society and its limits.

However, the ethics on the Internet (or the lack, thereof) follow different rules as it is rather different from the real world, even if it can be considered its extension. The inherent intractability of regulating the Internet due to its distributed nature, in conjunction with the unprecedented penetration of information and communication technologies in almost all aspects of modern life has enabled malicious actors to develop and deploy an unimaginable arsenal of illicit tactics to exploit digital platforms and harm users. This thesis aims to shed light on some novel and relatively unexplored deviant/malicious behaviors, through analyzing and untangling their mechanics, as well as proposing strategies for their detection where possible. It particularly focuses on the realm of social media (Part II), and the world of cybercrime (Part III). Based on this structure, this leading chapter provides a high-level overview of the topics explored and outlines the underlying motivation behind this work, and it is organized in two parts: the first is devoted to Deviant Online Behaviors in Social Media (Section 1.1), and the second one to Malicious Behaviors Beyond Social Media (Section 1.2).

It should be noted that a significant part of the present work could be characterized as *data-driven* research involving the collection of data produced by human subjects in the context of several online platforms. Provided that various aspects of the present dissertation involve the analysis of behaviors potentially associated with criminal activities, the researchers who undertook the original research took all the necessary measures and precautions to minimize their exposure to any illegal content. This involved removing the human-in-the-loop where applicable by leveraging automated means of visual data (e.g. broadcast) classification without exposing the visual content to the user, as described in Section 3.2.2. Further details regarding the ethical and legal compliance of the data collection approach can be found in [1].

1.1 Deviant Online Behaviors in Social Media

The recent advances in telecommunications have unleashed the potential of sharing and exchanging content, changing the way we interact with others online radically. By lifting many bandwidth barriers, users may generate and share arbitrary content and seamlessly disseminate it to millions of users instantly. As a result, we observe the dominance of Social Networks and Media in various aspects of our daily lives. This radical shift and penetration of mobile devices have led millions of people and youngsters to use them on a daily basis. Naturally, the veil of anonymity and the difficulty to regulate the formation of topical communities within the social media ecosystems, have enabled users to freely exhibit deviant behaviors, evading moderation mechanisms. Specifically, this work focuses on studying and analyzing adult-content related and sexually exploitative behaviors in social media.

Most social media platforms have a clear policy against such illicit practices and facilitate some mechanisms to detect and ban misbehaving users, either in the form of filters from the service provider or by peer reporting. In practice, as proven in the context of this research and the related literature, such mechanisms are largely inefficient in battling deviant behaviors. Moreover, while most social networks have specific policies disallowing their use by minors, they tend to bypass them by declaring fake ages to register to service providers and end up using the services as normal users. While this might not be noticed or be overseen by service providers, this is not the case for users. Unfortunately, there are thousands of users who maliciously target and exploit minors. Of specific interest is the case of *grooming*.

There is a steady increase of reports regarding the exploitation of social networks for grooming [2]. The problem, regardless of the age factor, is stigmatizing thousands of people.

To this end, of particular interest are Social Live Streaming Services (SLSS), where users can live stream parts of their daily lives. Such platforms provide a new level of interaction and hook their subscribers as the users become part of the daily life of others. Practically, users open up their cameras and share snapshots of what they do, what they think or live at the moment with others and interact with them via chat messages. Nonetheless, the emergence of these novel social networks allowing live streaming to potentially thousands of users along with traditional chatting and appraisal methods of traditional social networks, provide fertile ground for adult content and grooming related deviant behaviors [3].

The findings of this thesis clearly indicate that in the context of SLSS, the deviant behaviors of adult content production/consumption and sexual grooming are interrelated to a significant extent. Finally, this thesis examines the penetration and normalization of adult content in mainstream social media platforms through monetization means coined as “premium” accounts, which is exacerbated by the prevalence of “influencer” culture.

Concretely, within this thesis, the following topics are explored:

- **Chapter 3:** Detecting and characterizing users involved with adult content production and consumption in SLSS.
- **Chapter 4:** Analysis of grooming in SLSS.
- **Chapter 5:** Study of the adult content market of “premium” accounts layered on the top of popular social media platforms.

1.1.1 Deviant behaviors related to adult content production and consumption in SLSS

In this thesis two popular Social Live Stream Services are considered, namely, LiveMe¹ and Loops Live². Both operate as video chat apps in mobile phones and have millions of users that produce massive amounts of video content on a daily basis. Both these apps share many similarities regarding community policies, e.g., they explicitly forbid broadcasters from engaging in, or broadcasting, any sex-related content that promotes sexual activity, exploitation, and/or assault. Moreover, both apps prohibit violence and/or self-harm, bullying, harassment, hate

¹<https://www.liveme.com>

²<https://www.loopslive.com>

speech, on-screen substance use, posting of private contact information, prank calls to emergency authorities or hotlines, and solicitation or encouragement of rule-breaking. There is a variation in the user's age, as for the case of LiveMe, users have to be at least 18 while in Loops Live the users have to be at least 13 years old.

To counter possible violations of the aforementioned policies, both services have implemented reporting mechanisms, so that users can easily report a channel once they identify an underage user or detect suspicious behavior or violations of the service policies. On top of that, LiveMe employs a team of human moderators around the world, working 24/7 to respond to users' reports. Violators are subject to immediate suspension or ban from the app. Those safeguards are in place to protect young people since live streaming apps and sites can expose them to graphic and distressing content and can leave them vulnerable to bullying and online harassment.

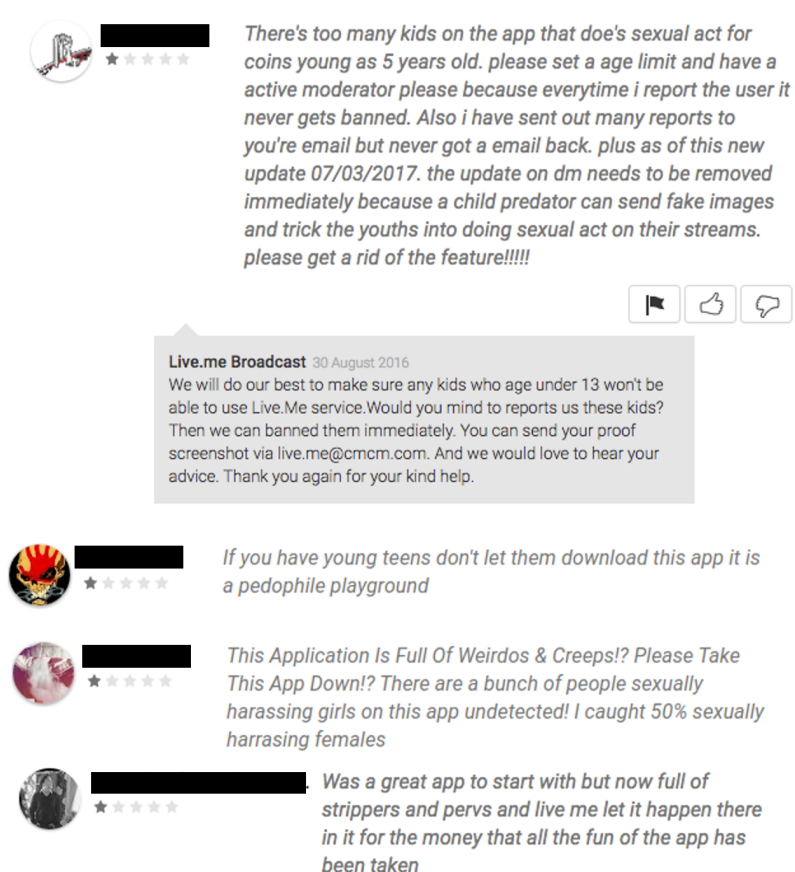


Figure 1.1: Some illustrative Google Play reviews of LiveMe, highlighting the problem of deviant behaviors in SLSS.

However, these mechanisms do not seem to be working as intended. Many users report in the app reviews that they are constantly witnessing violations of the aforementioned policies, as evident in Figure 1.1. This thesis investigates the interaction patterns between users producing or consuming adult content within the social graphs of LiveMe and Loops Live platforms, and proposes novel features that can be exploited for the effective detection of deviant users falling into the two aforementioned categories.

1.1.2 Grooming in Social Live Streaming Services

Grooming refers to the process by which an offender prepares a victim for sexually abusive behavior. More precisely, according to [4]:

[Grooming is]... a process by which a person prepares a child, significant others, and the environment for the abuse of this child. Specific goals include gaining access to the child, gaining the child's compliance, and maintaining the child's secrecy to avoid disclosure. This process serves to strengthen the offender's abusive pattern, as it may be used as a means of justifying or denying their actions...

Apparently, child grooming is of extreme importance due to the impact that it can have in children's lives. In fact, despite the measures that social networks might have already taken, they do not seem to be successful at all [5]. This fact also has been demonstrated by the findings of the present thesis. The moderation systems used by the LiveMe platform at that time were highly ineffective in suspending the accounts of deviant users producing adult content. Notably, the respective part of this work was published in the same year as, FOX 11; a major mainstream media outlet, reported that [3]:

A FOX 11 investigation has found that pedophiles are using the popular live streaming app LiveMe to manipulate underage girls into performing sexual acts, reward them with virtual currency, and then post screen captures or recordings of the girls online to be sold and distributed as child porn.

This work primarily aims to identify and disentangle the mechanics of grooming and predatory behaviors in the context of Social Live Streaming Services, by analyzing the behavioral and communication patterns of the viewers of live streams, as well as investigate the means employed by groomers to bypass the policies and filters set by social networks.

1.1.3 Inside the realm of “premium” social media accounts

In the world of social media, content creators play a central role in shaping global online culture. The content creators who raise in popularity can attain the status of online micro-celebrities, and they are commonly characterized as *influencers* [6]. The main objective of influencers is to produce digital content which attracts users’ attention and rapidly gains popularity, often becoming ‘viral’, in platforms such as Instagram and YouTube [7, 8]. In this regard, influencers leverage focused visual content and targeted communication techniques to capture and sustain the attention of social media users, thus building large follower bases and attaining organic social reach. Social media content creators can thus monetize their reach in various ways, such as using word-of-mouth marketing techniques and promoting brands and campaigns [9, 10].

One of the most prevalent strategies employed by influencers to entice followers towards heightened forms of emotional engagement is sexualized labor [11]. As previously discussed, all mainstream social media platforms explicitly prohibit accounts that promote or distribute pornographic content. Nonetheless, as demonstrated by this thesis, community guidelines cannot be effectively enforced to ban adult content in social media. As such, Performers who systematically violate community guidelines by posting overtly sexual content, have to use external means for managing transactions with their client base, as well as maintaining their digital presence in multiple social outlets, in case their accounts get suspended.

To this end, the present work analyzes *FanCentro*³, a platform where Performers can monetize their fan base via selling subscriptions to their private social media accounts. Accordingly, this thesis aims to shed light on the mechanics of the semi-illicit industry of premium social media subscriptions and services offered by Performers, in the context of adult content marketplaces.

1.2 Malicious Behaviors Beyond Social Media

Online malicious behaviors extend far beyond the realm of social media, and are inherently multi-sided and arbitrarily complex, as malicious actors leverage a wide variety of means to achieve their ends. In particular, the present thesis focuses on the study of the mechanics

³<https://fancentro.com>

and the detection of malicious behaviors practiced by cybercriminals within the context of Surface Web and the abuse of DNS and its extensions. With the continuous digitization of procedures, services, and products, crime has shifted towards the same direction. As a result, almost every aspect of a modern crime is facilitated by digital means, and consequently, almost every criminal investigation involves some sort of digital evidence. The above are the primary reasons why cybercrime has evolved into a multi-billion underground economy. Its economic impact is devastating [12], with FBI estimating the losses to be \$3.5 billion only within the USA [13]. In fact, this continuous rise is so threatening that it has become the second most-concerning risk for global commerce over the next decade, according to the World Economic Forum [14]. Concretely, if someone considered global cybercrime as a country, then its economy would have the 13th highest GDP in the world [15]. Consequently, in the past few years, there has been a significant increase in reported data leaks, online extortion schemes and credential trading, affecting a broad spectrum of online services and service providers, including retailers, payment processors, and government entities [16]. Moreover, the recent COVID-19 pandemic and the spike in usage of digital services has also resulted in an analogous increase of cybercrime activities as reported by multiple sources [17, 18], further exacerbating the situation.

To this end, the present thesis explores different facets of malicious behaviors associated with cybercriminal activities, involving:

- **Chapter 6:** Selling illegal products (i.e. breached data, stolen accounts, hacking/cracking tools/services, botnets, custom malware, etc.).
- **Chapter 7:** Detecting Algorithmically Generated Domains used by malware and botnets.
- **Chapter 8:** Novel and emerging threats in the context of blockchain-based DNS services.

1.2.1 Cybercrime Markets on the Surface Web

A staple of cybercrime has been the Internet communities centered on the exchange of knowledge and illicit services among malicious actors. In particular, forums have been shown to comprise the principal medium for cybercriminals to network, form communities, and operate online stolen data markets, despite numerous successful infiltrations by law enforcement

agencies [19]. Malicious actors gain internal access to sensitive data sources, and then acquire millions of credit and debit card details, user credentials, as well as sensitive data which can be used to identify individuals uniquely. The sheer quantity of data that can be acquired has given rise to a burgeoning market for actors who sell the information that they obtain, through, e.g. hacking and other forms of data theft, to other users. Participants in these illegally acquired data markets leverage various communication and networking methods, enabling them to freely form communities and interaction mechanisms. The most prevalent forms of such marketplaces, as identified in the literature, are Internet forums and Internet-Relay-Chat (IRC) channels [20, 21].

The dark web is considered the default place on the Internet in which such behaviors and actions flourish. Nonetheless, they are promoted in closed circles so that there is some “control” over who can access this information as well as to retain the anonymity of the perpetrators. However, should this information be openly disseminated in public channels, it implies that the promoted behavior is widely practiced and is considered a norm by some groups. This thesis tries to answer the question of whether such actions are so widely performed that they can be observed on the surface web. While dark web markets are still the key stakeholders when it comes to illegal trading, several surface web marketplaces have recently been repeatedly reported for trading leaked data [22, 23]. A very interesting characteristic, in this case, is that despite the trading of illicit products, the surface marketplaces have a very open form, e.g. they do not require any registration to access them, and the “loot” that is traded is advertised openly across the web. Currently, there are several such marketplaces operating with similar functionality; however, this work is mainly focused on Shoppy (<https://shoppy.gg/>) which appears to have the most users and products at the time of writing. Nonetheless, similar illicit trends have been found in the rest of the surface web marketplaces. As such, this thesis aims to provide an overview of what is actually being sold in such a marketplace, and leverage methods (e.g. machine learning) to automatically determine which are the illegal products and the main organizations affected.

1.2.2 Domain Generation Algorithms

The continuous arms race between malware authors and security researchers has pushed modern malware to evolve into highly sophisticated software, with present-day botnets having the capacity to infect millions of devices. The vast amount of sensitive information that can be extracted from compromised devices, coupled with the harnessing of their resources

and processing power, provides a wide range of monetization methods fuelling a flourishing worldwide underground economy.

Persistence and orchestration are the central objectives of cybercriminals employing botnets. An orchestrating entity, the botmaster, manages infected devices (bots) which in many cases can scale to the order of millions, creating a botnet [24]. The botmaster manages a Command and Control (C2) server that communicates with the bots. This communication must preserve some degree of unlinkability to thwart any attempts to identify the botmaster. To ensure unlinkability, and as a counter-measure against take-down operations, botnets frequently make use of domain fluxing [25, 26] through Domain Generation Algorithms (DGAs). DGAs produce a huge number of domain names which bots try to communicate with iteratively to find the actual C2 server. However, only a small part of them is registered and active, creating a hydra effect [27]. The botmaster may regularly pivot control between domains, thus hampering the task of seizing control of the botnet. This is helped by the fact that an outsider cannot determine which domains will be used, nor statically block all these requests. The latter stems from the fact that there are too many domains, and the seed yielding a particular sequence of domains might be unknown or change frequently. Currently, there are several families of DGAs employed by various malware with varying rates of requests and different characteristics. This heterogeneous landscape hinders the timely and accurate detection of an Algorithmically-Generated Domain (AGD) [28] request, which could serve as a precise indicator of compromise (IoC) of a host at the network level. Motivated by the continuous evolution of DGAs and their use by cybercriminals in the context of developing increasingly advanced and resilient botnets, this thesis elaborates the creation of a large-scale dataset comprising more than 95 million domains belonging to 105 unique DGA families, which is then leveraged for proposing a novel set of features relevant for the efficient and effective detection of DGAs, even adversarial ones, or hard to detect (dictionary-based).

1.2.3 Emerging threats in Blockchain-based DNS

One could argue that there is a periodic paradigm shift between centralization and decentralization in computer science. Although the Internet was in principle designed to be distributed and decentralized by nature, in reality, the control is placed onto a relatively limited number of stakeholders and the quest for further decentralization is becoming an imminent need. As such, in recent years, an increasing demand and creation of decentralized services is witnessed.

The most profound example is blockchain technology, which is being widely deployed in various and different fields [29]. In different forms, this decentralization shift is gradually being realized in traditionally centralized services, such as DNS. DNS is a distributed database with a centralized data governance model, primarily controlled by ICANN. While DNS is currently one of the oldest still working Internet application-level protocols, it has several drawbacks that mandate its replacement. For instance, the most profound one is that DNS does not support cryptographic primitives. Therefore, any query and response can be intercepted by anyone in the same network, implying many privacy issues. Furthermore, some regimes have reportedly exploited DNS to censor web pages and services that contain content that they do not approve of. Finally, during the past few years, DNS servers have been used in amplification denial of service attacks and their records have been poisoned.

One of the most promising solutions to the aforementioned issues is decentralised blockchain DNS which is already adopted by several chains, e.g. Ethereum, Namecoin, and Emercoin, or specific protocols. In fact, despite their infancy, blockchain domains have attracted the interest of several big players. A notable example is Alibaba which recently filed a patent for a blockchain-based management domain name management system [30]. A brief overview of blockchain DNS and some skepticism was initially provided [31]. So far, the proliferation of blockchain DNS projects and research proposals, are already being exploited by cybercriminals [32]. Adversaries are expected to take advantage of such systems by exploiting the lack of knowledge, experience and maturity of the users, as well as inherent flaws that are present in the early stages of new technology.

As such, it becomes evident that there is a pressing need for exploring threat models relating to novel blockchain solutions. To this end, the present thesis explores several novel threats relevant for the blockchain DNS ecosystem, and performs an investigative analysis unveiling how such treats have already given form to tangible risks by being exploited by malicious actors, particularly focusing on the Namecoin and Emercoin blockchains.

Chapter 2

Publications and Contributions

The research conducted in the course of the PhD led to the publication of several articles. More precisely, 6 articles have been published in journals and 3 in conferences.

2.1 Publications of the compendium and Contributions

This section presents the list of articles this thesis as a compendium of publications is based on. For each, the contributions of the PhD candidate are described:

N. Lykousas, C. Patsakis, and V. Gómez. “Adult Content in Social Live Streaming Services: Characterizing Deviant Users and Relationships”. *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*. Ed. by U. Brandes, C. Reddy, and A. Tagarelli. IEEE Computer Society, 2018, pp. 375–382. DOI: [10.1109/ASONAM.2018.8508246](https://doi.org/10.1109/ASONAM.2018.8508246)

- **Overall contributions:** This work performs an in depth analysis of two Social Live Streaming Services in order to understand and characterize deviant behaviors involving the production and consumption of adult content in these platforms, as presented in Chapter 3. It was the first quantitative study of deviant behaviors in SLSS. Furthermore, it resulted the collection and open-sourcing of a large scale dataset with user profile information and directed friendship links for the studied social platforms.

– **Contributions of the PhD candidate:**

- * First author of the article.
- * Conception of the main idea.
- * Implementation of the data collection pipeline.
- * Major contributions to the methodology and the presentation of the results.

N. Lykousas and C. Patsakis. “Large-scale analysis of grooming in modern social networks”. *Expert Systems with Applications*, 176, (2021), p. 114808. DOI: [10.1016/j.eswa.2021.114808](https://doi.org/10.1016/j.eswa.2021.114808)

- **Overall contributions:** This article primarily aims to identify and disentangle the mechanics of grooming and predatory behaviours in the context of Social Live Streaming Services, by analysing the behavioural and communication patterns of viewers, at a broadcast-level. The analysis performed in this work illustrates how predatory behaviors bypass the filters of service providers by, e.g. altering some “bad words”, or by using emojis. Notably, this was the first work highlighting the role of emojis in grooming. Furthermore, the analysis demonstrated that it was possible to identify illegal actions, such as the grooming of minors. The exact methodology and analysis is outlined in Chapter 4. Moreover, this work aims to facilitate research in this field and the generation of new filters and algorithms to detect such predatory behaviour through the release of a large-scale dataset of both verbal and non-verbal interactions (e.g. likes and rewards) in a Social Live Streaming Service [34].

– **Contributions of the PhD candidate:**

- * First author of the article.
- * Conception of the main idea.
- * Implementation of the data collection pipeline.
- * Design of the experimental set-up and training of the underlying ML models.
- * Major contributions to the methodology and the presentation of the results.

N. Lykousas, F. Casino, and C. Patsakis. “Inside the X-Rated World of “Premium” Social Media Accounts”. *Social Informatics - 12th International Conference, SocInfo 2020, Pisa, Italy, October 6-9, 2020, Proceedings*. Ed. by S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, and D. Pedreschi. Vol. 12467. Lecture Notes in Computer Science. Springer, 2020, pp. 181–191. DOI: [10.1007/978-3-030-60975-7_14](https://doi.org/10.1007/978-3-030-60975-7_14)

- **Overall contributions:** This article aims to shed light on the semi-illicit adult content market layered on the top of popular social media platforms and its offerings, as well as to profile the demographics, activity and content distributed by the adult content producers. The analysis is focused on the FanCentro platform, and the results are presented in Chapter 5.
- **Contributions of the PhD candidate:**
 - * First author of the article.
 - * Conception of the main idea.
 - * Implementation of the data collection pipeline.
 - * Major contributions to the methodology and the presentation of the results.

F. Casino, N. Lykousas, I. Homoliak, C. Patsakis, and J. Hernandez-Castro. “Intercepting Hail Hydra: Real-time detection of Algorithmically Generated Domains”. *Journal of Network and Computer Applications*, 190, (2021), p. 103135. DOI: [10.1016/j.jnca.2021.103135](https://doi.org/10.1016/j.jnca.2021.103135)

- **Overall contributions:** This work focuses on the study of Algorithmically-Generated Domains (AGD) and their detection, which are often employed by malware and bot-nets primarily for evading take-down attempts, thus enhancing their persistence on compromised systems. To this end, the HYDRAS dataset is developed, the most comprehensive and representative collection of AGDs to date. Based on the analysis of this dataset, a set of novel features useful for the detection of AGDs is introduced, which outperform the current state-of-the-art in terms of both classification performance and efficiency. The proposed approach is covered in Chapter 7.
- **Contributions of the PhD candidate:**
 - * Second author of the article.
 - * Implementation of the presented Machine Learning experiments.
 - * Profiling of the performance and the computational overhead of the proposed method.
 - * Major contributions to the methodology and the presentation of the results.

C. Patsakis, F. Casino, N. Lykousas, and V. Katos. “Unravelling ariadne’s thread: Exploring the threats of decentralised dns”. *IEEE Access*, 8, (2020), pp. 118559–118571. DOI: [10.1109/access.2020.3004727](https://doi.org/10.1109/access.2020.3004727)

- **Overall contributions:** This article presents the emerging threat landscape of blockchain-based decentralised DNS and provides an empirical validation of such threats with

real-world data. Specifically, a part of the blockchain DNS ecosystem is explored in terms of the browser extensions leveraging such technologies, the chain itself (Namecoin and Emercoin), the domains, and users who have been registered in these platforms. The findings reveal several potential domain extortion attempts and possible phishing schemes. Moreover, countermeasures are suggested for addressing the identified threats. Chapter 8 presents the most significant results of this study.

– **Contributions of the PhD candidate:**

- * Co-author of the article.
- * Contributions to the methodology and the presentation of the results.

F. Casino, N. Lykousas, V. Katos, and C. Patsakis. “Unearthing malicious campaigns and actors from the blockchain DNS ecosystem”. *Computer Communications*, 179, (2021), pp. 217–230. DOI: [10.1016/j.comcom.2021.08.023](https://doi.org/10.1016/j.comcom.2021.08.023)

- **Overall contributions:** The article performs a longitudinal analysis on the two major blockchain DNS solutions to date, namely the Namecoin and Emercoin, which have been repeatedly reported for malicious abuse, trying to identify and quantify the penetration of malicious actors in their ecosystems. To this end, a taint analysis on the metadata existing in these blockchains is performed, aiming to identify malicious acts. The analysis provides an automated validation methodology that supports the various reports about the wide-scale abuse of these solutions showing that malicious actors have already obtained an alarming and extensive share of these platforms. Some of the most important findings of this study are outlined in Chapter 8.

– **Contributions of the PhD candidate:**

- * Second author of the article.
- * Major contributions to the methodology and the presentation of the results.

2.2 Other publications and Contributions

This section presents the publications which have also been published in the course of this dissertation but are not included in the compendium, as their relevance to the main theme of the thesis is limited.

N. Lykousas, C. Patsakis, A. Kaltenbrunner, and V. Gómez. “Sharing emotions at scale: The vent dataset”. *Proceedings of the Thirteenth International Conference on*

Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019. AAAI Press, 2019, pp. 611–619

- **Overall contributions:** This work introduces the Vent dataset [40], the largest annotated dataset of text, emotions, and social connections to date. This dataset comprises more than 33 millions of posts by nearly a million of users together with their social connections, collected from the Vent social network¹. The dataset is studied to map *affective landscape* of Vent, finding agreements with existing (small scale) annotated corpus in terms of emotion categories and positive/negative valences. Moreover, statistical analysis describes the global patterns of activity in the Vent platform is performed, which reveals large heterogenities and certain remarkable regularities regarding the use of the different emotions.
- **Contributions of the PhD candidate:**
 - * First author of the article.
 - * Conception of the main idea.
 - * Implementation of the data collection pipeline.
 - * Major contributions to the methodology and the presentation of the results.

V. Koutsokostas, N. Lykousas, T. Apostolopoulos, G. Orazi, A. Ghosal, F. Casino, M. Conti, and C. Patsakis. “Invoice #31415 attached: Automated analysis of malicious Microsoft Office documents”. *Computers & Security*, 114, (2022), p. 102582. DOI: [10.1016/j.cose.2021.102582](https://doi.org/10.1016/j.cose.2021.102582)

- **Overall contributions:** This article explores the modern landscape of malicious Microsoft Office documents, exposing the means that malware authors use. To this end, this work introduces a public dataset [42] containing both benign and malicious documents that incorporate dynamic features such as VBA macros and DDE. Then, the relevant features are extracted with an automated analysis pipeline, for the purpose of efficiently and accurately classifying a document as benign or malicious using machine learning methods. The proposed approach is found to outperform the current state of the art detection algorithms.
- **Contributions of the PhD candidate:**
 - * Second author of the article.
 - * Major contributions to the methodology and the presentation of the results.

¹<https://www.vent.co>

F. Casino, N. Totosis, T. Apostolopoulos, N. Lykousas, and C. Patsakis. “Analysis and Correlation of Visual Evidence in Campaigns of Malicious Office Documents”. *Digital Threats*, (2022). ISSN: 2692-1626. DOI: [10.1145/3513025](https://doi.org/10.1145/3513025)

- **Overall contributions:** This article explores how visual elements used to lure users in enabling malicious payload on Microsoft (MS) Office documents can be used construct lightweight malware signatures that can be applied with minimal effort. The extensive tests from active malware campaigns can efficiently identify, correlate, and distinguish campaigns illustrating that some of them are either using the same tools or that there is some collaboration between them.
- **Contributions of the PhD candidate:**
 - * Co-author of the article.
 - * Major contributions to the data curation, clustering, and the presentation of the results.

Part II

Deviant Online Behaviors in Social Media

Chapter 3

Detecting and characterizing users involved with adult content production and consumption in SLSS

Social Live Stream Services (SLSS) exploit a new level of social interaction. One of the main challenges in these services is to detect and prevent deviant behaviors that violate community guidelines. This chapter focuses on the analysis of two popular SLSS services where adult content-related deviant behaviors have been prevalent: LiveMe (**LM**) and Loops Live (**LL**). First, a functional overview of the studied services is provided in Section 3.1. Then, the data collection and user labeling approach is outlined in Section 3.2. Afterwards, a characterization of users involved with adult content production/consumption and their relationships is discussed in Section 3.3. Finally, Section 3.3 presents and evaluates a set of novel graph-based features useful for the characterization and classification of adult content producers and consumers in SLSS.

3.1 LiveMe & Loops Live functional overview

Most of the features and functionality offered by those platforms are mobile-only, in that users wishing to actively participate in their communities need to own mobile devices such

Chapter 3 | Detecting and characterizing users involved with adult content production and consumption in SLSS

as smartphones and tablets running on Android or iOS.

The dynamics of both communities are based mostly on three possible actions performed by the users: **(a)** create real-time broadcasts and optionally associate hashtags representing thematic categories/user interests with them; **(b)** join broadcasts created by other users and interact with them as well as with the other viewers. Those interactions include exchanging chat messages with other viewers, and rewarding the broadcasters with “likes” and purchasable virtual gifts; and **(c)** follow other users and receive notifications when they are broadcasting. Contrary to other popular SLSS like Periscope [44], all the broadcasts in LM and LL are public. All active broadcasts are visible on a global public list. In both platforms, the concept of re-sharing/re-posting broadcasted content across different users is not present. Nevertheless, users are able to get shareable links to live shows that can be used for promoting broadcasters on other social media.

As already discussed, both platforms enable users to report community policy violators and underage users, who consequently get their accounts banned after their activity has been reviewed by moderators. Additionally, LM offers safety features to proactively protect its users, like the “Admin” feature, which enables broadcasters to allow other trusted users to be administrators for their broadcasts to block commenters on their behalf in real time.

Both platforms are equipped with more advanced features. Some significant examples are the ability to view currently popular/trending or “featured” broadcasts, either globally (both services), or by geographical region (LM), or by hashtag (both LM and LL) and the ability to find users or hashtags matching a search term. The mechanics of the broadcast featuring system are different for each platform, but in both cases factors such as the number of viewers, the amount of user interaction within the broadcast including likes, gifts and messages, and the duration of the live show are taken into account. Moreover, the popularity and experience of a user is reflected by their “level”, which is determined by their participation in activities such as broadcasting, joining broadcasts of others, sending and receiving gifts, chatting, etc. Leveling up enables users to receive various privileges such as discounts for buying virtual currency and access to premium gifts.

Broadcasters have the incentive to get their live shows featured, since this leads to better visibility within the app, thus, attracting a higher number of viewers who in turn can potentially reward them with virtual gifts. Once a broadcaster has received a certain amount of virtual gifts, they are able to cash them out for real money. Finally, both platforms offer a range of synchronous interaction features traditionally provided from the majority of OSNs

like direct messaging between users and the ability to “block” users.

Follow edges in LM and LL social graphs are directed; users can follow other users who do not follow them back. In addition, following someone does not require their permission. In the context of this study the focus is specifically on the user-specific attributes and following-follower social graphs of those platforms.

3.2 Data collection methodology

In this section, the methodology for collecting and labeling the data is described. The objective of the proposed data collection methodology is twofold: to identify adult content producers by analyzing the available broadcast replays and to sufficiently sample the portion of the social graphs where adult content production and consumption phenomena are predominant. To accomplish this, a novel data collection and labeling approach was developed which is detailed in the following subsections.

3.2.1 Sampling the social graphs

Both LM and LL applications communicate with their servers using an API with TLS-protected access. At the time of writing, no open-source clients for these services exist, hence, an approach similar to [45] and [46] is considered. For each platform, the network traffic between the app and the service is analyzed. More precisely, a TLS-capable man-in-the-middle proxy is leveraged, which uses proxies the network communication between a mobile device with the specific apps installed and the LM and LL services, acting as a transparent proxy. The proxy intercepts the HTTPS requests sent by the mobile device and pretends to be the server to the client and the client to the server, enabling the examination of the exchange of requests and responses between the client apps and the servers.

Using this setup, a set of APIs enabling the crawling of the social graph, extracting user profile and broadcast information, and using the search capabilities offered by the services is identified. Content-wise, both services use the HTTP Live Streaming (HLS) protocol [47] for hosting and delivering broadcast video replays, similar to other well-known live streaming services such as YouNow, Periscope, and Twitch. Figure 3.1 illustrates this architecture (steps 1-5).

Chapter 3 | Detecting and characterizing users involved with adult content production and consumption in SLSS

First, a set of *seed nodes* likely to be involved with the production of adult content is identified, in order to bootstrap a subsequent crawling procedure for sampling the social graphs. To this end, a seed node is defined as a user that (i) has a username that contains a pornographic term, (ii) has broadcast activity and (iii) has been banned by the system. For this, a list of adult keywords provided by [48] is used, which was also utilized in the context of their proposed deviant graph extraction procedure. This list contains 5,283 search keywords from professional adult websites.

Using these three criteria, 390 and 47 seed nodes for LM and LL, respectively, were identified. Figure 3.1 (steps 6-7) illustrates the seed identification step. Note that this step does not consider the network structure.

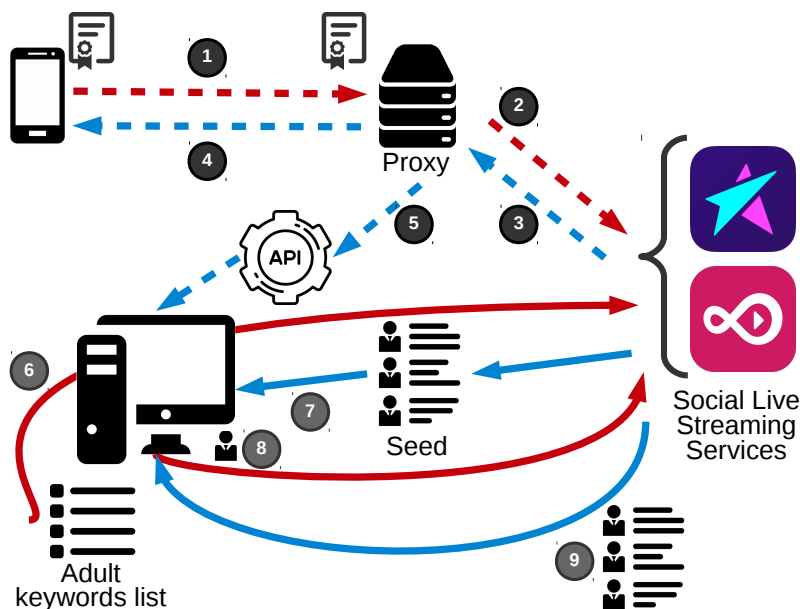


Figure 3.1: Data collection methodology. A proxy intercepts the messages from the smartphone to the SLSS. To decrypt the traffic and derive the API, a root certificate is installed in the smartphone (steps 1-6). Then, an adult keyword list is used to get an initial set of seed users, which are then queried to collect even more users. Based on their properties (e.g. banned, gifts) the dataset is constructed accordingly (steps 7-9).

The next step (denoted as 8-9 in Figure 3.1) consists of traversing and collecting profile information as well as broadcast video replays from each user, following the directed friendship links. A Breadth-First (BF) traversal limited to two hops away (undirected distance) from the seed nodes is followed. Thus, the constructed network consists of the union of the 2-hop ego-networks of all seed nodes. This union resulted in one single connected component in both platforms.

For computational reasons, the nodes in the boundary that appear as neighbors of a node with a degree higher than $10K$ are discarded. These nodes correspond to only 718 and 267 profiles for LM and LL, respectively, a very small proportion of the complete 2-hop ego-networks. It is emphasized that this work is not focused on capturing the entire network of users, but a tractable subset of tightly connected groups of users in which adult content is predominant. BF search covers satisfactorily small regions of a graph [49] and has been used in many analyses. Thus, it is expected that selection or sample bias will have a reduced impact in this work. During the data collection period, which lasted from January to November of 2017, this crawling procedure was repeated once per week on average. Based on the number of installations reported by LM on Google Play at the time of writing ($20M - 50M$ installations), the described approach managed to crawl roughly $5.8\% - 14.5\%$ of the total network. Similarly, for LL it collected $5.46\% - 27.3\%$ of the total network.

Table 3.1 summarizes the obtained networks for both platforms. As expected, the LM network is much larger than the LL network, containing approximately 10 times more users and 30 times more edges. The LL network is, however, one order of magnitude denser than the LM one.

	$ N $	$ E $	$ B $	$\langle k \rangle$	D	ρ
LM	2,942,407	37,440,992	142,345	25.4	4.32×10^{-6}	0.14
LL	273,177	1,193,780	114	8.73	1.59×10^{-5}	0.08

Table 3.1: Network statistics of the crawled graphs: number of nodes $|N|$, number of edges $|E|$, number of banned nodes $|B|$, average degree $\langle k \rangle$, density D , and reciprocity ρ .

The approach followed has three main limitations. Firstly, the set of replays captured includes only the available replays of past broadcasts at crawling time. Replays that were deleted in-between the crawls as well as all live broadcasts streamed during our crawls were not included. This is not a fundamental limitation, and can be fixed by using more sophisticated approaches [44]. Moreover, regardless of whether an account is banned (suspended) or active, none of the platforms provides metadata to determine the reason behind the account suspension. This means that the collected dataset includes false positives that were banned because of other unrelated policy violations. This limitation is addressed in the next subsection, in which the replays' content is analyzed to determine whether a user is deviant or not. Finally, there is a small probability of false negatives, a portion of deviant users are not retrieved by the proposed method. This can happen because moderators can only identify a

limited number of users engaging in inappropriate behavior [50] and those may lie isolated (more than two hops away) from the seed nodes.

3.2.2 labeling the users

Next, the labeling approach of the collected users is outlined.

This work considers three types of users: adult content producers, or simply *producers* (based on their broadcast activity), *consumers* (based on their relation with producers), and *normal* users that are not included in any of the two other categories.

Given the network, the criterion to establish whether a user is a producer is exclusively based on the images of the user's broadcast activity. This choice disregards indirect sources of information and does not require manual inspection, and hence, it can scale up efficiently. Alternative approaches are based on manual inspection of metadata only [48], which may not have been sufficient, or using crowdsourcing approaches for categorizing broadcasts [51], which would require a pool of crowd workers to be potentially exposed to offensive material.

To this end, the OpenNSFW¹, a deep neural network model pre-trained to detect pornographic images is used. Convolutional Neural Networks are the state of the art in image classification problems [52, 53]. OpenNSFW takes an image as input and provides a value representing confidence in an image's resemblance to pornography. The network is fed with frames sampled from the broadcasts at 1/3 Hz, and the *highest* confidence score for every broadcast replay is stored. This value represents the maximum probability a replay contains pornographic content. Then, at the end of the data collection period, every user was associated with the *highest* value provided by OpenNSFW over all of their replays in the dataset. The aforementioned value can be considered as a user's *adult content production score*. This value is set to zero for the users that no broadcast data was collected.

Figure 3.2 shows the cumulative distribution function (CDF) of the adult content production score for both LM and LL networks differentiating between seed users and all the users in our sample.

It is observed that a very small proportion of all users (only around 0.4%) scored above 0.5, indicating that the vast majority of users do not broadcast adult related material. On

¹<https://yahoeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for>

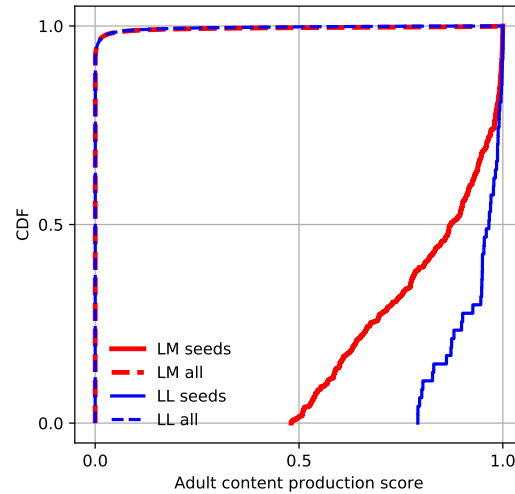


Figure 3.2: Cumulative distribution function of the adult content score values.

the other hand, the seed users have been assigned high scores (all starting at 0.5 for LM and around 0.8 for LL). This confirms that the choice of seed nodes and the outcome of the classifier agree to a large extent.

Next, whether a user is a *producer* is established using a thresholding approach. In particular, the probability distribution of the scores for both banned and non-banned users is considered, and a threshold is computed by the Bayesian decision rule that separates the two classes. This results in a threshold of 0.82 and 0.93 for LM and LL, respectively.

Moreover, whether a user is considered a *consumer* is based on the set of producers and the network structure. In particular, a user is labeled as a consumer if they follow *at least two* adult content producers. While following a single user (producer or not) can be expected by random chance, following two users of the producer class (given they only make up for a minor fraction of the total users), is much less likely to be by chance. In this work, the definition of consumer is stricter than the one of [48], which defines as a *passive consumer* a user that follows *at least one* single producer. In the case a user falls in both categories, i.e., producers that also followed at least two other producers, the producer label is considered more relevant. In practice, only 9 users of LM and 2 users of LL should have been labeled as both producers and consumers. Finally, users who do not fall into the above classes are labeled as *normal* users.

Table 3.2 summarizes the resulting labeling according to the proposed procedure. As

expected, one can observe that only a small proportion of the crawled networks are not labeled as normal users. Also, inside parentheses, it is shown how the seed nodes are distributed in the three categories. Recall that seed users are banned users with broadcast activity and with a adult-related username. Although most of them are labeled as producers, there is a significant proportion labeled as normal users. This can be explained by the fact that those users may exhibit other (non adult-related) deviant behaviors and thus not relevant, or because their score did not reach the specified threshold, as reported from the OpenNSFW classifier. Remarkably, none of them is labeled as a consumer, which already suggests that producers are not well connected between them.

Class	LiveMe	Loops Live
Producers	7,135 (228 seeds)	92 (33 seeds)
Consumers	30,872 (0 seeds)	1,243 (0 seeds)
Normal	2,904,400 (162 seeds)	271,842 (14 seeds)

Table 3.2: Distribution of users according to their class.

3.2.3 Effectiveness of SLSS moderation systems

Having identified the aforementioned user classes, next, the discussed labeling approach is compared to the moderation of each platform. Table 3.3 shows how banned users are distributed in each class. In the case of LM, it is observed that only 43.5% of the labeled pro-

Class	LiveMe	Loops Live
Producers	43.5% (3,109)	96.7% (89)
Consumers	9.6% (2,970)	0.08% (1)
Normal	4.6% (136,266)	0.008% (24)

Table 3.3: Proportion (total in parenthesis) of banned accounts in each class.

ducers have been banned. Since it is unlikely that the frames extracted from the broadcasts contained adversarial perturbations [54] against the OpenNSFW model, it can be safely assumed that moderation is highly ineffective in detecting such cases.

On the contrary, moderation of LL is consistent with the labeling outcome, with 96.7% of users placed in the producers class being banned. This consistency provides further confir-

mation of the decision to use a pre-trained deep learning classifier for detecting adult content. Finally, the high number of banned users placed in the *normal* class suggests the existence of a significant proportion of policy violators outside the context of the present work.

3.3 Profiling deviant users

In this section, the behavior of adult content producers, consumers and their relationships within the sampled networks are characterized. First, a set of features directly accessible from each user is considered, and their relevance for distinguishing between classes is analyzed. The classes are: normal users, producers, and consumers.

3.3.1 Features

Based on the available profile information collected from the two platforms, the following set of features is introduced:

- **Network features:**
 - Number of followers.
 - Number of followings.
 - Number of bidirectional friends.
- **User-based features**
 - Pornographic username (binary): *whether or not the username contains a pornographic term.*
 - Suspended/Banned (binary): *whether the account has been suspended by platform moderators.*
 - Replay count: *Number of past broadcasts available for replaying.*
 - Level: *An integer value reflecting the participation level of a user in various SLSS-specific activities.*
 - Praise (**only LM**): *Total number of likes received in all user's broadcasts.*
 - Income (**only LM**): *Total virtual currency value of gifts received in all of the user's broadcasts.*

The relative power of these features in discriminating the three user classes is assessed by computing the information gain of each feature. Table 4.5 reports the ranking of the top five

Chapter 3 | Detecting and characterizing users involved with adult content production and consumption in SLSS

most important features in differentiating the three user classes for each platform.

IG Rank	LM		LL	
	Feature	IG Score	Feature	IG Score
1	#Followings	0.31	#Followings	0.37
2	#Followers	0.25	#Friends	0.26
3	Praise	0.15	#Followers	0.19
4	#Friends	0.12	Banned	0.10
5	Income	0.06	Porn nickname	0.05

Table 3.4: Top 5 features for differentiating the three classes.

The number of followings, followers and friends are among the highest ranked features for both networks which indicates the importance of social relationships for characterizing the given classes. Also, for LM it is observed that the amount of likes (praise) and virtual gifts (income) are highly important as well. In order to get a deeper insight on how these features are distributed across the different classes, their cumulative distribution functions (CDFs) are plotted in Figure 4.10. Information about praise and income was not available for LL, preventing an 1-to-1 comparison between the two datasets. Instead, high importance for the banned and pornographic username attributes is observed. This is due to the fact that, as shown in Subsection 3.2.3, the banned LL users are almost exclusively adult content producers, and also a significant part of them have pornographic usernames (see also Subsection 3.2.1).

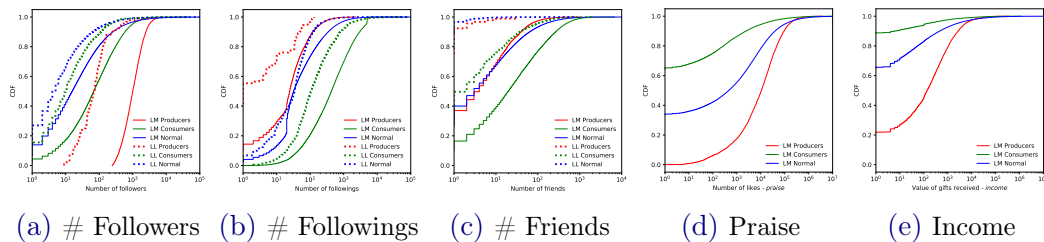


Figure 3.3: Cumulative distribution functions (CDFs) of the different user profile attributes.

From Figure 4.10c, it is evident that adult content producers tend to have many more followers than the other classes, and there exists a lower bound to the follower number of producers, approximately 10 and 250 for LL and LM, respectively. On the contrary, consumers in LM are found to have the least amount of followers among the three classes. For the number of followings (Figure 3.3b), however, the situation is reversed. Consumers dom-

inate over the other classes by following significantly higher amounts of users, while the producers come last in this aspect with around 41% (LL) and 10% (LM) of them not following any other users. The distribution of the friend number reveals that consumers are much more likely to form reciprocal relationships, while it appears to be almost identical for producers and normal users, as Figure 3.3c indicates. Furthermore, it can be observed that adult content producers tend to receive the highest amount of praise and income among the three classes. Additionally, consumers receive much less recognition for their broadcasting activities compared to both normal users and producers. In fact, while no producers with zero praise exist, approximately 65% of consumers and the 33% of normal users in the dataset fall in this “unpopular” category, see Figure 3.3d. This either means that they have not received any likes during their shows, or they have never broadcasted anything. A similar trend is observed for the total value of the virtual gifts received, represented by the income attribute and illustrated in Figure 3.3e. Only 21% of producers have not received gifts, while the same holds for the 88% of consumers and the 65% of normal users.

3.3.2 Deviant relationships

To determine the community structure of these networks, existing variants of the Louvain method [55] do not find well identifiable clusters of users. In both networks, it was observed that producers and consumers are distributed nearly uniformly across the clusters. Furthermore, the results vary significantly between different runs. Thus, a different approach was adopted to better understand the network structure, aiming to assess who in the sampled networks is significant with regards to their social relationships. To this end, the ranking HITS algorithm [56] is used to identify the hubs and authorities in the social graphs. The basic principle behind HITS algorithm is the following mutually reinforcing relationship between hubs and authorities: good hubs point to many good authorities and vice-versa. Interestingly, it appears that adult content consumers have the highest hub scores among all users in both networks, as shown in Figure 3.4b. Moreover, the highest authority scores in LM belong almost exclusively to producers. The latter could be correlated with the significance of the number of followers and followings to discriminate producers and consumers, from normal users, as previously shown. For LL, it is observed that most authoritative users do not belong to the producers class, and exhibit characteristics expected of prominent users in a social community such as the number of followers in the order of hundred of thousands. The reason behind this difference between LM and LL could be attributed to the very limited extent of the adult content production behavior in the later. Therefore legitimate popular users dominate

the authority scores in the sampled graph by being followed by consumers. Additionally, that the hub scores of the highly authoritative users are particularly low in both networks. This finding contradicts other studies on different social networks such as Twitter [57, 58], where researchers observed many well-connected users that have high scores as both authorities and hubs.

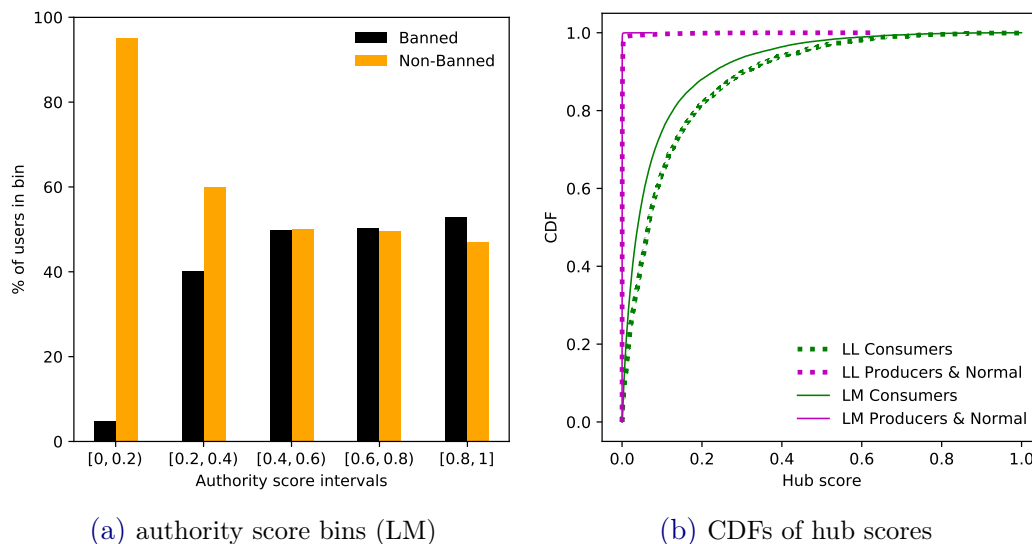


Figure 3.4: User relationship insights, provided by HITS.

A major finding was that in LM, the ratio of banned to non-banned users increases along with the increase of authority score. To better demonstrate this, the users are placed into bins based on their authority score, and the fractions of banned and non-banned users in each bin are calculated, as shown in Figure 3.4a. It is important to note that 99.5% of users fall into the first bin, having authority score less than 0.2. Thus, it can be concluded that banned users are more densely concentrated towards the higher end of the authority score spectrum, and the reason behind their suspension was likely the production of adult content, since they are followed by the consumers/hubs. A similar phenomenon is observed for LL, but with banned users mostly concentrated in the 0.02 – 0.35 authority score range, while the 97.7% of users have authority scores below 0.02.

Based on the arguments above, it is expected that consumers will follow multiple producers, a considerable portion of which will be banned. To quantify the relationship between the fractions of banned users and producers followed by consumers, their correlation is calculated using Spearman’s rank correlation coefficient ρ . Indeed, there exists a nearly perfect correlation for LL with $\rho = 0.96$, meaning that consumers do not follow almost any banned

users outside the producers class, and a moderately strong correlation in LM ($\rho = 0.63$).

Another dimension to examine is the connectivity within each class in the context of the sampled graphs. For this, the edge density is measured, computed as the ratio of edges between the users belonging in each class over the total number of possible edges between them. However, it is known that in real networks the number of nodes significantly impacts the density [59]. To account for this, the connectivity of the sampled graphs was compared with a null model that randomly rewires the edges while keeping the degree of each node unchanged, as described in [60]. Table 3.5 contains the link density comparison between the subgraphs induced by producers, consumers and normal users in each sampled network and the null model.

	Class	Crawled graph D	Null model D
LM	Producers	1.90×10^{-5}	4.55×10^{-6}
	Consumers	1.20×10^{-3}	4.46×10^{-6}
	Normal	2.18×10^{-7}	4.32×10^{-6}
LL	Producers	1.91×10^{-3}	0
	Consumers	3.28×10^{-3}	1.42×10^{-5}
	Normal	1×10^{-5}	1.59×10^{-5}

Table 3.5: Comparison in terms of link density D between the crawled networks of producers, consumers normal users and a corresponding random network.

It is observed that consumers, when compared to the null model, are several orders of magnitude more densely connected to each other in the sampled networks. On the contrary, the subgraphs of producers and normal users are much more sparse compared to consumers, with the producers being only slightly more densely connected than random for LM. This finding comes in contrast with the behavior of adult content producers in Tumblr and Flickr, where they are observed to form densely interconnected communities [48]. In LL the producers appear to have a density comparable to those of consumers, but given their limited number, this is due to the existence of producers who also exhibit consumer behavior.

In summary, the findings indicate that the closely-knit groups of consumers act as a “bridge” between the otherwise isolated producer nodes. Concretely, from a network perspective, the most effective way to reach adult content in the studied networks is by traversing the social links of consumer nodes that point to both producers and other consumers, enabling the reach of even more deviant nodes belonging in those two categories.

Since adult content producers are isolated in the network, it can be assumed that some of the consumers are actively monitoring the list of active broadcasts (see Subsection ??), and proceed to follow users who broadcast adult content, while also possibly sharing links to such live streams with other consumers. These “consumer leaders” are likely to become popular among their kin by being followed by many other consumers, thus serving as a means for the diffusion of information about producers, effectively compensating for the absence of the content reposting functionality in SLSS.

3.4 Modeling deviant behavior

Based on the findings discussed above, a set of novel graph-based features is proposed, useful for the characterization and classification of adult content producers and consumers in SLSS. These features are local, and exploit the social graph structure up to two hops away from each node, as well as the noisy account suspension signal of neighboring nodes.

3.4.1 Notation

Consider a follower graph G composed by a set of nodes N , a set of banned nodes $B \subseteq N$ and a set of directed edges $E \in N \times N$. When building the follower graph, an edge is drawn from node i to node j if i has followed j at any time, $i, j \in N$. The set of out-neighbors of node i is denoted as $\Gamma_{\text{out}}(i)$, while the set of in-neighbors as $\Gamma_{\text{in}}(i)$. Accordingly, the in-degree and out-degree of a user i are defined as $d_{\text{in}}(i) = |\Gamma_{\text{in}}(i)|$ and $d_{\text{out}}(i) = |\Gamma_{\text{out}}(i)|$, respectively.

3.4.2 Proposed features

Let $B(i) = |B \cap \Gamma_{\text{out}}(i)|$ denote the number of banned users followed by user i and let the approximate fraction of banned users that are followed by i be:

$$\text{pb}(i) = \frac{B(i)}{d_{\text{out}}(i) + 1}$$

The first feature, named *Deviant Rank*, aims to reflect the deviant behavior of a node $n \in N$ as either adult content producers or consumers. It is based on the intuition that

producers tend to be followed by users who follow many banned users, without exhibiting the same banned-following behavior themselves, while on the other hand *consumers* tend to follow banned users and at the same time be followed by other consumers. It is calculated as the average fraction of followed banned users over the followers of node n minus the proportion of banned users followed by node n :

$$\text{DevRank}(n) = \frac{1}{d_{\text{in}}(n) + 1} \sum_{i \in \Gamma_{\text{in}}(n)} (\text{pb}(i) - \text{pb}(n)). \quad (3.1)$$

Note that this measure takes values in the range $[-1, +1]$. For producers, the first term of Equation (3.1) will dominate the second one, while for consumers the second term will either be of comparable value to the first, or dominate it.

The second feature, *Deviant Authority Rank*, tries to characterize adult content producers from a different perspective. It is based on the idea that producers will tend to follow fewer banned users than their followers (consumers). Therefore, it is denoted as the sum of the ratios between the number of banned users of a follower i and the user n .

$$\text{DevAuthRank}(n) = \frac{1}{B(n) + 1} \sum_{i \in \Gamma_{\text{in}}(n)} (B(i) + 1). \quad (3.2)$$

Equation (3.2) can also be useful for the early detection of adult content production behavior in the context of SLSS moderation, provided that it is computed every time a user gets a follower. A sharp increase in its value (possibly above some threshold) for a user could signify that they started producing adult content, alerting the moderators for further inspection. Figure 3.5 shows the CDFs of the described features. It can be observed that they are robust and consistent in describing the different classes across both networks.

Next, the uncertainty on $\text{pb}(i)$ is quantified using a Bayesian estimate. For that, a Beta distribution² is considered, with parameters $\alpha = B(i) + 1$ and $\beta = d_{\text{out}}(i) - B(i) + 1$ (one is added so that none of the parameters is zero), which corresponds to modeling each outgoing link according to a Bernoulli process with probability $\text{pb}(i)$. The full posterior distribution of $\text{pb}(i)$ after having observed $B(i)$ banned users out of $d_{\text{out}}(i)$ events is fully described by the Beta distribution parameters. As such, it becomes possible to capture the difference between users with a low outdegree $d_{\text{out}}(i)$ (high uncertainty) and users with high outdegree $d_{\text{out}}(i)$ (high certainty) in the estimated probability of following a banned user.

²https://en.wikipedia.org/wiki/Beta_distribution

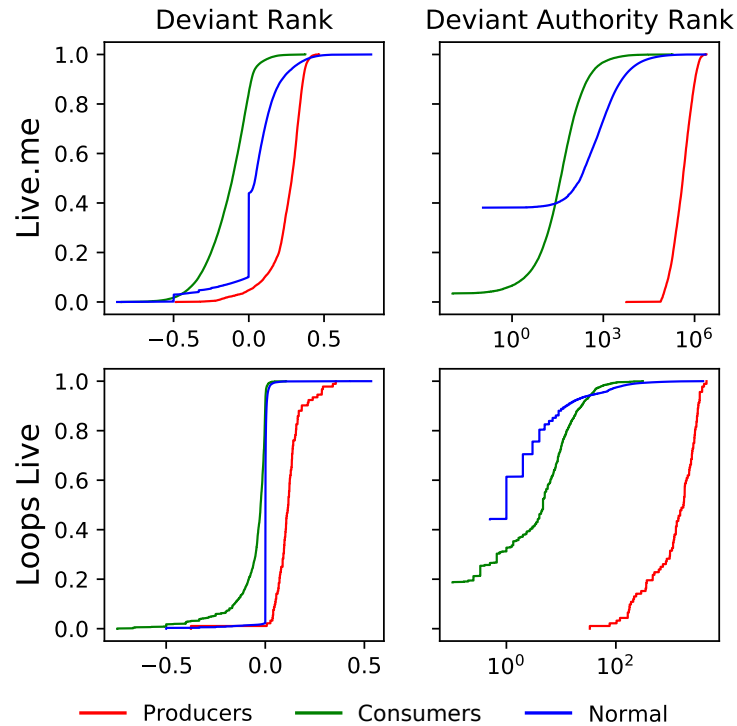


Figure 3.5: Cumulative distribution function (CDF) for the novel features for each dataset.

Accordingly, as additional features for characterizing consumers, the mean and the variance of the Beta distribution model is considered. For consumers the mean value is expected to be higher and the variance considerably lower compared to the other classes.

3.4.3 Evaluation

Finally, an evaluation of the proposed features in terms of discriminating among the three user classes is performed, by comparing them to a baseline of features that can be extracted from the users' profile information, as well as the global structure of the sampled social graphs. For this, a 3-class classification experiment was conducted using the datasets described in Table 3.2.

Three different configurations of features are considered: (a) A baseline set of features described in 3.3.1, without including the banned attribute (**BL**). (b) The proposed features calculated per user plus the baseline features (**P+BL**). (c) The hub and authority score values

for each user provided by the HITS algorithm, plus the baseline features (**H+BL**).

Using the above feature sets, a Random Forest classifier of size 30 with maximum depth for each tree of 15 is trained using 10-fold cross-validation. Specifically, the cross-validation was stratified in the sense that the proportion of samples from each class is approximately equal in each of the ten folds, in order to account for the class imbalance.

Table 3.6 shows the Precision (Pr), Recall (Rc), F-score (F1) performance indicators at the class level, as well as the macro-averaged F-score (macro-F1) across all the classes, for each model (M) on each dataset. Our comparative evaluation shows that the proposed features achieved significant improvements over the baseline model. More precisely, recall increased by 49.6% and 46.6% for producers, and by 90.1% and 82.1% for consumers in LM and LL datasets, respectively. The increase in precision was also substantial: 7.4% and 56.2% for producers, and 21% and 48.3% for consumers in LM and LL, respectively. These improvements are also reflected by the increased F-score, and the results are consistent across both datasets, demonstrating the robustness of our proposed features. Moreover, although in both cases the addition of HITS scores provides increased classification performance over the baseline set, the proposed features outperform them by a large margin in the case of LL, with precision increased by 9.6% and 39.2%, and recall by 26% and 68.4% for producers and consumers, respectively. For LM, P+BL and H+BL models performed equally well.

M	C	LM				LL			
		Pr	Rc	F1	MF1	Pr	Rc	F1	MF1
BL	P	0.728	0.232	0.352		0.323	0.119	0.174	
	C	0.745	0.057	0.106	0.484	0.495	0.147	0.227	0.466
	N	0.988	0.999	0.993		0.995	0.999	0.997	
P+BL	P	0.802	0.728	0.763		0.885	0.586	0.705	
	C	0.955	0.959	0.957	0.906	0.978	0.969	0.974	0.893
	N	0.998	0.999	0.998		0.999	0.999	0.999	
H+BL	P	0.813	0.753	0.782		0.789	0.326	0.461	
	C	0.960	0.962	0.961	0.914	0.586	0.285	0.384	0.614
	N	0.999	0.999	0.999		0.996	0.999	0.997	

Table 3.6: Classification Performance. P: Producers, C: Consumers, N: Normal, MF1: macro-F1.

It is important to note that HITS is a global ranking algorithm, and the descriptive power of HITS scores regarding the identified roles is a result of the social graph sampling approach used. In a global context, the existence of legitimate popular users with a very large number of links will have a critical impact on their performance for this task, as observed in LL, thus deeming them unreliable for scenarios where the entire network structure is known (i.e. SLSS moderation systems).

Chapter 4

Analysis of Grooming Behaviors in Social Live Streaming Services

This chapter aims to identify and disentangle the mechanics of grooming and predatory behaviors in the context of Social Live Streaming Services, by analyzing the behavioral and communication patterns of viewers, at a broadcast-level. In Section 4.1, the dataset used in the context of this research is presented, along with some descriptive statistics to better illustrate the dynamics of chatting in the context of SLSS broadcasts. Next, Section 4.2 identifies several characterizing attributes of grooming behavior as observed in the dataset, related to adversarial perturbations in chat text introduced by offenders for evading detection mechanisms, as well as the extensive use of emoji symbols in the grooming context. Finally, Section 4.3 describes the fitting of a topic model for clustering the chats, and analyzes the resulting topics in terms of signals indicative of different aspects of grooming behavior.

4.1 The dataset

This research produced a large-scale dataset of the public interactions between streamers and viewers during the live broadcasts of users identified as adult content producers in [33], from the LiveMe platform presented in Section 3.1. This dataset is made publicly available at [34]. In total, the dataset comprises 39,382,838 chat messages exchanged by 1,428,284 users, in the

context of 291,487 live broadcasts during a period of approximately two years, from July 2016 to June 2018. Each broadcast effectively functions as a temporary chatroom. The audience can interact with the streamers via text messages and reward them with virtual rewards, e.g. points, gifts, badges (some of which are purchasable) even virtual money. Apart from the chat messages, the dataset contains a wide range of user interactions along with metadata. The features are described below:

- **Metadata (broadcast)**
 - Total Viewers: total number of viewers who joined the livestream as viewers.
 - Duration: duration of stream in seconds
- **Metatadata (broadcaster)**
 - Country Code
- **Interactions**
 - Likes: Viewers who liked the broadcast & the number of likes given.
 - Follows: Viewers who followed the broadcaster during the livestream.
 - Gifts: Viewers who sent virtual gifts to the broadcaster, along with value (in virtual currency) for each gift.
 - Shares: Viewers who shared the broadcast (via a link so others can join).
 - Blocks: Viewers who have been blocked by the broadcaster (i.e. banned from a stream).

To better understand how the features mentioned above are distributed, their cumulative distribution functions (CDFs) are plotted in Figure 4.1. Notably, every broadcast in the dataset had viewers (143.5 on average) and a sizeable duration (31.4 minutes on average). While most of the broadcasts received likes (92%), the 55% did not receive any gifts (since they cost money, contrary to likes). Furthermore, 47% of the broadcasts did not generate any new followers for the broadcasters. At the same time, the interactions of sharing and blocking are relatively rare (i.e. they are zero for 0.67% and 0.86% of the broadcasts, respectively). Next, the distribution of the broadcasters per country of the whole dataset is plotted in Figure 4.2, focusing on the 15 countries with most broadcasters, to provide an understanding of the geographical distribution of adult content producers.

4.2 | Identifying the characterizing attributes of grooming behavior

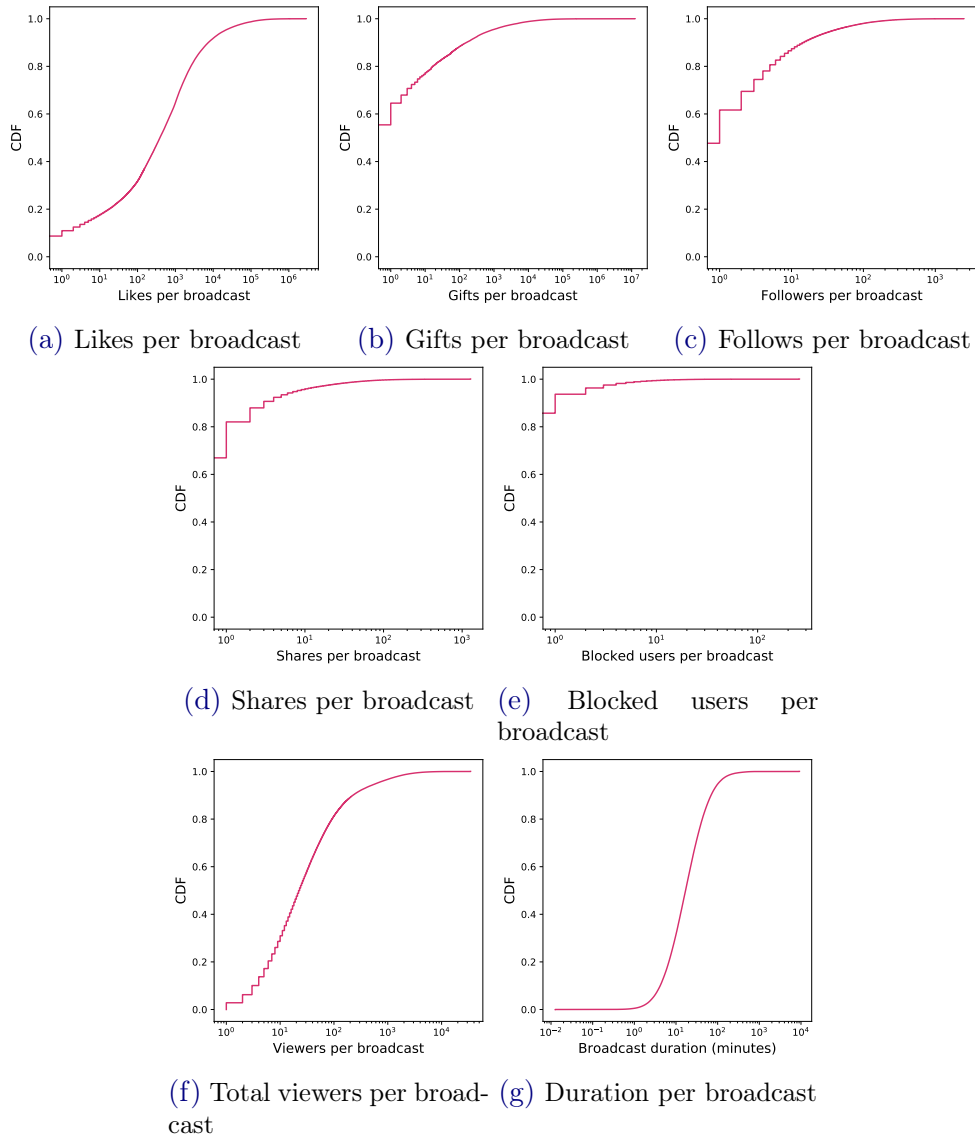


Figure 4.1: Cumulative distribution functions (CDFs) of broadcast metadata features and interactions.

4.2 Identifying the characterizing attributes of grooming behavior

As shown in Figure 4.3, around 82% of the broadcasts of adult content producers receive less than 100 chat messages. Moreover, out of the unique users chatting during these broadcasts,

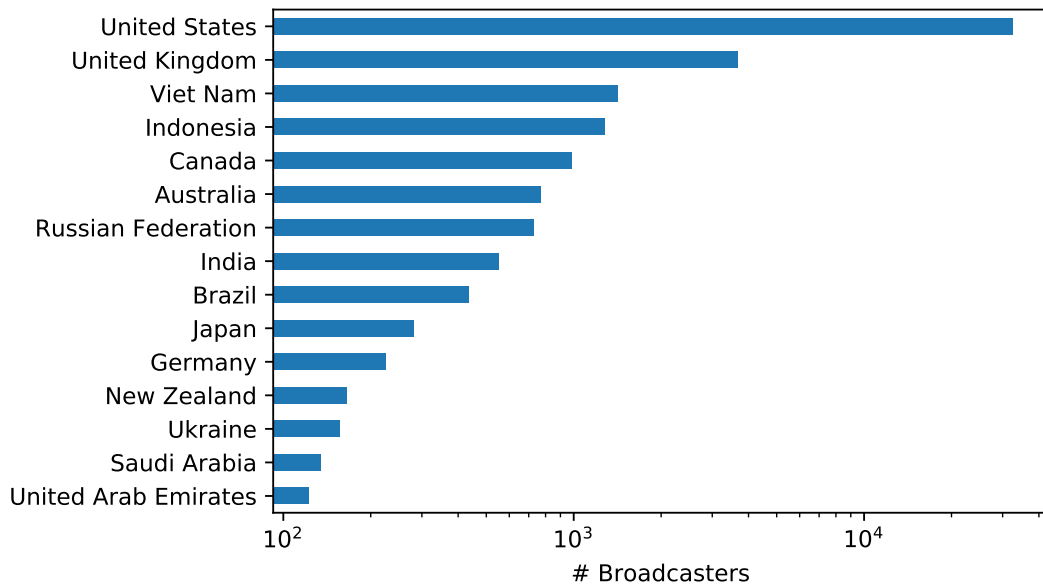


Figure 4.2: Broadcasters count per country

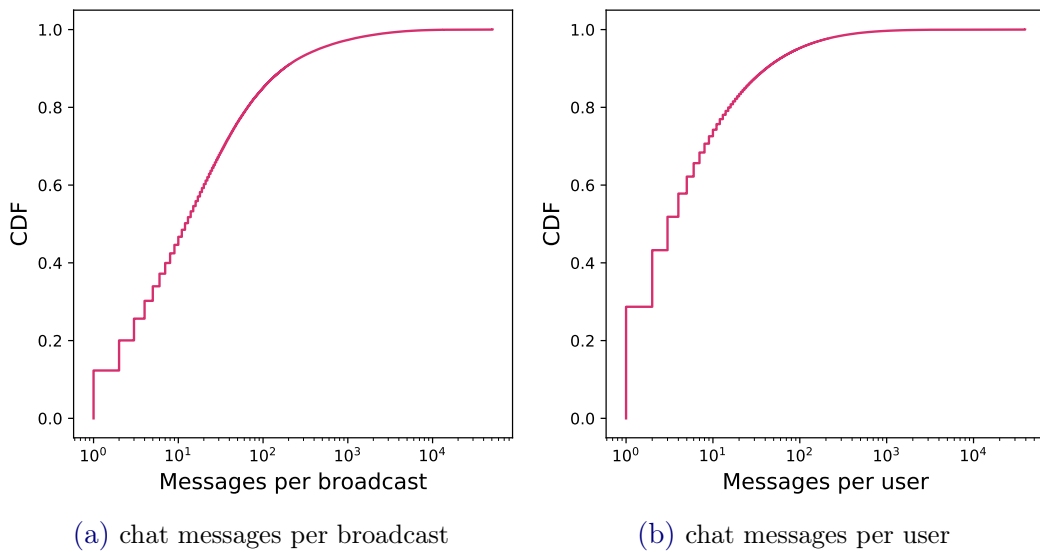


Figure 4.3: Cumulative distribution functions (CDFs) of the chat messages per broadcast and per user.

only 30% send more than ten messages in total. Both distributions are particularly heavy-tailed, meaning that the majority of chat messages in the dataset are exchanged during a few highly popular broadcasts.

4.2 | Identifying the characterizing attributes of grooming behavior

An approach followed by several authors in the most recent relevant works [61, 62, 63] was adopted to identify sexual grooming behavior in the chat messages: analyzing the Perverted-Justice Dataset (PJ), which although dated and relatively small-scale, was the only publicly available dataset of chats produced by online groomers at the time of writing, succeeded only by the dataset produced in the context of the present thesis. To this end, the chat messages comprising the collected dataset were searched for sexual content keywords defined in Linguistic Inquiry and Word Count (LIWC) corpus [64]. More precisely, the 2015 version of the LIWC dictionary for the sexual content variable comprises a total of 131 words. These include a wide range of terms about sexual matters, including sexual orientation (e.g. bi-sexual, heterosexual), sexual organs (e.g. penis*, vagin*, womb), slang terms, sexually transmitted diseases and infections, sexual violence and assault terms and sex enhancements. The most frequently occurring sexual terms in the PJ dataset, had a very low number of occurrences in the LiveMe chats (less than five exact matches in most occasions). The very low occurrence of such words implies the existence of an automated filtering mechanism in place. Nonetheless, relevant literature related to online chat has demonstrated that users with previous exposure to text-based automatic moderation techniques can easily circumvent them by introducing noise such as typos, grammatical errors, uncommon abbreviations and out-of-vocabulary words [65, 66]. To determine whether this is relevant in the dataset, Facebook's FastText library [67] was used to train subword-informed word representations on the LiveMe chats, which then were leveraged for identifying the semantically-similar adversarial misspellings of filtered terms (such as pussy, boobs, dick, etc.), by querying their nearest neighbors. The results indicate that indeed this is the case in LiveMe chats, as illustrated in Tables 4.2 and 4.3. To illustrate the sexual word misspellings better, the word cloud of the closest neighbors for the relevant LIWC terms is plotted in Figure 4.5.

Next, to understand the contexts where the aforementioned terms are used, Figure 4.6 plots the word cloud of their top collocates. It is evident that the 3 most frequently collocated words are the verbs *show* (13, 329 collocation occurrences), *open* (3, 032 collocation occurrences), and *see* (3, 028 collocation occurrences). To further investigate the imperative meaning of such words in the context of the grooming problem, Figure 4.7 plots the top collocates in the whole dataset of chat messages for the most frequent one: *show* (203, 230 total occurrences), clearly indicating the existence of sexually predatory behaviors. Similarly, the word *open* is most frequently collocated with words denoting positive politeness (such as please, plz) and endearment (e.g. baby, dear), as well as sexually connoted words, mostly related to clothing (e.g. underwear, clothes, top, shirt, pants, dress), and emojis representing clothing items (e.g. 👙, 👕, 👖, 👗).

While emojis are present in many published datasets, this is the first study to highlight their relevance in the context of grooming, especially the ones referring to clothing. To this end, the previously described embeddings-based approach was used to capture similar clothing terms, see Table 4.4. Using this method, a list of 300 unique terms is assembled, appearing in the chat messages of 45,086 live streams. Next, to examine the intentions underlying these messages, dependency parsing was performed on every chat message the clothing terms appear in, using the spaCy parser [68]. From the extracted parse trees, the simple and phrasal verbs were collected. Table 4.1 contains the 15 most frequently occurring simple and phrasal verbs in their base forms obtained using spaCy lemmatizer¹, after removing the verbs contained in the NLTK [69] stopwords list (e.g. be, can, do, have) to reduce noise in the results. Figure 4.8 plots a word cloud of the extracted verbs and verb phrases. The latter is a clear indication that predators are requesting streamers to perform inappropriate acts involving the removal of their clothes. These findings highlight the imperative nature of the predators' communications related to clothing items.

Additionally, the use of clothing-related emojis² was explored, occurring in 153,797 chats. Notably, 83.6% (128,604) of these messages contain only emojis, without any text. The emojis co-occurring with clothing-related emojis were deduplicated, since it has been shown that in text messages, emoji sequences tend to have a high level of repetition [70]. Plotting the 10 most frequent emojis, see Figure 4.4, one can observe that first one is “back-hand index pointing down” (👉) emoji with 28,248 occurrences, followed by the “tongue” emoji (👅) appearing 14,919 times. Considering the high co-occurrence of emojis depicting hand gestures, it can be argued that the use of such emoji combinations comprises a novel nonverbal communication pattern adopted by predators to convey to potential victims their requests for sexually inappropriate and suggestive acts, involving the removal of clothes.

4.3 Topical analysis

This section investigates the extent to which grooming behaviours can be modelled mainly using the textual content of chat messages in broadcasts. To this end, a class of probabilistic techniques called “topic models” is considered, comprising a method well suited to studying high-level relationships between text documents.

¹<https://spacy.io/api/lemmatizer>

²<https://unicode.org/emoji/charts-12.0/emoji-ordering.html#clothing>

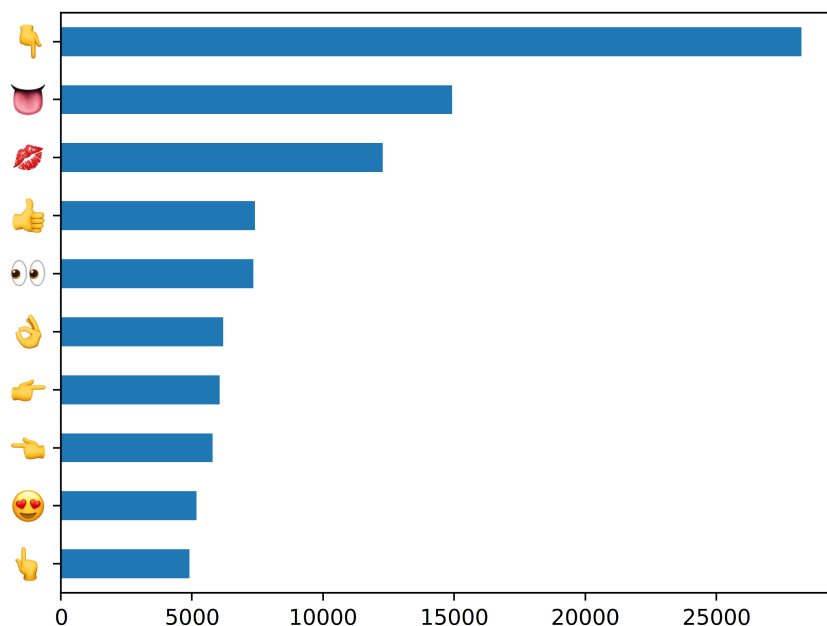


Figure 4.4: Top emoji collocations for clothing related emojis.

Specifically, this work employs Latent Dirichlet Allocation (LDA), a type of generative probabilistic model proposed by [71]. It comprises an endogenous NLP technique, which as highlighted in [72] “involves the use of machine-learning techniques to perform semantic analysis of a corpus by building structures that approximate concepts from a large set of documents” without relying on any external knowledge base. As the name implies, LDA is a latent variable model in which each item in a collection (e.g., each text document in a corpus) is modelled as a finite mixture over an underlying set of topics. Each of these topics is characterised by a distribution over item properties (e.g., words). LDA assumes that these properties are exchangeable (i.e., ordering of words is ignored, as in many other “bag of words” approaches in text modelling), and that the properties of each document are observable (e.g., the words in each document are known). The word distribution for each topic and the topic distribution for each document are unobserved; they are learned from the data.

In this study, all the chat messages sent by users during a broadcast are considered to represent a *document*, similar to the notion of chat log documents; described in [73]. The topics learned from LDA trained on the chat log documents from the dataset could highlight specific terms associated with latent communication patterns emerging within the broadcasts, facilitating the identification of the modus operandi of sexual groomers in the context of

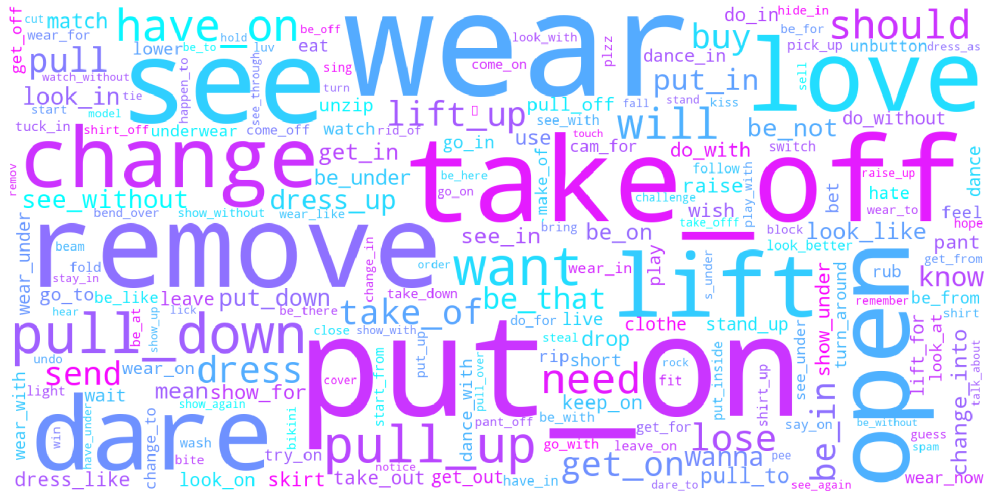


Figure 4.8: Verbs extracted from chats containing clothing terms.

are trained with 1,000 Gibbs sampling iterations and priors $\alpha = 5/k$ and $\beta = 0.01$.

For establishing the optimal k value in terms of topic coherence provided the current setting, the C_v score defined in [78] was adopted. This measure has been proven to have the the most considerable correlation to human interpretability. As such, the $C_v(k)$ metric is computed for each trained model using the implementation provided by Gensim library [79]. According to the C_v metric, it is observed that $k = 20$ is the optimal topic number ($C_v(20) = 0.52$), see Figure 6.6.

Table 6.4 presents the topics learned by the best LDA model, including the most relevant terms describing each topic and the number of chats where each topic is dominant. To obtain the most descriptive terms for topic interpretation, the approach of ranking individual terms within topics presented in [80] was adopted, and parametrized with $\lambda = 1$.

Verb	Count	Verb	Count
wear	6553	put_on	6083
show	5817	take_off	4229
remove	3947	pull_down	1930
see	3765	pull_up	1625
get	3157	have_on	1214
open	3105	take_of	731
like	2914	get_on	728
love	2913	lift_up	690
dare	2844	put_in	507
lift	2159	dress_up	433
change	2154	see_without	371
want	1811	look_in	336
take	1412	put_down	312
go	1267	look_like	295
say	1237	change_into	274

Table 4.1: Top 15 verbs (simple and phrasal) associated with clothing items.

Term	Distance	Count	#Broadcasts	#Users
pusy	0.828122	956	513	432
pus	0.768473	416	305	259
pushy	0.741119	267	185	158
bussy	0.799563	209	128	100
püussy	0.810713	198	133	101
puzzy	0.753680	195	122	113
pûussy	0.781377	184	110	90
pussycat	0.818996	169	138	141
piussy	0.702024	160	141	142
pssy	0.812888	135	103	79

Table 4.2: Top 10 nearest neighbors (cosine distance) of the word “pussy”.

Term	Distance	Count	#Chatrooms	#Users
bobs	0.752709	14728	5720	5754
boos	0.756812	670	490	444
booms	0.759904	638	305	189
boobes	0.868892	578	315	182
bobbs	0.794095	494	341	292
boops	0.803665	452	276	177
boody	0.784702	400	285	190
boobz	0.858590	389	256	161
bobss	0.787802	267	175	113
boobd	0.896997	159	146	150

Table 4.3: Top 10 nearest neighbors (cosine distance) of the word “boobs”.

Term	Count	#Chatrooms	#Users
shirt	36306	19070	16969
shorts	17449	7635	7635
dress	12319	7267	6154
pants	11693	6597	6479
short	10682	6379	6416
clothes	10504	5940	5905
underwear	6055	2490	3022
bottoms	4768	2997	3272
bikini	4621	1928	1993
socks	4563	1855	2581

Table 4.4: Top 10 most frequent clothing terms.

4.3.3 Interpretation and analysis of topics

From Table 6.4, it is evident that the most prevalent topic across all broadcasts, as dictated by the fitted LDA model, is topic #18, dominating the topic mixture proportions in 12,209 chat log documents (19% of the modelled documents). This topic is clearly related to sexual grooming, with key terms including *CLOTH_TERM*, *show*, *open*, *SEX_TERM*, and various other relevant terms previously identified in the grooming behavior analysis (e.g. *remove*,

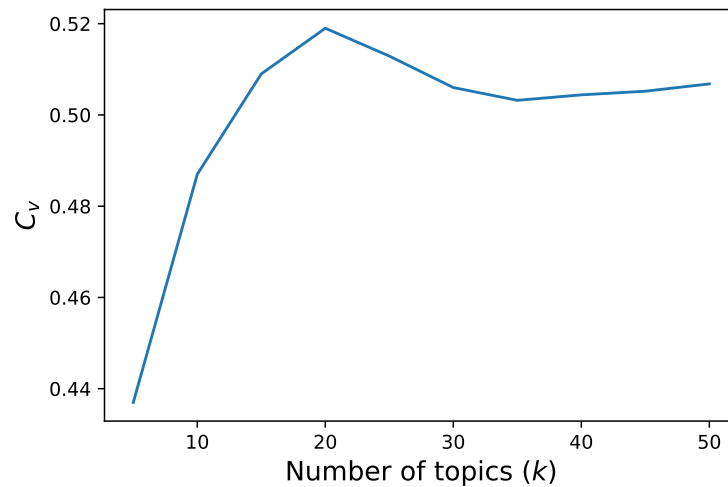


Figure 4.9: C_v metric according to no of topics.

wear, top - which in this case refers to a clothing item, etc.).

The second most dominant topic (#11) reflects flirtatious behaviors, including many endearment terms (e.g. love, nice, pretty, kiss, cute, gorgeous, hot), and words associated with appearance features (e.g. eye, lip, hair, smile, tattoo). Topic #11 is the most representative of 8,477 chat log documents (13% of modeled broadcasts). The rest of the topics describe a wide range of behaviors occurring in the context of live streams, including virtual currency and gifts of LiveMe (i.e. coin, coindrop, castle, diamond, wand), dancing, singing, eating, social media, etc. An interesting observation is the emergence of a topic containing mostly Spanish words (topic #2). It can be speculated that a proportion of the US viewers are using Spanish to communicate within the broadcasts, something not considered in the preprocessing stage. It should be noted that Spanish is the second most spoken language in the US and is widely used in some states. Nonetheless, provided that it dominates only 2,142 chats (3% of modelled broadcasts), it is expected that its impact will be negligible for the rest of the analysis.

Next, the degree to which user interactions other than chatting can be characteristic of grooming is assessed. For this, the interaction and metadata features of the dataset were used. Additionally, the interaction features are normalized by the total number of viewers of each stream, considering them as additional features. Next, the Mean Decrease Impurity (MDI) [81, 82] measure is employed, which is obtained in the process of random forest growing, in order to assess the importance of the described features for discriminating between the

broadcasts where topic #18 is dominant in the topic mixture and the rest.

Table 4.5 reports the ranking of the top five most important features according to the normalized MDI metric. To understand how these features are distributed across the two latent classes of broadcasts, their cumulative distribution functions (CDFs) are presented in Figure 4.10. Notably, the most characterizing feature is the fraction of viewers who started following the broadcaster during the stream (Fig. 4.10a), which in the case of the broadcasts where the grooming topic dominates is much higher than the ones where it does not. Moreover, in Fig. 4.10c it can be observed that only around 6% of the grooming broadcasts have not generated any followers for the broadcaster, while the same is true for 17% of the rest of the broadcasts. This behavior is in line with the findings of [33], where the adult content producers of LiveMe were found to have an exceptionally high number of followers which are characterized by their tendency to follow users who have broadcasted adult content systematically, labeled as *adult content consumers*. A possible explanation could be that in broadcasts where the grooming behavior is prevalent, broadcasters are coerced into performing sexual acts requested by the viewers, as previously outlined. This could justify why the number of new followers they gain in such broadcasts is significantly higher since the viewers might expect that the broadcasters will stream more nude/adult content in the future, and following them is the only way to be notified when they start a new broadcast. Similarly, the fraction of viewers who have liked a broadcast is higher when the grooming behavior is dominant (Fig. 4.10b, which is consistent with the findings of [33] where adult content producers are observed to have received higher amounts of praise than the users found in their ego-networks (i.e. followers and followees). This further exemplifies the predatory behavior of viewers who use likes/praise to coerce broadcasters into inappropriate acts or reward them when they have achieved their objective. Interestingly, for the *Chat messages per user* and *Total chat messages* features which were also found to be important (albeit considerably less impactful in a classification setting), the opposite behavior is observed: In grooming broadcasts users exchange fewer chat messages, both per-user and at the broadcast level.

4.3.4 Topic relatedness

The aim of this last section, is to explore the relatedness of the dominant grooming topic and other topics learned by LDA, which could unveil different aspects of this deviant behavior, beyond the initial analysis. To this end, a frequent itemset mining approach was used to examine the co-occurrence of prevalent topics within the chat log documents. More precisely,

Rank	Feature	MDI
1	New followers to viewers	0.36
2	Likers to viewers	0.16
3	Total new followers	0.10
4	Chat messages per user	0.10
5	Total chat messages	0.05

Table 4.5: Top 5 interaction features relevant for characterizing grooming broadcasts

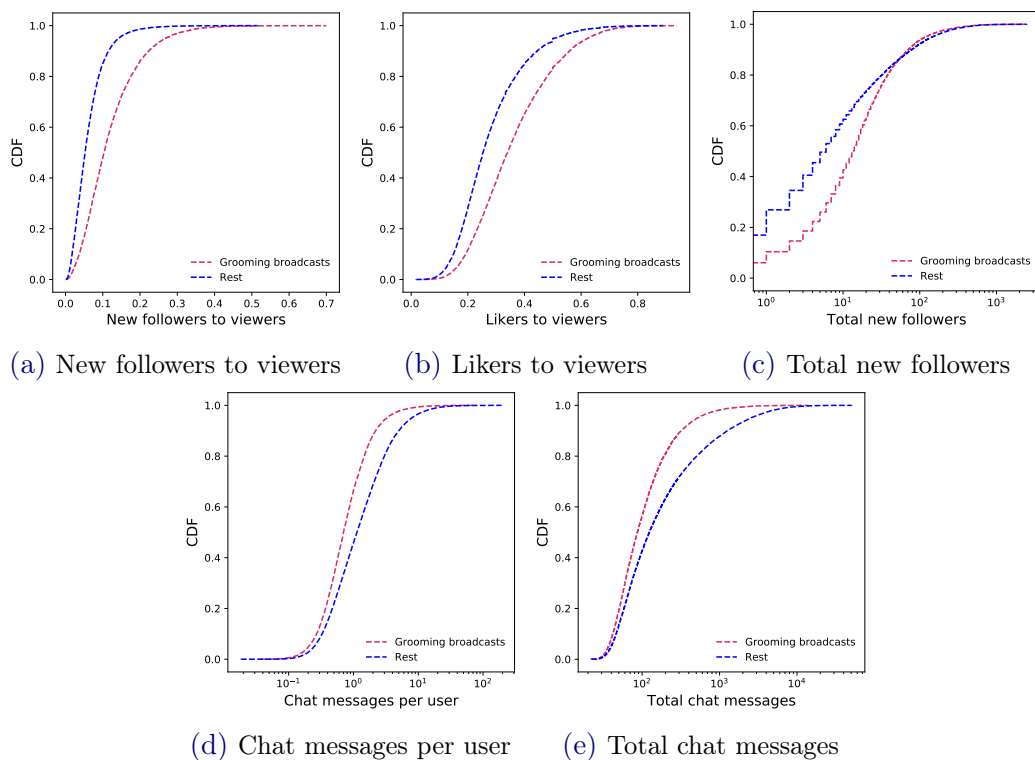


Figure 4.10: Cumulative distribution functions (CDFs) of the features of Table 4.5.

first, the three topics with the highest probability in the mixture assigned to each broadcast were selected. Then, FP-growth algorithm [83] was applied to discover frequent patterns of size two. As expected, the top result includes the two most prevalent topics in the mixture (#11 & #18). The second most frequent pattern includes the grooming topic (#18), and topic #7, which contains terms related to the (self) moderation of broadcasts (i.e. block, report, ban, shut), terms indicating young age (i.e. kid, young, child, girl, boy), terms of hostility (i.e. creep, perv, hater), words bearing negative sentiment according to LIWC (nasty, wrong,

fake, lie). Moreover, the key term that possibly contributes the most to the interpretation of this topic is **police**. Thus, it is expected that this topic is indicative of the criminal dimension of sexual grooming of minors in LiveMe, a large-scale deviant behavior, also attested by popular media [3].

To test this speculation, a portion of chat messages from broadcasts where Topic #7 is dominant was manually examined, and some illustrative examples are presented in Table 4.7. It can be observed that a part of the users expresses their discontent and anger towards the predators/groomers and their harassment targeting minors. The above clearly illustrates the extent of deviant behavior in SLSS, something that beyond the media is also reported by users in, e.g. their feedback for the app. Moreover, the high ranking of this pair indicates that such phenomena, despite the app's moderation mechanisms, are often and known to many users. Finally, the fourth pair (13,18), beyond the common keywords of both topics, shows that some users request further engagement through other platforms, and a primary phase of praise of clothing and body parts, possibly preceding the grooming phase.

Topic	Keywords	#Docs
18	CLOTH_TERM, show, open, SEX_TERM, nice, dare, dance, hot, stand, leg, put, kiss, turn, pull, wear, cam, camera, remove, foot, top, snapchat, gift, girl, rub, low, hand, lift, finger, message, tease	12,209
11	love, nice, pretty, kiss, cute, eye, girl, gorgeous, hot, SEX_TERM, sweet, lip, hair, dear, tattoo, smile, single, dance, friend, number, stand, beauty, lovely, cutie, face, boyfriend	8,477
1	sleep, phone, bed, tired, cool, car, wake, cold, drive, smoke, hour, fall, hear, talk, asleep, stay, high, long, house, game, guess, chill, goodnight, sound, money, iphone, fun, pay	5,731
19	talk, hear, happen, friend, leave, wrong, true, cool, mad, sad, care, sound, hurt, smile, fight, stay, fine, dude, person, funny, break, hard, nice, head, long, boy, army, problem, lose, girl, yep	4,432
7	block, admin, girl, message, leave, report, show, talk, account, kid, creep, young, shut, ban, rude, fake, nasty, send, perv, truth, boy, lie, hater, wrong, police, child, unblock	3,328
16	drink, cat, food, pizza, laugh, eat, funny, face, water, put, chicken, dead, hair, head, challenge, roast, cream, leave, SEX_TERM, apple, taco, chocolate, pet, bob, hand, candy, mouth, cheese, nose	3,180
12	cute, snapchat, send, instagram, rate, clown, dab, hot, number, insta, hair, play, text, friend, pretty, love, put, single, phone, kik, profile, eye, cutie, chat, ghost, boy, girl, fake, girlfriend	3,080
3	send, gift, spam, castle, diamond, share, top, level, broadcast, win, giveaway, broadcaster, wand, number, stream, boat, enter, entry, star, love, feature, join, porsche, coin, awesome, comment, fan	2,953
5	coin, drop, coindrop, follower, send, win, feature, shout, fan, castle, love, wand, dab, that, number, gift, shoutout, giveaway, diamond, stream, iphone, lag, goal, dude, pumpkin, andy, light, level	2,798
4	song, play, sing, love, voice, rap, singing, amazing, nice, dance, awesome, beat, put, hear, panda, listen, cool, singer, sound, closer, jiju, black, job, girl, talent, guitar, boy, heart, hit, drake	2,687
8	love, stream, friend, accent, talk, remember, guess, cool, speak, leave, sleep, skype, long, cute, funny, nice, number, meet, hair, mate, lot, person, dad, class, cat, joke, jenni, kat, join, change	2,682
20	light, turn, gang, love, stay, queen, squad, hit, chill, number, king, slay, fact, that, level, savage, rock, party, dead, boy, mad, play, homie, ight, lot, black, nun, show, petty, dope, top, sun	2,366
2	hola, manni, como, cute, hermosa, eres, show, spanish, amor, SEX_TERM, pretty, bella, donde, hot, bonita, bien, lip, kiss, speak, tienes, espanol, gorgeous, stand, rico, jada	2,142
13	girl, love, cute, play, blue, twin, pretty, red, hot, black, dance, snapchat, green, pink, makeup, hair, lady, white, friend, cool, color, game, face, team, texas, nice, CLOTH_TERM, batman, favorite	1,954
14	kate, love, kid, nice, awesome, cool, tree, country, santa, dad, boy, level, show, broadcast, send, amazing, hear, wolf, talk, lot, son, king, falcon, grim, happen, stream, matt, house, long, rock	1,831
9	beam, love, lag, send, king, cris, stream, castle, fetch, broadcast, show, level, dude, awesome, nick, game, amazing, feature, remember, joey, gift, beam, roll, diamond, join, happen, rip, rackbar	1,663
15	ready, love, spam, feature, stay, game, number, win, boy, tre, read, letter, chat, duck, turtle, gregg, cat, spanme, fun, red, ugh, play, controller, send, coin, hehe, cool, high, comment, gift, party	1,027
17	love, fan, favorite, youtube, shout, meet, dab, channel, pickle, song, shoutout, canada, movie, awesome, fav, twerk, vote, magic, food, subscribe, notice, tattoo, cool, texas, win, vid, hair, ily	908
6	race, love, family, human, unity, amen, put, country, draw, earth, whiskey, broadcast, peace, block, three, lucky, princess, spam, brit, join, general, respect, coin, barbie, send, level, lag, brit	642
10	president, kira, criticize, article, essay, literary, loco, fand, natur, fward, lag, foard, riot, ward, folard, killo, folhrd	14

Table 4.6: Topics

Chat message
A predator is a person who asks kids to undress in front of the camera
And his bio said he likes meeting young girls
Block foot fetish creep
Don't show the creeps anything
Everyone report chat police
Leave her alone creep
Pervs . This kid is like 12
Report that creep too the police
Report the users asking kids to undress; to authorities not LiveMe
Show your kids
So if they didn't ban people for nudity you would show?
This needs to be reported what sort of sick people are ye. She is only 11
YOUR MOM WILL NOW GET A CALL TO KNOW YOU TALK TO 40 years old creeps
You pervs are nasty as f***
block & report nasty stuff
creeps make kids do nasty stuff
he's following lots of young girls
pervs stop asking her to undress
report these pedos to police mate
she is a child stop asking that
she not letting you creeps or sick perv seeing her dress or undress ok
she's a kid perv
they can't ban you if you delete your video after you show
too young this is illegal and wrong lol
try not to undress on stream, it will draw in a lot of creeps
you have creeps who made you do nasty stuff
you look very young .there are lots of pedos on here. be careful

Table 4.7: Illustrative chat messages from broadcasts where Topic #7 is dominant. Key terms of Topic #7 are in bold.

Chapter 5

Inside the realm of “premium” social media accounts

In this chapter, the *FanCentro*¹ platform is analysed through a data-driven study. FanCentro is an online marketplace where users can sell adult content and subscriptions to private accounts in platforms like Snapchat and Instagram. This work aims to explore semi-illicit adult content market layered on the top of popular social media platforms and its offerings, as well as to profile the demographics, activity, and content produced by its users. First, an overview of the FanCentro platform and its user base is provided in Section 5.1. Next, in Section 5.2 the data collection methodology is outlined. Then, the performer profiles are studied in Section 5.3, and in Section 5.4 the purchasable content and the marketplace dynamics are explored. Finally, Section 5.5 presents the various offerings of “premium” social media accounts, their popularity, and the relevant monetisation models.

5.1 The FanCentro Platform

Posting sexualized images in social media is a popular form of self-presentation for young adults [84, 85, 86, 87], and it is outlined as the core tactic to attract followers for a particular type of influencers, which are categorized as “performers” in [11].

¹<https://fancentro.com>

This category of influencers includes adult performers/entertainers, sex workers and models. In all cases, after building an audience in mainstream social media, performers redirect their followers to external outlets for purchasing exclusive content, often pornographic in nature. Notable examples of such outlets are platforms such as OnlyFans² (effectively an ‘adult’ version of Instagram), and “premium” Snapchat accounts, offering a lucrative income stream for performers looking to monetize their online presence [88]. One of these outlets is FanCentro, a platform enabling performers to directly sell private content through a media feed, as well as chatting functionality between performers and their subscribers. As a requirement for opening an account in FanCentro, performers have to provide a digital copy of a government-issued ID for age verification purposes. After this verification step, FanCentro; for a fraction of the paid subscriptions, handles all of the necessary transactions and administrative activities.

There are two main reasons FanCentro was chosen over other similar platforms such as OnlyFans, which have gained wide mainstream media attention [89]. First, its primary focus is selling access to “premium” accounts in social platforms which, strictly, are not content marketplaces (i.e. Snapchat and Instagram). Second, FanCentro website provides a complete listing of performer profiles, enabling the collect data without having to employ sampling techniques. In contrast, OnlyFans platform does not have such functionality.

5.2 Data Collection

To perform this data-driven study, the first step was to collect a *complete* dataset with the profiles of performers registered in FanCentro as of April 5th, 2020. In Figure 5.1, an illustrative example of a performer’s profile page is provided. Note that only performers have public profiles and can post content, while regular users/subscribers can only interact with performers (i.e. follow, message, like/comment to their posts) and not other users. In total, the profile attributes, published content metadata, and offered products for 16,488 users were collected. For this, a crawler was implemented which consumes the API used by FanCentro’s website, enabling us to collect the relevant data. Despite the “public” nature of collected information, this work follows Zimmer’s approach [90]. In this regard, the data remains anonymized during all the steps of the analysis, and only aggregate findings are reported.

To measure the activity in FanCentro in terms of new registrations of performers, Fig-

²<https://onlyfans.com>

5.3 | Characterizing Performers

The screenshot shows a profile page for a performer named 'Serena' on the FanCentro platform. The page includes a navigation bar with 'FANCENTRO' and a search bar. Below the navigation bar are tabs for 'ABOUT', 'FEED (9)', 'TIP ME', and 'CLIPS (4)'. The main content area features a large profile picture of the performer, a bio, and three subscription options: 'Unlock my story' (starting from \$10.83 /mo), 'Follow my' (starting from \$10 /mo), and 'Subscribe to' (starting from \$0 /one-time). Below the main content are sections for 'My Links' (Private Instagram, Amazon Wishlist, My Fanclub, Pornhub, My Stripchat Profile, MyGirlFund) and 'Stats' (Date of birth: 07 Mar 1998, Languages: English, Gender: Female, Breast size: A, Weight: 92 lbs, Height: 5' 1", Hair color: Other). There are also tags for 'petite', 'petite blonde', 'naughty', 'milf', 'amateur', and 'small tits'.

Figure 5.1: An example performer's profile page in FanCentro.

Figure 5.2 plots the number of accounts created each week since the launch of the platform. From January of 2017 (FanCentro launch), the weekly registrations show an increasing trend until a peak was reached in November of 2018. Since then the registration rate has been generally sustained, until we observe a spike in registrations the last week of March 2020, followed by the first week of April 2020, with 196 and 161 new users, respectively. This sharp increase in new users towards the end of March 2020 is also reflected in other similar sites, and it can be linked to the coronavirus pandemic, the consequent lockdowns, and its implications for sex work [91, 92].

5.3 Characterizing Performers

In this section, the collected profiles are studied in terms of characterizing attributes. This includes self-reported demographic information (i.e. sexual identity and orientation, age),

Sexual Orientation	Sexual Identity			Total
	Female	Male	Trans	
Bisexual	2486	52	41	2579
Gay	202	34	5	241
Straight	3544	141	27	3712
Trans	18	4	44	66
Total	6250	231	117	6598

Table 5.1: Sexual identity and orientation

descriptive tags, and external links to other sites, as provided by performers. Table 5.1 reports the number of profiles per sexual identity and orientation. Notably, 9,879 profiles did not include this information. Nevertheless, after analyzing the rest of the profiles, it can be concluded that the majority of performers identify as straight females. Figure 5.3 depicts the age distribution for the profiles containing the birthdate attribute (4,526 profiles). It can be observed that the most common age group is 20-25 years (1,857 profiles), followed by 25-30 (1,347 profiles). The latter means that 70% of the performers who reported their birthday are within the range of 20 to 30 years. The next step of the analysis focused on the tags used by the performers. In this regard, Figure 5.4a shows a WordCloud representation of the most frequent tags used by performers (found in 4,558 profiles). It is evident that they mostly include pornographic terms, with “sexy” and “ass” being the most popular (1,472 and 928 occurrences, respectively). The outcomes of the analysis of the external links are depicted in In Figure 5.4b, where it can be observed that the most common external links from the collected profiles are Instagram and Twitter, closely followed by public Snapchat accounts. This indicates that performers orchestrate their online presence across multiple social out-

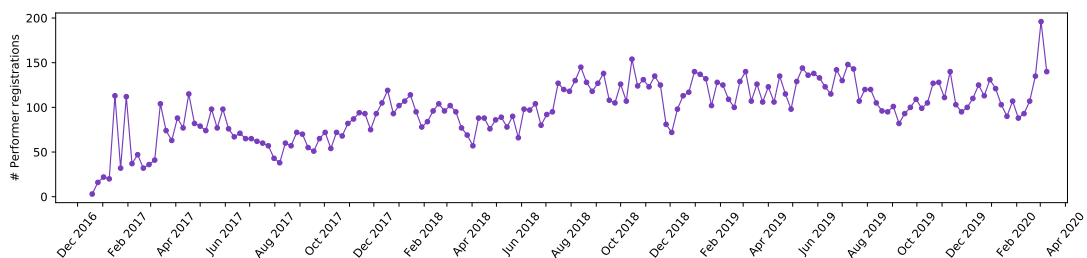


Figure 5.2: Weekly registrations

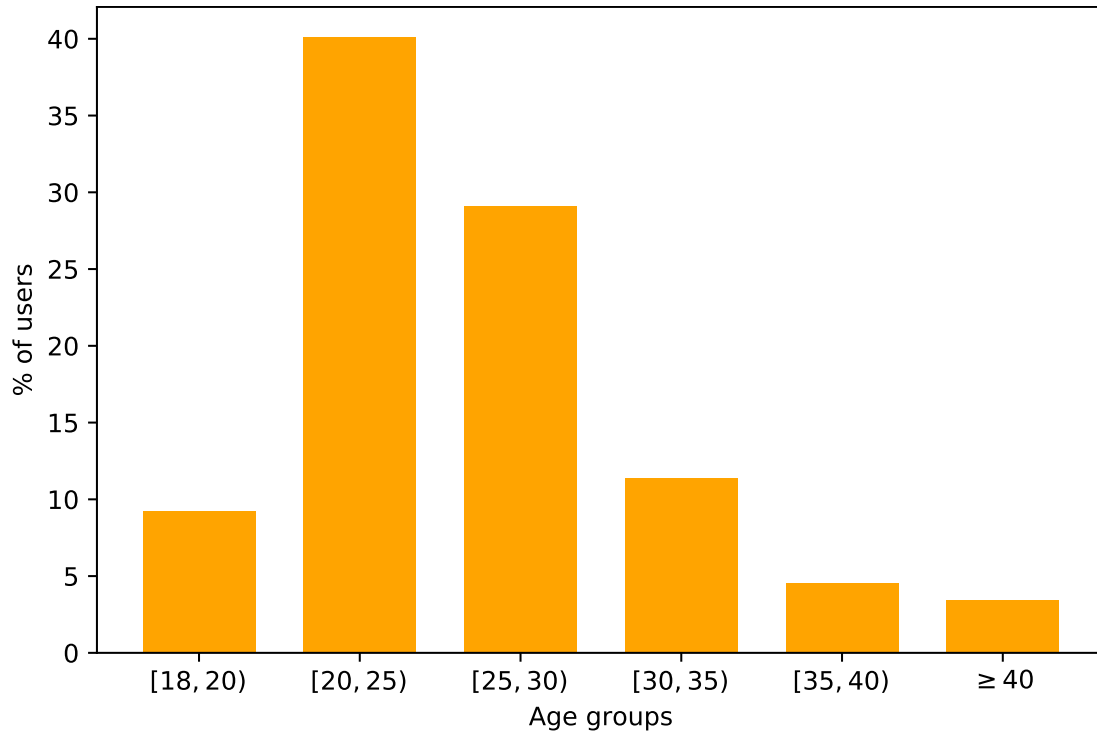
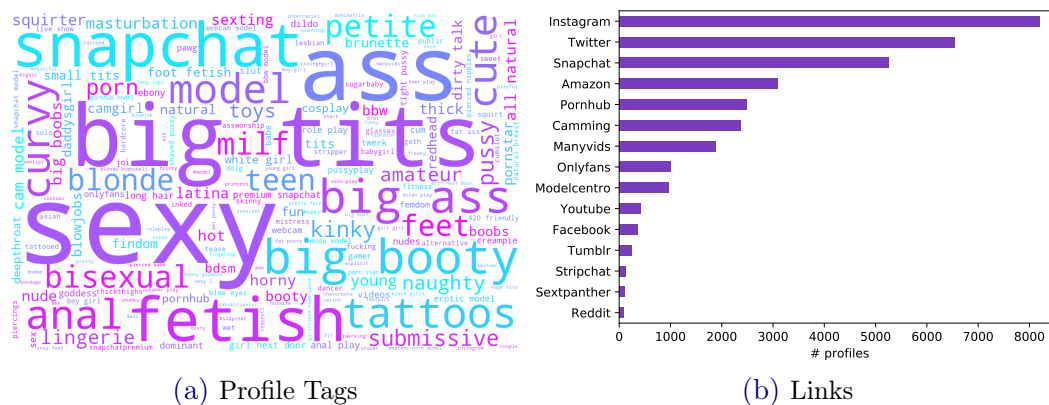


Figure 5.3: Age Distribution

lets, enabling them to reach and engage a diverse audience. Moreover, Amazon wish lists, webcam modeling (“camming”) platforms [93] and porn sites have a relevant representation.



(a) Profile Tags

(b) Links

Figure 5.4: Descriptive characteristics of performers profiles: tags and links to external sites.

5.4 Exploring the supply and demand

To get an insight into the activities performed in FanCentro, the metadata information related to the collected profiles is analyzed, including the amount of funds payable to the performer in the next payout (revenue), the followers and the content posted by performers.

The *revenue* reflects the monetary sum of recurring sales (i.e. subscriptions) at crawl time, plus any income from one-off payments (including gratuity/tips, video clip sales and ‘lifetime access’ services) that are on hold by FanCentro until the next payout to the content creator. FanCentro pays Influencers once a week after two weeks of the revenue generation date, according to their license agreement. Based on the dynamic nature of subscriptions and content produced by performers, revenue is a quantity that fluctuates due to a variety of reasons, including cancellation of subscriptions, chargebacks, external factors governing performers’ popularity, etc. To assess the extent to which the revenue fluctuates over time, this study uses a snapshot of FanCentro profiles that was collected on March 2nd, 2020. To this end, a two-tailed Kolmogorov–Smirnov test was performed, revealing no significant differences in performers’ revenues between two consecutive months ($p = 0.44$). It was found that the revenue distribution is extremely skewed, with the overwhelming majority of the performers (96.4%) generating zero revenue within the aforementioned period. Figure 5.5a plots the revenue cumulative distribution function (CDF) for the 602 revenue-earning performers (3.6% of profiles). It can be observed that 80% are below the minimum payout threshold of 100 USD³, meaning that only a negligible fraction of the performers in the dataset (0.8% approx.) would be certain to receive income by FanCentro during the next payout. Nonetheless, the revenues for the period between 23 March - 5 April 2020 period reach up to 12, 615 USD. In total, the gross earnings of performers amount to 73, 607 USD for the payout period captured in the dataset.

Next, Figure 5.5b shows the CDF of the number of followers. Contrary to the revenue, 78.5% of the profiles have followers. However, the revenue-generating performers have up to two orders of magnitude more followers than the rest. The statistical significance of this difference was also confirmed by a two-tailed Kolmogorov–Smirnov test ($p < 0.01$). In terms of posts, performers in total have uploaded 73, 233 photos, 43, 860 videos, and 4, 867 clips, with the first two being part of their media feeds, while the clips are sold separately. Figure 5.5c shows the CDF of the total number of posts. It is observed that performers earning income

³<https://centroprofits.com/faq>

5.4 | Exploring the supply and demand

have clearly more posts than the ones who do not, however, the majority of performers have less than ten posts (61% and 93% for the revenue and non-revenue generating users, respectively). Again, a two-tailed Kolmogorov–Smirnov test confirms that the difference between the distributions of the number of posts for revenue and non-revenue earning performers is significant ($p < 0.01$). The low number of posts indicates that performers generally prefer to share their content in outlets different than FanCentro.

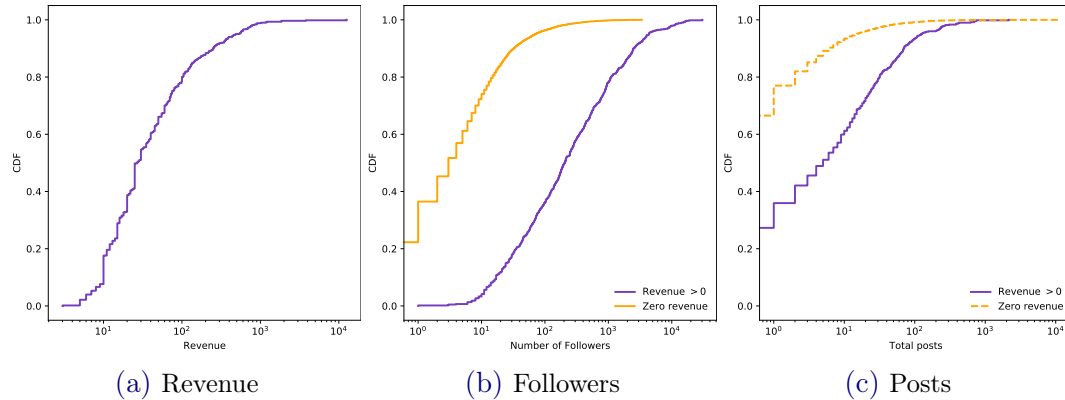


Figure 5.5: Cumulative distribution functions (CDFs) of (non-zero) revenue, number of followers and posts.

5.4.1 FanCentro content

To get a provide a better insight on the content performers upload in FanCentro, their media feeds are analyzed which, in terms of access, can contain two kinds of posts: *private* (only accessible by paying subscribers to their media feed) and *public* (freely accessible). In the collected dataset, the majority (89%) of posts are private (104, 737 posts), while the rest are public (12, 356 posts).

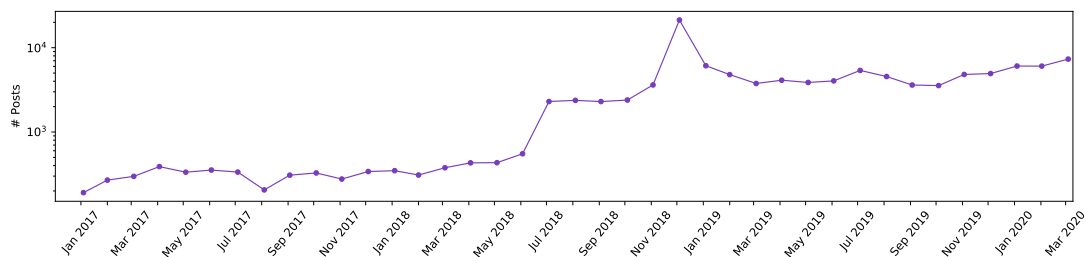


Figure 5.6: Monthly posting activity

Figure 5.6 depicts the number of posts per month. A consistently increasing trend in the

number of posts can be observed, with a spike of 21,300 posts in December 2018, followed by March 2020 (7,325 posts), which is the second most active month in terms of posting activity. Next, the characteristics of performers’ posts are examined in terms of text content (titles) and user reactions, depicted in Figure 5.7. In the dataset, user reactions to performers’ posted content are relatively scarce, with 79% and 92% of the posts receiving zero likes and comments, respectively. This behavior can be observed in Figure 5.7a, which shows the CDFs of the reactions per post.

Notably, the majority of these posts received just one reaction, while the most popular post in the dataset has 316 likes and 55 comments. The low number of reactions in posts comes in contrast with the relatively large numbers of followers that performers attract, as showcased previously. In fact, there exists only a moderate correlation between the number of reactions per post and the total number of a performer’s followers (Spearman’s $\rho = 0.48$). Figure 5.7b presents a WordCloud of the post titles. It is apparent that, apart from terms of endearment and sexual terms, the phrase “subscriber benefits” is prevalent, which could explain our previous observation: to a significant extent, performers might use FanCentro media feed posts as an additional means to promote their premium content in other channels.

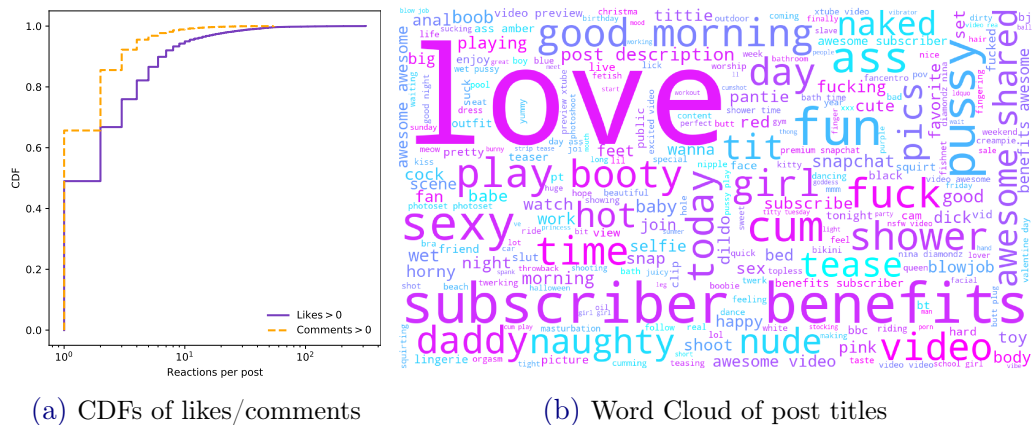


Figure 5.7: Characteristics of posts in terms of text content and reactions (likes, comments).

Finally, the characteristics of the video clips uploaded by performers are studied, which are sold separately. The 4,867 clips in the dataset were produced by 920 performers, 285 (31%) of which had non-zero revenue. A subset of 1,078 clips is categorized as “free for followers”, meaning that the performers’ followers can view these clips for free. This could justify the high numbers of followers that some of the performers attract since this is a characterizing behavior of the consumers of adult content [33]. The clips captured by the dataset have a mean duration of approximately 8 minutes and an average price of 11 USD per clip,

5.5 | Monetization means of premium social media accounts

while clip duration and price are moderately correlated (Spearman's $\rho = 0.45$).

5.5 Monetization means of premium social media accounts

This data-driven analysis is concluded by examining the different payment models for accessing the different channels used by performers to distribute their private content. In FanCentro, the purchasable services include access to “premium” Snapchat and Instagram accounts and the platform’s private media feed. In the collected data three separate payment models were identified for accessing these services: *one-time*, *recurring* and *free trial*. The first two refer to one-off and recurring payments to access new content, respectively, while the “free trial” model allows customers to have a month of free access to the specific service, before reverting to recurring subscription payment. Table 5.2 describes the distribution of the different payment models for the offered services. Private Snapchat is by far the most popular premium service, and the majority of performers prefer offering their services as subscriptions.

Premium Service	Payment model			Total
	Recurring	One-time	Free Trial	
Snapchat	11635	1153	41	12829
FanCentro	4716	0	5	4721
Instagram	1741	191	0	1932
Total	18092	1344	46	19482

Table 5.2: Premium services

The mean price of the performers selling their services under one-off payments is 30 USD for Snapchat and 32 USD for Instagram. To get a deeper insight into the recurring payment model adopted by the majority of performers, Figure 5.8 plots the distribution of subscription offerings, and Figure 5.9 shows the monthly subscription price distribution per service and total subscription duration. For simplicity, only subscription periods with more than 100 occurrences in the dataset were considered. Note that performers can offer their services at discounted rates as a means of promotion (similar to free trial access), which comprise a small fraction of the total offerings (2,004 in total).

In Figure 5.8, it can be observed that the most popular service is the yearly Snapchat subscription, offered by 5,892 performers, followed by monthly Snapchat subscription (3,828 offerings) and yearly access to FanCentro feed (2,921 offerings). While three-month and half-year subscriptions exist, they are not common, accounting only for 25% of total offerings. The subscription fee is calculated on the total subscription period. As such, the monthly price generally decreases as the subscription duration increases. The monthly subscription to performers’ premium accounts, which is the pricier option in all cases, on average costs 21.7 USD and 58 USD for Snapchat (Figure 5.9a) and Instagram (Figure 5.9c) accounts, respectively. Notably, in the first case, the price can go up to 5,000 USD, and in the second case up to 8,000 USD. In this regard, the lowest priced service is access to FanCentro media feed ($\mu = 17.3$ USD), which can cost up to 500 USD monthly (Figure 5.9b). Nevertheless, the most common subscription duration is one year, priced on average 10 USD/month for Snapchat and FanCentro feed, and 14 USD/month for Instagram.

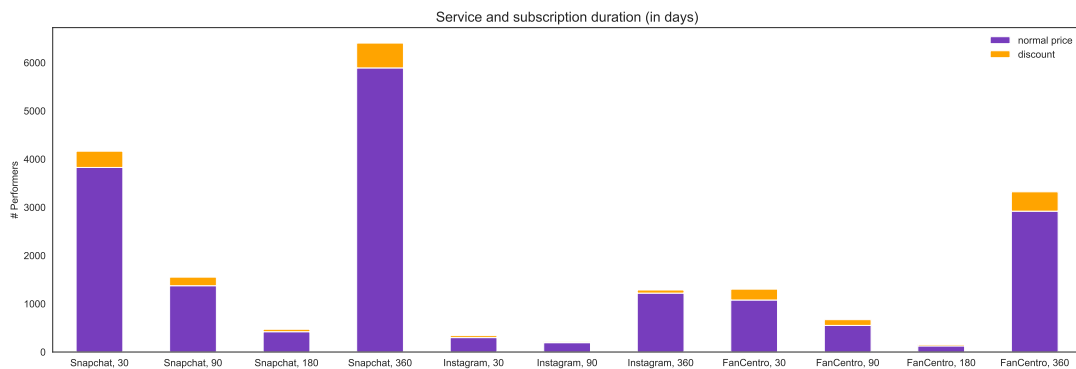


Figure 5.8: Number of offerings per service and subscription duration.

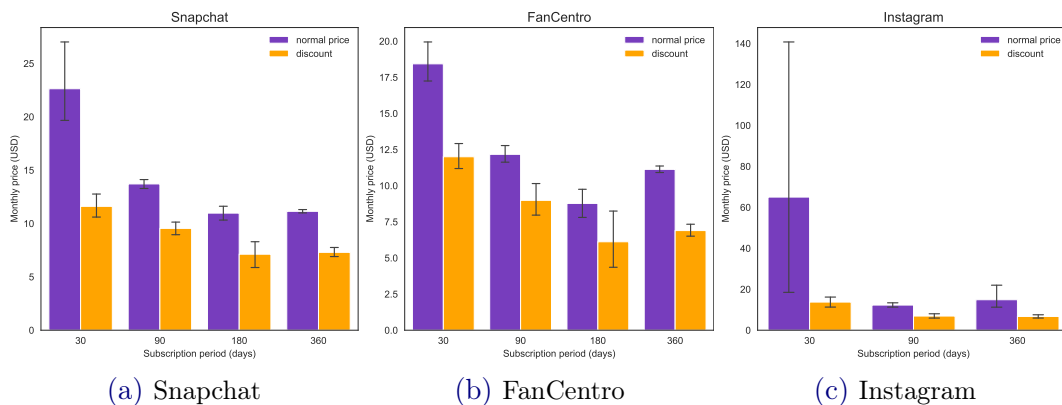


Figure 5.9: Bar plots of monthly subscription price (normal and discounted) per service and subscription duration.

Part III

Malicious Behaviors Beyond Social Media

Chapter 6

Analysis of a cybercriminal marketplace on the Surface Web

The chapter heading the third part of this thesis, is dedicated to the study of *Shopyy*, a marketplace on the surface web used by cybercrime vendors. Normally, such marketplaces reside in the dark web, commonly behind Tor [94], and are referred to as “Darknet markets”. Darknet markets are popular among criminals since they enable them to anonymously trade illegal goods and services extending well beyond stolen data. The latter was discussed by Thomas et al. [95] by pointing out the complex value chain of the underground market economy at scale. These marketplaces comprise the essential pillars of this global-scale cybercrime economy and thus have become the key information source for investigating the cybercriminal ecosystem. An extensive body of literature has explored the darknet marketplaces [96], the involved stakeholders and their communication patterns [97], and their modus operandi [98]. The complexities characterizing the ecosystem of illegal online markets extend beyond the convoluted dynamics in play, to an equally diverse range of offerings, services and products relevant for a variety of illicit topics, such as underground drug economies, data breaches, and cyberwarfare related tools and services. To this end, *Shopyy* provides a unique opportunity to explore the extent of the cybercrime-related services and tools that are being marketed and sold “in plain sight” through a platform operating in the realm of the surface web.

First, an overview of the *Shopyy* platform is provided in Section 6.1. Next, Section 6.2

outlines the approach for collecting data from Shoppo, leveraging information from well-known hacking forums. Then, Section 6.3 explores the collected dataset, and outlines the approach followed for characterizing products and services. Finally, Section 6.4 focuses on the analysis of Shoppo offerings related to cybercrime.

6.1 The case of Shoppo

Shoppo¹ is a shop hosting service that provides the opportunity to individual vendors to sell their products, allows payments in different forms, and a set of APIs to, e.g. advertise one's products in forums etc. In practice, it is commonly used by cybercriminal entrepreneurs to advertise and sell their products and illegal services. As depicted in Figure 6.1, there is strong evidence of this on well-known hacking-related forums such as `blackhatworld`² and `cracked.to`³, which comprise an accessible and “reliable” source for identifying the emergence of novel marketplaces and platforms supporting illegal activities. In this regard, Shoppo was widely used in these forums to monetize some of the reported activities, which additionally were advertised by cybercriminal vendors. A typical store in Shoppo selling illicit products is shown in Figure 6.2.

Therefore, this section of the thesis focuses on describing and implementing a methodology to analyze what types of activities and products were being sold in Shoppo. A crucial difference between Shoppo and the underground marketplaces studied in the literature is that the former does not offer a centralized listing of the sold vendors and products. Each vendor obtains a unique URL where they can host their shop without providing any means for a user to look for similar shops of products offered by different vendors, a common feature in e-commerce platforms [99].

6.2 Data Collection

The decentralized architecture of Shoppo hinders the extraction of knowledge, and thus, the data collection methodology that was developed, specifically aims to discover shops associated with illicit offerings and services, given the context established by focusing on hacking-

¹<https://shoppo.gg/>

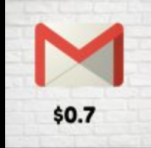
²<https://www.blackhatworld.com/>

³<https://cracked.to/>

Aug 12, 2020 #1

Regular Member Joined: Feb 7, 2019
Messages: 242
Reaction score: 59
Website: shopyy.gg

Selling Gmail Accounts With Instant Delivery.
Account Are Minimum 1 Week Old.
We Removed Phone Number After Registration.
We Add Recovery Email For Every Account.




CLICK HERE TO BUY

Price \$0.7 Each.
Minimum Purchase 11 Accounts.
To Buy More Than 11 Click + Button (Multiple Only)

No Refund/Cancel After You Sent The Payment.
You Have 24 Hours To Report Any Failed Login And We Will Give Replacement.
After 24 Hours Our Transaction Is Fully Done No Report For Bad Login Or Anything.

All Transaction Only On Our Store [Https://Shopyy.Gg/](https://Shopyy.Gg/)
Payment Method: Bitcoin, Litecoin And Ethereum.

(a) A thread in *blackhatworld.com* forum advertising Gmail accounts sold in Shopyy.

MY SHOPPY:  <https://shopyy.gg/>

Chegg Study@Chegg Study Pack :
<https://shopyy.gg/product/>
<https://shopyy.gg/product/>

Making money is my second plan, we need friends. I do my job honestly.

If you want to see special products, check out my market..

LIFESELECTOR TOKENS :<https://shopyy.gg/>

1-Like my posts
2-Eventally Rep me if you like it

Reviews of friends who bought products from shopyy:<https://shopyy.gg/>/feedback

DONATE BTC

(b) A user's signature in *cracked.to* forum, advertising their Shopyy store.

Figure 6.1: Screenshots from hacking forums indicative of the use of Shopyy for selling illicit products.

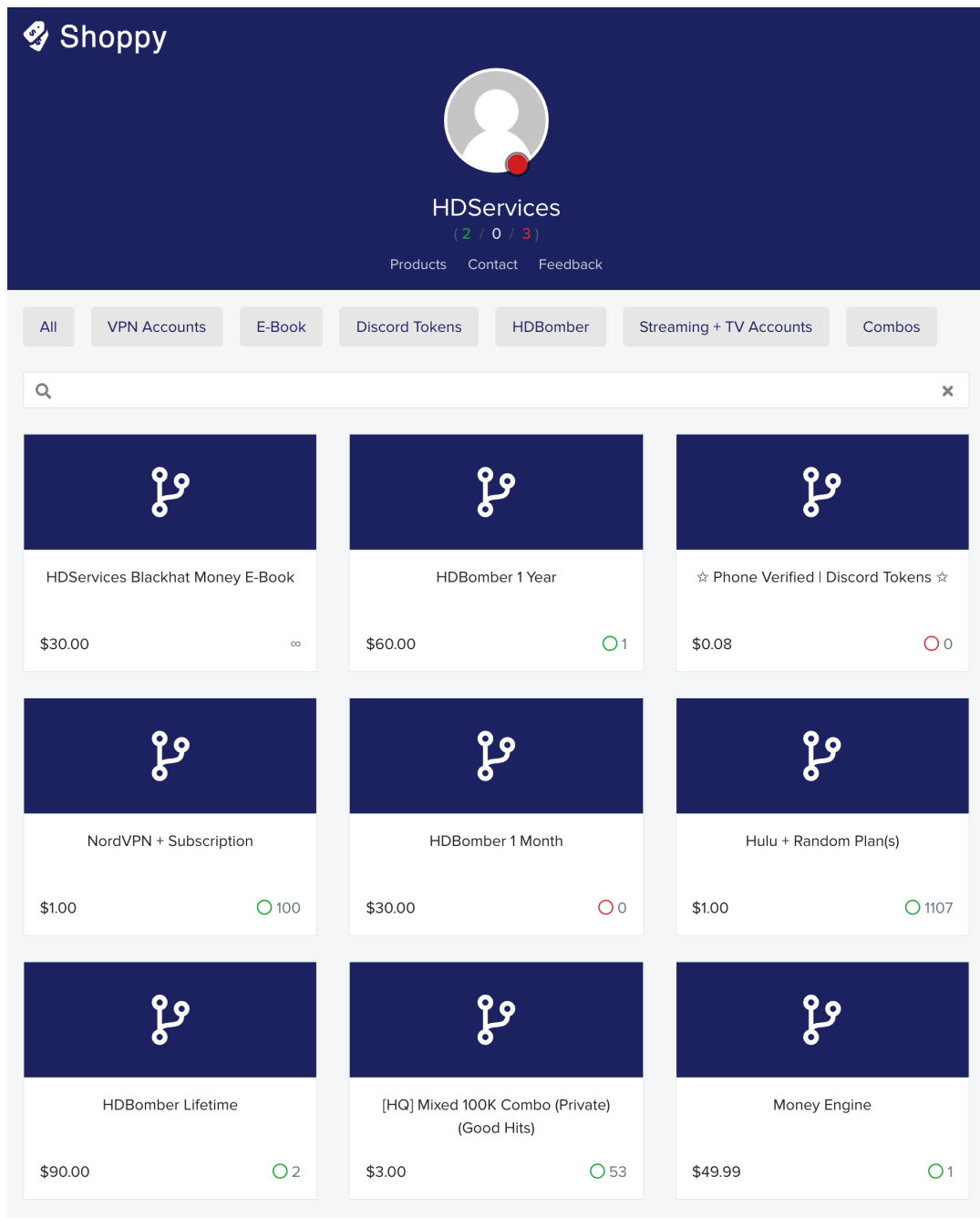


Figure 6.2: A store of a cybercriminal vendor hosted by Shoppy platform.

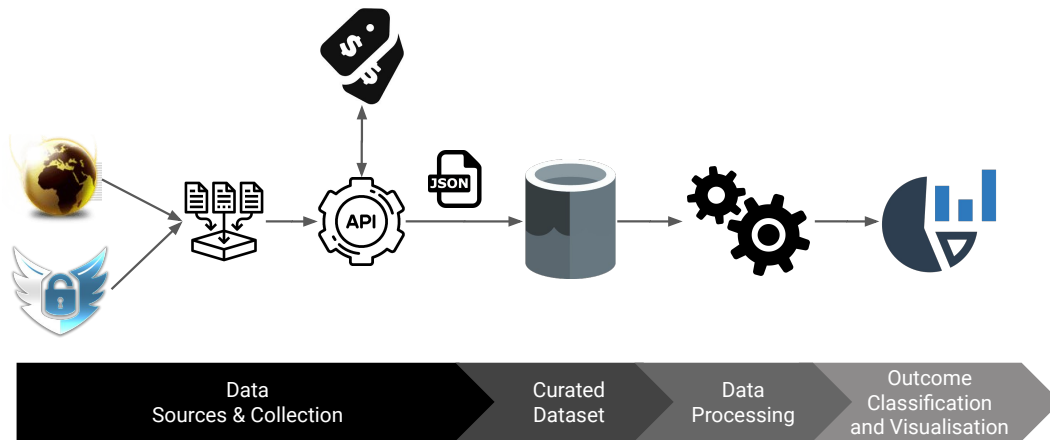


Figure 6.3: Workflow of the *Shoppo* data collection and analysis pipeline implemented.

related forums. The methodology adopted to address this challenge consists of several steps, as depicted in Figure 6.3. First, the blackhatworld and cracked forums were crawled and scrapped, collecting usernames, as well as references to Shoppo accounts in posts and user signatures. Given the size of these two communities, the crawling was focused only to the “Marketplace” forums. To this end, the architecture of the Structure-driven Incremental Forum crawler (SInFo) [100] is adopted, which enabled the efficient data collection and crawling of the aforementioned forums. Nevertheless, one limitation introduced by adopting SInFo crawler is that authenticated user accounts are not supported. Such accounts could potentially allow access to even more content, restricted to authenticated users [101]. Next, the extent to which the collected usernames and Shoppo account data could be correlated with existing shops in the Shoppo ecosystem was examined. The data collection process lasted from March to April of 2020. A total of 68,045 usernames, and Shoppo links from forum post signatures was collected, 2,906 of which were linked to existing Shoppo shops at the time of crawling. The results are summarised in Table 6.1. Notably, a large fraction of the links to Shoppo accounts found in post signatures, that did not resolve to existing shops, indicating that accounts in Shoppo may be banned, deleted, or renamed.

With the collected data, the open Shoppo API was used to retrieve all the information associated with these shops, including products, prices, and their corresponding metadata to create a curated dataset.

Source	#	Valid
blackhatworld - usernames	24,658	827
blackhatworld - signatures	660	359
cracked.to - usernames	41,890	1,230
cracked.to - signatures	837	490
Total (unique)	64,726	2,906

Table 6.1: Collected usernames and *Shopyy* links.

6.3 Shopyy Dataset Exploration

6.3.1 Shopyy in numbers

In this subsection, a quantitative analysis of the collected Shopyy shops and advertised products is provided, as well as highlights on the particular behaviors of vendors. In total, the collected dataset contains 64,726 products advertised by 2,906 vendors. Shopyy provides vendors with the ability to categorize their products as accounts, services or files. The distribution of product categories in the collected dataset is provided in Table 6.2. “Account” is the default category, which evidently dominates the other two by a large margin.

Type	Count
Account	52,850
Service	8,708
File	3,168
Total	64,726

Table 6.2: Shopyy products per category

Figure 6.4 describes the cumulative distribution functions (CDFs) of the number of items per shop (Fig 6.4a) and the product prices, in USD (Figure 6.4b). It is observed that, while around 40% of the shops have less than ten items listed, there are shops with thousands of items. As seen in Figure 6.4b, the price distribution of products is remarkably well described by a lognormal distribution ($\mu = 1.6$, $\sigma = 1.58$), highlighting that the prices of approximately 62% of the products fall within a small range comprised between 1 to 10\$. Moreover, the

median price of ShoppY products is 5 USD, and, as observed in the dataset, the prices can reach up to 10,000 USD.

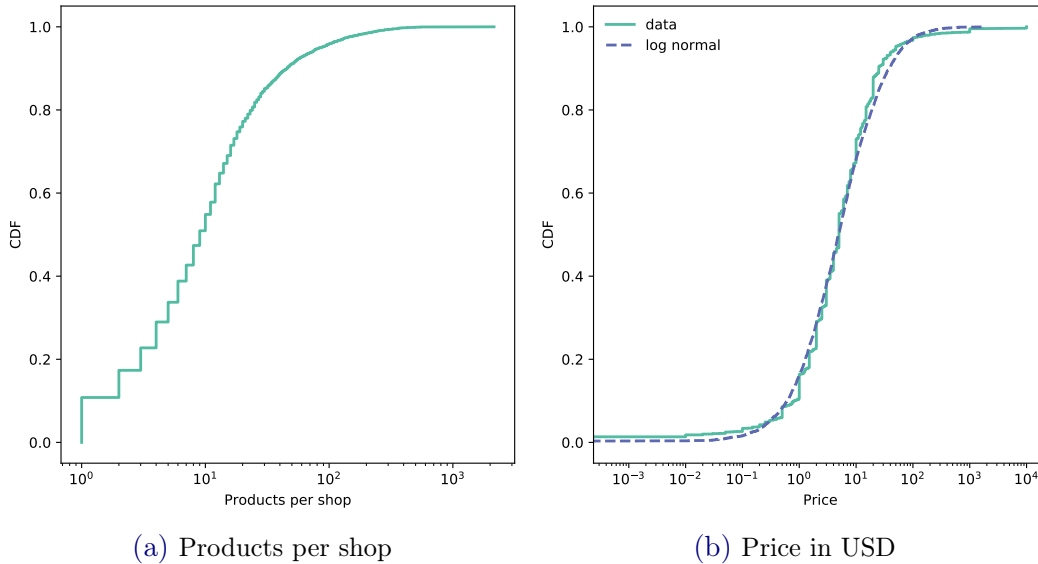


Figure 6.4: CDFs of products and prices offered by *ShoppY* cybercrime vendors.

To get a deeper insight on how different types of products are priced, focusing on possible outliers priced well above the median of 5 USD, the products are binned based on their price, and the fractions of each product type placed in each bin are calculated, as shown in Figure 6.5. Notably, while the lowest price bin is dominated by accounts, the fractions of services per bin follow a consistently increasing trend as the prices increase. In contrast, the relative representation of accounts is inversely proportional to the price, ultimately making services the predominant product category (approx. 70% of total) for the last bin reflecting the highest-priced offerings ($\geq 500\$$). The fractions of the file type products, which as previously shown comprise only a small fraction of the total offerings, are generally sustained, accounting for less than 10% of the products in each bin. It is worth noting that the initial observation related with the use of default categories is reflected in Figure 6.5 showing that, for instance, account products are well represented within all the range of possible prices. The latter behavior seems quite unrealistic in a real and competitive market scenario and is further supported by the experiments described in the following subsection.

The high priced services dominating the upper price bracket were manually examined, and some illustrative examples are provided in Table 6.3. Evidently, these items are false products and rather contain information such as merchants' terms of service, notes regarding

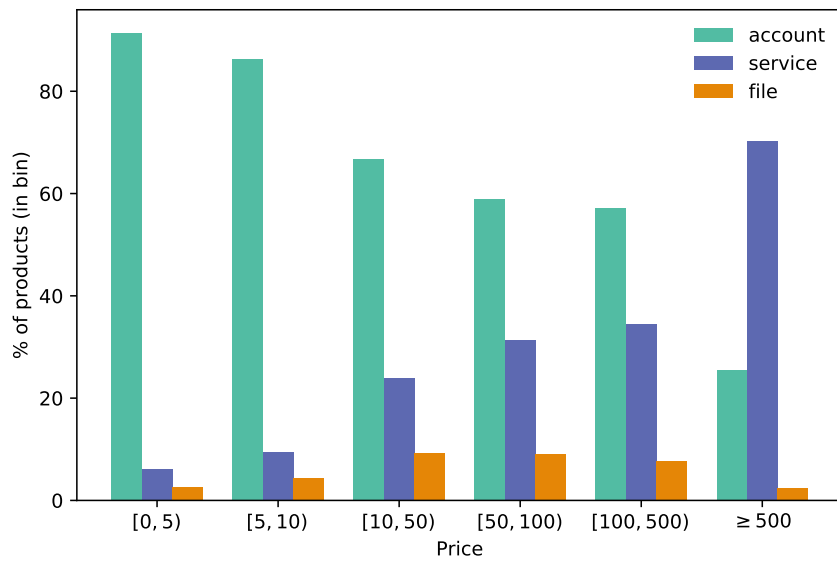


Figure 6.5: Fractional price bins of Shopyy products.

provided shop feedback, support information, and links to Discord servers and Telegram channels maintained by the merchants. This behavior has been highlighted in recent literature by arguing that the unregulated and anonymous nature of platforms such as Telegram and Discord, makes them the perfect habitats for scammers and cybercriminals [102, 103].

Product Title
Terms of Service. READ BEFORE BUYING
Terms of Service & General Information
Discord - DONT BUY
Come Join! Premium Town Our Newest Discord Server
Discord, Telegram, Skype Support Group (Special Discount Codes)
Terms Of Service / Warranty Information
Attention Contact me for support on telegram
Discord & Telegram server Links + Website Link [Click Me]
Discord Server Join For Support IMPORTANT - NEW DISCORD 13.04.2020 banned again

Table 6.3: Some illustrative false “services”, priced ≥ 500 \$.

6.3.2 Characterizing products and services

In this subsection, a topic-based characterization of the offered products is presented, by analyzing their titles. To this end, Latent Dirichlet Allocation (LDA) is leveraged, following an approach similar to the one described in Section 4.3.

In the context of Shopsy, we consider as a document the aggregate titles of the offered products in each of the 2906 shops in the collected dataset. For training LDA models on the generated documents, the LDA implementation provided by Machine Learning for Language Toolkit (MALLET) ⁴ was employed. To obtain the most coherent topic model for our data, the number of topics (k) was iterated within the range from 5 to 50, with a step of 5. The corresponding LDA models were trained with 1,000 Gibbs sampling iterations and priors $\alpha = 5/k$, $\beta = 0.01$. For each trained model, the $C_v(k)$ metric was computed [79]. This metric combines the indirect cosine measure with the normalized point-wise mutual information (PMI) and the boolean sliding window technique, to determine the number of optimal topic classes according to data distribution [104]. According to Figure 6.6, the value yielding the highest C_v corresponds to $C_v(20) = 0.621$ and thus, for the following analysis, the number of topics k to 20.

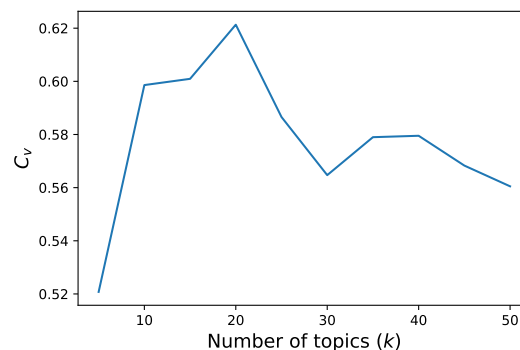


Figure 6.6: C_v metric according to the number of topics.

In Table 6.4, the topics learned by the best LDA model are presented, including the most relevant terms describing each topic and the number of shops where each topic is dominant. To obtain the most descriptive terms for topic interpretation, the approach of ranking individual terms within topics was adopted, described in [80].

To provide insight into the products sold by the shops classified in each topic, Table 6.5

⁴<http://mallet.cs.umass.edu/>

Topic	Key Terms	Documents
4	skin, fortnite, trooper, black, knight, renegade, galaxy, raider, ghoul, recon, expert, ikonik, skull, psn, linkable, aerial, random	466
2	spotify, premium, netflix, nordvpn, vpn, family, hulu, pornhub, upgrade, nord, crunchyroll, grammarly, expire, bulk, uhd, disney	344
9	discord, month, paypal, service, btc, day, buy, week, server, term, read, bot, youtube, twitch, contact, apple, cheap	219
14	method, amazon, free, balance, paypal, store, acc, guide, money, carding, make, google, ebook, check, work, ebay, refund	205
8	access, full, minecraft, nfa, minecon, cape, sfa, hypixel, acces, optifine, semi, rank, roblox, unmigrated, vip, mvp, ufa	197
10	key, lifetime, steam, origin, game, uplay, window, license, edition, office, pro, battlefield, standard, global, microsoft, crypto, fifa	178
5	config, capture, openbullet, checker, fast, cpm, full, ultra, package, proxy, api, onetap, instagram, aimware	153
11	premium, hulu, disney, hbo, live, monthly, pass, plan, yearly, commercial, showtime, ad, nba, gold, directv, starz, tidal	135
13	private, day, combo, pack, email, mail, pass, combolist, usa, fresh, access, valid, hit, list, user, domain, guarantee, database	130
3	point, dominos, balance, wing, wild, buffalo, free, reward, pizza, usa, subscription, jersey, sonic, drive, mike, payment, amc	128
6	account, random, gta, shop, dork, crack, site, move, box, source, red, shopy, mixed, high, target, game	119
1	premium, year, subscription, vpn, auto, lifetime, renewal, adult, security, monthly, avast, pro, membership, kaspersky, renew	96
16	method, follower, include, depop, credit, balance, subway, free, deliveroo, footasylum, pizza, voucher, refundable, wowcher, tesco, attach, guide, disney	77
17	good, order, android, web, ios, pack, iptv, basic, complete, learn, facebook, website, deezer, test, video, theme, virtual, country	77
12	account, random, level, skin, champion, inactive, euw, legend, champs, verify, league, valorant, unverified, lvl, eune, active, unverified, lol, region	73

Table 6.4: Different topic classes and their corresponding key terms, sorted according to the number of documents found.

presents some indicative examples per topic, with respect to the number of topic-relevant terms contained in their titles.

The latter further allows the characterization of each one of the learned topics in a qualitative manner. Topics #1 and #2 describe “premium” accounts for a variety of online services and software products including streaming and VPN services. Topic #3 describes accounts associated with popular restaurants and fast food companies. Topic #4 reflects accounts associated with in-game items and collectibles for the popular online game Fortnite. This topic is

found to be dominant in most shops, in comparison to the other topics, with 466 occurrences (i.e. 16% of all shops). Although selling game accounts can be perceived as an innocuous activity, provided the context of our data collection, these selling activities could be linked with money laundering schemes, based on the idea of converting stolen money to virtual currencies which are used to purchase in-game items [105, 106]. Topic #5 focuses on OpenBullet configurations. OpenBullet is a brute-forcing tool used to perform credential stuffing attacks against online services [107], which are described by configuration files “*configs*”, offering features such as checking multiple credentials simultaneously (advertised by metrics such as CPM, standing for “Checks Per Minute”) and bypassing rate-limiting. Topic #6 contains several classes associated with a broad spectrum of products ranging from game accounts to hacking and reconnaissance tools such as dorks. Topic #7 includes mainly subscriptions to various sports and video streaming services. Topic #8 highlights accounts, hacking tools and in-game items for the popular video game Minecraft. Topic #9 models the false products previously described (cf Table 6.3), containing information regarding vendor’s terms of service and links to external Discord servers, Telegram channels, and etc. Topic #10 includes product licences and keys for a variety of software packages, games and operating systems. Topic #11 describes subscription plans for streaming services, similar to Topic #7. Topic #12 involves accounts for the popular game League of Legends. Topic #13 describes selling leaked user data from security breaches, in the form of *combo lists*, i.e. combinations of usernames/emails and passwords [108], which can be used for compromising accounts with the same credentials in other services, by means of credential stuffing attacks, as seen in Topic #5. Topic #14 involves mostly guides and e-books regarding carding and other methods of financial fraud. Topic #15 contains discount codes and accounts containing redeemable credits for various online shops and e-commerce platforms. Topic #16 is closely related to topics #14 and #15 and includes vouchers for online purchases in various venues as well as methods to perform a fraud or to scam sellers. Topic #17 mainly includes subscriptions for online services and products with a focus on mobile apps. Topic #18 is related to serial numbers for computer peripherals such as monitors, keyboards, etc. Topic #19 provides assorted “*random*” accounts for various social media and sites. Finally, Topic #20 is related to products such as redeemable gift cards, mainly for restaurants and food suppliers.

Chapter 6 | Analysis of a cybercriminal marketplace on the Surface Web

Topic	Sample Products	Topic	Sample Products
1	Avast Premier Premium Security 2 Year 1 Device	11	Hulu Premium Plan - No Commercials, ShowTime, Live TV, HBO, Cinemax, STARZ, Entertainment Add-on.
1	HMA VPN PREMIUM AUTO RENEWAL MONTHLY/YEARLY KEY/ACCOUNT	11	NBA League Pass Premium Monthly Subscription 1 Week Warranty
1	Spotify Premium Account + [Auto-renewal]	11	Disney+ Bundle Monthly Plan Hulu, ESPN+
2	Nord Vpn Premium Expire - 2021	12	EUNE Verfied Inactive (IRON) Level 30 Account Random Champs & Skins
2	Netflix Premium Accounts (UHD)	12	EUW Level 30+ Verfied [Silver Kayle] Account Random Champs & Skins
2	Spotify Family Owner Premium	12	NA Mystery Account Level 30 Inactive verified RANDOM Everything 0-Max Champs 0-900 Skins
3	Jersey Mike's Free Regular Sub - Wrap - Tub USA	13	151k USA Valid Mail Access Combolist HQ Private
3	Dominos 2 Free Pizza (USA)	13	1.1 Million USA Domain Valid Mail Access Combolist Private
3	Buffalo Wild Wings (USA) (1500-2000pts)	13	110GB BRAZZERS USA DATABASE HQ [USER:PASS] COMBO
4	SPECIAL OFFER RENEGADE RAIDER+160SKINS	14	[EBOOK] Amazon Carding Giftcards Pro Method 100% Work
4	Fortnite 1 Renegade Raider + Recon Expert + Black Knight	14	Paypal: Double Your Balance [Method][Guide]
4	Fortnite account with 3 EPIC SKIN skull trooper + ikonik + the ace Warranty 100/100	14	Free .RDP for Paypal Carding Method
5	CONFIG INTERMARCHE FR API FULL CAPTURE ULTRA FAST CPM (socks4/5)	15	HelloFresh \$20 Discount Code (PayPal)
5	Subway Config + Full Capture for OpenBullet (Fast CPM)	15	DISCOUNT CODE 20% - 25% ADIDAS US
5	[OpenBullet] STREAMATE API CONFIG [FULL CAPTURE]	15	Starbucks. ACCOUNT with \$6.50, 200 stars
6	== Depop Account 10k Followers == [HQ SHOP]	16	Deliveroo Refundable- £30.00- £34.99 -48HR (Method Included)
6	GTA V Account (Cr4ck3d)	16	Deliveroo Free Food Method (In Depth)
6	10x Hulu Account Random Subscription	16	FootAsylum - Account Balance - £8+
7	Sling Orange & Blue + Sports Extra + NBA League Pass 6 Months Warranty	17	Scribd Read Books, Audiobooks & Magazines - 1 year warranty [Web/iOS/Android]
7	DAZN USA 1 Year Warranty	17	SkillShare.com Premium - 3 months warranty [Android/iOS/Web]
7	Hulu Premium 1 year warranty (Package: No Commercials)	17	Deezer [Android/iOS/Web] - 1 year warranty
8	Minecraft unmigrated full access account - With optifine cape	18	Logitech PRO Wireless Gaming Mouse (Serial Number)
8	DMC 2.1 - Minecraft Checker / VIP HYPixel, CAPE OPTIFINE, CAPE MINECON, SECURED, INSECURED	18	HP EliteDisplay E223 21.5-inch Monitor Serial Number
8	Hypixel VIP+ Account [Lifetime] Minecraft Non-Full-Access	18	SteelSeries Arctis Pro Wireless Serial
9	Minecraft FA Account (To buy with Paypal contact us on Discord!)	19	PlayStation Account 5-10 Random Games
9	Discord Token Checker [BOT] FREE READ DESC	19	Instagram Random Account
9	Contact me / discord server	19	Twitter Random Account x5
10	Borderlands 3 Standard Edition Epic Games Key	20	Chipotle Gift Card \$10-\$20 [Pin less]
10	Microsoft Office 2019 Pro Plus (1 PC License)	20	Round Table Pizza \$40 Gift card + PIN
10	Windows 8 PRO Digital License Key 32 & 64 Bit	20	Farrelli's Pizza Gift Card 50\$ Giftcard

Table 6.5: Sample products for each topic.

6.4 Products and services related to cybercrime

When examining the Shopsy in terms of identifying illicit activity, of particular interest are the Topics #5 and #13, which as highlighted above, model products related with cybercriminal practices such as selling breached credential dumps and using tools for automating the compromise of accounts in different online services. To provide better insight on such services, this study leverages the term-salience metric defined in [109], which given the set of representative terms per topic, ranks them according to their distinctiveness, i.e. how informative a specific term is for determining the generating topic, versus a randomly-selected term. Subsequently, the top-3 most salient terms for topics #5 (*config*, *openbullet*, *capture*) and #13 (*combo*, *database*, *records*), are selected and used to query product titles, to identify the most prevalent products modeled by these topics. For Topic #13, the term *db* is additionally included, which is a common abbreviation for the term *database*.

As previously reported (Table 6.5), Topic #13 models leaked data from online data breaches, which are sold in the form of username/email and password combinations, along with other personal information. Such listings usually advertise the number of the breached records, as well as the source of the leak. In Table 6.7 some of the largest account dumps found in Shopsy dataset are presented, including their prices. It is observed that popular password breaches checker platforms, such as <https://haveibeenpwned.com>, list the majority of the account database dumps sold on Shopsy. Moreover, this could explain the relatively low price tag for leaks, including up to millions of records, as the respective breaches have already been made public.

In Table 6.6, some illustrative products with titles including at least one of the selected salient terms for Topic #5 are listed. These products represent configurations for software such as OpenBullet [110], BlackBullet [111] and Storm [112]. Such tools can be used to automate credential stuffing attacks [113], versus various online services, as shown from the product titles. Sellers of such “configs” often advertise features such as CPM (checks per minute) and capturing functionality offered, i.e. the ability to capture specific information associated with a compromised account, such as saved credit cards and payment methods, reward points, etc.

From the above, it becomes possible to infer the modus operandi of the account sellers of Shopsy and other cybercriminal markets: A malicious actor is able to purchase massive quantities of breached credentials, and by exploiting the password reuse behavior exhibited

Title	Price
Custom Config for Openbullet	999
New Nectar Card capture Site #2 config for Blackbullet	450
nintendo captchaless open bullet config	250
Spotify Config [OpenBullet]	200
Skout Dating Site OpenBullet config [Anom]	100
Badoo.com OpenBullet config (Fixed on 30.1.2020)	100
Benaughty & Naughtydate Openbullet Configs(2 configs)	100
NINTENDO SWITCH CONFIG [WITH FULL GAME CAPTURE PAYMENT METHOD AND BALANCE]	100
Luminati.io OB config	50
[OPENBULLET] COOP UK CONFIG WITH FULL CAPTURE	50
CONFIG MYCANAL API FULL CAPTURE + CHECK MAIL ACCESS ULTRA FAST CPM	40
PSN Captchaless API (50K CPM) Config Full Capture	40
Btc.com BlackBullet Config with CAPTURE	35
Custom Config (We code your Configs, Web applications, Scrapers, Bruteforcers, Everything related)	20
[OPENBULLET] KrispyKreme Config With Detailed Captures	20
[OPENBULLET] Grubhub Config CAPTURES CC, PAYPAL, AND GC BALANCE	15
NordVPN Config + Expiration Capture	15
[OPENBULLET] Papa John's Config CAPTURES POINTS	15
CodeCademy API Checker +2.6k CPM Capture: Pro	15
Apple Valid Emails Checker By OPENBULLET GROUP	15
[CCShop] streetcc.pw *.loli for OpenBullet [Capture Balance]	15
Facebook Config Capture (check if its ad account or not) Very Fast	15
[OB] Config ICams.com With Capture Balance	10
[OPENBULLET] McDonalds USA Config CAPTURES CC	10
SkinHub +8k CPM Captures: [Balance,RefBalance,Country,TotalWithdrawals,...]	10
[OPENBULLET] Wiki Mining BTC Config + Capture	5
[OB] PicArt With Capture Followers	5
[openbullet] Shipt - [High CPM] Orders + Cards + Rewards Program Capture	5
[STORM] CONFIG NETFLIX + FULL CAPTURE [WORKING] & FAST	3
[STORM] Hotstar Config + Capture (Fast)	2

Table 6.6: Indicative products modeled by Topic #5 in descending price order.

by many users [114], they could compromise user accounts with same credentials in other online services by using credential stuffing tools with different configurations.

6.4 | Products and services related to cybercrime

Title	# Records	Price
Combo List 528M Yahoo.com	528,000,000	400
Combo List 376M Hotmail.com	376,000,000	200
Facebook - 267 Million Records Breach [FULL DB]	267,000,000	500
Combo List 258M Gmail.com	258,000,000	250
Zynga - 213 Million Records	213,000,000	250
Dubsmash Full Database 162 million hashed	162,000,000	120
DubSmash - 162 Million Records (FULL SQL DB)	162,000,000	35
MyFitness Pal - 4.62GB (144 Million Records)	144,000,000	25
Xiaomi - 144 Million Records	144,000,000	25
Canva - 137 Million Records	137,000,000	80
111 MILLION-RECORD PEMIBLANC USA DATABASE COMBO LIST	111,000,000	30
MyHeritage - 92.2 Million Records	92,200,000	75
Houzz - 49 Million Records	49,000,000	300
Facebook DB - 45 Million	45,000,000	265
Chegg - 29 Million Records (Dehashed)	29,000,000	50
Evony - Multiplayer Game : 28.7 Million + 13.8 Million Records	28,700,000	24
Hautelook - 28 Million Records (Full DB)	28,000,000	100
24 Million LUMINATI PROXY DATABASE HQ [EmailPass] Combo	24,000,000	40
YouNow - 18.2 Million Records	18,200,000	50
8tracks - 18 Million Records	18,000,000	41
500PX - Full DB [14.9 million records]	14,900,000	250
Dubsmash 12 million lines Private Combo	12,000,000	25
CouponMom / Armor Games - 11 Million Records	11,000,000	41
Cafepress *NEW* 11 Million Records	11,000,000	41
Bitly - 9.3 Million Records	9,300,000	41
BlankMediaGames - 7.6 Million Records Breach	7,600,000	41
GAMESTOP.COM Database 7.6M UHQ LINES 100K Splits Mail:Pass Good for EVERYTHING	7,600,000	10
StockX - Full 6.8 Million Records DB	6,800,000	65
SNAPCHAT.COM DATABASE LEAK 4,6M LINES PHONE NUMBERS, USERNAMES	6,000,000	2
5.7M Facebook Profiles w/Email	5,700,000	20
Stronghold Kingdoms - 5.1 Million Records	5,100,000	50
5M BITLY DATABASE HQ COMBOLIST (Netflix,Hulu,Spotify,PSN,& More)	5,000,000	20
4.6 Million Snapchat.com Databases Private Combos	4,600,000	20
Game Salad [Dehashed] - 1.8 Million Records	1,800,000	65
PRIVAT HQ 800k Yahoo.com USA Combolist	800,000	20
Hookers.nl (Dutch prostitution forum) - 291K Records	291,000	50
240k Sbcglobal.net Domain HQ private Combolist 100%	240,000	8
225k Icloud.com Domain HQ Private Combolist 100% HQ	225,000	12
211k Cox.net Domain HQ Combolist 100% Private Base	211,000	12
Coinmama dehashed db 209k mail:pass. exclusive	209,000	700
naughtyamerica.com [PORN] Mail:Pass 114K Database	114,000	0
110k Sbcglobal.net Domain HQ private Combolist 100%	110,000	8

Table 6.7: Known breaches sold through Shopsy, as identified by Topic #13's most salient terms.

Chapter 7

Detecting Algorithmically Generated Domains

A crucial technical challenge for cybercriminals is to keep control over the potentially millions of infected devices that build up their botnets, without compromising the robustness of their attacks. This chapter addresses the detection of algorithmically generated domains employed by modern malware to avoid having a single, fixed C2 server which can be trivially detected either by binary or traffic analysis and immediately sink-holed or taken-down by security researchers or law enforcement. As such, Botnets often use Domain Generation Algorithms (DGAs), primarily to evade take-down attempts. DGAs can enlarge the lifespan of a malware campaign, thus potentially enhancing its profitability. They can also contribute to hindering attack accountability. [Figure 7.1](#) illustrates the *modus operandi* of a typical DGA-powered botnet.

To this end, motivated by the continuous evolution of DGAs, the research conducted in the context of this thesis led to the creation of the novel HYDRAS dataset comprising real-world domains produced by DGAs, which is presented in [Section 7.1](#). The dataset consists of more than 95 million domains that belong to 105 unique DGA families. Next, [Section 7.2](#) presents a novel feature set that is designed based on information learned from the analysis of HYDRAS dataset, including lexical and statistical features over the collected DGAs, as well as English gibberish detectors. Finally, in [Section 7.3](#) an evaluation of the proposed features for detecting generated domains is performed, showing promising results.

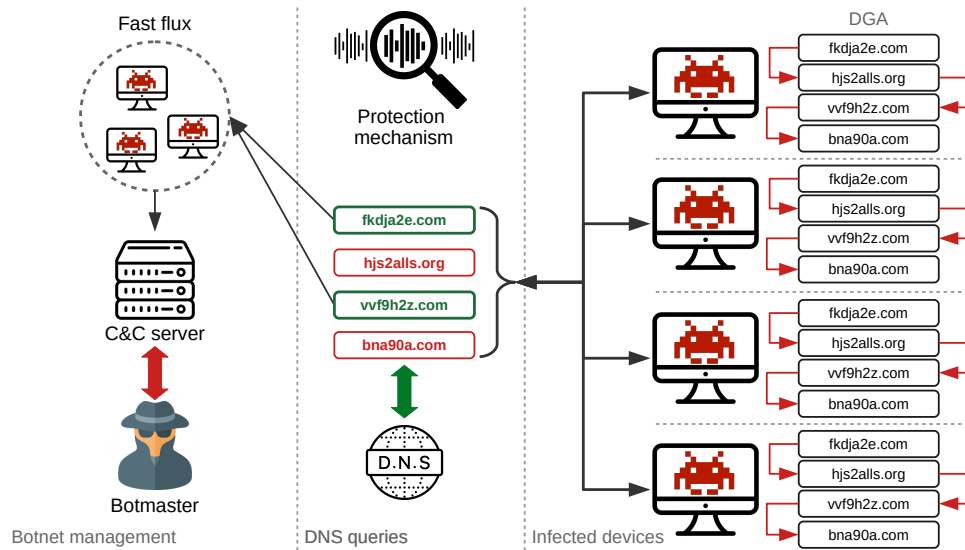


Figure 7.1: The modus operandi of a typical DGA-powered botnet [115].

7.1 The HYDRAS Dataset

In this section, the HYDRAS dataset is introduced, which consists of a collection of benign and algorithmically generated (AGD) domains, both real-world as well as adversarial. The name of the dataset originates from the insightful parallelism suggested by Nadji et al. [27] between DGA-powered botnets and the mythical ancient Greek monster.

Benign domains are sampled from the Alexa 1M dataset, but since the Alexa dataset contains sites, not domains, it had to be preprocessed. First, all top-level domain names (e.g., *.com*, *.org*) are removed from each entry and only the SLDs are kept. Then, the duplicates were pruned since some web pages have multiple entries in the dataset (e.g., *google.com* and *google.co.in*) or been subdomains of identical services (e.g., various blogs of *blogspot.com*). Finally, all internationalised domain names were removed, since they are encoded using Punycode¹ representation. After preprocessing the 1M Alexa dataset, the final dataset contains 915,994 unique domains.

The use of small and unrepresentative datasets, unfortunately, is very frequent in the literature and leads to several biases and other issues that can easily result in wrong analysis and misleading conclusions. For instance, the public feed of DGAs provided by the Network Security Research Lab at 360² as well as the DGArchive [116] provide real-world datasets with millions of samples from many DGA families. Nonetheless, despite the numerous samples

¹<https://tools.ietf.org/html/rfc3492>

²<https://data.netlab.360.com/dga>

Table 7.1: Distribution of records per DGA in our dataset. DGAs in green denote those which were frequently underrepresented, so they were run to create more samples, while purple indicates adversarial ones.

Class	Support	Class	Support	Class	Support	Class	Support
bamital	86892	feodo	247	omexo	41	sisron	2580
banjori	439423	fobber	2000	padcrypt	246096	sphinx	174726
bedep	7814	fobber_v1	298	pandabanker	32484	suppobox	98304
beebone	72	fobber_v2	299	pitou	74314	sutra	3295
bigviktor	999	gameover	22723000	pizd	16384	symmi	65
blackhole	732	geodo	576	post	66000	szribi	20661
bobax/kraken/oderoor	30459	gozi	163529	prosliekfan	218399	tempedreve	13323
ccleaner	12000	goznym	364	pushdo	380427	tinba	72719
chinad	750312	gspy	100	pushdotid	6000	tinynuke	52832
chir	100	hesperbot	16512	pykspace	1996763	tofsee	2100
conficker	2082010	infy	5220	pykspace_v1	44688	torpig	18716
corebot	20931	khaos	10000	pykspace_v2_fake	798	tsifiri	59
cryptolocker	368196	kingminer	252	pykspace_v2_real	198	ud2	491
cryptowall	56624	locky	994381	pykspace2	1248	ud3	20
darkshell	49	madmax	4850	pykspace2s	9960	ud4	70
deception	149854	makloader	256	qadars	630127	vawtrak	17807
deception2	149908	matsnu	40050	qakbot	4579999	vidro	62567
diamondfox	279	mirai	2716	qhost	23	vidrotid	101
dircrypt	11210	modpack	107	qsnatch	1246482	virut	23669176
dmsniff	70	monerodownloader	2995	ramdo	6000	volatilecedar	498
dnschanger	1499578	monerominer	364271	ramnit	150662	wd	32172
dromedan	10000	murofet	13824213	ranbyus	578080	xshellghost	12001
dyre	2046998	murofetweekly	600000	redyms	91	xxhex	1900
ebury	2000	mydoom	2599	rovnix	207996	zloader	29992
ekforward	3649	nekurs	12751075	shifu	2554		
emotet	431048	nymaim	700102	shiotob/urlzone/bebloh	37031	Total	95,325,598
enviserv	500	nymaim2	110511	sinda	24345		

in both these datasets, many malware families are significantly underrepresented. A demonstrative example is the xshellghost family in the 360 dataset which contains only a single sample at the time of writing. Thankfully, the researchers at 360 have reversed engineered the code of this DGA³.

Since the provision of many samples is required to perform an adequate evaluation of any detection technique, the code of poorly represented DGAs was leveraged to enlarge HYDRAS dataset. The dataset was initialised with several public DGA repositories, e.g., J. Bader’s [117], A. Abakumov’s [118], and P. Chaignon’s [119]. In the cases of underrepresented DGA families where the original code of DGA generation was utilized, a few random seeds and/or an extended date range to obtain new samples was used.

³<https://github.com/360netlab/DGA/blob/master/code/xshellghost/dga.py>

In the cases the DGA generation code was used, the added domains have identical characteristics to original ones and might occur in the real-world. Thus, these AGDs could have been collected in a real setting. Moreover, the SLDs of three adversarially designed DGAs, namely `deception`, `deception2` [120] and `khaos` [121] were added.

In summary, the introduced HYDRAS dataset consists of 95,325,598 AGDs belonging to a total of 109 families, from which 105 are unique. It should be noted that a few DGAs are used by multiple families. The families included, along with their corresponding number of collected samples, are reported in [Table 7.1](#). The dataset can be found at [122].

7.2 Proposed Features for AGD detection

The throughout analysis of AGDs and the exploration of ideas behind existing AGD detection approaches conducted in the context of this thesis, lead to some important observations:

- The basic strategy for detecting non-wordlist-based DGAs is to take advantage of the fact that they, in general, make little effort to be human-memorable, as they are typically randomly generated.
- If the generated domains show a high correlation with readable words in terms of vowel/consonant usage, etc., they are expected to contain zero to only a few words having a short length.

7.2.1 Feature Extraction Approach

A general description of the proposed approach is as follows: On receiving a domain name, it is first cached to discover correlations with previous ones. Then, it is attempted to determine whether the SLD matches some specific patterns, e.g., it is a hex value, its combination of vowels/consonants, length, etc. Later on, after removing all digits, the remaining characters are split into words. Within these words, the short ones (e.g., stop words, articles) are pruned, and the remaining terms are examined to determine whether they are real words or just gibberish. Moreover, the entropy of the domain is computed and a subset of the patterns created during the correlation process. All the above led to the design of several features that can be efficiently used to determine whether a domain name is benign or not, without the need for

7.2 | Proposed Features for AGD detection

external information (e.g., WHOIS) or waiting for the DNS resolution revealing whether it is an NXDomain. In this way, a significant number of requests are pruned, regardless of their outcome.

Table 7.2: Features used in the proposed approach and their corresponding description.

Feature Set	Notation	Description
Alphanumeric Sequences	<i>Dom</i>	Domain without TLD
	<i>Dom-D</i>	<i>Dom</i> without digits
	<i>Dom-3G</i>	Set of 3-grams of <i>Dom</i>
	<i>Dom-4G</i>	Set of 4-grams of <i>Dom</i>
	<i>Dom-5G</i>	Set of 5-grams of <i>Dom</i>
	<i>Dom-W</i>	Domain concatenated words
	<i>Dom-WS</i>	Domain concatenated words with spaces
	<i>Dom-WD</i>	<i>Dom-D</i> concatenated words
	<i>Dom-WDS</i>	<i>Dom-D</i> concatenated words with spaces
	<i>Dom-W2</i>	Domain concatenated words of length > 2
	<i>Dom-W3</i>	Domain concatenated words of length > 3
Statistical Attributes	<i>L-HEX</i>	The domain name is represented with hexadecimal characters
	<i>L-LEN</i>	The length of <i>Dom</i>
	<i>L-DIG</i>	The number of digits in <i>Dom</i>
	<i>L-DOT</i>	The number of dots in the raw domain
	<i>L-CON-MAX</i>	The maximum number of consecutive consonants <i>Dom</i>
	<i>L-VOW-MAX</i>	The maximum number of consecutive vowels <i>Dom</i>
	<i>L-W2</i>	Number of words with more than 2 characters in <i>Dom</i>
	<i>L-W3</i>	Number of words with more than 3 characters in <i>Dom</i>
Ratios	<i>R-CON-VOW</i>	Ratio of consonants and vowels of <i>Dom</i>
	<i>R-Dom-3G</i>	Ratio of benign grams in <i>Dom-3G</i>
	<i>R-Dom-4G</i>	Ratio of benign grams in <i>Dom-4G</i>
	<i>R-Dom-5G</i>	Ratio of benign grams in <i>Dom-5G</i>
	<i>R-VOW-3G</i>	Ratio of grams that contain a vowel in <i>Dom-3G</i>
	<i>R-VOW-4G</i>	Ratio of grams that contain a vowel in <i>Dom-4G</i>
	<i>R-VOW-5G</i>	Ratio of grams that contain a vowel in <i>Dom-5G</i>
	<i>R-WS-LEN</i>	<i>Dom-WS</i> divided by <i>L-LEN</i>
	<i>R-WD-LEN</i>	<i>Dom-WD</i> divided by <i>L-LEN</i>
	<i>R-WDS-LEN</i>	<i>Dom-WDS</i> divided by <i>L-LEN</i>
	<i>R-W2-LEN</i>	<i>Dom-W2</i> divided by <i>L-LEN</i>
	<i>R-W2-LEN-D</i>	<i>Dom-W2</i> divided by <i>Dom-D</i>
	<i>R-W3-LEN</i>	<i>Dom-W3</i> divided by <i>L-LEN</i>
	<i>R-W3-LEN-D</i>	<i>Dom-W3</i> divided by <i>Dom-D</i>
Gibberish Probabilities	<i>GIB-1-Dom</i>	Gibberish detector 1 applied to <i>Dom</i>
	<i>GIB-1-Dom-WS</i>	Gibberish detector 1 applied to <i>Dom-WS</i>
	<i>GIB-1-Dom-D</i>	Gibberish detector 1 applied to <i>Dom-D</i>
	<i>GIB-1-Dom-WDS</i>	Gibberish detector 1 applied to <i>Dom-WDS</i>
	<i>GIB-1-Dom-W2</i>	Gibberish detector 1 applied to <i>Dom-W2</i>
	<i>GIB-1-Dom-W3</i>	Gibberish detector 1 applied to <i>Dom-W3</i>
	<i>GIB-2-Dom</i>	Gibberish detector 2 applied to <i>Dom</i>
	<i>GIB-2-Dom-WS</i>	Gibberish detector 2 applied to <i>Dom-WS</i>
	<i>GIB-2-Dom-D</i>	Gibberish detector 2 applied to <i>Dom-D</i>
	<i>GIB-2-Dom-WDS</i>	Gibberish detector 2 applied to <i>Dom-WDS</i>
	<i>GIB-2-Dom-W2</i>	Gibberish detector 2 applied to <i>Dom-W2</i>
	<i>GIB-2-Dom-W3</i>	Gibberish detector 2 applied to <i>Dom-W3</i>
	Entropy	<i>E-Dom</i>
<i>E-Dom-WS</i>		Entropy of <i>Dom-WS</i>
<i>E-Dom-D</i>		Entropy of <i>Dom-D</i>
<i>E-Dom-WDS</i>		Entropy of <i>Dom-WDS</i>
<i>E-Dom-W2</i>		Entropy of <i>Dom-W2</i>
<i>E-Dom-W3</i>		Entropy of <i>Dom-W3</i>

Using the insights from the analysis of DGA families in the dataset, several features were

engineered, defined in Table 7.2. The first set of parameters is computed when trying to identify valid n-grams and words. For the former, an n-gram model is trained with Alexa n-grams and lengths three, four and five. For the latter, the wordninja⁴ word splitter was used, which probabilistically analyses its input using NLP based on the unigram frequencies of the English Wikipedia. Hence, the domain is split into meaningful words, according to a minimum word-length w . Therefore, only terms which contain at least w characters are considered as significant. Then, the percentage of the domain characters which are meaningful is computed, by calculating the ratio γ between characters belonging to words and the domain's total length. Next, two more sets of features are computed according to statistical attributes as well as ratios using the previously calculated features.

7.2.1.1 Gibberish Detection

In addition, a Gibberish detection layer is used, which consists of two methods. The first one is a 2-character Markov chain Gibberish detector⁵, which is trained with English text to determine how often characters appear next to each other. Therefore, a text string is considered valid if it obtains a value above a certain threshold for each pair of characters. The second is a Gibberish classifier.⁶ In this case, the method checks mainly three features of the text: whether (i) the amount of unique characters is within a typical range, (ii) the number of vowels is within a standard range and (iii), the word to char ratio is in a healthy range. Finally, the entropy of a subset of the alphanumeric sequences is computed, to enrich the feature set. An exemplified overview of the feature extraction process is illustrated in Figure 7.2.

7.3 Classification of AGDs

As both empirical and theoretical results have shown that a combination of models (in an ensemble) can increase classification performance [123, 124, 125, 126] even in the case of imbalanced datasets [125], the classification model primarily considered was Random Forest, which is a non-parametric ensemble classifier. Random Forest has previously achieved outstanding performance results in DGA classification tasks [127, 128, 129].

⁴<https://github.com/keredson/wordninja>

⁵<https://github.com/rrenaud/Gibberish-Detector>

⁶<https://github.com/ProgramFOX/GibberishClassifier-Python>

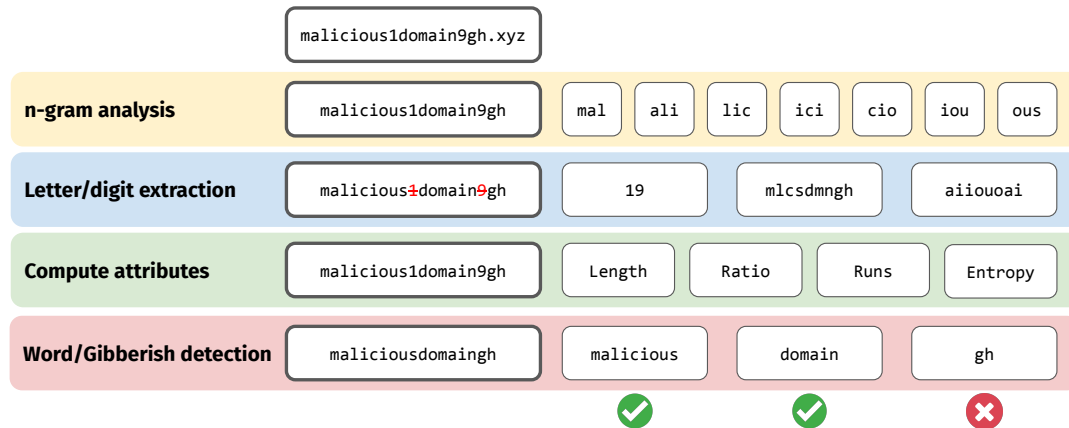


Figure 7.2: Exemplified overview of the feature extraction process.

The hyperparameters of the Random Forest algorithm were tuned with grid search, to maximise classification performance in the task of distinguishing between benign and malicious domains over a subset of our dataset. It was determined that best performance is achieved using an ensemble of 100 decision trees with unlimited depth and bootstrap aggregation (i.e., bagging), where each new tree is fitted from a bootstrap sample of the training data [130]. Additionally, standard 10-fold cross-validation was applied to avoid overfitting and get a roughly unbiased estimate of the performance of the trained models. The performance of the trained classifiers is evaluated using the standard classification metrics of Precision, Recall, F_1 score and the area under the curve (AUC).

This thesis mainly focuses on a binary classification setting (i.e., per DGA detection) using the optimal feature weights per DGA family. In a binary classification, this is a justified setting since feature weights for a particular DGA family are not expected to vary in time. However, this is opposed to multi-class classification (requiring frequent feature/weight tuning), which is not convenient for DGA detection since it deals with the more challenging classification problem. (i.e., intuitively, it is more difficult to find accurate separating hyperplanes among multiple classes than between two classes.)

7.3.1 Binary Classification using the HYDRAS Dataset

For the binary classification setting, several binary Random Forest classifiers that correspond to DGA family detectors were cross-validated – each detector is represented by a single such classifier. To build an input sub-dataset of each DGA family detector, random sampling was

employed without replacement on AGDs from the corresponding family (or benign samples) to fit a 1:1 ratio with the benign domains, resulting in a balanced sub-dataset. To ensure the statistical significance of the results, each cross-validation execution was repeated 100 times (with different randomly selected sample subsets). Note that for each of the 100 runs of cross-validation, different sample sets were randomly selected from the Alexa dataset as benign class representatives. Also, note that due to the particular format in which several families are available in their reverse-engineered form, as well as in the AGD repositories used to initialise our dataset, the $L - DOT$ feature could not be used homogeneously, and thus, was excluded from this experiment.

The averaged outcomes of the binary classification can be seen in [Table 7.3](#). From the results, it is evident that most of the DGA families are classified almost without errors, obtaining precision and recall metrics above 99.9%. Even in the case of families with small representation (e.g., `darkshell`, `omexo`, `qghost`, and `ud3`), the classifier can discern between benign and malicious domains in almost 100% of cases, with only a few exceptions. Moreover, the standard deviation of the F_1 (i.e., σ_{F_1}) achieves values $\sim 1\%$ (in most cases only $\leq 0.01\%$), which showcases the robustness of both the classifier and the proposed feature set.

The lowest accuracy was obtained by `bigviktor` (with a precision of 91.07% and a recall of 76.44%) followed by `suppobox`, `gozi`, `matsnu`, `khaos` and `symmi` with F_1 scores ranging between 95% and 98%. This is due to the fact that most of these families use a composition of English dictionary words to create AGDs, so the extracted lexical features are not always able to properly differentiate them from our benign dataset or are adversarial. In the case of such dictionary-based families, a further enhancement based on probabilistic-based methodologies is an interesting but challenging future research direction. In the case of adversarially designed DGAs, the accuracy obtained is close to the one reported for the dictionary-based DGAs, which showcases the difficulty of capturing such families. A more detailed comparison and analysis of adversarially designed DGAs is later presented in [subsection 7.3.2](#). Overall, the outstanding detection rates showed in [Table 7.3](#) by using the same feature set across such a big dataset proves the robustness and adaptability of the proposed approach. Note that the more divergent families (and samples) are, the more difficult is to select a common set of features which can capture them accurately.

Table 7.3: Performance measures for binary classification (in percentage).

Class	Prec.	Recall	F_1	σ_{F_1}	AUC	Class	Prec.	Recall	F_1	σ_{F_1}	AUC	Class	Prec.	Recall	F_1	σ_{F_1}	AUC
bamital	100	100	100	0	100	gspy	99.67	100	99.83	0.29	100	qakbot	100	99.97	99.98	0.01	99.99
banjori	99.99	99.74	99.87	0.01	99.87	hesperbot	100	99.85	99.92	0.01	99.93	qghost	100	100	100	0	100
bedep	100	100	100	0	100	infy	100	99.97	99.98	0.01	99.98	qsnatch	99.77	99.86	99.81	0.02	99.81
beebone	100	99.07	99.53	0.40	99.54	khaos	99.47	96.47	97.95	0.10	97.98	ramdo	100	100	100	0	100
bigviktor	91.07	76.44	83.11	0.32	87.85	kingminer	100	97.62	98.8	<0.01	98.81	ramnit	100	99.95	99.97	0.02	99.98
blackhole	100	99.86	99.93	<0.01	99.93	locky	100	99.79	99.9	0.04	99.9	ranbyus	100	99.99	100	0	100
bobax/ /kraken	100	99.62	99.81	0.01	99.81	madmax	100	99.98	99.99	<0.01	99.99	redyms	100	99.63	99.82	0.32	99.82
/oderoor																	
ccleaner	100	100	100	0	100	makloader	100	100	100	0	100	rovnix	100	100	100	0	100
chinaad	100	100	100	0	100	matsnu	95.73	97.74	96.72	0.2	96.69	shifu	100	99.63	99.82	0.01	99.82
chir	100	100	100	0	100	mirai	100	99.96	99.98	<0.01	99.98	shiotob/ /urlzone /bebloh	100	99.99	99.99	<0.01	99.99
conficker	99.99	99.64	99.81	0.02	99.82	modpack	100	100	100	0	100	simda	99.99	99.66	99.83	0.01	99.83
corebot	100	99.98	99.99	<0.01	99.99	monerodownloader	100	100	100	0	100	sisron	100	100	100	0	100
cryptolocker	100	99.98	99.99	<0.01	99.99	monerominer	100	100	100	0	100	sphinx	100	100	100	0	100
cryptowall	100	99.87	99.93	0.02	99.94	murofet	100	99.99	100	<0.01	100	suppobox	96.84	98.3	97.57	0.02	97.55
darkshell	100	97.5	98.73	0	98.75	murofetweekly	100	100	100	0	100	sutra	100	99.97	99.98	<0.01	99.98
deception	99.03	97.00	98.00	0.07	98.02	mydoom	100	99.6	99.8	0.01	99.8	symmi	100	93.85	96.83	<0.01	96.92
deception2	98.25	96.15	97.19	0.1	97.22	nekurs	100	99.89	99.95	0.02	99.95	szribi	99.98	99.68	99.83	0.03	99.83
diamondfox	100	97.13	98.55	<0.01	98.57	nymaim	100	99.6	99.8	0.03	99.8	tempedreve	100	99.56	99.78	0.02	99.78
dircrypt	100	99.94	99.97	<0.01	99.97	nymaim2	98.35	97.99	98.17	0.07	98.17	tinba	100	99.92	99.96	0.02	99.96
dmsniff	100	95.71	97.81	<0.01	97.86	omexo	100	100	100	0	100	tinynuke	100	100	100	0	100
dnschanger	100	99.93	99.96	0.02	99.97	padcrypt	100	100	100	0	100	tofsee	99.94	99.92	99.93	0.02	99.95
dromedan	100	100	100	0	100	pandabanker	100	100	100	0	100	torpig	100	99.79	99.89	0.01	99.9
dyre	100	100	100	0	100	pitou	100	99.89	99.94	0.02	99.95	tsifiri	100	100	100	0	100
ebury	100	100	100	0	100	pizd	99.43	99.62	99.52	0.09	99.52	ud2	100	100	100	0	100
ekforward	100	100	100	0	100	post	100	100	100	0	100	ud3	100	100	100	0	100
emotet	100	100	100	0	100	proslifeban	100	99.63	99.81	0.04	99.82	ud4	100	96.19	98.06	0.43	98.1
enviserv	100	100	100	0	100	pushdo	99.94	98.99	99.46	0.02	99.46	vawtrak	99.92	99.44	99.68	0.01	99.68
feodo	100	100	100	0	100	pushdotid	100	99.62	99.81	0	99.81	vidro	100	99.78	99.89	0.03	99.89
fobber	100	99.85	99.92	<0.01	99.93	pykspa	100	99.7	99.85	0.01	99.85	vidrotid	100	96.04	97.98	<0.01	98.02
fobber_v1	100	100	100	0	100	pykspa_v1	100	99.28	99.64	0	99.64	virut	99.97	99.99	99.98	<0.01	99.98
fobber_v2	100	99.78	99.89	0.1	99.89	pykspa_v2_fake	100	99.77	99.89	0.01	99.89	volatilecedar	99.93	100	99.97	0.06	100
gameover	100	100	100	0	100	pykspa_v2_real	100	99.77	99.88	0.01	99.88	wd	100	100	100	0	100
geodo	100	100	100	0	100	pykspa2	100	99.12	99.56	0.06	99.56	xshellghost	100	99.93	99.97	0.01	99.97
gozi	95.28	95.93	95.6	0.11	95.59	pykspa2s	100	97.47	98.72	<0.01	98.74	xxhex	100	99.96	99.98	0.02	99.98
gozonym	100	99.27	99.63	0.16	99.63	qadars	100	99.98	99.99	0.01	99.99	zloader	100	100	100	0	100

7.3.2 Classification of Adversarially Designed AGDs

To further assess the quality of the defined features, they were used to detect three especially “hard to detect” DGAs. These DGAs, *deception*, *deception2* [120], and *khaos* [121] are specially crafted, using machine learning methods, to evade detection. While the proposed features are generic and not targeted towards identifying any particular set of these families, they manage to model well these adversarially designed AGDs, providing significantly better classification performance than in previous works. In detail, the precision achieved by the novel features defined in the context of this thesis, is by 15% to 30% better. Similarly, the recall and F1 score are by more than 10% better in almost all cases. It may also be observed that in some cases, the detection rates slightly vary if the ratio of malicious to benign samples is increased. Nonetheless, the F1 score is at least 92.48%, which indicates that the described

Table 7.4: A detection of unknown DGA families represented by adversarially designed DGAs (leave-one-out experiment).

DGA	Precision	Recall	F1
khaos	100	85.40	92.12
deception	99.99	84.71	91.72
deception2	99.99	73.08	84.44

approach is very effective even when applied against specially crafted DGAs – a challenge that is very close to represent the most challenging scenario.

7.3.3 Detection of Unknown Families

The capability of the defined approach for detecting previously unknown DGA families was assessed using a leave-one-out experiment, with the same configuration as in the binary classification experiments (i.e., 10-fold cross validation) but in this case, the target family was completely hidden to the training phase. Concretely, the aim of this experiment is to predict whether a set of AGDs is benign or malicious without previous knowledge of the DGA family generating them. The outcome of this experiment is reported in Table 7.4. As it can be observed in the table, the described approach is able to correctly classify most of the samples, achieving a slightly lower F_1 -score than the one reported for the binary classification setting (see Table 7.5), due to a general decrease of the recall values. Note that the most affected family is `deception2`, yet the proposed approach still outperforms the original works in which these families were proposed, thus further exemplifying the robustness of novel features presented in this thesis.

Table 7.5: Binary classification against adversarially designed DGAs. First row of each family denotes the reported results in the original work.

DGA	Method	Precision	Recall	F1
khaos	Yun et al. [121]	68.00	98.00	80.30
	Proposed approach - ratio 1:1	99.47	96.47	97.95
	Proposed approach - ratio 1:10	96.08	90.73	93.32
	Proposed approach - ratio 1:100	96.55	89.63	92.96
deception	Spooren et al. [120]	84.40	87.10	85.72
	Proposed approach - ratio 1:1	99.03	97.00	98.00
	Proposed approach - ratio 1:10	96.21	93.86	95.02
	Proposed approach - ratio 1:100	96.12	93.29	94.68
deception2	Spooren et al. [120]	77.50	81.50	79.45
	Proposed approach - ratio 1:1	98.25	96.15	97.19
	Proposed approach - ratio 1:10	94.44	91.29	92.84
	Proposed approach - ratio 1:100	94.56	90.50	92.48

Chapter 8

Emerging threats in Blockchain-based DNS

Blockchain-based DNS alternatives have been receiving an ever-increasing attention in recent years [31]. Such solutions claim to solve many limitations of traditional DNS. However, this does not come without security concerns and issues, as any introduction and adoption of a new technology does - let alone a disruptive one such as blockchain. The last chapter of the current part of the thesis explores a number of associated and emerging threats in the field of blockchain DNS, and attempts to validate some of them through real-world data. Specifically, Section 8.1 explores a part of the blockchain DNS ecosystem. Next, Section 8.2 provides a detailed presentation of the emerging threats and how they could be amplified. Finally, Section 8.3 goes further than speculating future threats, by performing an in-depth cyberthreat-based analysis of Namecoin and Emercoin ecosystems, where it becomes clear that these threats have already given form to tangible risks.

8.1 Blockchain-based DNS

Currently, there are several relevant and widely adopted blockchain-DNS projects. Handshake ¹ is one of the most widely supported technologies, which aims at creating an alter-

¹<https://handshake.org/>

native to existing certificate authorities. Therefore, Handshake aims to replace the root zone file and the DNS name resolution and registration services worldwide. The Ethereum name service (ENS)² uses smart contracts to manage the .eth registrar by means of bids and recently added the support for .onion addresses. Namecoin³ is a cryptocurrency, based on Bitcoin, with additional features such as decentralized name system management, mainly of the .bit domain. It was the first project to provide a solution to Zooko's triangle since their system is secure, decentralized and human-meaningful. Nevertheless, contrary to well-established blockchains like Bitcoin, Namecoin's main drawback is its insufficient computing power which makes it more vulnerable to the 51% attack. Practically, if an adversary manages to get a slight majority of the computing power, they may rewrite the whole chain. Blockstack [131] is a well-known blockchain-based naming and storage system that overcomes the main drawbacks of Namecoin. Blockstack's architecture separates control and data planes, enabling seamless integration with the underlying blockchain. EmerDNS⁴ is a system for decentralized domain names supporting a full range of DNS records. EmerDNS operates under the "DNS" service abbreviation in the Emercoin NVS. Nebulis⁵ is a globally distributed directory that relies on the Ethereum ecosystem and smart contracts to store, update, and resolve domain records. Moreover, Nebulis proposes the use of off-chain storage (i.e. IPFS) as a replacement to HTTP. OpenNIC⁶ deserves a special mention since it is a hybrid approach in which a set of peers manages namespace registration, yet the name resolving task is fully decentralized. OpenNIC provides DNS namespace, and resolution of a set of domains, some of them agreed with Blockchain solutions such as EmerDNS and New Nations⁷, the latter being a TLD provider for nation-states that have not received a country code top-level domain (ccTLD). Moreover, OpenNIC resolvers have recently added access to domains administered by ICANN. In addition to namespace registrar, users can also create their own TLD on request. Table 8.1 summarises the main features of the most relevant Blockchain-DNS systems.

Internet users can reach the TLDs offered by Namecoin, OpenNIC, New Nations, and EmerDNS (e.g. .bit, .coin, .emc, .lib and .bazar) through various browser extensions such as peername, blockchain-DNS and friGate [132]. Their modus operandi is described in Figure 8.1.

²<https://ens.domains>

³<https://www.namecoin.org>

⁴<https://emercoin.com/en/documentation/blockchain-services/emerdns/emerdns-introduction>

⁵<https://www.nebulis.io>

⁶<https://www.opennic.org>

⁷<http://www.new-nations.net>

8.2 | Threats in the context of Blockchain-based DNSs

Table 8.1: Technical characteristics of the most relevant DNS systems. Although Blockstack is blockchain agnostic, it is mainly used with Bitcoin blockchain.

Method	Pedigree Platform	Registrar and Resolution Management	TLD Examples
ICANN	Network of Servers and resolvers	Centralised	.com .net .org
OpenNIC	Decentralised Servers	Hybrid	.bbs .pirate .libre
ENS	Ethereum	Decentralised	.eth .onion
Handshake	Bitcoin	Decentralised	unrestricted
Blockstack	Blockchain agnostic	Decentralised	.id .podcast .helloworld
Emercoin	Bitcoin	Decentralised	.coin .bazar .emc
Namecoin	Bitcoin and Peercoin	Decentralised	.bit

8.2 Threats in the context of Blockchain-based DNSs

As previously stated, blockchain-based DNSs provide a set of characteristics, which are summarised in Table 8.2. In this regard, one could argue that the traditional DNS seems to be outdated, compared to the novel Blockchain DNSs. Nevertheless, traditional DNS proved their reliability and scalability from early 80's until today with modest adjustments. Moreover, blockchain-based DNSs exhibit a set of potential threats and attack vectors that must be considered [32, 133, 134, 135].

In the following sections, the most well-known threats are analyzed, and a few novel ones are identified. Moreover, their possible impact for the system and the final users is discussed. A summary of the emerging threats due to the adoption of blockchain DNSs is depicted in Figure 8.2.

Table 8.2: Main characteristics of blockchain DNSs.

Property	Description
Trust	Verifiable and robust consensus mechanisms
Decentralization	The network is totally distributed with no central entities
Availability	The availability of the network depends on multiple peers and not on a single entity.
Censorship-resistant	Access to information and domain name resolution are not subject to borders or bans
Robustness	Resilient to attacks that affect centralized DNS systems such as MiM, spoofing, cache poisoning, cracking.
Unlimited Resources	A high number of simultaneous users sharing their assets.
Namespace Freedom	Registration of new SLDs and TLDs
Automated Management	Auctions to register domain names, fast and transparent ownership control

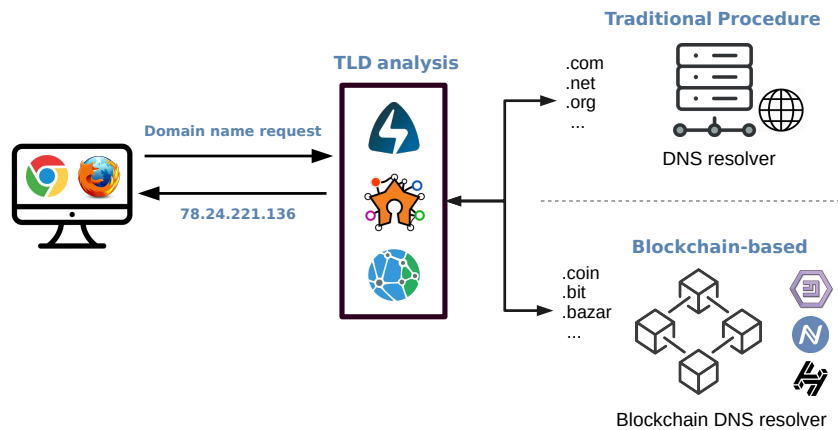


Figure 8.1: Workflow of the browser extensions procedure to enable resolution of EmerDNS, Namecoin, New Nations and OpenNIC domains. The extension analyses the TLD extension of the requested domain and directs the query to the corresponding DNS system.

8.2.1 Malware

In the case of malware, blockchain-based DNSs offer considerable potential. Employing such technologies unlock the capability to register a substantial number of domains with low entropy, which were not available in the market. Currently, as previously stated, malware authors are using DGAs to generate domain names (i.e. AGDs); however, since most short and meaningful domain names are not available, they resorted to the use of long and random-looking domain names. Therefore, a compromised host which uses a DGA to resolve the C2 server issues many Non-Existent Domain (NXDomain) requests which can be analyzed and the attribution to the proper DGA can be made efficiently.

With the use of blockchain-based DNS systems, the conventional NXDomain requests will not be issued, hence hindering the detection mechanisms. Moreover, by using domain names with lower entropy, many filtering and machine learning algorithms are rendered useless. The latter practice is exposed in Section 8.3.

Even more, the use of blockchain-based DNSs implies further issues for malware analysts. When performing static analysis of the malware and its reverse-engineered code, the analyst and the tools that she uses must be aware of the new domains. Traditional filters for domain names will fail, for instance, to reveal calls to .bit domain as the resolution mechanism is completely different. However, requests to traditionally benign domains, e.g. to google.com, may resolve to a completely different IP and the same applies for case sensitive

8.2 | Threats in the context of Blockchain-based DNSs

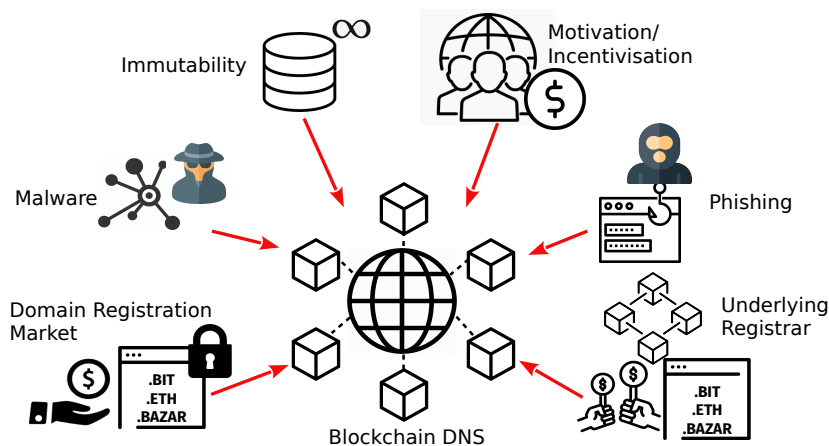


Figure 8.2: An overview of main threats of blockchain DNSs.

domains, e.g. to [GoOgle.com](https://www.google.com), or the use of spaces, e.g. [goog_le.com](https://www.google.com). While Handshake, for instance, may have already taken some precautionary measures for the highly visited domain names, this does not prevent the use of existing domain names with less visibility from being exploited to serve malware. Unfortunately, as discussed in the next section, our data indicate that this attack vector is something that will be used in the near future. Finally, it should be noted that an adversary could easily perform fast fluxing and change the IP addresses that are used whenever deemed necessary. As reported by the Spamhaus Project [136], more than 100 domains registered in blockchain DNS registrars were used by C2 servers in 2018 implying that their use is actively being exploited by cybercriminals.

8.2.2 Underlying registrar mechanism

The primary methodology to register domains in blockchain DNSs is to perform bids or auctions, being the first-request, first-served an outdated strategy. Nevertheless, in the case of a vulnerability in the underlying bid system, users can get control of domains as recently happened [137]. Moreover, most blockchain DNS systems such as Emercoin allow the registration of case sensitive domains, something infeasible in traditional systems. The latter, if paired with some other unrestricted practices such as the use of spaces, non UTF-8 or ASCII characters, may end up with an uncontrollable amount of alternative domain names which are indistinguishable from the legitimate ones. Although this may be a target scenario for malicious actors, this situation may have an impact to the trust and the will of users towards the system. Note that these practices could be prevented and reverted in traditional

DNS, but not in blockchain DNSs. Therefore, careful design of the methodology, as well as a proper implementation of the underlying smart contracts, should be carried out to prevent this kind of behaviour. In the case of systems that offer the use of DNS name resolving services and the registrar of some TLD, a way to prevent this is blacklisting them, although being a controversial strategy. This threat is critical in systems like Handshake and others that may arise, aiming at a full substitution of traditional mechanisms such as ICANN, since legitimate names that are owned by an organization in a conventional ICANN-supported DNS may end up being controlled by malicious users.

In essence, an uncontrolled and fully decentralized DNS type of service may lead to having *parallel* Internets. Note that each blockchain DNS system enables the registration of arbitrary sets of TLDs, which may overlap with existing ones. Therefore, the same domain would resolve to different IPs, depending on the blockchain DNS system used. For instance, even if not used, the domain [google.com](https://www.google.com) is registered in Emercoin in block 252362⁸. This opens a whole new scenario of possibilities in which users can have access to a myriad of contents without restriction, yet in most cases, they could be owned by a malicious entity. The latter problem, as discussed in the following paragraphs, is exacerbated by other properties such as immutability.

8.2.3 Domain registration market

In the least sinister scenario, the case of one registering the domain name of an existing, legitimate webpage is considered. Since the blockchain TLDs are not known to the vast majority of people, it is expected that many people will rush to buy such names requesting a good payment in exchange for the name. As discussed in more detail in Section 8.3, this is not only a hypothesis but a real case. Block 160356 of Emercoin⁹ illustrates such requests were the fees range from \$600 to \$20,000. The existence of ICANN and intermediates, e.g. registrars, allows in many cases the arbitration or even the shutdown or handing over of a domain name; however, the use of blockchain systems does not allow for such mechanisms to be applied. In fact, at the time of writing, one can register a name for an arbitrary amount of time in Emercoin. For instance, there are many domain names in Emercoin which are registered for thousands of years, e.g. there are domains registered up to 5014 and 12012 in

⁸<https://explorer.Emercoin.com/block/252362>

⁹<https://explorer.emercoin.com/block/160358>

blocks 200590 and 380209, respectively¹⁰.

8.2.4 Phishing

Phishing is a fraudulent practice which targets an audience to obtain valuable personal information by using impersonation of entities, persons and more techniques. According to State of the phish 2019 by Proofpoint[138], the number of compromised accounts by these attacks varied from 38% to 65% from 2017 to 2018. This type of cyber-attack leverages socially engineering methods to trick users into performing activities that in some way; most usually monetary, will benefit the attacker [139]. Email is by far the most widely used method to date to perform phishing is email is the most popular avenue for a phishing attack, with more than 90% of successful cyber-attacks/security breaches starting from a spoofed email[140]. In fact, the automated nature of this attack, coupled with the incapacity of users to determine a phishing attack [141] makes the threat even more dangerous. There are many factors that augment this threat and most reside on the human-side aspect of the problem. For instance, the timing of the attack, the authoritarian writing, as well as the exploitation of common practices in an organization, may significantly bias the user into accepting the email as legitimate. Clearly, the use of spoofed or compromised email accounts further complicates the situation.

In the context of blockchain DNSs, the problem is amplified. The users are accustomed to visiting specific web pages and sending emails to particular accounts. If these accounts are pointing to a similar address, e.g. changing the TLD, many users are for sure expected to be tricked. The use of punycodes for phishing or the use of different TLDs can be considered a norm in phishing. With the introduction of blockchain DNSs, an adversary has far more options as there is a wide range of domains that are becoming available at a minimum cost. Practically, this means that not only the phishing sites may have a similar domain name with legitimate ones, but with the use of, e.g. a Let's Encrypt¹¹ certificate, the fraudulent web pages may have valid and trusted HTTPS support. Therefore, the phishing page may have all the distinctive elements, from the UI, the HTTPS support and the valid domain name, making it very difficult for a common user to distinguish the original from the phishing page.

¹⁰<https://explorer.emercoin.com/block/200590> and <https://explorer.emercoin.com/block/380209>

¹¹<https://letsencrypt.org>

8.2.5 Lack of motivation

Motivation under the blockchain DNS paradigm is clearly related to the features offered by such a system, including censorship resistance as one of the main attractions. Nevertheless, these desirable features come at a cost, since decentralized systems totally rely on their network of nodes and their participation. Therefore, keeping the user's interest in blockchain DNSs is critical.

Unarguably, blockchain's adoption in a myriad of scenarios is a reality [29]. Nevertheless, not all blockchain-based projects succeed. In this regard, according to statistics retrieved from Deadcoins¹² there are approximately 1000 dead cryptocurrencies and more than 660 fraudulent cryptocurrency attempts. Interestingly enough, as of 2018, ICO scams have already raised more than 1,000 million dollars [142]. Despite the existence of some awareness campaigns such as HoweyCoin¹³, the lack of a specific and interoperable framework to pursue such deviant behavior enables the persistence of these practices. In the case of blockchain, this may hinder the creation of new projects as well as the persistence of well-known and established ones. One of the main problems that could arise is an unbalanced/unstable computational power, which could compromise the underlying consensus mechanisms and trigger, for instance, a 51% attack. Note that this attack may be applied regardless of the number of users that use a blockchain DNS solution, as the attack is targeted towards the nodes that store the blockchain which, depending on the rewards they have, their participation may decrease over time. The latter may allow an adversary to control the blockchain and compromise its integrity without having to exploit any, e.g. software vulnerability of the system.

8.2.6 Immutability

The immutability property of blockchains, although standing as one of the main beneficial features, may also be abused for malicious purposes. Well-known blockchains such as Bitcoin Satoshi Vision (BSV)¹⁴ and Bitcoin Blockchain have suffered from illegal data storage than cannot be deleted [143, 144]. The lack of verifiable deletion mechanisms also enables DFS systems such as IPFS and IndImm [145] to host and disseminate illegal content [146]. Therefore, neither contents nor domain names are subject to a take-down mechanism. More-

¹²<https://deadcoins.com>

¹³<https://www.howeycoins.com>

¹⁴<https://bitcoinsv.io>

over, strategies such as blacklisting domains are unpractical if the number of domains is high enough.

From a legal perspective, the GDPR does not consider the immutable nature of blockchains and DFS. In this sense, novel decentralized technologies implement features that are not aligned with current regulations and their requirements, which prevents the possibility to apply requests such as the right to be forgotten [147, 148]. Thus, the aforementioned facts make the combination of blockchain DNS and DFS systems a fertile playground for malicious practices. For instance, at the moment of writing, Emercoin supports I2P links; well-known for their anonymity, however, given the continuous rise of IPFS and other DFS solutions, blockchain DNS systems may support IPFS in the near future. The support of a permanent and distributed storage, like IPFS, with blockchain DNS, would actually make a permanent link that cannot be taken down. It is evident that the combination of both would be ideal for the distribution of illegal content as the content would become permanently available for everyone who has access to the link. It should be noted that there are already initiatives that are making this bridge available, not for illegal purposes, e.g. Unstoppable Domains¹⁵.

8.3 Analysis of real-world data

To assess the extent of the aforementioned threats, an analysis of real-world data was conducted. To this end, a dump of the Namecoin and Emercoin blockchains was performed to collect all the domain names and the IPs that have been used by them. Contrary to traditional DNS systems, in blockchain-based DNS all the history of a domain, including the IPs that were used to provide the content is recorded and publicly accessible. Namecoin was the first widely used Blockchain DNS, becoming a reference point for more recent approaches such as Emercoin and Blockstack. This blockchain manages the registrar of the `.bit` TLD through a straightforward procedure, in which a registrant specifies the SLD that they wish to register (which is subsequently appended with the `.bit` TLD), as well as the resolving IP and other secondary parameters. The Emercoin blockchain is one of the most well-known services for domain registration. Surprisingly enough, although the naming requirements of Emercoin specify that only lowercase alphanumeric ASCII characters are allowed, the chain contains case-sensitive domains not only for the advertised TLDs but for traditional TLDs like `.com`. The implemented analysis pipeline is detailed in Figure 8.3.

¹⁵<https://unstoppabledomains.com>

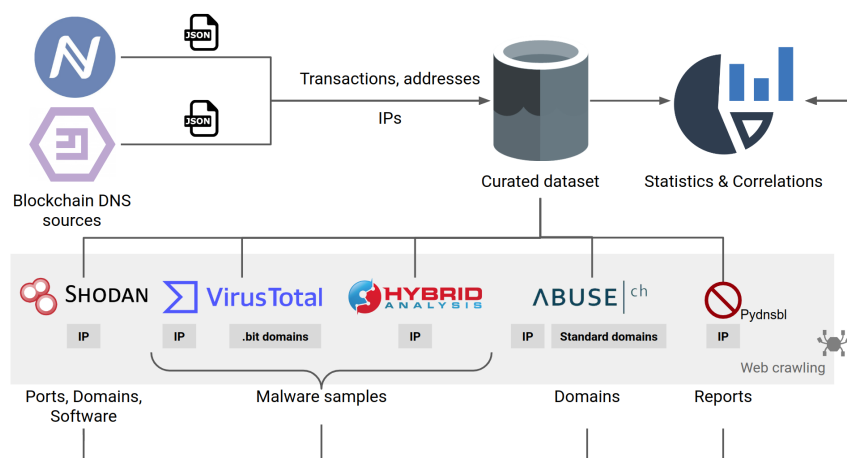


Figure 8.3: Outline of the methodology for analyzing blockchain DNS data.

8.3.1 Data collection and labeling

For the purposes of this research, all the data was collected from the two most widely used chains supporting blockchain DNS, which at the time of writing are Emercoin and Namecoin, in the form of JSON files. From these files, a subset of relevant information was extracted, namely domain names, IPs and emails (by using the *value* field), and the wallets associated to each domain, to create a curated dataset. Based on this, the resulting dataset consists of 5,130 public IPs being used in Namecoin, 919 in Emercoin, and 55 IPs are in both chains. In addition, the dataset contains 2469 Emercoin wallets and 61,357 Namecoin wallets, which are related to these IPs in distinct ways. Finally, the number of domains related to these IPs are 4,452 in the case of Emercoin, and 27,403 in the case of Namecoin. Nonetheless, not all of them are valid domain names. There are multiple domains which do not conform to the DNS format, e.g. they contain non-allowed characters, have registered the same domain with combinations of upper and lower case characters etc. Consequently, the resulting numbers of domains are 2,675 for Emercoin and 27,261 for Namecoin. Then, using VirusTotal the collected domains were queried, of which, only 661 were recorded and 195 were reported malicious. Notably, these malicious domains were associated with 576 unique public IP addresses, implying that almost all of them have been updated several times. For the domains that are reported as malicious, the associated wallets that have registered them, and the IPs that have hosted them are accordingly flagged as malicious.

Next, all the extracted IPs from Namecoin and Emercoin were submitted to VirusTotal,

Hybrid Analysis, and Shodan, and collected the information that each platform has about them. The unique IPs to which domains have been mapped were queried in *VirusTotal* and *Hybrid Analysis* to determine how many of them are linked with malware samples that they have analyzed. Notably, 25.9% of the total IPs are reported malicious in the two platforms as they are correlated with 32,340 unique samples. Moreover, using intelligence from the different sources provided by Abuse¹⁶, some more IPs were identified as being malicious, reaching to 26.18% of the total. Finally, VirusTotal was queried for malicious activity other than malware, e.g. spamming, phishing etc, which resulted to the 34.32% of all IPs to be flagged as malicious.

Moreover, Pydnsbl¹⁷, an aggregator of blacklists of IPs, was used to determine how many of the IPs have been blacklisted. The insights obtained from there pushed the total fraction of malicious IPs to 50.78%, meaning that, effectively, more than half of the total IPs to which domains backed by blockchain DNS are redirecting to or are known to be malicious in some sense.

8.3.2 Representation of malicious activities in Emercoin and Namecoin ecosystems

The next phase of the analysis focused on the identification of possible relationships between the different objects existing in these blockchain systems. More precisely, the correlations between wallets were analyzed. For this purpose, a hop-based association approach was developed, as described in Algorithm 1. More concretely, if a wallet or a domain contains a malicious IP, the IPs associated with such wallet or domain are tagged as *suspicious*. Moreover, the approach considers additional information from the curated dataset to find further relationships between such domains and wallets (e.g. wallets using the same email). In this case, the IP addresses of the additional wallets are added to the suspicious list. Following this approach, it is assumed that if a wallet has used an IP reported in a malicious campaign, the rest of the associated IPs can potentially be used for similar purposes. Note that a suspicious state can only be updated by a malicious one if a specific IP is found to be malicious according to our ground truth, and that suspicious IPs do not spread their status further.

In the case of Emercoin, by applying the proposed hop-based association method to iden-

¹⁶<https://abuse.ch>

¹⁷<https://github.com/dmippolitov/pydnsbl>

Algorithm 1 Hop-based Association

```

1: function ComputeSuspiciousIPs( Dict ip_to_wallet, Dict wallet_to_mail, Dict
   ip_to_domain, List malicious_ips)
2:   Dict status_ips = { };
3:   while (ip in malicious_ips) do
4:     status_ips[ip] = malicious                                ▷ Store {key, value} pair.
5:     wallet_list = GetWallets (ip_to_wallet, ip)                ▷ Wallets associated with
   malicious IP.
6:     domain_list = GetDomains (ip_to_domain, ip)                ▷ Domains associated with
   malicious IP.
7:     associated_ips = GetIPs (wallet_to_mail, wallet_list, domain_list) ▷ Get
   IPs of associated wallets and domains
8:     status_ips = UpdateDict (associated_ips)                    ▷ Update benign IPs with
   suspicious value
9:   end while
10: return status_ips                                             ▷ Dict with classified IPs
11: end function

```

tify suspicious IPs, it was found that only 8% of the IPs did not present any connection with malicious activities. To demonstrate the results of the labeling approach, a relationship graph between the IPs of Emercoin domains was constructed, in which nodes represent IPs, colored by their classification (i.e. **benign**, **malicious** and **suspicious**) provided by the hop-based approach, and the edge connecting two IPs represents a commonly shared interrelation in the form of, e.g. a wallet, an email, a domain or a combination of them. The result is presented in Figure 8.4a).

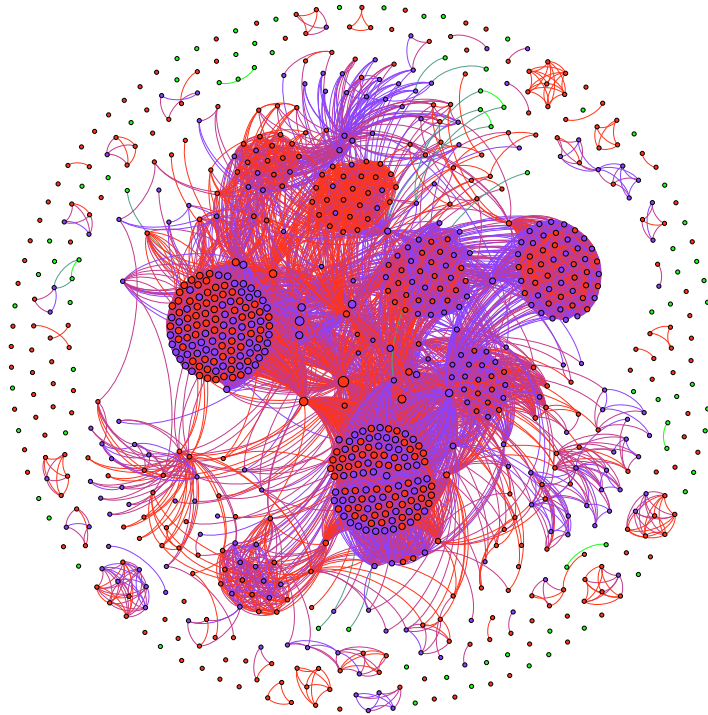
In the case of malicious and suspicious clusters, their connections and all the associations can be clearly identified, showcasing the relevance of the hop-based procedure to find new, potentially malicious groups of IPs. As for benign IPs, it can be observed that they are mostly isolated (cf Figure 8.4a), since they have a very small representation in Figure 8.4b.

Finally, by using Algorithm 1, the set of suspicious IP addresses contained in Namecoin was computed. In this case, in addition to the 2,577 malicious reported IPs, 1118 more were classified as suspicious ones, leaving 1,431 as benign (i.e. only a 28% of the IPs were not connected to maliciously reported IPs). After computing such statistics, the graph representation of the Namecoin ecosystem is depicted in Figure 8.5a. As in the case of Emercoin, nodes represent the IPs, and edges represent a common value (e.g. wallet, email, domain) shared between them. By comparing the representations depicted in Figure 8.5a and Figure 8.5b, it can be observed that the number of benign nodes is substantially reduced in the latter, since

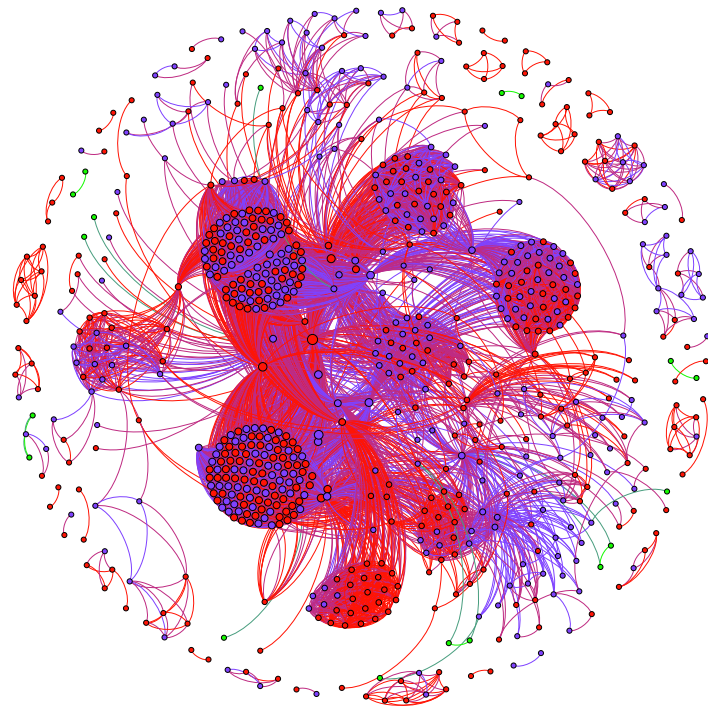
most of them appear to be isolated. In the case of suspicious nodes, they are correlated with malicious ones, exhibiting clearly identifiable clusters. Moreover, there are different sizes of malicious clusters, yet well represented due to the high connectivity between malicious IPs.

For both cases, considering structural factors of the constructed graph representations such as the centrality distributions and average clustering coefficient (explored in depth in [38]), it is clear that the highly connected clusters of suspicious and malicious nodes can be potentially related to a specific malware campaign, orchestrated by one or several users using a closed set of IPs, wallets, emails and domains. Moreover, the nodes of the malicious clusters are highly interconnected between them and, in some cases, to other clusters. Therefore, in some occasions, the same assets (i.e. wallets, emails, IPs, or domains) have been used in more than one campaign, probably triggered by the same entities.

From the above, it becomes evident that Namecoin and Emercoin are currently primarily used for malicious purposes since a huge share of the IPs registered in Emercoin and Namecoin can be directly associated with malicious activities. Such statistics hinder the adoption of blockchain DNS systems and the trust of the community towards them. Therefore, the emergence of novel solutions to overcome the main drawbacks of blockchain DNS is required.

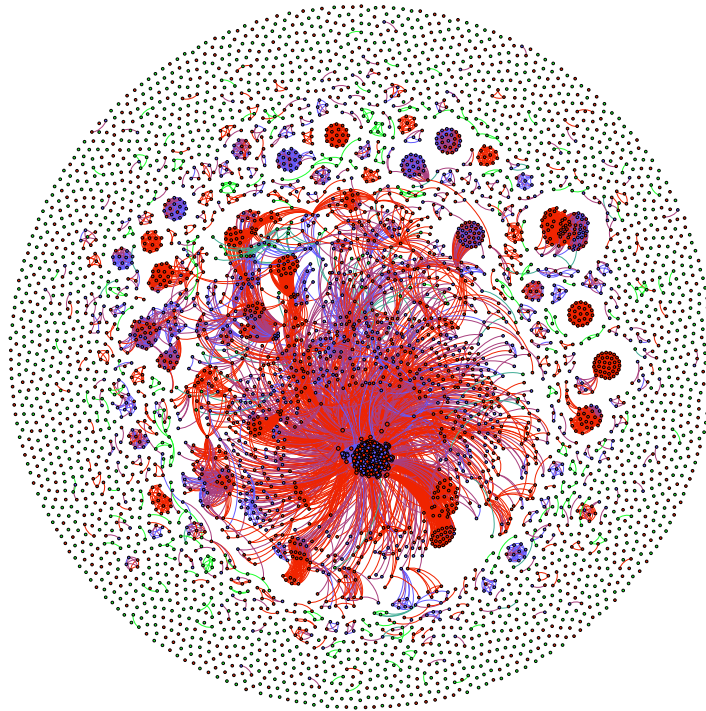


(a) Emercoin representation including all isolated nodes, where each node represents an IP, and their size is weighted according to their connectivity. The edges represent commonly shared data between nodes, such as wallets, emails or domains.

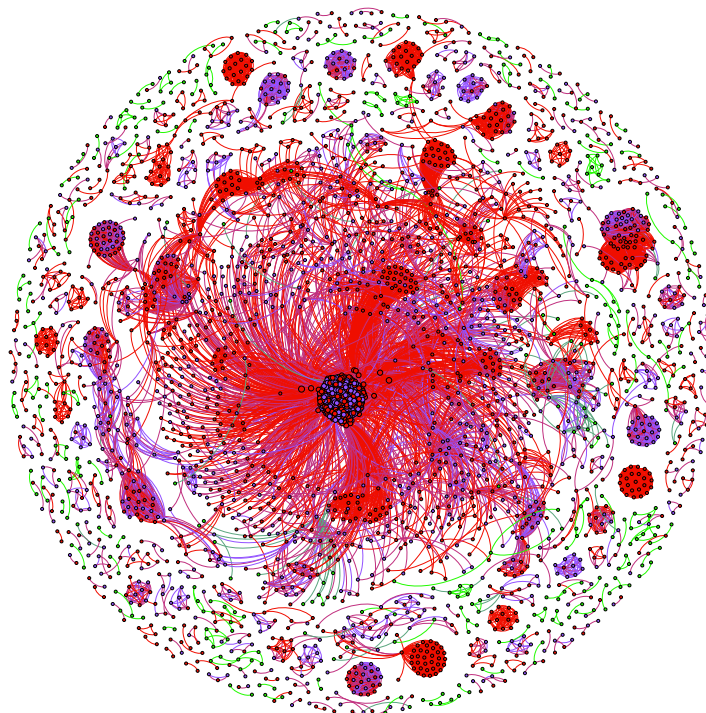


(b) Emercoin graph representation excluding isolated nodes. It can be observed that only a reduced number of benign nodes are present.

Figure 8.4: Graph-based representation of the Emercoin ecosystem.



(a) Namecoin representation including all isolated nodes, where each node represents an IP, and their size is weighted according to their connectivity. The edges represent commonly shared data between nodes, such as wallets, emails or domains.



(b) Namecoin graph representation excluding isolated nodes. It can be observed that the amount of benign nodes is substantially reduced.

Figure 8.5: Graph-based representation of the Namecoin ecosystem.

Part IV

Closure

Chapter 9

Conclusions and Future Work

This final chapter presents a series of conclusions reached after all the research conducted in the course of this dissertation, with the goal of providing useful insight for future researchers working on related topics concerning the study of deviant behaviors online. Moreover, research directions for future work are suggested.

9.1 Conclusions

Throughout the different chapters of this dissertation, different facets of deviant and criminal online behaviors have been explored, in the context of social media and beyond. Chapter 3, demonstrates the prevalence of deviant behaviors in Social Live Streaming Services, and the limitations of current moderation mechanisms in terms of detecting adult content consumption and production. The proposed approach overcomes scalability issues that appear when a large number of humans are needed to report content offending the community guidelines, at the cost of relying on the accuracy of automatic image classification. The findings can lead to speculation that the inefficiency of moderation can be partially attributed to a *voyeur* phenomenon. Many adult content producers are not reported to moderators as the consumers like the content, so their accounts are not suspended, allowing them to continue broadcasting inappropriate content. Moreover, although consuming any kind of content, including adult, is not explicitly prohibited by the community guidelines of these platforms, suspending the ac-

counts of the users who intentionally seek adult content would be meaningful, due to the law of *supply and demand*. It is therefore necessary to incorporate effective, real-time detection mechanisms of deviant behaviors in the existing moderation systems, in order to maintain the SLSS communities safe, especially for the younger audience.

Next, in Chapter 4 the problem of sexual grooming in the context of Social Live Streaming Services is studied, based on the work presented in the previous chapter. Obviously, grooming is not only performed in SLSS, nor it is the only deviant behavior on such platforms. Nonetheless, the different possible user interactions coupled with the live streaming nature, create a novel and less explored field. This chapter provides an in-depth analysis of the chats of thousands of users and identifies characteristics of the grooming behavior in the verbal and non-verbal context. Moreover, the analysis demonstrates how users bypass the word filters that service providers use for blocking offensive behaviors, using adversarial typos and emojis. The work presented in this chapter constitutes a significant contribution towards understanding predatory behaviors on social networks. The latter should be considered in the light of the role that social networks have in our daily lives and the potentials that the emergence of SLSS have. The findings imply that additional risks and dangers (especially for the younger audience) exist due to the inefficiency of moderation mechanisms. Therefore, further measures must be taken to secure the content of what is broadcast, from whom, and to whom. Undoubtedly, due to the size and rate of exchanged information, moderation mechanisms may be difficult to be performed in real-time. However, this chapter illustrates how grooming and other predatory behaviors can be detected effectively without resolving to the use of multimedia which require heavy processing.

Given the scale of the demand for adult content combined with the inefficacy of moderation mechanisms, deviant behaviors involving the production and consumption of adult content have spread far beyond SLSS, into the territory of mainstream social media. To this end, Chapter 5 presented the first quantitative analysis of the semi-illicit adult content market layered on the top of popular social media platforms like Snapchat and Instagram. Specifically, the demographics and activity of the selling users in FanCentro are analyzed. The existence of sites like FanCentro where influencers can openly sell and promote “premium” social media accounts, indicates that the industry built on the inefficacy of social media platforms to enforce community guidelines for effectively banning adult content is here to stay. This inefficacy is exploited and monetized in large scale, exacerbated by the fact the explicit content is staying “hidden” in private accounts, access to which is sold through the different models studied. Moreover, the findings indicate that the coronavirus-induced lockdowns

have accelerated the growth of such markets. This phenomenon is also reflected by the rise of other influencer-centric adult content markets, such as OnlyFans, which observed a major increase in traffic during the coronavirus pandemic [91, 92]. In part, this is due to the fact that a large number of sex workers lost their original revenue streams because of the virus; in addition, an increasing number of influencers transition to online sex work as a means to adapt to the economic downturn which caused companies to reduce marketing budgets, that would have been otherwise used for sponsored content [89]. The strong online presence of Performers across multiple popular social media sites where they openly promote their paid content signals the shift of the online adult content industry towards an increasingly mainstream, gig economy. Nonetheless, the proliferation of adult content flowing unobstructed through social media, diffused and being promoted via users with large followings, might pose a serious threat to the safety of mainstream online communities, especially for the younger users.

After exploring the aforementioned deviant behaviors, the present work shifts its focus to other kinds of malicious and criminal behaviors beyond the realm of social media.

The research presented in Chapter 6 showed that most of the activities that are leveraged in the dark web are also taking place on the Surface Web and yet, no effective mechanisms or take-down measures are taking place. This claim is supported by the throughout analysis of Shoppy, a marketplace used for selling illicit digital products and services, including breached data/credentials, hacking tools, etc. The findings of the study evince the cyber-criminal nature of a myriad of shops and users in the Shoppy ecosystem, and their strong links with popular hacking communities. Moreover, it is highlighted that while the proposed analysis provides clear evidence of the illicit activities taking place in plain sight, malicious actors' antics are resilient to take-down attempts. This is due to the fact that they only use such platforms as a contact point, redirecting all of their activities to other external channels such as Telegram or Discord. Consequently, it becomes clear that there is pressing need for robust investigation protocols and more support from law enforcement towards the prosecution of such activities.

Of course, the mechanics of the global cybercrime economy are far more complex than selling illicit services and products. One of the staples of the underground economy is the deployment of botnets, capable of infecting millions of devices through hard-to-detect and resilient malware campaigns. To this end, the work presented in Chapter 7 aims to contribute towards enabling faster and more accurate botnet detection, and to speed-up take-down operations via a novel DGA detection approach using machine learning. In this regard, a dataset with more than 95 million AGDs is constructed and shared, providing the extracted features

to the community, allowing future research on DGA detection to benefit from a significantly richer baseline dataset than the ones already existing in the literature, both in terms of number of families and samples. Using this dataset, the proposed approach leverages a novel set of comprehensive features capable to outperform the current state-of-the-art methods in DGA detection, by achieving an almost optimal detection rate in the binary classification problem for the broadest possible set of DGA families. Moreover, the proposed approach was able to detect adversarially designed DGA families, including cases where the employed classification models were not trained to detect such families (i.e. assuming no previous knowledge).

Finally, the present thesis concludes with an analysis of the threats lurking in the fabric of current blockchain-based DNS alternatives. To this end, in Chapter 8 several different subsets of challenges applicable to blockchain DNS systems were identified. These challenges can be mainly classified into (i) the registration procedure and users behavior, (ii) the extraction of information flows and their links with external threat analysis systems, and (iii) the security of the underlying blockchain platform and proactive measures. From the analysis of Emercoin and Namecoin ecosystems, it becomes clear that there is an urgent need to improve the robustness and security of the registration procedures in blockchain DNS systems, since they can be easily abused for malicious activities. As observed in the studied systems and due the possibility of having other potential indicators, it is evident that exploring and assessing the different data managed by such systems is crucial to design the proper mitigation strategies. For example, parameters such as the amount of suspicious domains registered (e.g., domain squatting [149], or artificially generated domains as discussed in Chapter 7), the number of wallet updates, the IPs and domains registered, and the connectivity of the nodes are features that can be used to identify potentially harmful user behaviors. The latter can be augmented by the proposed hop-based approach as well as similar methods following blacklisting policies, enhancing the reliability and trust of blockchain DNS while reducing the impact of malicious campaigns.

9.2 Future work

In its essence, this thesis can be considered as a pioneering effort to shed light on some of the lesser-studied facets of several deviant and malicious online behaviors. Thus, in the course of this work, several points were identified where the underlying research can be extended, while considering other promising approaches. In the context of the adult content problem in SLSS, the resemblance between consumers and lurkers [150, 151] should be investigated further,

to better understand the proportion of users consuming adult content passively, and users participating in predatory behaviors by, e.g., providing praise and currencies to producers, or by publicly chatting with the producers that promote specific behaviors and content.

To this end, the use of behavioral features of user interactions should be explored thoroughly, as proposed in the recent work by Milon-Flores et al. [152]. Interactions such as chatting, viewing streams, following, sending likes and gifts, etc., provide additional context regarding the behavior of each user and can be leveraged to decipher their intentions and strategies. The present work highlighted that predators employ cunning means to evade moderation mechanisms and safeguards. Nonetheless, the careful study of their behavioral patterns, can provide valuable insights towards crafting robust and resilient detection mechanisms. Due to the complexity of identifying malicious/predatory interactions, a promising approach for the automation of the moderation task is to consider ensemble methods as proposed in [153], since such approaches can effectively leverage different machine learning models for exploiting the diverse features emerging from the behavioral analysis of interaction and communication patterns between predators and victims. Furthermore, future efforts for the detection of predatory behaviors could greatly benefit from the large-scale dataset of both verbal and non-verbal interactions (e.g. likes and rewards) in a Social Live Streaming Service, collected in the course of the this work [34] and analyzed in Chapter 4.

A different dimension of this phenomenon that should be explored is the proliferation of pornographic content production and consumption throughout the COVID pandemic and its implications for the online safety of younger users. The confinement due to lockdowns led to an unprecedented increase in consumption of explicit material [154, 155] and in this case, as demonstrated by the present work, *supply meets demand* (see Section 5.2). One of the most alarming cases of such content is undoubtedly the production and exchange of Child Sexual Abuse/Exploitation Material (CSAM/CSEM). Notably, Europol's 2021 Annual Internet Organised Crime Threat Assessment (IOCTA) [2], reported a surge of self-generated explicit material often captured in the victims' bedrooms, exchanged through social media platforms by abusers, a phenomenon that can be directly linked with the sexual grooming taking place in SLSS platforms as outlined by the current study. As such, it becomes imperative to address this issue.

Going beyond the boundaries of deviant behaviors in social platforms, several other cybercriminal behaviors including money laundering and the financing of other, probably more dangerous activities, can be just happening in plain sight [156, 157, 158], and there is pressing need to further study and disentangle the modus operandi of cybercrime vendors operating in

the Surface Web. To this end, the topic model-based analysis presented in Chapter 6 suggests that while the product descriptions in Shoppy are short in length and provide limited context, they could still be effectively leveraged by more advanced machine learning approaches. As such, more effort should be devoted to the development of robust detection/classification methods for the automated identification of postings selling illicit products/services, or promoting communication channels used by cybercrime vendors.

Finally, as shown in this work, blockchain-based DNS systems comprise an alluring landscape for cybercriminals who abuse them for malicious purposes. Provided the novelty of such systems and the adaptability of cybercriminals, more research should be devoted towards the exploration of other blockchain-based DNS systems and the elaboration of ontologies and security models to overcome the main drawbacks of such systems, with the aim to provide a reliable and sustainable decentralized DNS landscape. A decisive step towards achieving this, is establishing a holistic end-to-end approach, possibly through integrating smart contracts with revocation mechanisms [159, 160], to manage the registration procedure as well as to protect blockchain DNS systems from misuse. Moreover, while security and privacy initiatives should be supported, the accountability perspective, especially when it comes to critical Internet infrastructures such as DNS must also be taken into consideration. Additionally, with the continuous rise of blockchain-backed DNS schemes, the quest for information about their domains and their interconnections becomes even more necessary. The timely collection of quality intelligence is crucial to detect cybercriminal campaigns and may lead to their prevention, since methodologies like the one outlined in Chapter 8 rely on such information to establish ground truth. Therefore, more efforts should be devoted to the active monitoring of the blockchain DNS ecosystem, including both their domains and IPs, in an automated way.

In the case of blockchain features, they are often recalled in their beneficial form, yet some of them can leverage malicious opportunities. The clearest example of this is immutability. In this regard, the impossibility of deleting records guarantees traceability and auditability of the modus operandi of malicious campaigns, and enables mitigation actions. For instance, proactive security in blockchains can be implemented, with, e.g. active checks focusing on the behavior of the users, as well as the information associated with each wallet. The latter can be used to detect future campaigns by using, e.g. the proposed approach as well as similar methods following blacklisting policies, enhancing the reliability and trust of blockchain DNS while reducing the impact of malicious campaigns. Nevertheless, the impossibility of deleting, e.g. malicious records or illegal information, is a clear disadvantage. In this regard,

there is still much work ahead to enable efficient blockchain deletion mechanisms [161, 148], since actual practices mainly rely on forks, and long block consolidation mechanisms, which add prohibitive overhead to blockchain systems. Aligned with the idea of forks, well-known systems such as Bitcoin and Ethereum have opted for forks as a solution to security issues or required protocol changes to enable further functionalities [162, 163]. Therefore, fork-based strategies, including novel and robust functionalities, could help in recovering the trust in Namecoin and Emercoin.

Bibliography

- [1] N. Lykousas and C. Patsakis. “Large-scale analysis of grooming in modern social networks”. *Expert Systems with Applications*, 176, (2021), p. 114808. doi: [10.1016/j.eswa.2021.114808](https://doi.org/10.1016/j.eswa.2021.114808) (cit. on pp. 4, 14).
- [2] Europol. *Internet Organised Crime Threat Assessment (IOCTA 2021)*. European Union Agency for Law Enforcement Cooperation (Europol), 2021 (cit. on pp. 4, 121).
- [3] B. Melugin. *Pedophiles using app to manipulate underage girls into sexual acts, sell recordings as child porn*. Ed. by F. L. Angeles. <https://www.foxla.com/news/pedophiles-using-app-to-manipulate-underage-girls-into-sexual-acts-sell-recordings-as-child-porn>. [Online; last accessed 12-March-2020]. May 2018 (cit. on pp. 5, 7, 55).
- [4] S. Craven, S. Brown, and E. Gilchrist. “Sexual grooming of children: Review of literature and theoretical considerations”. *Journal of sexual aggression*, 12 (3), (2006), pp. 287–299 (cit. on p. 7).
- [5] BBC. *Instagram biggest for child grooming online - NSPCC finds*. <https://www.bbc.com/news/uk-47410520>. 2019 (cit. on p. 7).
- [6] S. Khamis, L. Ang, and R. Welling. “Self-branding, ‘micro-celebrity’ and the rise of Social Media Influencers”. *Celebrity studies*, 8 (2), (2017), pp. 191–208 (cit. on p. 8).
- [7] A. R. Gómez. “Digital Fame and Fortune in the age of Social Media: A Classification of social media influencers”. *aDResearch: Revista Internacional de Investigación en Comunicación*, (19), (2019), pp. 8–29 (cit. on p. 8).
- [8] V. Nandagiri and L. Philip. “Impact of Influencers from Instagram and Youtube on their followers”. *International Journal of Multidisciplinary Research and Modern Education*, 4 (1), (2018), pp. 61–65 (cit. on p. 8).
- [9] T. Terranova. “Attention, economy and the brain”. *Culture Machine*, 13, (2012) (cit. on p. 8).
- [10] C. Lou and S. Yuan. “Influencer marketing: how message value and credibility affect consumer trust of branded content on social media”. *Journal of Interactive Advertising*, 19 (1), (2019), pp. 58–73 (cit. on p. 8).

- [11] J. Drenten, L. Gurrieri, and M. Tyler. “Sexualized labour in digital culture: Instagram influencers, porn chic and the monetization of attention”. *Gender, Work & Organization*, 27 (1), (2020), pp. 41–66 (cit. on pp. 8, 59).
- [12] D. S. Thomas. *Cybercrime Losses: An Examination of US Manufacturing and the Total Economy*. 2020 (cit. on p. 9).
- [13] I. C. C. C. (IC3). *2019 INTERNET CRIME REPORT*. https://pdf.ic3.gov/2019_IC3Report.pdf. 2019 (cit. on p. 9).
- [14] W. E. Forum. *Wild Wide Web Consequences of Digital Fragmentation*. <https://reports.weforum.org/global-risks-report-2020/wild-wide-web/>. Accessed: 2021-09-30. 2020 (cit. on p. 9).
- [15] GlobeNewswire. *Hyper-Connected Web of Profit Emerges, As Global Cybercriminal Revenues Hit \$1.5 Trillion Annually*. <https://www.globenewswire.com/news-release/2018/04/20/1482411/0/en/Hyper-Connected-Web-of-Profit-Emerges-As-Global-Cybercriminal-Revenues-Hit-1-5-Trillion-Annually.html>. Accessed: 2021-09-30. 2018 (cit. on p. 9).
- [16] D. Décary-Héту and A. Leppänen. “Criminals and signals: An assessment of criminal performance in the carding underworld”. *Security Journal*, 29 (3), (2016), pp. 442–460 (cit. on p. 9).
- [17] Europol. *Covid-19 sparks upward trend in cybercrime*. <https://www.europol.europa.eu/newsroom/news/covid-19-sparks-upward-trend-in-cybercrime>. Accessed: 2021-09-30. 2020 (cit. on p. 9).
- [18] Interpol. *Interpol report shows alarming rate of cyberattacks during COVID-19*. <https://www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19>. Accessed: 2021-09-30. 2020 (cit. on p. 9).
- [19] M. Yip, N. Shadbolt, and C. Webber. “Why forums? An empirical analysis into the facilitating factors of carding forums”. *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013, pp. 453–462 (cit. on p. 10).
- [20] K. K. Peretti. “Data breaches: what the underground world of carding reveals”. *Santa Clara Computer & High Tech. LJ*, 25, (2008), p. 375 (cit. on p. 10).
- [21] V. Benjamin, W. Li, T. Holt, and H. Chen. “Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops”. *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE. IEEE, 2015, pp. 85–90 (cit. on p. 10).
- [22] Z. Doffman. *Facebook Dark Web Deal: Hackers Just Sold 267 Million User Profiles For \$540*. <https://www.forbes.com/sites/zakdoffman/2020/04/20/facebook-users-beware-hackers-just-sold-267-million-of-your-profiles-for-540/>. Accessed: 2021-09-30. 2020 (cit. on p. 10).

- [23] C. Cimpanu. *Hacker selling data of 538 million Weibo users*. <https://www.zdnet.com/article/hacker-selling-data-of-538-million-weibo-users/>. Accessed: 2021-09-30. 2020 (cit. on p. 10).
- [24] M. Singh, M. Singh, and S. Kaur. “Issues and challenges in DNS based botnet detection: A survey”. *Computers & Security*, 86, (2019), pp. 28–52. issn: 0167-4048 (cit. on p. 11).
- [25] R. Perdisci, I. Corona, and G. Giacinto. “Early Detection of Malicious Flux Networks via Large-Scale Passive DNS Traffic Analysis”. *IEEE Transactions on Dependable and Secure Computing*, 9 (5), (2012), pp. 714–726. issn: 1545-5971 (cit. on p. 11).
- [26] X. Zang, J. Gong, S. Mo, A. Jakalan, and D. Ding. “Identifying Fast-Flux Botnet With AGD Names at the Upper DNS Hierarchy”. *IEEE Access*, 6, (2018), pp. 69713–69727 (cit. on p. 11).
- [27] Y. Nadji, M. Antonakakis, R. Perdisci, D. Dagon, and W. Lee. “Beheading hydras: performing effective botnet takedowns”. *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 2013, pp. 121–132 (cit. on pp. 11, 88).
- [28] S. Yadav and A. L. N. Reddy. “Winning with DNS Failures: Strategies for Faster Botnet Detection”. *Security and Privacy in Communication Networks*. Ed. by M. Rajarajan, F. Piper, H. Wang, and G. Kesidis. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 446–459. isbn: 978-3-642-31909-9 (cit. on p. 11).
- [29] F. Casino, T. K. Dasaklis, and C. Patsakis. “A systematic literature review of blockchain-based applications: Current status, classification and open issues”. *Telematics and Informatics*, 36, (2019), pp. 55–81 (cit. on pp. 12, 106).
- [30] A. Group. *Domain Name Management Scheme For Cross-Chain Interactions In Blockchain Systems*. <https://domainnamewire.com/wp-content/alibaba-blockchain-domain-patent.pdf>. Accessed: 2021-09-30. 2019 (cit. on p. 12).
- [31] E. Karaarslan and E. Adiguzel. “Blockchain Based DNS and PKI Solutions”. *IEEE Communications Standards Magazine*, 2 (3), (2018), pp. 52–57 (cit. on pp. 12, 99).
- [32] R. Amado. *How Cybercriminals are using Blockchain DNS: From the Market to the .Bazar*. <https://www.digitalshadows.com/blog-and-research/how-cybercriminals-are-using-blockchain-dns-from-the-market-to-the-bazar/>. 2018 (cit. on pp. 12, 101).
- [33] N. Lykousas, C. Patsakis, and V. Gómez. “Adult Content in Social Live Streaming Services: Characterizing Deviant Users and Relationships”. *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*. Ed. by U. Brandes, C. Reddy, and A. Tagarelli. IEEE Computer Society, 2018, pp. 375–382. doi: [10.1109/ASONAM.2018.8508246](https://doi.org/10.1109/ASONAM.2018.8508246) (cit. on pp. 13, 39, 53, 66).

- [34] N. Lykousas and C. Patsakis. *Large-scale analysis of grooming in modern social networks*. Dec. 2019. doi: [10.5281/zenodo.3560365](https://doi.org/10.5281/zenodo.3560365). url: <https://doi.org/10.5281/zenodo.3560365> (cit. on pp. 14, 39, 121).
- [35] N. Lykousas, F. Casino, and C. Patsakis. “Inside the X-Rated World of “Premium” Social Media Accounts”. *Social Informatics - 12th International Conference, SocInfo 2020, Pisa, Italy, October 6-9, 2020, Proceedings*. Ed. by S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, and D. Pedreschi. Vol. 12467. Lecture Notes in Computer Science. Springer, 2020, pp. 181–191. doi: [10.1007/978-3-030-60975-7_14](https://doi.org/10.1007/978-3-030-60975-7_14) (cit. on p. 14).
- [36] F. Casino, N. Lykousas, I. Homoliak, C. Patsakis, and J. Hernandez-Castro. “Intercepting Hail Hydra: Real-time detection of Algorithmically Generated Domains”. *Journal of Network and Computer Applications*, 190, (2021), p. 103135. doi: [10.1016/j.jnca.2021.103135](https://doi.org/10.1016/j.jnca.2021.103135) (cit. on p. 15).
- [37] C. Patsakis, F. Casino, N. Lykousas, and V. Katos. “Unravelling ariadne’s thread: Exploring the threats of decentralised dns”. *IEEE Access*, 8, (2020), pp. 118559–118571. doi: [10.1109/access.2020.3004727](https://doi.org/10.1109/access.2020.3004727) (cit. on p. 15).
- [38] F. Casino, N. Lykousas, V. Katos, and C. Patsakis. “Unearthing malicious campaigns and actors from the blockchain DNS ecosystem”. *Computer Communications*, 179, (2021), pp. 217–230. doi: [10.1016/j.comcom.2021.08.023](https://doi.org/10.1016/j.comcom.2021.08.023) (cit. on pp. 16, 111).
- [39] N. Lykousas, C. Patsakis, A. Kaltenbrunner, and V. Gómez. “Sharing emotions at scale: The vent dataset”. *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*. AAAI Press, 2019, pp. 611–619 (cit. on p. 16).
- [40] N. Lykousas, C. Patsakis, A. Kaltenbrunner, and V. Gómez. *Dataset for paper “Sharing emotions at scale: The Vent dataset”*. Jan. 2019. doi: [10.5281/zenodo.2537838](https://doi.org/10.5281/zenodo.2537838). url: <https://doi.org/10.5281/zenodo.2537838> (cit. on p. 17).
- [41] V. Koutsokostas, N. Lykousas, T. Apostolopoulos, G. Orazi, A. Ghosal, F. Casino, M. Conti, and C. Patsakis. “Invoice #31415 attached: Automated analysis of malicious Microsoft Office documents”. *Computers & Security*, 114, (2022), p. 102582. doi: [10.1016/j.cose.2021.102582](https://doi.org/10.1016/j.cose.2021.102582) (cit. on p. 17).
- [42] V. Koutsokostas, N. Lykousas, G. Orazi, T. Apostolopoulos, A. Ghosal, F. Casino, M. Conti, and C. Patsakis. *Malicious MS Office documents dataset*. Feb. 2021. doi: [10.5281/zenodo.4559436](https://doi.org/10.5281/zenodo.4559436). url: <https://doi.org/10.5281/zenodo.4559436> (cit. on p. 17).
- [43] F. Casino, N. Totosis, T. Apostolopoulos, N. Lykousas, and C. Patsakis. “Analysis and Correlation of Visual Evidence in Campaigns of Malicious Office Documents”. *Digital Threats*, (2022). issn: 2692-1626. doi: [10.1145/3513025](https://doi.org/10.1145/3513025) (cit. on p. 18).
- [44] B. Wang, X. Zhang, G. Wang, H. Zheng, and B. Y. Zhao. “Anatomy of a personalized livestreaming system”. *Proceedings of the 2016 ACM on Internet Measurement Conference*. ACM, 2016, pp. 485–498 (cit. on pp. 22, 25).

- [45] M. Siekkinen, E. Masala, and T. Kämäräinen. “A First Look at Quality of Mobile Live Streaming Experience”. *Proceedings of the 2016 ACM on Internet Measurement Conference*. ACM. 2016, pp. 477–483. isbn: 9781450345262. doi: [10.1145/2987443.2987472](https://doi.org/10.1145/2987443.2987472) (cit. on p. 23).
- [46] A. Papageorgiou, M. Strigkos, E. Politou, E. Alepis, A. Solanas, and C. Patsakis. “Security and privacy analysis of mobile health applications: the alarming state of practice”. *Ieee Access*, 6, (2018), pp. 9390–9403 (cit. on p. 23).
- [47] R. Pantos and W. May. *HTTP live streaming*. Tech. rep. 2017 (cit. on p. 23).
- [48] M. Coletto, L. M. Aiello, C. Lucchese, and F. Silvestri. “On the Behaviour of Deviant Communities in Online Social Networks.” *10th International Conference on Web and Social Media*. 2016, pp. 72–81. isbn: 9781577357582 (cit. on pp. 24, 26, 27, 33).
- [49] M. Kurant, A. Markopoulou, and P. Thiran. “Towards unbiased BFS sampling”. *IEEE Journal on Selected Areas in Communications*, 29 (9), (2011), pp. 1799–1809. issn: 07338716. doi: [10.1109/JSAC.2011.111005](https://doi.org/10.1109/JSAC.2011.111005) (cit. on p. 25).
- [50] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. “Antisocial Behavior in Online Discussion Communities”. *9th International Conference on Web and Social Media, ICWSM*. 2015, pp. 61–70. isbn: 978-1-57735-733-9 (cit. on p. 26).
- [51] M. B. Friedländer. “And Action! Live in front of the Camera: An Evaluation of the Social Live Streaming Service YouNow”. *International Journal of Information Communication Technologies and Human Development*, 9 (1), (2017), pp. 15–33. doi: [10.4018/IJICTHD.2017010102](https://doi.org/10.4018/IJICTHD.2017010102) (cit. on p. 26).
- [52] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. *Advances in neural information processing systems*. 2012, pp. 1097–1105 (cit. on p. 26).
- [53] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 26).
- [54] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. “Intriguing properties of neural networks”. *arXiv preprint arXiv:1312.6199*, (2013). arXiv: [1312.6199](https://arxiv.org/abs/1312.6199) (cit. on p. 28).
- [55] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. “Fast unfolding of communities in large networks”. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10), (2008). issn: 17425468. arXiv: [0803.0476](https://arxiv.org/abs/0803.0476) (cit. on p. 31).
- [56] J. M. Kleinberg. “Authoritative sources in a hyperlinked environment”. *Journal of the ACM*, 46 (5), (1999), pp. 604–632. issn: 00045411. doi: [10.1145/324133.324140](https://doi.org/10.1145/324133.324140) (cit. on p. 31).
- [57] K. Lee, P. Tamilarasan, and J. Caverlee. “Crowdturfers, campaigns, and social media: tracking and revealing crowdsourced manipulation of social media”. *Seventh International AAAI Conference on Weblogs and Social Media*. 2013, pp. 331–340 (cit. on p. 32).

- [58] A. Java, X. Song, T. Finin, and B. Tseng. “Why We Twitter: Understanding Microblogging Usage and Communities”. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. WebKDD/SNA-KDD '07. San Jose, California: ACM, 2007, pp. 56–65. isbn: 978-1-59593-848-0. doi: [10.1145/1348549.1348556](https://doi.org/10.1145/1348549.1348556) (cit. on p. 32).
- [59] J. Leskovec, J. Kleinberg, and C. Faloutsos. “Graphs over time: densification laws, shrinking diameters and possible explanations”. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 177–187 (cit. on p. 33).
- [60] S. Xiao, G. Xiao, T. H. Cheng, S. Ma, X. Fu, and H. Soh. “Robustness of scale-free networks under rewiring operations”. *EPL (Europhysics Letters)*, 89 (3), (2010), p. 38002. issn: 0295-5075. doi: [10.1209/0295-5075/89/38002](https://doi.org/10.1209/0295-5075/89/38002) (cit. on p. 33).
- [61] M. Drouin, R. L. Boyd, J. T. Hancock, and A. James. “Linguistic analysis of chat transcripts from child predator undercover sex stings”. *The Journal of Forensic Psychiatry & Psychology*, 28 (4), (2017), pp. 437–457 (cit. on p. 43).
- [62] N. Lorenzo-Dus and A. Kinzel. “‘So is your mom as cute as you?’: examining patterns of language use by online sexual groomers”. *Journal of Corpora and Discourse Studies*, 2 (1), (2019), pp. 1–30 (cit. on p. 43).
- [63] N. Lorenzo-Dus, A. Kinzel, and M. Di Cristofaro. “The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use”. *Journal of Pragmatics*, 155, (2020), pp. 15–27 (cit. on p. 43).
- [64] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. *The development and psychometric properties of LIWC2015*. Tech. rep. University of Texas at Austin, 2015 (cit. on p. 43).
- [65] E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. “Conversational networks for automatic online moderation”. *IEEE Transactions on Computational Social Systems*, 6 (1), (2019), pp. 38–55 (cit. on p. 43).
- [66] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. “Deceiving google’s perspective api built for detecting toxic comments”. *arXiv preprint arXiv:1702.08138*, (2017) (cit. on p. 43).
- [67] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. “Enriching word vectors with subword information”. *Transactions of the Association for Computational Linguistics*, 5, (2017), pp. 135–146 (cit. on p. 43).
- [68] M. Honnibal and M. Johnson. “An improved non-monotonic transition system for dependency parsing”. *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 1373–1378 (cit. on p. 44).
- [69] E. Loper and S. Bird. “NLTK: The Natural Language Toolkit”. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 2002, pp. 63–70 (cit. on pp. 44, 47).

- [70] G. McCulloch and L. Gawne. “Emoji grammar as beat gestures”. *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media*. 2018 (cit. on p. 44).
- [71] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent dirichlet allocation”. *Journal of machine Learning research*, 3 (Jan), (2003), pp. 993–1022 (cit. on p. 45).
- [72] E. Cambria and B. White. “Jumping NLP curves: A review of natural language processing research”. *IEEE Computational intelligence magazine*, 9 (2), (2014), pp. 48–57 (cit. on p. 45).
- [73] A. R. M. Basher and B. C. Fung. “Analyzing topics and authors in chat logs for crime investigation”. *Knowledge and information systems*, 39 (2), (2014), pp. 351–381 (cit. on p. 45).
- [74] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. “Detection of harassment on web 2.0”. *Proceedings of the Content Analysis in the WEB*, 2, (2009), pp. 1–7 (cit. on p. 47).
- [75] C. Doll, A. Sykosch, M. Ohm, and M. Meier. “Automated Pattern Inference Based on Repeatedly Observed Malware Artifacts”. *Proceedings of the 14th International Conference on Availability, Reliability and Security*. 2019, pp. 1–10 (cit. on p. 48).
- [76] L. Hong and B. D. Davison. “Empirical study of topic modeling in twitter”. *Proceedings of the first workshop on social media analytics*. 2010, pp. 80–88 (cit. on p. 48).
- [77] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. “Comparing twitter and traditional media using topic models”. *European conference on information retrieval*. Springer. 2011, pp. 338–349 (cit. on p. 48).
- [78] M. Röder, A. Both, and A. Hinneburg. “Exploring the space of topic coherence measures”. *Proceedings of the eighth ACM international conference on Web search and data mining*. Ed. by X. Cheng, H. Li, E. Gabrilovich, and J. Tang. ACM, 2015, pp. 399–408 (cit. on p. 49).
- [79] R. Řehůřek and P. Sojka. “Software Framework for Topic Modelling with Large Corpora”. English. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50 (cit. on pp. 49, 79).
- [80] C. Sievert and K. Shirley. “LDAvis: A method for visualizing and interpreting topics”. *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014, pp. 63–70 (cit. on pp. 49, 79).
- [81] L. Breiman. “Random forests”. *Machine learning*, 45 (1), (2001), pp. 5–32 (cit. on p. 52).
- [82] L. Breiman. “Manual on setting up, using, and understanding random forests v3. 1”. *Statistics Department University of California Berkeley, CA, USA*, 1, (2002), p. 58 (cit. on p. 52).

- [83] J. Han, J. Pei, Y. Yin, and R. Mao. “Mining frequent patterns without candidate generation: A frequent-pattern tree approach”. *Data mining and knowledge discovery*, 8 (1), (2004), pp. 53–87 (cit. on p. 54).
- [84] E. A. Daniels. “Sexiness on social media: The social costs of using a sexy profile photo”. *Sexualization, Media, & Society*, 2 (4), (2016), p. 2374623816683522 (cit. on p. 59).
- [85] S. E. Baumgartner, S. R. Sumter, J. Peter, and P. M. Valkenburg. “Sexual self-presentation on social network sites: Who does it and how is it perceived?” *Computers in Human Behavior*, 50, (2015), pp. 91–100 (cit. on p. 59).
- [86] J. M. van Oosten, J. Peter, and I. Boot. “Exploring associations between exposure to sexy online self-presentations and adolescents’ sexual attitudes and behavior”. *Journal of Youth and Adolescence*, 44 (5), (2015), pp. 1078–1091 (cit. on p. 59).
- [87] J. M. van Oosten and L. Vandenbosch. “Sexy online self-presentation on social network sites and the willingness to engage in sexting: A comparison of gender and age”. *Journal of adolescence*, 54, (2017), pp. 42–50 (cit. on p. 59).
- [88] L. Clarke. *The x-rated world of premium Snapchat has spawned an illicit underground industry*. <https://www.wired.co.uk/article/premium-snapchat-adult-models>. Accessed: 04/05/2020. 2019 (cit. on p. 60).
- [89] C. Downs. *OnlyFans, Influencers, And The Politics Of Selling Nudes During A Pandemic*. <https://www.elle.com/culture/a32459935/onlyfans-sex-work-influencers/>. Accessed: 14/05/2020. 2020 (cit. on pp. 60, 119).
- [90] M. Zimmer. ““But the data is already public”: on the ethics of research in Facebook”. *Ethics and information technology*, 12 (4), (2010), pp. 313–325 (cit. on p. 60).
- [91] G. Drolet. *Sex Work Comes Home*. <https://www.nytimes.com/2020/04/10/style/camsoda-onlyfans-streaming-sex-coronavirus.html>. Accessed: 04/05/2020. 2020 (cit. on pp. 61, 119).
- [92] A. Lee. *Coronavirus is bad news for Big Porn but great news for OnlyFans*. <https://www.wired.co.uk/article/coronavirus-porn-industry-onlyfans>. Accessed: 04/05/2020. 2020 (cit. on pp. 61, 119).
- [93] M. V. Henry and P. Farvid. ““Always hot, always live’: Computer-mediated sex work in the era of ‘camming’.” *Women’s Studies Journal*, 31 (2), (2017) (cit. on p. 63).
- [94] R. Dingedine, N. Mathewson, and P. Syverson. *Tor: The second-generation onion router*. Tech. rep. Naval Research Lab Washington DC, 2004 (cit. on p. 71).
- [95] K. Thomas, D. Y. Huang, D. Y. Wang, E. Bursztein, C. Grier, T. Holt, C. Kruegel, D. McCoy, S. Savage, and G. Vigna. “Framing Dependencies Introduced by Underground Commoditization”. *14th Annual Workshop on the Economics of Information Security, WEIS 2015, Delft, The Netherlands, 22-23 June, 2015*. 2015 (cit. on p. 71).
- [96] F. Wehinger. “The Dark Net: Self-regulation dynamics of illegal online markets for identities and related services”. *2011 European Intelligence and Security Informatics Conference*. IEEE. IEEE, 2011, pp. 209–213 (cit. on p. 71).

- [97] A. Hutchings and T. J. Holt. “A crime script analysis of the online stolen data market”. *British Journal of Criminology*, 55 (3), (2015), pp. 596–614 (cit. on p. 71).
- [98] X. Wang, P. Peng, C. Wang, and G. Wang. “You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces”. *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. Ed. by J. Kim, G. Ahn, S. Kim, Y. Kim, J. López, and T. Kim. ACM, 2018, pp. 431–442 (cit. on p. 71).
- [99] J. B. Schafer, J. A. Konstan, and J. Riedl. “E-commerce recommendation applications”. *Data mining and knowledge discovery*, 5 (1-2), (2001), pp. 115–153 (cit. on p. 72).
- [100] M. Pavkovic and J. Protic. “SInFo–Structure-Driven Incremental Forum Crawler That Optimizes User-Generated Content Retrieval”. *IEEE Access*, 7, (2019), pp. 126941–126961 (cit. on p. 75).
- [101] N. Tavabi, N. Bartley, A. Abeliuk, S. Soni, E. Ferrara, and K. Lerman. “Characterizing activity on the deep and dark web”. *Companion Proceedings of The 2019 World Wide Web Conference*. Ed. by S. Amer-Yahia, M. Mahdian, A. Goel, G. Houben, K. Lerman, J. J. McAuley, R. Baeza-Yates, and L. Zia. ACM, 2019, pp. 206–213 (cit. on p. 75).
- [102] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, M. Tesconi, and E. Ferrara. “Charting the Landscape of Online Cryptocurrency Manipulation”. *IEEE Access*, 8, (2020), pp. 113230–113245 (cit. on p. 78).
- [103] K. Turk, S. Pastrana, and B. Collier. “A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments”. *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. IEEE, 2020, pp. 428–437 (cit. on p. 78).
- [104] M. Röder, A. Both, and A. Hinneburg. “Exploring the Space of Topic Coherence Measures”. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM ’15. Shanghai, China: Association for Computing Machinery, 2015, 399–408. isbn: 9781450333177 (cit. on p. 79).
- [105] J. G. Cloward and B. L. Abarbanel. “In-Game Currencies, Skin Gambling, and the Persistent Threat of Money Laundering in Video Games”. *UNLV Gaming LJ*, 10, (2020), p. 105 (cit. on p. 81).
- [106] A. Moiseienko and K. Izenman. *Gaming the System: Money Laundering Through Online Games*. https://rusi.org/sites/default/files/20191011_newsbrief_vol39_no9_moiseienko_and_izenman_web.pdf. 2019 (cit. on p. 81).
- [107] P. Kirkbride, M. A. A. Dewan, and F. Lin. “Game-Like Captchas for Intrusion Detection”. *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCCom/CyberSciTech)*. IEEE. IEEE, 2020, pp. 312–315 (cit. on p. 81).

- [108] T. Malderle, M. Wübbeling, S. Knauer, A. Sykosch, and M. Meier. “Gathering and analyzing identity leaks for a proactive warning of affected users”. *Proceedings of the 15th ACM International Conference on Computing Frontiers, CF 2018, Ischia, Italy, May 08-10, 2018*. Ed. by D. R. Kaeli and M. Pericàs. ACM, 2018, pp. 208–211 (cit. on p. 81).
- [109] J. Chuang, C. D. Manning, and J. Heer. “Termite: Visualization techniques for assessing textual topic models”. *Proceedings of the international working conference on advanced visual interfaces*. ACM, 2012, pp. 74–77 (cit. on p. 83).
- [110] openbullet. *OpenBullet*. <https://github.com/openbullet/openbullet>. Accessed: 2021-09-30 (cit. on p. 83).
- [111] Wapack Labs. *BlackBullet Credential Stuffing*. <https://redskyalliance.org/x/industry/blackbullet-credential-stuffing>. Accessed: 2021-09-30 (cit. on p. 83).
- [112] Netacea. *STORM Cracker - Credential Stuffing Tool*. <https://www.netacea.com/blog/storm-cracker-tool/>. Accessed: 2021-09-30 (cit. on p. 83).
- [113] S. Rees-Pullman. “Is credential stuffing the new phishing?” *Computer Fraud & Security*, 2020 (7), (2020), pp. 16–19 (cit. on p. 83).
- [114] P. Poornachandran, M Nithun, S. Pal, A. Ashok, and A. Ajayan. “Password reuse behavior: how massive online data breaches impacts personal data in web”. *Innovations in Computer Science and Engineering*. Springer, 2016, pp. 199–210 (cit. on p. 84).
- [115] C. Patsakis, F. Casino, and V. Katos. “Encrypted and covert DNS queries for botnets: Challenges and countermeasures”. *Computers & Security*, 88, (2020), p. 101614 (cit. on p. 88).
- [116] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla. “A Comprehensive Measurement Study of Domain Generating Malware”. *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, 2016, pp. 263–278. isbn: 978-1-931971-32-4 (cit. on p. 88).
- [117] J. Bader. *Domain Generation Algorithms (DGAs) of Malware reimplemented in Python*. https://github.com/baderj/domain_generation_algorithms. 2020 (cit. on p. 89).
- [118] A. Abakumov. *DGA repository*. <https://github.com/andrewaeva/DGA>. 2020 (cit. on p. 89).
- [119] P. Chaignon. *DGA Collection*. <https://github.com/pchaigno/dga-collection>. 2020 (cit. on p. 89).
- [120] J. Spooren et al. “Detection of Algorithmically Generated Domain Names Used by Botnets: A Dual Arms Race”. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. SAC ’19. Limassol, Cyprus: ACM, 2019, pp. 1916–1923 (cit. on pp. 90, 95, 97).

- [121] X. Yun, J. Huang, Y. Wang, T. Zang, Y. Zhou, and Y. Zhang. “Khaos: An Adversarial Neural Network DGA With High Anti-Detection Ability”. *IEEE Transactions on Information Forensics and Security*, 15, (2020), pp. 2225–2240 (cit. on pp. 90, 95, 97).
- [122] F. Casino, N. Lykousas, I. Homoliak, C. Patsakis, and J. Hernandez-Castro. *HYDRA dataset*. Version 1.0. July 2020. url: <https://doi.org/10.5281/zenodo.3965397> (cit. on p. 90).
- [123] T. G. Dietterich. “Ensemble methods in machine learning”. *International workshop on multiple classifier systems*. Springer. 2000, pp. 1–15 (cit. on p. 92).
- [124] G. Valentini and F. Masulli. “Ensembles of learning machines”. *Italian workshop on neural nets*. Springer. 2002, pp. 3–20 (cit. on p. 92).
- [125] R. Barandela, R. M. Valdovinos, and J. S. Sánchez. “New applications of ensembles of classifiers”. *Pattern Analysis & Applications*, 6 (3), (2003), pp. 245–256 (cit. on p. 92).
- [126] L. I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014 (cit. on p. 92).
- [127] M. Alaeiyan, S. Parsa, V. P., and M. Conti. “Detection of algorithmically-generated domains: An adversarial machine learning approach”. *Computer Communications*, (2020). issn: 0140-3664 (cit. on p. 92).
- [128] H. S. Anderson, J. Woodbridge, and B. Filar. “DeepDGA: Adversarially-Tuned Domain Generation and Detection”. *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*. AISEC ’16. Vienna, Austria: ACM, 2016, pp. 13–21. isbn: 978-1-4503-4573-6 (cit. on p. 92).
- [129] J. Selvi, R. J. Rodríguez, and E. Soria-Olivas. “Detection of algorithmically generated malicious domain names using masked N-grams”. *Expert Systems with Applications*, 124, (2019), pp. 156–163. issn: 0957-4174 (cit. on p. 92).
- [130] L. Breiman. “Bagging predictors”. *Machine learning*, 24 (2), (1996), pp. 123–140 (cit. on p. 93).
- [131] M. Ali, J. Nelson, R. Shea, and M. J. Freedman. “Blockstack: A Global Naming and Storage System Secured by Blockchains”. *2016 USENIX Annual Technical Conference (USENIX ATC 16)*. Denver, CO: USENIX Association, June 2016, pp. 181–194. isbn: 978-1-931971-30-0 (cit. on p. 100).
- [132] Emercoin. *Emercoin Links & Resources*. <https://emercoin.com/en/documentation/links-resources>. 2019 (cit. on p. 100).
- [133] J. Sanders. *Blockchain-based Unstoppable Domains is a rehash of a failed idea*. <https://www.techrepublic.com/article/blockchain-based-unstoppable-domains-is-a-rehash-of-a-failed-idea/>. 2019 (cit. on p. 101).
- [134] F. Hassan, A. Ali, S. Latif, J. Qadir, S. Kanhere, J. Singh, and J. Crowcroft. “Blockchain And The Future of the Internet: A Comprehensive Review”. *arXiv preprint arXiv:1904.00733*, (2019) (cit. on p. 101).

- [135] R. Rasmussen and P. Vixie. “Surveying the DNS threat landscape”. *Technical Report Internet Identity*, (2013) (cit. on p. 101).
- [136] T. S. Project. *The most abused top-level domains in 2018*. <https://www.spamhaus.com/resource-center/the-most-abused-top-level-domains-in-2018>. 2019 (cit. on p. 103).
- [137] W. Foxley. *Ethereum Name Service Auction Exploited to Grab Apple Domain – And It Can’t Be Undone*. <https://www.coindesk.com/ethereum-name-service-auction-exploited-to-grab-apple-domain-and-it-cant-be-undone>. 2019 (cit. on p. 103).
- [138] Proofpoint. *2019 State of the Phish Report*. <https://www.proofpoint.com/us/resources/threat-reports/state-of-phish>. 2019 (cit. on p. 105).
- [139] M. Khonji, Y. Iraqi, and A. Jones. “Phishing detection: a literature survey”. *IEEE Communications Surveys & Tutorials*, 15 (4), (2013), pp. 2091–2121 (cit. on p. 105).
- [140] Phishme. *Enterprise Phishing Susceptibility and Resiliency Report*. https://www.infosecurityeurope.com/__novadocuments/351537?v=636276130024130000. 2016 (cit. on p. 105).
- [141] C. Iuga, J. R. Nurse, and A. Erola. “Baiting the hook: factors impacting susceptibility to phishing attacks”. *Human-centric Computing and Information Sciences*, 6 (1), (2016), p. 8 (cit. on p. 105).
- [142] S. Malwa. *ICO Scams Have Raised More Than \$1 Billion, Report Claims*. <https://www.ccn.com/ico-scams-have-raised-more-than-1-billion-report-claims>. 2018 (cit. on p. 106).
- [143] BBC. *Child abuse images hidden in crypto-currency blockchain*. https://www.bbc.com/news/technology-47130268?ocid=socialflow_twitter. 2019 (cit. on p. 106).
- [144] R. Matzutt, J. Hiller, M. Henze, J. H. Ziegeldorf, D. Müllmann, O. Hohlfeld, and K. Wehrle. “A quantitative analysis of the impact of arbitrary blockchain content on bitcoin”. *Proceedings of the 22nd International Conference on Financial Cryptography and Data Security (FC)*. Springer. 2018, pp. 420–438 (cit. on p. 106).
- [145] A. de Candia. *IndImm: a threat to the Ripple blockchain*. <https://en.cryptonomist.ch/2019/07/29/indimm-ripple-blockchain>. 2018 (cit. on p. 106).
- [146] C. Patsakis and F. Casino. “Hydras and IPFS: a decentralised playground for malware”. *International Journal of Information Security*, 18 (6), (2019), pp. 787–799. issn: 1615-5270. doi: [10.1007/s10207-019-00443-0](https://doi.org/10.1007/s10207-019-00443-0). url: <https://doi.org/10.1007/s10207-019-00443-0> (cit. on p. 106).
- [147] E. Politou, E. Alepis, and C. Patsakis. “Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions”. *Journal of Cybersecurity*, 4 (1), (2018), tyy001 (cit. on p. 107).

- [148] E. Politou, F. Casino, E. Alepis, and C. Patsakis. “Blockchain Mutability: Challenges and Proposed Solutions”. *IEEE Transactions on Emerging Topics in Computing*, (2019), pp. 1–1 (cit. on pp. 107, 123).
- [149] P. Xia, H. Wang, Z. Yu, X. Liu, X. Luo, and G. Xu. “Ethereum Name Service: the Good, the Bad, and the Ugly”. *arXiv preprint arXiv:2104.05185*, (2021) (cit. on p. 120).
- [150] A. Tagarelli and R. Interdonato. “Lurking in social networks: topology-based analysis and ranking methods”. *Social Network Analysis and Mining*, 4 (1), (2014), p. 230 (cit. on p. 120).
- [151] D. Perna, R. Interdonato, and A. Tagarelli. “Identifying users with alternate behaviors of lurking and active participation in multilayer social networks”. *IEEE Transactions on Computational Social Systems*, 5 (1), (2018), pp. 46–63 (cit. on p. 120).
- [152] D. F. Milon-Flores and R. L. Cordeiro. “How to take advantage of behavioral features for the early detection of grooming in online conversations”. *Knowledge-Based Systems*, 240, (2022), p. 108017 (cit. on p. 121).
- [153] M. A. Fauzi and P. Bours. “Ensemble method for sexual predators identification in online chats”. *2020 8th international workshop on biometrics and forensics (IWBF)*. IEEE, 2020, pp. 1–6 (cit. on p. 121).
- [154] G. Mestre-Bach, G. R. Blycker, and M. N. Potenza. “Pornography use in the setting of the COVID-19 pandemic”. *Journal of behavioral addictions*, 9 (2), (2020), pp. 181–183 (cit. on p. 121).
- [155] H. A. Awan, A. Aamir, M. N. Diwan, I. Ullah, V. Pereira-Sanchez, R. Ramalho, L. Orsolini, R. de Filippis, M. I. Ojeahere, R. Ransing, et al. “Internet and pornography use during the COVID-19 pandemic: presumed impact and what can be done”. *Frontiers in psychiatry*, 12, (2021), p. 220 (cit. on p. 121).
- [156] A. Irwin and J. Slay. “Detecting money laundering and terrorism financing activity in Second Life and World of Warcraft”. *International Cyber Resilience conference*. Mar. 2012 (cit. on p. 121).
- [157] A. Mikhaylov and R. Frank. “Cards, money and two hacking forums: An analysis of online money laundering schemes”. *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2016, pp. 80–83 (cit. on p. 121).
- [158] B. DaCosta and S. Seok. “Cybercrime in Online Gaming”. *Encyclopedia of Criminal Activities and the Deep Web*. IGI Global, 2020, pp. 881–892 (cit. on p. 121).
- [159] Y. Yu, Y. Zhao, Y. Li, X. Du, L. Wang, and M. Guizani. “Blockchain-Based Anonymous Authentication With Selective Revocation for Smart Industrial Applications”. *IEEE Transactions on Industrial Informatics*, 16 (5), (2020), pp. 3290–3300 (cit. on p. 122).
- [160] J. P. Cruz, Y. Kaji, and N. Yanai. “RBAC-SC: Role-Based Access Control Using Smart Contract”. *IEEE Access*, 6, (2018), pp. 12240–12251 (cit. on p. 122).

- [161] K. Maeda, M. Ohtani, Y. Oishi, C. Yasumoto, and J. Zhu. *Deletion of blocks in a blockchain*. US Patent 10,739,997. 2020 (cit. on p. [123](#)).
- [162] F. Schär. “Blockchain Forks: A Formal Classification Framework and Persistency Analysis”. *Munich Personal RePEc Archive*, (2020) (cit. on p. [123](#)).
- [163] T. Neudecker and H. Hartenstein. “An empirical analysis of blockchain forks in bitcoin”. *International Conference on Financial Cryptography and Data Security*. Springer. 2019, pp. 84–92 (cit. on p. [123](#)).

Part V

Publications

Adult content in Social Live Streaming Services: Characterizing deviant users and relationships

Nikolaos Lykousas*, Vicenç Gómez†, Constantinos Patsakis*

*University of Piraeus, Greece nikos.lykousas@gmail.com, kpatsak@unipi.gr

†Universitat Pompeu Fabra, Barcelona, Spain vicen.gomez@upf.edu

Abstract—Social Live Stream Services (SLSS) exploit a new level of social interaction. One of the main challenges in these services is how to detect and prevent deviant behaviors that violate community guidelines. In this work, we focus on adult content production and consumption in two widely used SLSS, namely Live.me and Loops Live, which have millions of users producing massive amounts of video content on a daily basis. We use a pre-trained deep learning model to identify broadcasters of adult content. Our results indicate that moderation systems in place are highly ineffective in suspending the accounts of such users. We create two large datasets by crawling the social graphs of these platforms, which we analyze to identify characterizing traits of adult content producers and consumers, and discover interesting patterns of relationships among them, evident in both networks.

I. INTRODUCTION

The wide adoption of mobile technologies have completely redesigned the way we consume and produce information as well as the way we interact with people. This shift and cultural change has led to the emergence of many new Social Media platforms that focus on features and topics that the traditional ones like Facebook and Twitter are lacking, with typical examples including Snapchat, Periscope, and musical.ly. Many of these platforms operate solely on mobile devices.

Social Live Streaming Services (SLSS) are examples of this new type of platforms in which users can actually live stream parts of their daily lives. These services provide a new level of interaction and hook their subscribers as the users become part of the daily life of others. Practically, users decide when to open up their cameras and share snapshots of what they do, what they think or live at the moment with others and interact with them via chat messages.

In this work, we analyze a gray area of these services: adult content production and consumption. Clearly, most SLSS have a clear policy against adult content and facilitate some mechanisms to detect and ban the misbehaving users, either in the form of filters from the service provider or by peer reporting.

To this end, we consider two SLSS, Live.me (www.liveme.com) and Loops Live (www.loopslive.com), from now on LM and LL, respectively. Both operate as video chat apps in mobile phones and have millions of users that produce massive amounts of video content on a daily basis. To quantify the latter, Cheetah Mobile’s CEO; the company which owns Live.me, reported that more than 200,000 hours of live video

are broadcast daily on Live.me¹. Both platforms are very successful, especially in young users². LM was ranked as the top grossing social app in the U.S. on Google Play since August 2016 and on the top five social apps on App Store.

Both these apps share many similarities regarding community policies, e.g., they explicitly forbid broadcasters from engaging in, or broadcasting, any sex-related content that promotes sexual activity, exploitation and/or assault. Moreover, both apps prohibit violence and/or self-harm, bullying, harassment, hate speech, on-screen substance use, posting of private contact information, prank calls to emergency authorities or hotlines and solicitation or encouragement of rule-breaking. There is a variation on the user’s age, as in LM users have to be at least 18 while in LL the users have to be at least 13 years old.

To counter possible violations of the aforementioned policies, both services have implemented reporting mechanisms, so that users can easily report a channel once they identify an underage user or detect suspicious behavior, or violations of the service policies. On top of that, LM employs a team of human moderators around the world, working 24/7 to respond to users’ reports. Violators are subject to immediate suspension or ban from the app. Those safeguards are in place to protect young people, since live streaming apps and sites can expose them to graphic and distressing content and can leave them vulnerable to bullying and online harassment [1].

However, these mechanisms do not seem to be working as intended. Many users report in the app reviews that they are constantly witnessing violations of the aforementioned policies. It is therefore a challenge to design detection mechanisms of deviant behavior that scale up to the massive amounts of streamed video data produced in these services.

Main Contributions: In this work we perform an in depth analysis of two SLSS to understand and characterize deviant behaviors involving the production and consumption of adult content in these platforms. To the best of our knowledge, this is the first quantitative study of deviant behaviors in SLSS. First, we collect two large datasets with user profile information and directed friendship links for LM and LL, following a sampling scheme that enables us to sufficiently cover the relevant part of social graphs. Next, we use a deep learning classifier to automatically identify producers of adult content from the available broadcast replays, and compare our findings with

¹<https://seekingalpha.com/article/4075406-cheetah-mobiles-cmcm-ceo-fu-sheng-q1-2017-results-earnings-call-transcript>

²<https://seekingalpha.com/article/4025223-cheetah-mobiles-cmcm-ceo-fu-sheng-q3-2016-results-earnings-call-transcript>

the moderation (banned users) of each platform. While our results are consistent with the moderation of LL, we observe many cases of undetected deviant behavior in LM. Moreover, we characterize adult content producers and consumers based on their profile attributes, and analyze their relationships to discover interesting patterns.

A. Ethical considerations

Clearly our methodology has the capacity to collect large bodies of data, including streams, messages and metadata exchanged between individuals around the world. There are therefore certain privacy considerations that must be taken into account. To anonymize users, we allocated a new unique random identifier for every user whose data we collected, obfuscating her platform-wide identity (user ID). We highlight that the terms of both services underlines that all data (and metadata for LM) and activity are by default public. Despite their “public” nature, we follow Zimmer’s approach [2]. In this regard, the data remains anonymized during all the steps of our analysis, and we report only aggregated information. The collected datasets are publicly available online ³.

II. RELATED WORK

As the adult content problem on SLSS has not been studied in the literature, we loosely categorize prior work into two main categories, reflecting the fundamental concepts present in this study. Finally, we provide a functional overview of the two platforms that we study.

A. Social Live Streaming Services

In SLSS users are able to stream their own live shows in real time as broadcasters, and to join the live shows of other users as viewers/audience. The audience is able to interact with the streamers through a chat and reward them with virtual rewards, e.g., points, gifts, badges (some of which are purchasable), or money. Also, various SLSS give broadcasters the opportunity to monetize part of the virtual gifts they receive from the audience during their broadcasts. Users of SLSS employ their own mobile devices (e.g. smartphones, tablets) or their PCs and webcams for broadcasting. In contrast to other social media, SLSS are mostly synchronous [3], [4], but they can also support asynchronous interactions between users, like direct messages and comments on broadcast video replays.

We differentiate between two kinds of SLSS: **General live streaming services** (without any thematic limitation), e.g. YouNow, Twitters Periscope, Cheetah Mobile’s Live.me, (now-defunct) Meerkat Streams, YouTube live or IBM’s Ustream, and **Topic-specific live streaming services**, e.g. Twitch (games), or Picarto (art).

Since SLSS are quite new, the literature in the field is rather limited. Some of these studies investigate the performance of such services, e.g., Meerkat and Periscope [5], [6], Periscope [7] and Twitch [8]. Human factors and user experience were studied in [9]. Having access to a large dataset of Inke, a Chinese SLSS, [10] identified several patterns in the users, e.g., fast interest shifts, user dedication to broadcasters as well as the locality bonds between users.

[11] analyzed traffic patterns and user characteristics of YouNow. [12] crawled Inke and identified that the main reasons that users are hooked in these services are the follower-follower model, the awards incentivisation, and the multi-dimensional interaction between broadcasters and viewers. Similar results, but with real users, were also reported by [13] for the case of Facebook Live, Periscope, and Snapchat.

Legal and ethical questions about SLSS were raised by [14]. Recently, [15] performed an empirical study on law infringements in several SLSS. While the focus was not on adult content, the researchers found that around 17.9% of their sample, consisting of more than 7,500 streams, somehow violated a law, e.g., copyright, road traffic, insult, etc. Different information behaviors of users, focusing on the assessment of streamers’ behavior with emphasis on produced content and motivations, as well as demographics, were studied in [3], [4]. The copyright aspect is also studied in [16], but in terms of broadcasting sport events.

B. Adult content in Social Media

In the computer science literature, adult content consumption has mostly been studied in the context of adult websites, several of which incorporate social networking functionality and features. Examples include the work by [17] that provides an overview of behavioral aspects of users in the PornHub social network, a recent paper [18] on the detection of fake user profiles in the same network, and various studies on the categorization of content, frequency of use, and analysis of user behavior in such platforms [19], [20]. To the best of our knowledge, the only other work that studies the production and consumption of adult content in general-purpose online social networks is a recent article by [21]. The authors perform a large-scale analysis of the adult content diffusion dynamics in Tumblr and in Flickr, while also examining and comparing the demographics of adult content producers and consumers across these platforms. A wider corpus of research has been produced by social and behavioral scientists, mostly based on surveys of relatively small numbers of individuals.

C. Live.me & Loops Live functional overview

This study uses data collected from LM and LL platforms, introduced previously. Most of the features and functionality offered by those platforms are mobile-only, in that users wishing to actively participate in their communities need to own mobile devices such as smartphones and tablets running on Android or iOS.

The dynamics of both communities are based mostly on three possible actions performed by the users: **(a)** create real-time broadcasts and optionally associate hashtags representing thematic categories/user interests with them; **(b)** join broadcasts created by other users and interact with them as well as with the other viewers. Those interactions include exchanging chat messages with other viewers, and rewarding the broadcasters with “likes” and purchasable virtual gifts; and **(c)** follow other users and receive notifications when they are broadcasting.

Contrary to other popular SLSS like Periscope [5], all the broadcasts in LM and LL are public. All active broadcasts are visible on a global public list. In both platforms, the concept of

³<https://github.com/nlykousas/asonam2018>.

re-sharing/re-posting broadcasted content across different users is not present. Nevertheless, users are able to get shareable links to live shows that can be used for promoting broadcasters on other social media.

As already discussed, both platforms enable users to report community policy violators and underage users, who consequently get their accounts banned after their activity has been reviewed by moderators. Additionally, LM offers safety features to proactively protect its users, like the “Admin” feature, which enables broadcasters to allow other trusted users to be administrators for their broadcasts to block commenters on their behalf in real time.

Both platforms are equipped with more advanced features. Some significant examples are the ability to view currently popular/trending or “featured” broadcasts, either globally (both services), or by geographical region (LM), or by hashtag (both LM and LL) and the ability to find users or hashtags matching a search term. The mechanics of the broadcast featuring system are different for each platform, but in both cases factors such as the number of viewers, the amount of user interaction within the broadcast including likes, gifts and messages, and the duration of the live show are taken into account. Moreover, the popularity and experience of a user is reflected by their “level”, which is determined by their participation in activities such as broadcasting, joining broadcasts of others, sending and receiving gifts, chatting, etc. Leveling up enables users to receive various privileges such as discounts for buying virtual currency and access to premium gifts.

Broadcasters have the incentive to get their live shows featured, since this leads to a better visibility within the app, thus attracting a higher number of viewers who in turn can potentially reward them with virtual gifts. Once a broadcaster has received a certain amount of virtual gifts, they are able cash them out for real money. Both platforms offer a range of synchronous interaction features traditionally provided from the majority of OSNs like direct messaging between users and the ability to “block” users. Finally, edges in LM and LL social graphs are directed; users can follow other users who do not follow them back. In addition, following someone does not require their permission. In the context of this study, we focus specifically on the user-specific attributes and following-follower social graphs of those platforms.

III. DATA COLLECTION METHODOLOGY

In this section, we first describe our methodology for collecting and labeling the data. The objective of our data collection methodology is twofold: to identify adult content producers by analyzing the available broadcast replays and to sufficiently sample the portion of the social graphs where adult content production and consumption phenomena are predominant. To accomplish this, we develop a novel data collection and labeling approach which we detail in the following subsections.

A. Sampling the social graphs

Both LM and LL applications communicate with their servers using an API with SSL-protected access. To the best of our knowledge, no open-source clients for these services exist at the time of writing, hence, we follow a similar method

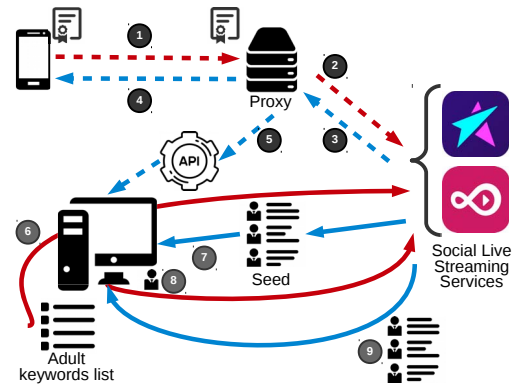


Fig. 1: Data collection methodology. A proxy intercepts the messages from the smartphone to the SLSS. To decrypt the traffic and derive the API, a root certificate is installed in the smartphone (steps 1-6). Then, we use an adult keyword list to get an initial set of seed users, which are then queried to collect even more users. Based on their properties (e.g. banned, gifts) we build our dataset (steps 7-9).

as in [6]. For each platform, we analyze the network traffic between the app and the service. More precisely, we set up a so-called SSL-capable man-in-the-middle proxy between a mobile device with the specific apps installed and the LM and LL services that acts as a transparent proxy. The proxy intercepts the HTTPS requests sent by the mobile device and pretends to be the server to the client and the client to the server, enabling us to examine and log the exchange of requests and responses between the client apps and the servers.

We select a set of APIs that allow us to crawl the social graph edges, extract user profile and broadcast information, and use the search capabilities offered by the services. Content-wise, both services use the HTTP Live Streaming (HLS) protocol [22] for hosting and delivering broadcast video replays, similar to other well known live streaming services such as YouNow, Periscope and Twitch. Figure 1 illustrates this architecture (steps 1-5).

We first identify a set of *seed nodes* likely to be involved with the production of adult content, in order to bootstrap a subsequent crawling procedure for sampling the social graphs. To accomplish this, we take advantage of the aforementioned search APIs. A seed node is defined as a user that satisfies *the three following conditions*: (i) having a username that contains a pornographic term, (ii) having broadcasted activity, and (iii) being banned by the system⁴. For the first condition, we use the list of adult keywords provided by [21] in the context of their proposed deviant graph extraction procedure. This list contains 5,283 search keywords from professional adult websites.

Using these three criteria, we were able to identify 390 and 47 seed nodes for LM and LL, respectively. Figure 1 (steps 6-7) illustrates the seed identification step. Note that this step does not consider the network structure. The next step (denoted as 8-9 in Figure 1) consists in traversing and collecting profile information as well as broadcast video replays from each user, following the friendship links. We follow a Breadth-First (BF)

⁴In both services, although the accounts of banned users are deactivated, their past activity in the platform is still retained, thus enabling us to perform the described analysis.

TABLE I: Network statistics of the crawled graphs: number of nodes $|N|$, number of edges $|E|$, number of banned nodes $|B|$, average degree $\langle k \rangle$, density D , and reciprocity ρ .

	$ N $	$ E $	$ B $	$\langle k \rangle$	D	ρ
LM	2,942,407	37,440,992	142,345	25.4	4.32×10^{-6}	0.14
LL	273,177	1,193,780	114	8.73	1.59×10^{-5}	0.08

traversal limited to two hops away (undirected distance) from the seed nodes. Thus, our network consists of the union of the 2-hop ego-networks of all seed nodes. This union resulted in one single connected component in both platforms. For computational reasons, we discard those nodes in the boundary that appear as neighbors of a node with degree higher than $10K$. These nodes correspond to only 718 and 267 profiles for LM and LL, respectively, a very small proportion of the complete 2-hop ego-networks.

We emphasize that our interest is not in capturing the entire network of users, but a tractable subset of tightly connected groups of users in which adult content is predominant. BF search covers satisfactorily small regions of a graph [23] and has been used in many analyses.

During the data collection period, which lasted from Jan. to Nov. of 2017, we repeated this crawling procedure once per week on average. Based on the number of installations reported by LM on Google Play ($20M - 50M$ installations), we managed to crawl roughly $5.8\% - 14.5\%$ and $5.46\% - 27.3\%$ of the entire LM and LL networks, respectively.

Table I summarizes the obtained networks for both platforms. As expected, the LM network is much larger than the LL network, containing approximately 10 times more users and 30 times more edges. The LL network is, however, one order of magnitude more dense than the LM one.

The approach we followed has three main limitations. Firstly, the set of replays that we captured includes only the available replays of past broadcasts at crawling time. Replays that were deleted in-between our crawls as well as all live broadcasts streamed during our crawls were not included. This is not a fundamental limitation, and can be fixed by using more sophisticated approaches [5]. Moreover, while we can determine whether an account is banned (suspended) or active, none of the platforms provides metadata to determine the reason behind the account suspension. This means that our dataset includes false positives that were banned because of other unrelated policy violations. This limitation is addressed in the next subsection, in which we consider the replay’s content to determine whether a user is deviant or not. Finally, there is a small probability of false negatives, a portion of deviant users that are not retrieved by our method. This can happen because moderators can only identify a limited number of users engaging in inappropriate behavior [24] and those may lie isolated (more than two hops away) from the seed nodes.

B. Labeling the users

Having described our procedure to identify an adult content related network, we now describe how we label the users within this network. We differentiate between three types of users: adult content producers, or simply *producers* (based on their broadcast activity), *consumers* (based on their relation

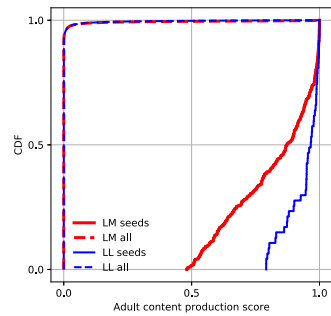


Fig. 2: Cumulative distribution function of the adult content score values.

with producers), and *normal* users that are not included in any of the two other categories.

One could argue that these consumers are lurkers. The lurking phenomenon in social networks has been studied in great depth [25], [26]. In general, lurkers are passive users who do not contribute to the community. While consumers in our scenario could also be lurkers, we argue that, despite the obvious resemblance, they are not. Indeed, their behavior seems passive as they do not create content. Nevertheless, they have actual interactions by, e.g., providing praise and currencies to producers, or by publicly chatting with the producers, that promote specific behaviors and content.

Given the network, our criterion to establish whether a user is a producer is exclusively based on the images of the user’s broadcast activity. This choice disregards indirect sources of information and does not require manual inspection, allowing us to scale up the method efficiently. Alternative approaches are based on manual inspection of metadata only [21], which may not be sufficient for our purposes, or using crowdsourcing approaches for categorizing broadcasts [4], which would require a pool of crowd workers to be potentially exposed to offensive material.

To this end, we use OpenNSFW⁵, a deep neural network model pre-trained to detect pornographic images. Convolutional Neural Networks are the state of the art in image classification problems [27], [28]. OpenNSFW takes an image as input and provides a value representing confidence in an image’s resemblance to pornography. We feed the network with frames sampled from the broadcasts at 1/3 Hz, and keep the *highest* confidence score for every broadcast replay. This value represents the maximum probability a replay contains pornographic content. Then, at the end of our data collection period, we can associate every user with the *highest* value provided by OpenNSFW over all of their replays in the dataset. The aforementioned value can be considered as a user’s *adult content production score*. For the users we were unable to collect any broadcast data, we set this value to zero.

Figure 2 shows the cumulative distribution function (CDF) of the adult content production score for both LM and LL networks differentiating between seed users and all the users in our sample. We observe that a very small proportion of all users (only around 0.4%) scored above 0.5, indicating that the

⁵<https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for>.

TABLE II: Distribution of users according to their class.

Class	Live.me	Loops Live
Producers	7,135 (228 seeds)	92 (33 seeds)
Consumers	30,872 (0 seeds)	1,243 (0 seeds)
Normal	2,904,400 (162 seeds)	271,842 (14 seeds)

vast majority of users do not broadcast adult related material. On the other hand, the seed users have been assigned high scores (all starting at 0.5 for LM and around 0.8 for LL). This confirms that our choice of seed nodes and the outcome of the classifier agree to a large extent.

We establish whether a user is a *producer* using a thresholding approach. In particular, we consider the probability distribution of the scores for both banned and non-banned users and choose as a threshold the Bayesian decision rule that separates the two classes. This results in a threshold of 0.82 and 0.93 for LM and LL, respectively. Although the described approach does not require any human supervision, in order to further evaluate the “goodness” of the threshold, we manually inspected the frames for 100 LM and 50 LL random producers. All of them contained either nudity or semi-nudity, suggesting the validity of our thresholding method.

We establish whether a user is a *consumer* based on the set of producers and the network structure. In particular, we label a user as a consumer if the user follows *at least two* adult content producers. While following a single user (producer or not) can be expected by random chance, following two users of the producer class (given they only make up for a minor fraction of the total users), is much less likely to be by chance. Our definition of consumer is stricter than the one of [21], which defines as a *passive consumer* a user that follows *at least one* single producer. In our analysis, for those users that fall in both categories, i.e., producers that also followed at least two other producers, the producer label is considered more relevant. In practice, only 9 users of LM and 2 users of LL fall in both categories. Finally, users who do not fall into the above classes are labeled as *normal* users.

Table II summarizes the resulting labeling according to our proposed procedure. As expected, we observe that only a small proportion of the crawled networks are not labeled as normal users. We also show in parenthesis how the seed nodes are distributed in the three categories. Recall that seed users are banned users with broadcast activity and with a adult-related username. Although most of them are labeled as producers, there is a significant proportion labeled as normal users. This can be explained by the fact that those users may exhibit other (non adult-related) deviant behaviors and thus not relevant for our analysis, or because their score did not reach our threshold, as reported from the OpenNSFW classifier. Remarkably, none of them are labeled as consumers, which already suggests that producers are not well connected between them.

C. Effectiveness of SLSS moderation systems

Having identified the aforementioned user classes, we proceed to examine how our labeling approach compares to the moderation of each platform. Table III shows how banned users are distributed in each class. In the case of LM, we observe that only 43.5% of the labeled producers have been banned. Since it is unlikely that the frames extracted from the

broadcasts contained adversarial perturbations [29] against the OpenNSFW model, we can safely assume that moderation is highly ineffective in detecting such cases.

On the contrary, moderation of LL is consistent with our labeling outcome, with 96.7% of users placed in the producers class being banned. This consistency provides further confirmation of our decision to use a pre-trained deep learning classifier for detecting adult content. Finally, the high number of banned users placed in the *normal* class suggest the existence of a significant proportion of policy violators outside the context of our study.

IV. PROFILING DEVIANT USERS

In this section, we present our efforts to characterize adult content producers, consumers and their relationships in the sampled networks. We first consider a set of features directly accessible from each user and analyze their relevance for distinguishing between classes: normal users, producers, and consumers. We then look at the network structure to gain understanding about the relations between consumers and producers.

A. Features

Based on the available profile information we collected from the two platforms, we define a set of features that can be grouped as follows:

- **Network features:** Number of followers, number of followings, number of bidirectional friends.
- **User-based features**
 - Pornographic username (binary): *whether the username contains a pornographic term.*
 - Suspended/Banned (binary): *whether the account has been suspended by platform moderators.*
 - Replay count: *Number of past broadcasts available for replaying.*
 - Level: *An integer value reflecting the participation level of a user in various SLSS-specific activities.*
 - Praise (**only LM**): *Total number of likes received in all user’s broadcasts.*
 - Income (**only LM**): *Total virtual currency value of gifts received in all of user’s broadcasts.*

We assessed the relative power of these features in discriminating the three user classes by using the Mean Decrease Impurity (MDI) metric, where a higher score implies a more important feature. Table IV reports the ranking of the top five most important features in differentiating the three user classes for each platform.

The number of followings, followers and friends are among the highest ranked features for both networks, which suggests relevance of social relationships for characterizing the given

TABLE III: Proportion (total in parenthesis) of banned accounts in each class.

Class	Live.me	Loops Live
Producers	43.5% (3,109)	96.7% (89)
Consumers	9.6% (2,970)	0.08% (1)
Normal	4.6% (136,266)	0.008% (24)

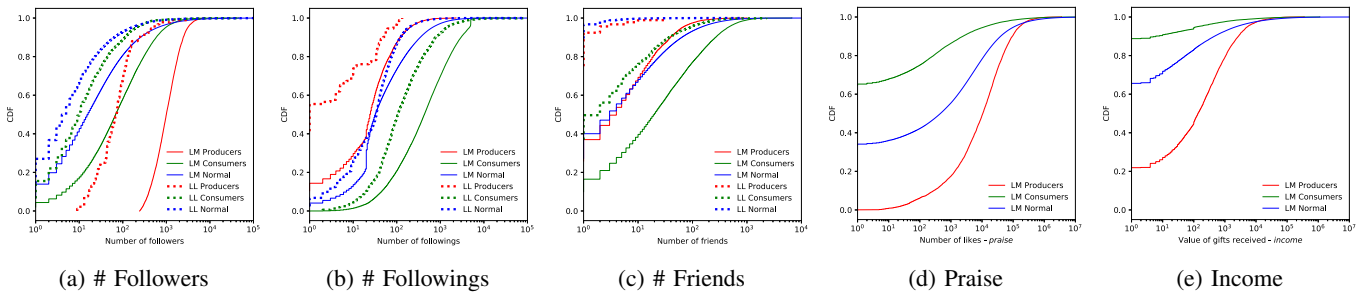


Fig. 3: Cumulative distribution functions (CDFs) of the different user profile attributes.

TABLE IV: Top 5 features for differentiating the three classes.

Rank	LM		LL	
	Feature	MDI	Feature	MDI
1	#Followings	0.31	#Followings	0.37
2	#Followers	0.25	#Friends	0.26
3	Praise	0.15	#Followers	0.19
4	#Friends	0.12	Banned	0.10
5	Income	0.06	Porn nickname	0.05

classes. Also, for LM we observe that the amount of likes (praise) and virtual gifts (income) are highly important as well. To get a deeper insight on how these features are distributed across the different classes, we plot their cumulative distribution functions (CDFs) in Figure 3. Information about praise and income was not available for LL, preventing us from performing a 1-to-1 comparison between the two datasets. Instead, we observe a high importance for the banned and pornographic username attributes. This is due to the fact that, as shown in Subsection III-C, the banned LL users are almost exclusively adult content producers, and also a significant part of them have pornographic usernames (see Subsection III-A).

From Figure 3a, we observe that producers tend to have many more followers than the other classes, and there exists a lower bound to the follower number of producers, approximately 10 and 250 for LL and LM, respectively. In contrast, consumers in LM are found to have the least amount of followers among the three classes. For the number of followings (Figure 3b), however, the situation is reversed. Consumers dominate over the other classes by following significantly more users, while the producers come last in this aspect with around 41% (LL) and 10% (LM) of them not following any other users. The distribution of the friend number reveals that consumers are much more likely to form reciprocal relationships, while it appears to be almost identical for producers and normal users, as Figure 3c indicates. Furthermore, we found that adult content producers tend to receive the highest amount of praise and income among the three classes. We note that, although the higher (undirected) degree of consumers and producers is explained by the criteria used in the seed selection, the edge directionality can not be fully attributed to our sampling method, which is blind with respect to it.

Additionally, consumers receive much less recognition for their broadcasting activities compared to both normal users and producers. In fact, while no producers with zero praise exist, approximately 65% of consumers and the 33% of normal users in our dataset fall in this “unpopular” category, see Figure 3d. This either means that they have not received any likes during their shows, or they have never broadcasted anything. A similar

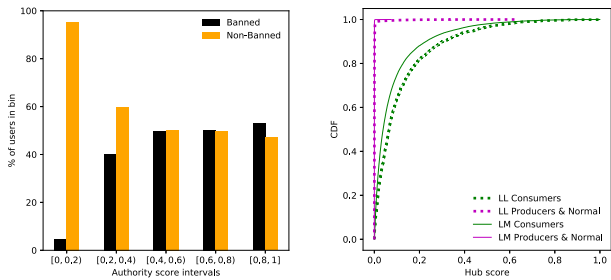
trend is observed for the total value of the virtual gifts received, represented by the income attribute and illustrated in Figure 3e. Only 21% of producers have not received gifts, while the same holds for the 88% of consumers and the 65% of normal users.

B. Deviant relationships

To determine the community structure of these networks, existing variants of the Louvain method [30] do not find well identifiable clusters of users. In both networks, we observe that producers and consumers are distributed nearly uniformly across the clusters. Further, the results vary significantly between different runs. We thus adopt a different approach in order to better understand the network structure.

In particular, we examine who in the sampled networks is significant with regards to their social relationships. We use the ranking HITS algorithm [31] to identify the hubs and authorities in the social graphs. The basic principle behind HITS algorithm is the following mutually reinforcing relationship between hubs and authorities: good hubs point to many good authorities and vice-versa. Interestingly, it appears that adult content consumers have the highest hub scores among all users in both networks, as shown in Figure 4b. Moreover, the highest authority scores in LM belong almost exclusively to producers. The latter could be correlated with the significance of the number of followers and followings to discriminate producers and consumers, from normal users, as previously shown. For LL, we observe that most authoritative users do not belong to the producers class, and exhibit characteristics expected of prominent users in a social community such as the number of followers in the order of hundred of thousands. The reason behind this difference between LM and LL could be attributed to the very limited extent of the adult content production behavior in the later. Therefore legitimate popular users dominate the authority scores in the sampled graph by being followed by consumers. Additionally, we notice that the hub scores of the highly authoritative users are particularly low in both networks. This finding contradicts other studies on different social networks such as Twitter [32], [33], where researchers observed many well-connected users that have high scores as both authorities and hubs.

An interesting finding was that in LM, the ratio of banned to non-banned users increases along the increase of authority score. To better demonstrate this, we bin the users based on their authority score and we calculate the fractions of banned and non-banned users in each bin, as shown in Figure 4a. We observe that 99.5% of users in our sample fall into the first bin, having authority score less than 0.2. We can thus conclude



(a) authority score bins (LM) (b) CDFs of hub scores
 Fig. 4: User relationship insights, provided by HITS.

TABLE V: Comparison in terms of link density D between the crawled networks of producers, consumers normal users and a corresponding random network.

	Class	Crawled graph D	Null model D
LM	Producers	1.90×10^{-5}	4.55×10^{-6}
	Consumers	1.20×10^{-3}	4.46×10^{-6}
	Normal	2.18×10^{-7}	4.32×10^{-6}
LL	Producers	1.91×10^{-3}	0
	Consumers	3.28×10^{-3}	1.42×10^{-5}
	Normal	1×10^{-5}	1.59×10^{-5}

that banned users are more densely concentrated towards the higher end of the authority score spectrum, and the reason behind their suspension was likely the production of adult content, since they are followed by the consumers/hubs. A similar phenomenon is observed for LL, but with banned users mostly concentrated in the 0.02 – 0.35 authority score range, while the 97.7% of users have authority scores below 0.02.

Based on the arguments above, we expect that consumers will follow multiple producers, a considerable portion of which will be banned. To quantify the relationship between the fractions of banned users and producers followed by consumers, we calculate their correlation using Spearman’s rank correlation coefficient ρ . Indeed, there exists a nearly perfect correlation for LL with $\rho = 0.96$, meaning that consumers do not follow almost any banned users outside the producers class, and a moderately strong correlation in LM ($\rho = 0.63$).

Another dimension to examine is the connectivity within each class in the context of the sampled graphs. For this we measure the edge density, computed as the ratio of edges between the users belonging in each class over the total number of possible edges between them. To account for the differences in sizes of the subnetworks [34], we resort to a comparison of the connectivity of the sampled graphs with a null model that randomly rewires the edges while keeping the degree of each node unchanged, as described in [35].

Table V contains the link density comparison between the subgraphs induced by producers, consumers and normal users in each sampled network and the null model. We observe that consumers, when compared to the null model, are several orders of magnitude more densely connected to each other in the sampled networks. On the contrary, the subgraphs of producers and normal users are much more sparse compared to consumers, with the producers being only slightly more

dense connected than random for LM. This finding comes in contrast with the behavior of adult content producers in Tumblr and Flickr, where they are observed to form densely interconnected communities [21]. In LL the producers appear to have a density comparable to those of consumers, but given their limited number, this is due to the existence of producers who also exhibit consumer behavior.

In summary, we can conclude that the closely knit groups of consumers act as a “bridge” between the otherwise isolated producer nodes. Concretely, from a network perspective, the most effective way to reach adult content in the studied networks is by traversing the social links of consumer nodes that point to both producers and other consumers, enabling the reach of even more deviant nodes belonging in those two categories. Since adult content producers are isolated in the network, we speculate that some of the consumers are actively monitoring the list of active broadcasts (see Subsection II-C), and proceed to follow users who broadcast adult content, while also possibly sharing links to such live streams with other consumers. These “consumer leaders” are likely to become popular among their kin by being followed by many other consumers, thus serving as a means for diffusion of information about producers, effectively compensating for the absence of the content reposting functionality in SLSS.

V. DISCUSSION

With the continuous growth of SLSS, an increase in deviant behaviors on social media is expected. In our work, we show that current moderation mechanisms may have important limitations when addressing the detection of adult content consumption and production. Our approach overcomes scalability issues that appear when a large number of humans are needed to categorize the content, at the cost of relying on the accuracy of automatic image classification. Image classification is the primary application domain for machine learning [36], reaching human-level performance in many tasks. Our results could be further improved by replacing or accommodating the OpenNSWF classifier with more effective models.

The inefficiency of moderation can be partially attributed to a *voyeur* phenomenon. Many adult content producers are not reported to moderators as the consumers like the content, so their accounts are not suspended, allowing them to continue broadcasting inappropriate content. Moreover, although consuming any kind of content, including adult, is not explicitly prohibited by the community guidelines of these platforms, suspending the accounts of the users who intentionally seek adult content would be meaningful, due to the law of *supply and demand*. It is therefore necessary to incorporate effective, real-time detection mechanisms of deviant behaviors in the existing moderation systems, in order to maintain the SLSS communities safe, especially for the younger audience.

In future work, we will investigate quantitatively the identification between consumers and lurkers. Moreover, we plan to develop graph-based features for the detection and classification of adult content producers and consumers in SLSS by exploiting the characteristics of deviant behavior presented in this paper, as well as study other available data from broadcast-related user interactions in SLSS (chat messages, likes, gift exchange), to further analyze the nature of deviant behaviors in such platforms.

ACKNOWLEDGMENTS

We thank Andreas Kaltenbrunner for his helpful comments and suggestions. This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the Practicies project (Grant Agreement no. 740072), and by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502).

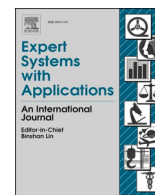
REFERENCES

- [1] S. Bearne, “Is live streaming your life good business or dangerous?” <http://www.bbc.com/news/business-39778550>, 2017.
- [2] M. Zimmer, ““But the data is already public”: on the ethics of research in Facebook,” *Ethics and information technology*, vol. 12, no. 4, pp. 313–325, 2010.
- [3] K. Scheibe, K. J. Fietkiewicz, and W. G. Stock, “Information behavior on social live streaming services,” *Journal of Information Science Theory and Practice*, vol. 4, no. 2, pp. 6–20, 2016.
- [4] M. B. Friedländer, “And Action! Live in front of the Camera: An Evaluation of the Social Live Streaming Service YouNow,” *International Journal of Information Communication Technologies and Human Development*, vol. 9, no. 1, pp. 15–33, 2017.
- [5] B. Wang, X. Zhang, G. Wang, H. Zheng, and B. Y. Zhao, “Anatomy of a personalized livestreaming system,” in *Proceedings of the 2016 ACM on Internet Measurement Conference*. ACM, 2016, pp. 485–498.
- [6] M. Siekkinen, E. Masala, and T. Kämäräinen, “A First Look at Quality of Mobile Live Streaming Experience,” in *Proceedings of the 2016 ACM on Internet Measurement Conference*. ACM, 2016, pp. 477–483.
- [7] L. Favario, M. Siekkinen, and E. Masala, “Mobile live streaming: Insights from the periscope service,” in *Multimedia Signal Processing (MMSP), 2016 IEEE 18th International Workshop on*, 2016, pp. 1–6.
- [8] J. Deng, G. Tyson, F. Cuadrado, and S. Uhlig, “Internet scale user-generated live video streaming: The Twitch case,” in *International Conference on Passive and Active Network Measurement*. Springer, 2017, pp. 60–71.
- [9] J. C. Tang, G. Venolia, and K. M. Inkpen, “Meerkat and periscope: I stream, you stream, apps stream for live streams,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 4770–4780.
- [10] M. Ma, L. Zhang, J. Liu, Z. Wang, W. Li, G. Hou, and L. Sun, “Characterizing user behaviors in mobile personal livecast,” in *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*, ser. NOSSDAV’17. ACM, 2017, pp. 43–48.
- [11] D. Stohr, T. Li, S. Wilk, S. Santini, and W. Effelsberg, “An analysis of the YouNow live streaming platform,” in *2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops)*. IEEE, oct 2015, pp. 673–679.
- [12] J. Zhao, M. Ma, W. Gong, L. Zhang, Y. Zhu, and J. Liu, “Social media stickiness in mobile personal livestreaming service,” in *Quality of Service (IWQoS), 2017 IEEE/ACM 25th International Symposium on*. IEEE, 2017, pp. 1–2.
- [13] O. L. Haimson and J. C. Tang, “What makes live events engaging on Facebook Live, Periscope, and Snapchat,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 48–60.
- [14] C. Faklaris, F. Cafaro, S. A. Hook, A. Blevins, M. O’Haver, and N. Singhal, “Legal and ethical implications of mobile live-streaming video apps,” in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 2016, pp. 722–729.
- [15] F. Zimmer, K. J. Fietkiewicz, and W. G. Stock, “Law infringements in social live streaming services,” in *International Conference on Human Aspects of Information Security, Privacy, and Trust*, 2017, pp. 567–585.
- [16] M. Edelman, “From Meerkat to Periscope: Does intellectual property law prohibit the live streaming of commercial sporting events?” *Columbia Journal of Law & the Arts*, vol. 39, no. 4, 2016.
- [17] G. Tyson, Y. Elkhatib, N. Sastry, and S. Uhlig, “Are people really social in porn 2.0?” in *9th International Conference on Web and Social Media, ICWSM*. AAAI Press, 2015, pp. 236–444.
- [18] W. Magdy, Y. El-khatib, G. Tyson, S. Joglekar, and N. R. Sastry, “Fake it till you make it: Fishing for catfishes,” in *ASONAM*. ACM, 2017, pp. 497–504.
- [19] M. Schuhmacher, C. Zirn, and J. Völker, “Exploring youporn categories, tags, and nicknames for pleasant recommendations,” in *Search and Exploration of X-rated Information : WSDM’13 Workshop Proceedings*. ACM, 2013, pp. 27–28.
- [20] G. Tyson, Y. Elkhatib, N. Sastry, and S. Uhlig, “Demystifying porn 2.0,” in *Proceedings of the 2013 conference on Internet measurement conference*. ACM Press, 2013, pp. 417–426.
- [21] M. Coletto, L. M. Aiello, C. Lucchese, and F. Silvestri, “On the behaviour of deviant communities in online social networks,” in *10th International Conference on Web and Social Media*, 2016, pp. 72–81.
- [22] R. Pantos and W. May, “HTTP live streaming,” <https://tools.ietf.org/html/rfc8216>, Tech. Rep., 2017.
- [23] M. Kurant, A. Markopoulou, and P. Thiran, “Towards unbiased BFS sampling,” *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1799–1809, 2011.
- [24] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Antisocial Behavior in Online Discussion Communities,” in *9th International Conference on Web and Social Media, ICWSM*, 2015, pp. 61–70.
- [25] A. Tagarelli and R. Interdonato, “Lurking in social networks: topology-based analysis and ranking methods,” *Social Network Analysis and Mining*, vol. 4, no. 1, p. 230, 2014.
- [26] D. Perna, R. Interdonato, and A. Tagarelli, “Identifying users with alternate behaviors of lurking and active participation in multilayer social networks,” *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 46–63, 2018.
- [27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [29] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, dec 2013.
- [30] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, mar 2008.
- [31] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, sep 1999.
- [32] K. Lee, P. Tamilarasan, and J. Caverlee, “Crowdturfers, campaigns, and social media: tracking and revealing crowdsourced manipulation of social media,” in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013, pp. 331–340.
- [33] A. Java, X. Song, T. Finin, and B. Tseng, “Why we twitter: Understanding microblogging usage and communities,” in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, ser. WebKDD/SNA-KDD ’07. New York, NY, USA: ACM, 2007, pp. 56–65.
- [34] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graphs over time: densification laws, shrinking diameters and possible explanations,” in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 177–187.
- [35] S. Xiao, G. Xiao, T. H. Cheng, S. Ma, X. Fu, and H. Soh, “Robustness of scale-free networks under rewiring operations,” *EPL (Europhysics Letters)*, vol. 89, no. 3, p. 38002, feb 2010.
- [36] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015.



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Large-scale analysis of grooming in modern social networks

Nikolaos Lykousas^a, Constantinos Patsakis^{a,b,*}^a Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece^b Information Management Systems Institute, Athena Research Center, Artemidos 6, Marousi 15125, Greece

ARTICLE INFO

Keywords:

Online grooming
Social networks
LDA
Text analysis
Emoji

ABSTRACT

Social networks are evolving to engage their users more by providing them with more functionalities. One of the most attracting ones is streaming. Users may broadcast part of their daily lives to thousands of others world-wide and interact with them in real-time. Unfortunately, this feature is reportedly exploited for grooming. In this work, we provide the first in-depth analysis of this problem for social live streaming services. More precisely, using a dataset that we collected, we identify predatory behaviours and grooming on chats that bypassed the moderation mechanisms of the LiveMe, the service under investigation. Beyond the traditional text approaches, we also investigate the relevance of emojis in this context, as well as the user interactions through the gift mechanisms of LiveMe. Finally, our analysis indicates the possibility of grooming towards minors, showing the extent of the problem in such platforms.

1. Introduction

The recent advances in telecommunications have unleashed the potentials of sharing and exchanging content, changing radically the way we interact with others online. By lifting many bandwidth barriers, users may generate and share arbitrary content and disseminate it instantly to millions of users. As a result, we see Social Networks and Media's dominance in various aspects of our daily lives.

This radical shift and penetration of mobile devices have led millions of people and youngsters to use them on a daily basis. While most social networks have specific policies about use from minors, in practice, this policy is bypassed. Minors declare fake ages to register to service providers and end up using the services as normal users. While this might not be noticed or overseen by service providers, this is not the users' case. Unfortunately, thousands of users maliciously target minors. Of specific interest is the case of *grooming*. Grooming refers to the process by which an offender prepares a victim for sexually abusive behaviour. More precisely, according to Craven et al. (2006):

[Grooming is]...a process by which a person prepares a child, significant others, and the environment for the abuse of this child. Specific goals include gaining access to the child, gaining the child's compliance, and maintaining the child's secrecy to avoid disclosure. This process serves to

strengthen the offender's abusive pattern, as it may be used as a means of justifying or denying their actions...

Apparently, child grooming is of extreme importance due to the impact that it can have in the children's lives. In fact, despite the measures that social networks might have already taken, they do not seem to be successful at all.¹ To this end, it is necessary to investigate how grooming in social networks works and how groomers manage to bypass the policies and filters set by social networks. In terms of verbal content, currently, there is only one available dataset from the Perverted Justice website.² The organisation behind this website, Perverted Justice Foundation, Inc., has recruited volunteers to carry out sting operations. They appear as minors to several online services and record the interactions with them. Their operations have made a tremendous positive impact as they have led to the conviction of more than 620 offenders. While undoubtedly, this is a huge contribution, the problem persists, and the provided dataset is rather old to be used for modern filters.

1.1. Motivation

The past few years, there is a steady increase of reports in mainstream media and officials³ regarding the exploitation of social networks for grooming. The problem regardless of the age factor is rather big and

* Corresponding author.

E-mail addresses: nlykousas@unipi.gr (N. Lykousas), kpatsak@unipi.gr (C. Patsakis).¹ <https://www.bbc.com/news/uk-47410520>.² <http://perverted-justice.com/>.³ <https://www.nspcc.org.uk/what-we-do/news-opinion/3000-new-grooming-offences/>.<https://doi.org/10.1016/j.eswa.2021.114808>

Received 1 April 2020; Received in revised form 15 December 2020; Accepted 28 February 2021

Available online 6 March 2021

0957-4174/© 2021 Elsevier Ltd. All rights reserved.

stigmatises the life of thousands of people. The emergence of new social networks, allowing live streaming to potentially thousands of users along with traditional chatting and appraisal methods of traditional social networks can be further exploited for grooming.

The findings discussed in Lykousas et al. (2018) demonstrated that the moderation systems used by the LiveMe platform at that time were highly ineffective in suspending the accounts of deviant users producing adult content. Notably, in the same year, FOX 11; a major mainstream media outlet, reported that (Melugin, 2018):

A FOX 11 investigation has found that pedophiles are using the popular live streaming app LiveMe to manipulate underage girls into performing sexual acts, reward them with virtual currency, and then post screen captures or recordings of the girls online to be sold and distributed as child porn.

As such, it is reasonable to assume that the adult content problem and the sexual grooming behaviours identified by FOX 11 are related to some extent. In this work, we aim to unveil communication patterns of sexual groomers in the context of social live streaming services.

1.2. Main contributions

Our work primarily aims to identify and disentangle the mechanics of grooming and predatory behaviours in the context of Social Live Streaming Services, by analysing the behavioural and communication patterns of viewers, at a broadcast-level. Therefore, user-level detection of groomers falls beyond the scope of our work. It has to be noted though that a distinctive difference of grooming in Social Live Streaming Services is that is not performed through one-to-one interaction with the victim, but many-to-one.

Based on the above, the contributions of this work are multifold. First, we facilitate research in this field and the generation of new filters and algorithms to detect such predatory behaviour through the release of a large-scale dataset of both verbal and non-verbal interactions (e.g. likes and rewards) in a Social Live Streaming Service. Due to its nature, the dataset is available only to researchers and law enforcement agencies upon request via Zenodo.⁴ Second, we analyse the basic characteristics of the verbal content. Our analysis illustrates how such predatory behaviour bypasses the filters of service providers by, e.g. altering some “bad words”, or by using emojis. Notably, to the best of our knowledge, this is the first work highlighting the role of emojis in grooming. Then, based on our analysis, we manage to identify chats where grooming is performed. Moreover, we analyse non-verbal interactions between users that differentiate chats where grooming is performed from the others. Finally, our analysis shows that it is possible to identify illegal actions, such as the grooming of minors.

1.3. Organisation of the article

The rest of this work is organised as follows. In the next section, we provide an overview of the related work on the detection of deviant behaviour and grooming. Then, we provide the legal and ethical justification of collecting such data and GDPR compliance assessment. Section 4 provides an overview of our dataset. In Section 5, we analyse our dataset and provide some insight into it. Afterwards, in Section 6, we investigate possible modelling of grooming behaviours using both verbal and non-verbal features. Finally, the article concludes summarising our contributions and discussing ideas for future work.

2. Related work

Coletto et al. (2017) aimed at going beyond previous studies that

considered deviant groups in isolation by observing them in context. In particular, they attempted to answer questions relevant to the deviant behaviours related with pornographic material in the social media context, such as i) how much deviant groups are structurally secluded from the rest of the social network, and what are the characteristics of their subgroups who build ties with the external world; ii) how the content produced by a deviant community spreads and what is the entity of the diffusion which reaches users outside the boundaries of the deviant community who voluntarily or inadvertently access the adult content, and iii) what is the demographic composition of producers and consumers of deviant content and what is the potential risk that young boys and girls are exposed to it. Very interestingly, they find that while deviant communities may have limited size, they are tightly connected and structured in subgroups. Moreover, the content which is first shared in these groups soon reaches a broad audience of not previously considered deviant users.

The proliferation of the Internet has transformed child sexual abuse into a crime without geographical boundaries. Child sex offenders turning to the Internet as a means of creating and distributing child pornography has allowed the creation of a network of support groups for child sex offenders, when historically, this was an offence that occurred in isolation (Westlake et al., 2016). This concern was echoed by Mitchell et al. (2010), who recognised that a small percentage of offenders used social networking sites (SNS) to distribute child pornography. While there is scientific debate on whether the online predator is a new type of child sex offender (Quayle et al., 2000) or if those with a predisposition to offend are responding to the opportunities afforded by the new forms of social media (Cooper, 1998), empirical evidence points to the problem of Internet-based paedophilia as endemic. Recent work, such as Winters and Jeglic (2017), Zambrano et al. (2019), shows that nearly half of the offenders who had committed one or more contact offences, i. e., they had directly and physically abused children, had displayed so-called “grooming behaviour”.

However, when investigating the possibility of developing automated methods to detect grooming online, researchers are confronted with many issues. First, only one benchmark dataset contains (English) chat conversations written by child sex offenders, the PAN 2012 Sexual Predator Identification dataset, which leverages data from PJ. Concretely, PJ data comprises a single class of chats in the context of PAN 2012 data. Yet, because the victims were actually adult volunteers posing as children, it is likely that these conversations are not entirely representative of online predator-victim communications (Pendar, 2007). Moreover, since the seduction stage often shows similar characteristics with adults’ or teenagers’ flirting, initial studies trying to detect predatory behaviour directly on the user level typically resulted in numerous false positives when they were applied to non-predatory sexually-oriented chat conversations in the PAN 2012 dataset (Inches and Crestani, 2012).

For machine learning algorithms to identify online sexual predators effectively, they need to be trained with both illegal conversations between offenders and their victims and sexually-oriented conversations between consenting adults (Pendar, 2007). Since such data are rarely made public, initial studies (Pendar, 2007; McGhee et al., 2011) only experimented with the PJ data. The k-NN classification experiments based on word token n-grams performed in Pendar (2007) achieved up to 93.4% F-score (trigrams with $k = 30$) when identifying the predators from the pseudo-victims. Miah et al. were the first to include additional corpora in the non-predatory class (Miah et al., 2011). They included 85 conversations containing adult descriptions of sexual fantasies and 107 general non-offensive chat logs from websites like <http://www.fugly.com> and <http://chatdump.com>. When distinguishing between 200 PJ conversations and these additional chat logs, the Naïve Bayes classifier outperformed the Decision Tree and the Regression classifier, which resulted in an F-score of 91.7% for the PJ class. In Bogdanova et al. (2014), Peersman et al. (2012), Morris and Hirst (2012), Hidalgo and Díaz (2012), the researchers used a corpus of cybersex chat logs and the

⁴ <https://zenodo.org/record/3560365>.

Naval Postgraduate School (NPS) chat corpus and experimented with new feature types such as emotional markers, emoticons and imperative sentences and computed sex-related lexical chains to detect offenders directly in the PJ dataset automatically. Their Naïve Bayes classifier yielded an accuracy of 92% for PJ predators vs NPS and 94% for PJ predators vs cybersex based on their high-level features. However, both [Miah et al. \(2011\)](#) and [Bogdanova et al. \(2014\)](#) did not filter out any cues that were typical of the social media platforms from which the additional corpora were extracted, which could entail that their models were (to some degree) trained on detecting these cues rather than the grooming content. Moreover, because the high-level features described by [Bogdanova et al. \(2014\)](#) were (partially) derived from the PJ dataset itself, these experiments may have resulted in overestimated accuracy when detecting predators from the same dataset.

Recently, the detection of Internet child sex offenders has been extensively investigated in the framework of the PAN 2012 competition, during which efforts have been made to pair the PJ data with a whole range of non-predatory data, including cybersex conversations between adults ([Inches and Crestani, 2012](#)). Because the PAN 2012 benchmark dataset was heavily skewed towards the non-predatory class, most participants applied a two-stage classification framework in which they combined information on the conversation level to the user level ([Vilatoro-Tello et al., 2012](#)). Moreover, apart from one submission that used character-gram features, all other studies used (combinations of) lexical (e.g., token unigrams) and “behavioural” features (e.g., the frequency of turn-taking or the number of questions asked). [Morris and Hirst \(2012\)](#) achieved the best results using a Neural Network classifier combined with a binary weighting scheme in a two-stage approach to first identify the suspicious conversations and, secondly, distinguish between the predator and the victim. Their system achieved an F-score of 87.3%. However, during their study, they assumed that “predators usually apply the same course of conduct pattern when they are approaching a child” ([Morris and Hirst, 2012](#)), which is in contrast with research by [Gottschalk \(2011\)](#), which resulted in three different types of predators and, hence, of grooming approaches. Moreover, the PJ dataset was also not cleansed of platform-specific cues, which could again have led to overestimated F-scores during the competition. A more detailed overview of the PAN 2012 International Sexual Predator Identification Competition results can be found in [Inches and Crestani \(2012\)](#).

Concerning the content of predatory chat conversations, McGhee et al. were the first to investigate the possibility to detect different stages in the grooming process automatically ([McGhee et al., 2011](#)). Based on an expanded dictionary of terms they applied a rule-based approach, which categorised a post as belonging to the stage of gaining personal information, grooming (which included lowering inhibitions or re-framing and sexual references), or none. Their rule-based approach outperformed the machine learning algorithms they tested and reached up to 75.1% accuracy when categorising posts from the PJ dataset into one of these stages. A similar approach was used by Michalopoulos and Mavridis whose Naïve Bayes classifier achieved a 96% accuracy when categorising predatory PJ posts as belonging to either the gaining access, the deceptive relationship or the sexual affair grooming stage ([Michalopoulos and Mavridis, 2011](#)). The second task of the PAN 2012 competition consisted of detecting the specific posts that were most typical of predatory behaviour from the users that were labelled suspicious during the first task. To this end, most participants either created a dictionary-based filter containing suspicious terms ([Morris and Hirst, 2012](#); [Parapar et al., 2012](#)) or used their post-level predictions from the predator identification task ([Kontostathis et al., 2012](#); [Hidalgo and Díaz, 2012](#)). The best F-score was achieved by [Peersman et al. \(2012\)](#), who used a dictionary-based filter highlighting the utterances that referred to one of the following grooming stages: sexual stage, re-framing, approach, requests for data, isolation from adult supervision and age- and child-related references. Their approach resulted in a 35.8% precision, a 26.1% recall and a 30.2% F-score. Finally, [Elzinga et al. \(2012\)](#) proposed a method based on Temporal Concept Analysis using Temporal

Relational Semantic Systems, conceptual scaling and nested line diagrams to analyse PJ chat conversations. Their transition diagrams of predatory chat conversations seemed to be useful for measuring the level of threat each offender poses to his victim based on the presence of the different grooming stages.

Although these studies showed promising results, the issue remains that these methods are applied to a corpus that contains conversations between offenders and pseudo-victims. Hence, the adult volunteers that were posing as children could not accede to requests for “cammin”, sending pictures, etc. As a result, the PJ dataset contains hardly any conversations by groomers, because this type of offender typically does not invest much time in the seduction process and switches to a different victim when his needs are not fulfilled quickly. Moreover, it is highly likely that children would have responded differently to the grooming utterances than the adult volunteers did, which could have influenced the offenders’ language use.

2.1. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) is a type of generative probabilistic model proposed by [Blei et al. \(2003\)](#). It comprises an endogenous NLP technique, which as highlighted in [Cambria and White \(2014\)](#) “involves the use of machine-learning techniques to perform semantic analysis of a corpus by building structures that approximate concepts from a large set of documents” without relying on any external knowledge base. As the name implies, LDA is a latent variable model in which each item in a collection (e.g., each text document in a corpus) is modelled as a finite mixture over an underlying set of topics. Each of these topics is characterised by a distribution over item properties (e.g., words). LDA assumes that these properties are exchangeable (i.e., ordering of words is ignored, as in many other “bag of words” approaches in text modelling), and that the properties of each document are observable (e.g., the words in each document are known). The word distribution for each topic and the topic distribution for each document are unobserved; they are learned from the data.

Since LDA is an unsupervised topic modelling method, there is no direct measure to identify the optimal number of topics to include in a model. What LDA does is to assign to documents probabilities to belong to different topics (an integer number k provided by the user), where these probabilities depend on the occurrence of words which are assumed to co-occur in documents belonging to the same topic (Dirichlet prior assumption). This exemplifies the main idea behind all unsupervised topic models, that language is organised by latent dimensions that actors may not even be aware of [McFarland et al. \(2013\)](#). Thus, LDA exploits that even if a word belongs to many topics, occurring in them with different probabilities, they co-occur with neighbouring words in each topic with other probabilities that help define the topics better. The best number of topics is the number of topics that helps the most human interpretability of the topics. This means that if the topics given by LDA can be well-distinguished by humans, then the corresponding number of topics is acceptable. Researchers have recommended various approaches to establish the optimal k (e.g. [Cao et al., 2009](#); [Arun et al., 2010](#); [Deveaud et al., 2014](#); [Röder et al., 2015](#); [Zhao et al., 2015](#)). These approaches provide a good range of possible k values that are mathematically plausible. However, according to [DiMaggio et al. \(2013\)](#), when topic modelling is used to identify themes and assist in interpretation (like in the present study), rather than to predict a knowable state or quantity, there is no statistical test for the optimal number of topics or the quality of a solution. A simple way to evaluate topic models is to look at the qualities of each topic and discern whether they are reasonable ([McFarland et al., 2013](#)). In addition, the topic number selection was guided by the model’s ability to identify a number of substantively meaningful and analytically useful topics. In fact, the increase in fit is sometimes at the expense of interpretability due to overfitting ([Dyer et al., 2017](#)). Increasing the number of topics, producing ever-finer partitions can result in a less useful model because it becomes almost

impossible for humans to differentiate between many of the topics (Chang et al., 2009). Ultimately, the choice of models must be driven by the questions being analysed. DiMaggio et al. (2013) suggest that the process is empirically disciplined, in that, if the data are inappropriate for answering the analysts' questions, no topic model will produce a useful reduction of the data. To the best of our knowledge, the topic coherence measure with the most considerable correlation to human interpretability is the C_v score defined in Röder et al. (2015), which we also adopt in this study to establish the optimal number of topics, see Section 6.

3. Ethical and legal compliance

Data scraping from the web is extensively used by academic researchers to track the web, and companies to gain information about their customers. The philosophy of crawling is to index the web and the Internet as a whole, to make information available to the public, and to extract information for different business and research purposes. Yet, due to the invasive practices used for extracting large amounts of information, there is an ongoing debate on the ethical and legal aspects of web data crawling.

According to Internet advocates, if web crawling were to be unethical, then the whole web would not have been discoverable since the entire expansion of the Internet is based on web crawling. As a matter of fact, web scraping has benefited the web so much that virtually everyone on the net is directly or indirectly involved in web scraping. Even big service providers like Google scrap the Internet to be able to provide qualified and verified data in the search results. However, for web data crawling to be ethical, there must be some rules to be followed (like those imposed in the `robots.txt` file of every web site) to not infringe on the security and the rights of the users. In fact, there are already several professional web scraping service providers who abide by the general rules and regulations to get adequate and appropriate authorisation from the concerned web resource.

As a matter of fact, many scholars advocate that it is the application of the data that have been scrapped and not the web scraping *per se*, that may be unethical or illegal. For instance, there might be issues when data that are not meant to be made public are scraped and reused for commercial or other purposes. The legal issues of web scraping are widely discussed in the context of the copyrighted and data protection law. The latter is expressed in the EU by the GDPR, which defines the privacy and data protection rights and the rules to be respected when the processing of personal data takes place. While the GDPR is applicable even for research purposes, it states that for meeting "the specificities of processing personal data for scientific research purposes, specific conditions should apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes" (recital 159). Inevitably, when web crawling collects the personal data of web users to facilitate specific research purposes, this processing needs to be aligned with the data protection principles enshrined in the GDPR.

The GDPR requires a specific lawful basis for the processing of the personal data of individuals, with the consent to be the most commonly advertised among them. Beyond consent, however, the GDPR defines some other bases so as the processing of the personal data to be lawful: when the processing is necessary to protect the vital interests of the data subject or of another natural person (Article 6(1)(d)); when the processing is necessary for the performance of a task carried out in the public interest (Article 6(1)(e)); or when the processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party (Article 6(1)(f)). Therefore, while research is not explicitly designated as its own lawful basis for processing, in some cases it may qualify under Articles 6(1)(d)(e)(f) compatible with some of the already foreseen lawful bases. When a controller collects personal data under a lawful basis, Article 6(4) allows it to process the data for a secondary research purpose. Thus, while the GDPR explicitly permits re-purposing collected data for research, it also may permit a controller to collect

personal data initially for research purposes, without requiring the data subject's consent.

Furthermore, although research is not mentioned explicitly as a lawful basis for personal data processing, Recital 157 identifies the benefits associated with personal data research, subject to appropriate conditions and safeguards. These benefits include the potential for new knowledge when researchers "obtain essential knowledge about the long-term correlation of a number of social conditions". The results of the research "obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people, and improve the efficiency of social services."

Moreover, the GDPR foresees derogations for the secondary processing of personal data for research purposes as long as there is a lawful basis for such processing (Article 5, Recital 50). Article 89 sets out the "appropriate safeguards" that controllers must implement to further process personal data for research. It mandates controllers explicitly to put in place "technical and organisational measures" to ensure that they process only the personal data necessary for the research purposes, in accordance with the principle of data minimisation outlined in Article 5 (c). Article 89(1) provides that one way for a controller to comply with the mandate for technical and organisational measures is through the deployment of "pseudonymisation." Pseudonymisation is "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution to an identified or identifiable individual" (Article 4(3b)).

Taken the above into consideration, one may consider that the case of personal data scraped from social media sites without the consent of the user that added them, may raise serious concerns regarding its ethical and legal consequences. Yet, these concerns can be easily removed, as we demonstrate below, when data scrapping is performed by researchers to facilitate the mitigation of malevolent uses such as those of pedophile and sex exploitations. More specifically, a team of researchers scraped a well-known social media site that attracts millions of teenagers (even if the web site's terms of service forbid its use by people under the age of 18). They found that the exchanged text chats among its participants include numerous instances of discussions involving sexual harassment and pedophile actions, all covered up under seemingly innocent words and terminologies that are impossible to be tracked by conventional software tailored to identify specific words for sex abuse. To facilitate research on advanced and innovative ways of tracking down suspicious cases of child abuse and harassment, the researchers, after scrapping the chats on the site referring to the coded malevolent conversations, published a dedicated corpus including these suspicious words, strings and emoticons. All user data, namely the user's nickname, have been anonymised with masking techniques whereas every single user was always masked with the same string. Taking into account that the identification of the users could be potentially possible when additional information (held by the researchers) is used, this masking technique is, in fact, a pseudonymisation in GDPR terms. Since pseudonymised data are still personal, they still fall under the scope of the GDPR. Therefore, researchers had to ensure that the processing of the personal data contained in the scrapped chats is compatible with the data protection provisions of the GDPR, and in particular with at least one of the six lawful purposes of processing enshrined in GDPR Article 6. Taking into account that the undertaken data crawling of the personal data can protect the vital interests of the children participating in the social media site so as not to be fooled by pedophile users, as well as that this processing is beyond any doubt carried out in the public interest, the data scrapping and subsequent analysis of the concerned data by the researcher are in accordance with the GDPR.

Particular attention should be paid for the processing of users data, given that the processed information most likely refers to the sexual preferences of the data subjects, a piece of information considered to be

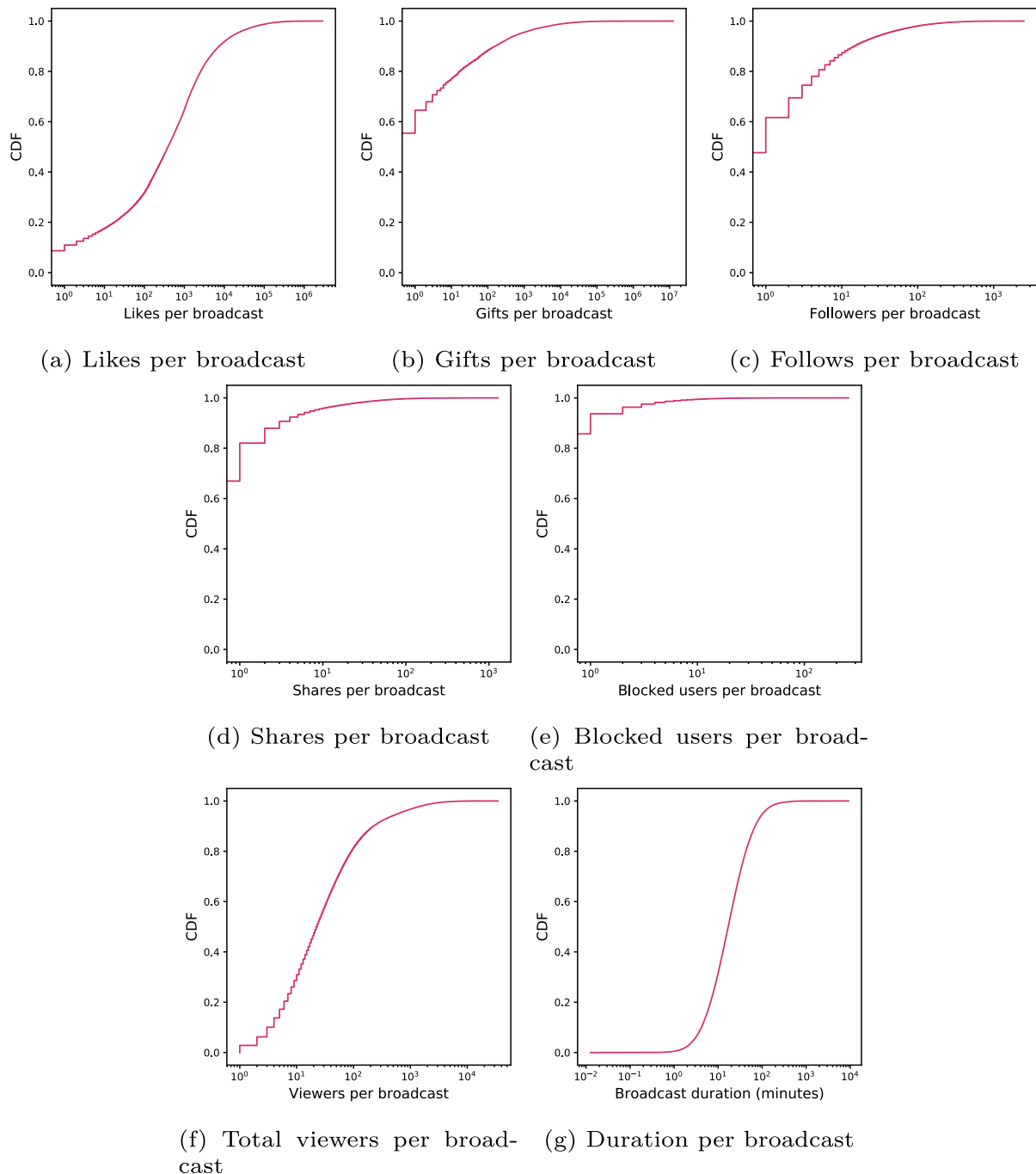


Fig. 1. Cumulative distribution functions (CDFs) of broadcast metadata features and interactions.

among the special categories of personal data referred to as “sensitive” for which stricter provisions apply (Article 9). Yet, derogating from the prohibition on processing special categories of personal data “should also be allowed when provided for in Union or Member State law and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where it is in the public interest to do so” (recital 52, Article 9(2)(j)). And beyond any doubt, protecting children from pedophile actions and sexual harassment is, above all, of substantial public interest and has been foreseen to all domestic legislations. Therefore, provided that the processing of the personal data of the data subjects is proportionate to the aim pursued, respects the essence of the right to data protection and provides for suitable and specific measures, i.e. pseudonymisation, to safeguard the fundamental rights and the interests of the data subject, the derogations for processing sensitive information under the Article 9

(2) are fulfilled.

Finally, the GDPR Article 12(1) requires controllers to “take appropriate measures” to inform data subjects of the nature of the processing activities and the rights available to them. Controllers are required to provide this information in all circumstances, regardless of whether consent is the basis for processing, “in a concise, transparent, intelligible and easily accessible form, using clear and plain language” (Article 12(1)). Nevertheless, a researcher may be exempted from the notice requirement if she received the personal data from someone other than the data subject, such as where the data came from a publicly available source. Article 14 exempts controllers in these circumstances, if “the provision of such information proves impossible or would involve a disproportionate effort,” which “could in particular be the case” in the research context (Recital 62). A researcher also may claim an exemption if providing

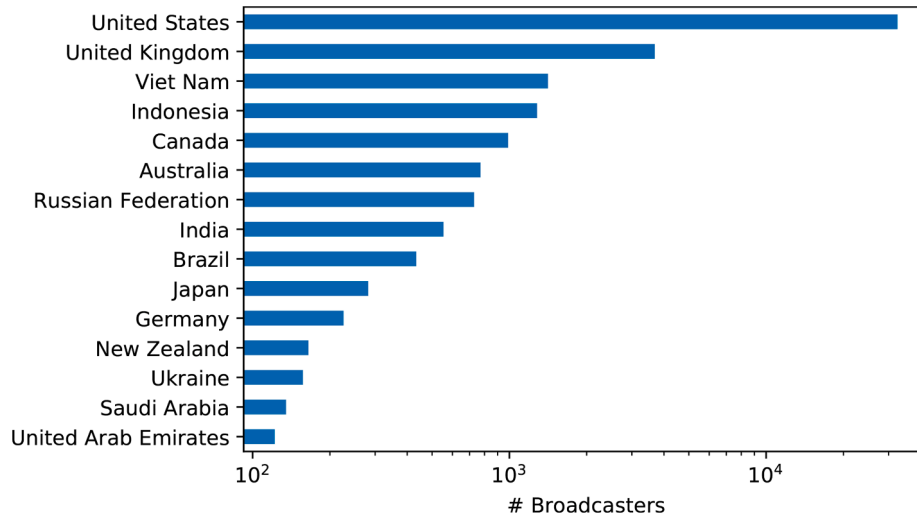


Fig. 2. Broadcasters count per country.

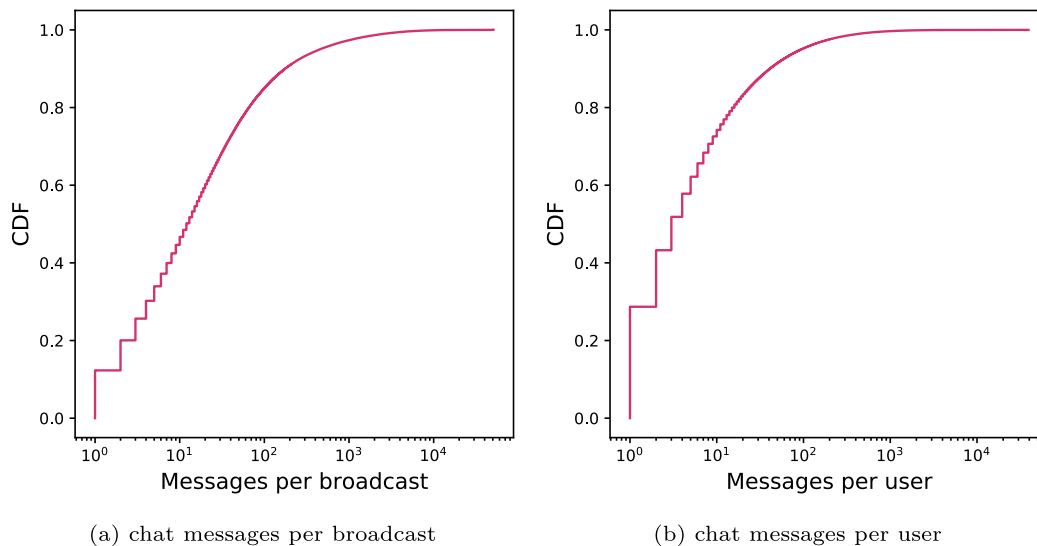


Fig. 3. Cumulative distribution functions (CDFs) of the chat messages per broadcast and per user.

notice would be “likely to render impossible or seriously impair the achievement of the [research] objectives,” provided there are appropriate safeguards in place, “including making the information publicly available” (Article 14(5)(b)).

In summary, scrapping personal data from social media sites and publishing them in pseudonymised form for research purposes is legal and ethical as long as it is performed to protect the vital interests of the data subjects or others and it is in the public interest to do so.

4. The dataset

In what follows, we analyse a large-scale dataset that we created based on the public interactions between streamers and viewers during the live broadcasts of users identified as adult content producers in Lykousas et al. (2018), from the LiveMe⁵ platform, a major Social Live Streaming Service (SLSS). The dataset comprises 39,382,838 chat messages exchanged by 1,428,284 users, in the context of 291,487 live broadcasts during a period of approximately two years, from July 2016

to June 2018. Each broadcast effectively functions as a temporary chatroom. The audience can interact with the streamers via text messages and reward them with virtual rewards, e.g. points, gifts, badges (some of which are purchasable) even virtual money. Apart from the chat messages, the dataset contains a wide range of user interactions along with metadata. We describe the features below:

- **Metadata (broadcast)**

- Total Viewers: total number of viewers who joined the livestream as viewers.
- Duration: duration of stream in seconds

- **Metatadata (broadcaster)**

- Country Code

- **Interactions**

- Likes: Viewers who liked the broadcast & the number of likes given.

⁵ <https://www.liveme.com/>.

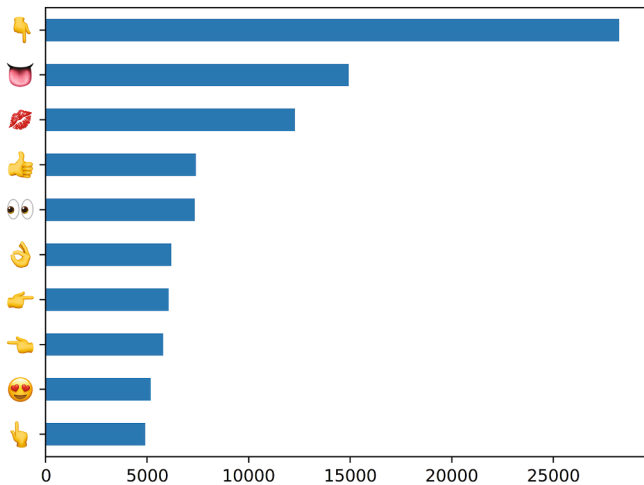


Fig. 4. Top emoji collocations for clothing related emojis.

- Follows: Viewers who followed the broadcaster during the livestream.
- Gifts: Viewers who sent virtual gifts to the broadcaster, along with value (in virtual currency) for each gift.
- Shares: Viewers who shared the broadcast (via a link so others can join).
- Blocks: Viewers who have been blocked by the broadcaster (i.e. banned from a stream).

To better understand how the features mentioned above are distributed, we plot the cumulative distribution functions (CDFs) in Fig. 1. We observe that every broadcast in our dataset had viewers (143.5 on average) and sizeable duration (31.4 min on average). While most of the broadcasts received likes (92%), the 55% did not receive any gifts (since they cost money, contrary to likes). Furthermore, 47% for the broadcasts did not generate any new followers for the broadcasters. At the same time, the interactions of sharing and blocking are relatively rare in our dataset (i.e. they are zero for 0.67% and 0.86% of the broadcasts, respectively). Next, to understand the geographical distribution of adult content producers, we plot the distribution of the broadcasters per country of the whole dataset, focusing on the 15 countries with most broadcasters, in Fig. 2.

5. Large-scale grooming analysis

By plotting the CDFs of the chat messages per broadcast and per user in Fig. 3,4, we notice that around 82% of the broadcasts of adult content producers receive less than 100 chat messages. Moreover, out of the unique users chatting during these broadcasts, only 30% send more than ten messages in total. Both distributions are particularly heavy-tailed, meaning that the majority of chat messages in our dataset are exchanged during a few highly popular broadcasts.

To identify sexual grooming behaviour in the chat messages, we adopted the approach followed by several authors in the most recent relevant works (Drouin et al., 2017; Lorenzo-Dus and Kinzel, 2019; Lorenzo-Dus et al., 2020) analysing the Perverted-Justice Dataset (PJ), which although dated and relatively small-scale, was the only publicly available dataset of chats produced by online groomers to date. To this end, we search the chat messages comprising our dataset for sexual content keywords defined in Linguistic Inquiry and Word Count (LIWC) corpus (Pennebaker et al., 2015). More precisely, the 2015 version of the LIWC dictionary for the sexual content variable comprises a total of 131 words. These include a wide range of terms about sexual matters, including sexual orientation (e.g. bi-sexual, heterosexual), sexual organs (e.g. penis*, vagin*, womb), slang terms, sexually transmitted diseases

Table 1
Top 15 verbs (simple and phrasal) associated with clothing items.

Verb	Count
Wear	6553
Show	5817
Remove	3947
See	3765
Get	3157
Open	3105
Like	2914
Love	2913
Dare	2844
Lift	2159
Change	2154
Want	1811
Take	1412
Go	1267
Say	1237
Put_on	6083
Take_off	4229
Pull_down	1930
Pull_up	1625
Have_on	1214
Take_of	731
Get_on	728
Lift_up	690
Put_in	507
Dress_up	433
See_without	371
Look_in	336
Put_down	312
Look_like	295
Change_into	274

Table 2
Top 10 nearest neighbors (cosine distance) of the word “pussy”.

Term	Distance	Count	#Broadcasts	#Users
Pusy	0.828122	956	513	432
Pus	0.768473	416	305	259
Pushy	0.741119	267	185	158
Bussy	0.799563	209	128	100
püussy	0.810713	198	133	101
Puzzy	0.753680	195	122	113
püussy	0.781377	184	110	90
Pussycat	0.818996	169	138	141
Pissy	0.702024	160	141	142
Pssy	0.812888	135	103	79

and infections, sexual violence and assault terms and sex enhancements. The most frequently occurring sexual terms in the PJ dataset, had a very low number of occurrences in the LiveMe chats (less than five exact matches in most occasions). The very low occurrence of such words implies the existence of an automated filtering mechanism in place. Nonetheless, relevant literature about online chat has demonstrated that users with previous exposure to text-based automatic moderation

Table 3
Top 10 nearest neighbors (cosine distance) of the word “boobs”.

Term	Distance	Count	#Chatrooms	#Users
bobs	0.752709	14728	5720	5754
boos	0.756812	670	490	444
booms	0.759904	638	305	189
boobes	0.868892	578	315	182
bobbs	0.794095	494	341	292
boops	0.803665	452	276	177
boody	0.784702	400	285	190
boobz	0.858590	389	256	161
bobss	0.787802	267	175	113
boobd	0.896997	159	146	150

Table 5
Top 5 interaction features relevant for characterising grooming broadcasts.

Rank	Feature	MDI
1	New followers to viewers	0.36
2	Likers to viewers	0.16
3	Total new followers	0.10
4	Chat messages per user	0.10
5	Total chat messages	0.05

etc.).

The second most dominant topic (#11) reflects flirtatious behaviours, including many endearment terms (e.g. love, nice, pretty, kiss, cute, gorgeous, hot), and words associated with appearance features (e.g. eye, lip, hair, smile, tattoo). Topic #11 is the most representative of 8,477 chat log documents (13% of modelled broadcasts). The rest of the topics describe a wide range of behaviours occurring in the context of live streams, including virtual currency and gifts of LiveMe (i.e. coin, coindrop, castle, diamond, wand), dancing, singing, eating, social media, etc. An interesting observation is the emergence of a topic containing mostly Spanish words (topic #2). We speculate that a proportion of the US viewers are using Spanish to communicate within the broadcasts, something we did not consider in the preprocessing stage. It should be noted that Spanish are the second most spoken language in the US and widely used in some states. Nonetheless, provided that it dominates only 2142 chats (3% of modelled broadcasts), we expect that its impact will be negligible for the rest of our analysis.

Next, we assess the degree to which user interactions other than chatting can be characteristic of grooming. For this, we leverage the interaction and metadata features of our dataset. Additionally, we normalise the interaction features by the total number of viewers of each stream, considering them as additional features. Next, we employ the Mean Decrease Impurity (MDI) (Breiman, 2001; Breiman, 2002) measure obtained in the process of random forest growing to assess the importance of the described features for discriminating between the broadcasts where topic #18 is dominant in the topic mixture and the rest.

Table 5,6 reports the ranking of the top five most important features according to the normalised MDI metric. To understand how these features are distributed across the two latent classes of broadcasts, we plot their cumulative distribution functions (CDFs) in Fig. 10. We note that the most characterising feature is the fraction of viewers who started following the broadcaster during the stream (Fig. 10a), which in the case of the broadcasts where the grooming topic dominates is much higher than the ones where it does not. Moreover, in Fig. 10c we observe that only around 6% of the grooming broadcasts have not generated any followers for the broadcaster, while the same is true for 17% of the rest of broadcasts. This behaviour is in line with the findings of Lykousas et al. (2018), where the adult content producers of LiveMe were found to have an exceptionally high number of followers which are characterised by their tendency to follow users who have broadcasted adult content systematically, labelled as *adult content consumers*. A possible explanation could be that in broadcasts where the grooming behaviour is prevalent, broadcasters are coerced into performing sexual acts requested by the viewers, as previously outlined. This could justify why the number of new followers they gain in such broadcasts is significantly higher since the viewers might expect that the broadcasters will stream more nude/adult content in the future, and following them is the only way to be notified when they start a new broadcast. Similarly, the fraction of viewers who have liked a broadcast is higher when the grooming behaviour is dominant (Fig. 10b, which is consistent with the findings of Lykousas et al. (2018) where adult content producers are observed to have received higher amounts of praise than the users found in their ego-networks (i.e. followers and followees). This further exemplifies the predatory behaviour of viewers who use likes/praise to coerce broadcasters into inappropriate acts or reward them when they have

Table 6
Topics.

Topic	Keywords	#Docs
18	CLOTH_TERM, show, open, SEX_TERM, nice, dare, dance, hot, stand, leg, put, kiss, turn, pull, wear, cam, camera, remove, foot, top, snapchat, gift, girl, rub, low, hand, lift, finger, message, tease	12,209
11	Love, nice, pretty, kiss, cute, eye, girl, gorgeous, hot, SEX_TERM, sweet, lip, hair, dear, tattoo, smile, single, dance, friend, number, stand, beauty, lovely, cutie, face, boyfriend	8477
1	Sleep, phone, bed, tired, cool, car, wake, cold, drive, smoke, hour, fall, hear, talk, asleep, stay, high, long, house, game, guess, chill, goodnight, sound, money, iphone, fun, pay	5731
19	Talk, hear, happen, friend, leave, wrong, true, cool, mad, sad, care, sound, hurt, smile, fight, stay, fine, dude, person, funny, break, hard, nice, head, long, boy, army, problem, lose, girl, yep	4432
7	Block, admin, girl, message, leave, report, show, talk, account, kid, creep, young, shut, ban, rude, fake, nasty, send, perv, truth, boy, lie, hater, wrong, police, child, unblock	3328
16	Drink, cat, food, pizza, laugh, eat, funny, face, water, put, chicken, dead, hair, head, challenge, roast, cream, leave, SEX_TERM, apple, taco, chocolate, pet, bob, hand, candy, mouth, cheese, nose	3180
12	Cute, snapchat, send, instagram, rate, clown, dab, hot, number, insta, hair, play, text, friend, pretty, love, put, single, phone, kik, profile, eye, cutie, chat, ghost, boy, girl, fake, girlfriend	3080
3	Send, gift, spam, castle, diamond, share, top, level, broadcast, win, giveaway, broadcaster, wand, number, stream, boat, enter, entry, star, love, feature, join, porsche, coin, awesome, comment, fan	2953
5	Coin, drop, coindrop, follower, send, win, feature, shout, fan, castle, love, wand, dab, thot, number, gift, shoutout, giveaway, diamond, stream, iphone, lag, goal, dude, pumpkin, andy, light, level	2798
4	Song, play, sing, love, voice, rap, singing, amazing, nice, dance, awesome, beat, put, hear, panda, listen, cool, singer, sound, closer, juju, black, job, girl, talent, guitar, boy, heart, hit, drake	2687
8	Love, stream, friend, accent, talk, remember, guess, cool, speak, leave, sleep, skype, long, cute, funny, nice, number, meet, hair, mate, lot, person, dad, class, cat, joke, jenni, kat, join, change	2682
20	Light, turn, gang, love, stay, queen, squad, hit, chill, number, king, slay, fact, thot, level, savage, rock, party, dead, boy, mad, play, homie, ight, lot, black, nun, show, petty, dope, top, sum	2366
2	Hola, mami, como, cute, hermosa, eres, show, spanish, amor, SEX_TERM, pretty, bella, donde, hot, bonita, bien, lip, kiss, speak, tienes, espanol, gorgeous, stand, rico, jada	2142
13	Girl, love, cute, play, blue, twin, pretty, red, hot, black, dance, snapchat, green, pink, makeup, hair, lady, white, friend, cool, color, game, face, team, texas, nice, CLOTH_TERM, batman, favorite	1954
14	Kate, love, kid, nice, awesome, cool, tree, country, santa, dad, boy, level, show, broadcast, send, amazing, hear, wolf, talk, lot, son, king, falcon, grim, happen, stream, matt, house, long, rock	1831
9	Beam, love, lag, send, king, cris, stream, castle, fletch, broadcast, show, level, dude, awesome, nick, game, amazing, feature, remember, joey, gift, beam, roll, diamond, join, happen, rip, rackbar	1663
15	Ready, love, spam, feature, stay, game, number, win, boy, tre, read, letter, chat, duck, turtle, greg, cat, spamme, fun, red, ugh, play, controller, send, coin, hehe, cool, high, comment, gift, party	1027
17	Love, fan, favorite, youtube, shout, meet, dab, channel, pickle, song, shoutout, canada, movie, awesome, fav, twerk, vote, magic, food, subscribe, notice, tattoo, cool, texas, win, vid, hair, ily	908
6	Race, love, family, human, unity, amen, put, country, draw, earth, whiskey, broadcast, peace, block, thre, lucky, princess, spam, britt, join, general, respect, coin, barbie, send, level, lag, brit	642
10	President, kira, criticize, article, essay, literary, loco, fard, natur, fward, lag, foard, riot, ward, folard, kilo, follrd	14

achieved their objective. Interestingly, for the *Chat messages per user* and *Total chat messages* features which were also found to be important (albeit considerably less impactful in a classification setting), we observe the opposite behaviour: In grooming broadcasts users exchange fewer chat messages, both per-user and at the broadcast level.

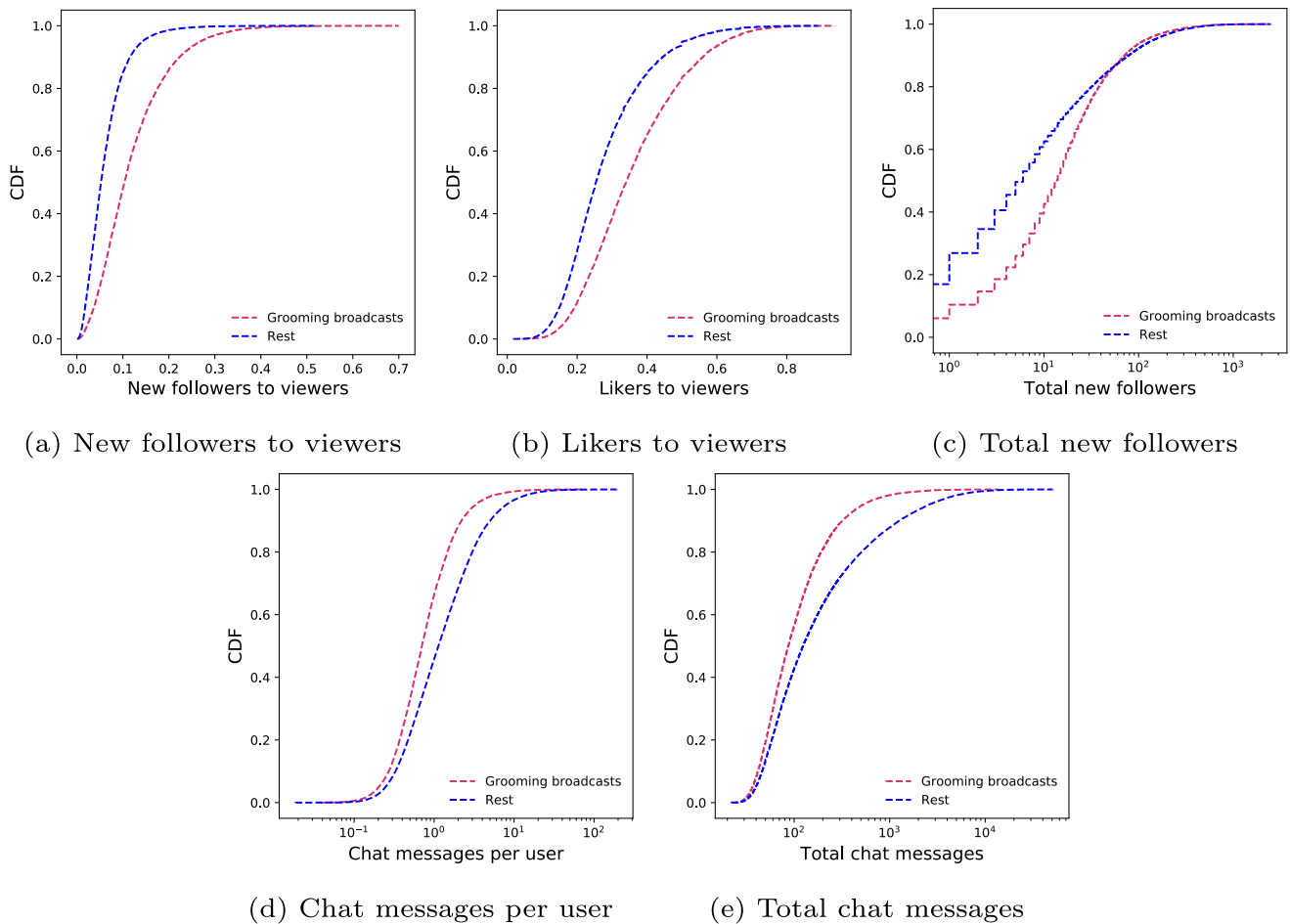


Fig. 10. Cumulative distribution functions (CDFs) of the features of Table 5.

Table 7

Illustrative chat messages from broadcasts where Topic #7 is dominant. Key terms of Topic #7 are in bold.

Chat message
A predator is a person who asks kids to undress in front of the camara
And his bio said he likes meeting young girls
Block foot fetish creep
Don't show the creeps anything
Everyone report chat police
Leave her alone creep
Pervs . This kid is like 12
Report that creep too the police
Report the users asking kids to undress; to authorities not LiveMe
Show your kids
So if they didn't ban people for nuditys you would show?
This needs to be reported what sort of sick people are ye. She is only 11
YOUR MOM WILL NOW GET A CALL TO KNOW YOU TALK TO 40 years old creeps
You pervs are nasty as f***
block & report nasty stuff
creeps make kids do nasty stuff
he's following lots of young girls
pervs stop asking her to undress
report these pedos to police mate
she is a child stop asking that
she not letting you creeps or sick perv seeing her dress or undress ok
she's a kid perv
they can't ban you if you delete your video after you show
too young this is illegal nd worng lol
try not to undress on stream, it will draw in a lot of creeps
you have creeps who made you do nasty stuff
you look very young .there are lots of pedos on here. be careful

6.4. Topic relatedness

In this section, we aim to explore the relatedness of the dominant grooming topic and other topics learned by LDA, which could unveil different aspects of this deviant behaviour, beyond our initial analysis. To this end, we use a frequent itemset mining approach to examine the co-occurrence of prevalent topics within the chat log documents. More precisely, we first selected the three topics with the highest probability in the mixture assigned to each broadcast. Then, we applied the FP-growth algorithm (Han et al., 2004) to discover frequent patterns of size two. In Table 8, we show the 10 most frequent patterns extracted following the described approach. As expected, the top result includes the two most prevalent topics in the mixture. Notably, the second most frequent pattern includes the grooming topic, and topic #7, which contains terms related to the (self) moderation of broadcasts (i.e. block,

Table 8

10 most frequent prevalent-topic patterns.

Topic pattern	Occurrences
(11, 18)	11,150
(7, 18)	5536
(1, 19)	4892
(13, 18)	3843
(1, 16)	3416
(1, 11)	3357
(7, 11)	3338
(3, 5)	3331
(2, 11)	3269
(12, 18)	2987

report, ban, shut), terms indicating young age (i.e. kid, young, child, girl, boy), terms of hostility (i.e. creep, perv, hater), words bearing negative sentiment according to LIWC (nasty, wrong, fake, lie). Moreover, the key term that possibly contributes the most towards the interpretation of this topic is **police**. Thus, we expect this topic to be indicative of the criminal dimension of sexual grooming of minors in LiveMe, a large-scale deviant behaviour, also attested by popular media (Melugin, 2018).

To test this speculation, we manually examined a portion of chat messages from broadcasts where Topic #7 is dominant where the aforementioned key terms appear, and we present some illustrative examples in Table 7,8. What we observe is that a part of the users expresses their discontent and anger towards the predators/groomers and their harassment targeting minors. We argue that the above illustrates the extent of deviant behaviour in SLSS, something that beyond the media is also reported by users in, e.g. their feedback for the app. Moreover, the high ranking of this pair indicates that such phenomena, despite the app's moderation mechanisms, are often and known to many users. Finally, the fourth pair (13,18), beyond the common keywords of both topics, shows that some users request further engagement through other platforms, and a primary phase of praise of clothing and body parts, possibly preceding the grooming phase.

7. Conclusions

Social live streaming services due to the continuous use of live streams and immediate user interaction are continuously expanding their user base. As expected, these platforms have attracted the interest of deviant users which try to exploit the new features on these platforms. Obviously, grooming is not only performed in SLSS, nor it is the only thing done on these platforms. Nonetheless, the different possible user interactions coupled with the live streaming nature, create a novel and less explored field.

This work performs an in-depth analysis of the chats of thousands of users and identifies characteristics of the grooming behaviour in the verbal and non-verbal context. To facilitate further research in the field, we responsibly share a massive dataset and provide ethical and legal justification for the collection and processing of such a dataset. Moreover, we illustrate in an automated way how users bypass the word filters that service providers use in their platforms. We also highlight the importance of emojis for the first time in the context of grooming. Finally, our work illustrates that more deviant behaviours may be performed on these platforms.

We believe that this scientific work constitutes a significant contribution towards understanding the deviant behaviours on social networks. The latter should be considered in the light of the role that social networks have in our daily lives and the potentials that the emergence of SLSS have. Our work implies that additional risks exist due to the inefficiency of current moderation mechanisms. Therefore, further measures must be taken to secure the content of what is broadcast, from whom, and to whom. Undoubtedly, due to the size and rate of exchanged information, moderation mechanisms may be difficult to be performed in real-time. However, our work illustrates how deviant behaviours can be detected effectively without resolving to the use of multimedia which require heavy processing. Therefore, we believe that the grooming and other predatory actions will be soon identified better and addressed more effectively by service providers.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRedit authorship contribution statement

Nikolaos Lykousas: Conceptualization, Methodology, Investigation, Software, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Constantinos Patsakis:** Methodology, Investigation, Data curation, Validation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the projects CyberSection 4Europe (<https://www.cybersec4europe.eu>) (Grant Agreement No. 830929) and LOCARD (<https://locard.eu>) (Grant Agreement No. 832735). The authors would also like to thank NVIDIA Corporation for their GPU donation supporting their research.

The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

References

- Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391–402). Springer.
- Basher, A. R. M., & Fung, B. C. (2014). Analyzing topics and authors in chat logs for crime investigation. *Knowledge and Information Systems*, 39(2), 351–381.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (Jan), 993–1022.
- Bogdanova, D., Rosso, P., & Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. *Computer Speech & Language*, 28(1), 108–120.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L. (2002). *Manual on setting up, using, and understanding random forests*, v3 p. 1. CA, USA: Statistics Department University of California Berkeley, 58.
- Cambria, E., & White, B. (2014). Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48–57.
- Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7–9), 1775–1781.
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296.
- Coletto, M., Aiello, L. M., Lucchese, C., & Silvestri, F. (2017). Adult content consumption in online social networks. *Social Network Analysis and Mining*, 7(1), 28.
- Cooper, A. (1998). Sexuality and the internet: Surfing into the new millennium. *CyberPsychology & Behavior*, 1(2), 187–193.
- Craven, S., Brown, S., & Gilchrist, E. (2006). Sexual grooming of children: Review of literature and theoretical considerations. *Journal of Sexual Aggression*, 12(3), 287–299.
- Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 17(1), 61–84.
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6), 570–606.
- Doll, C., Sykosch, A., Ohm, M., & Meier, M. (2019). Automated pattern inference based on repeatedly observed malware artifacts. In *Proceedings of the 14th international conference on availability, reliability and security* (pp. 1–10).
- Drouin, M., Boyd, R. L., Hancock, J. T., & James, A. (2017). Linguistic analysis of chat transcripts from child predator undercover sex stings. *The Journal of Forensic Psychiatry & Psychology*, 28(4), 437–457.
- Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, 64 (2–3), 221–245.
- Elzinga, P., Wolff, K. E., & Poelmans, J. (2012). Analyzing chat conversations of pedophiles with temporal relational semantic systems. In *2012 European intelligence and security informatics conference* (pp. 242–249). IEEE.

- Westlake, G. B. & Bouchard, M. (2016). Criminal careers in cyberspace: Examining website failure within child exploitation networks. *Justice Quarterly*, 33 (7), 1154–1181.
- Gottschalk, P. (2011). A dark side of computing and information sciences: Characteristics of online groomers. *Journal of Emerging Trends in Computing and Information Sciences*, 2(9).
- Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8 (1), 53–87.
- Hidalgo, J. M. G. & Díaz, A. A. C. (2012). Combining predation heuristics and chat-like features in sexual predator identification. In *CLEF (Online Working Notes/Labs/Workshop)*. Citeseer.
- Hong, L. & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80–88).
- Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373–1378).
- Hosseini, H., Kannan, S., Zhang, B. & Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138.
- Inches, G., & Crestani, F. (2012). Overview of the international sexual predator identification competition at pan-2012. *CLEF (Online working notes/labs/workshop)* (Vol. 30).
- Kontostathis, A., Garron, A., Reynolds, K., West, W. & Edwards, L. (2012). Identifying predators using chatcoder 2.0. In *CLEF (Online Working Notes/Labs/Workshop)*.
- Loper, E. & Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pp. 63–70.
- Lorenzo-Dus, N., & Kinzel, A. (2019). 'So is your mom as cute as you?': Examining patterns of language use by online sexual groomers. *Journal of Corpora and Discourse Studies*, 2(1), 1–30.
- Lorenzo-Dus, N., Kinzel, A., & Di Cristofaro, M. (2020). The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. *Journal of Pragmatics*, 155, 15–27.
- Lykousas, N., Gómez, V., & Patsakis, C. (2018). Adult content in social live streaming services: Characterizing deviant users and relationships. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 375–382). IEEE.
- McCulloch, G., & Gawne, L. (2018). Emoji grammar as beat gestures. In *Proceedings of the 1st international workshop on Emoji understanding and applications in social media*.
- McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, 41(6), 607–625.
- McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., & Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3), 103–122.
- Melugin, B. (2018). Pedophiles using app to manipulate underage girls into sexual acts, sell recordings as child porn. URL: <https://www.foxla.com/news/pedophiles-using-app-to-manipulate-underage-girls-into-sexual-acts-sell-recordings-as-child-porn>. [Online; last accessed 12-March-2020].
- Miah, M. W. R., Yearwood, J., & Kulkarni, S. (2011). Detection of child exploiting chats from a mixed chat dataset as a text classification task. *Proceedings of the Australasian Language Technology Association Workshop, 2011*, 157–165.
- Michalopoulos, D., & Mavridis, I. (2011). Utilizing document classification for grooming attack recognition. In *2011 IEEE symposium on computers and communications (ISCC)* (pp. 864–869). IEEE.
- Mitchell, K. J., Finkelhor, D., Jones, L. M., & Wolak, J. (2010). Use of social networking sites in online sex crimes against minors: An examination of national incidence and means of utilization. *Journal of Adolescent Health*, 47(2), 183–190.
- Morris, C., & Hirst, G. (2012). Identifying sexual predators by svm classification with lexical and behavioral features. *CLEF (Online Working Notes/Labs/Workshop)*, 12, page 29.
- Papegnies, E., Labatut, V., Dufour, R., & Linares, G. (2019). Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems*, 6 (1), 38–55.
- Parapar, J., Losada, D. E. & Barreiro, A. (2012). A learning-based approach for the identification of sexual predators in chat logs. In P. Forner, J. Karlgren & C. Womser-Hacker (Eds.), *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of CEUR Workshop Proceedings. CEUR-WS.org.
- Peersman, C., Vaassen, F., Van Asch, V., & Daelemans, W. (2012). Conversation level constraints on pedophile detection in chat rooms. *CLEF (Online Working Notes/Labs/Workshop)*, 1–13.
- Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)* (pp. 235–241). IEEE.
- Pennebaker, J. W., Boyd, R. L., Jordan, K. & Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin.
- Quayle, E., Holland, G., Linehan, C., & Taylor, M. (2000). The internet and offending behaviour: A case study. *Journal of Sexual Aggression*, 6(1–2), 78–96.
- Řehurek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta. ELRA.
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408).
- Sievert, C. & Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).
- Villatoro-Tello, E., Juarez-Gonzalez, A., Escalante, H. J., y Gomez, M. M. & Villasenor, L. (2012). A two-step approach for effective detection of misbehaving users in chats. In P. Forner, J. Karlgren & C. Womser-Hacker (Eds.), *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, Rome, Italy*.
- Winters, G. M., & Jeglic, E. L. (2017). Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters. *Deviant Behavior*, 38(6), 724–733.
- Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. In *Proceedings of the content analysis in the WEB* (pp. 1–7).
- Zambrano, P., Torres, J., & Flores, P. (2019). How does grooming fit into social engineering?. In *Advances in computer communication and computational sciences* (pp. 629–639). Springer.
- Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y. & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics* (Vol. 16, p. S8). Springer.
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338–349). Springer.



Inside the X-Rated World of “Premium” Social Media Accounts

Nikolaos Lykousas^{1(✉)}, Fran Casino¹, and Constantinos Patsakis^{1,2}

¹ University of Piraeus, Piraeus, Greece

{nlykousas, fran.casino, kpatsak}@unipi.gr

² Athena Research Center, Marousi, Greece

Abstract. During the last few years, there has been an upsurge of social media influencers who are part of the adult entertainment industry, referred to as *Performers*. To monetize their online presence, Performers often engage in practices which violate community guidelines of social media, such as selling subscriptions for accessing their private “premium” social media accounts, where they distribute adult content. In this paper, we collect and analyze data from FanCentro, an online marketplace where Performers can sell adult content and subscriptions to private accounts in platforms like Snapchat and Instagram. Our work aims to shed light on the semi-illicit adult content market layered on the top of popular social media platforms and its offerings, as well as to profile the demographics, activity and content produced by Performers.

Keywords: Influencers · Marketplace · Performers · Adult content · Premium accounts · Community guidelines

1 Introduction

In the world of social media, content creators play a central role in shaping a global online culture. The content creators who raise in popularity can attain the status of online micro-celebrities, and they are commonly characterized as *influencers* [9]. The main objective of influencers is to produce digital content which attracts users’ attention and rapidly gains popularity, often becoming ‘viral’, in platforms such as Instagram and YouTube [7, 13]. In this regard, influencers leverage focused visual content and targeted communication techniques to capture and sustain the attention of social media users, thus building large follower bases and attaining organic social reach. Social media content creators can thus monetize their reach in various ways, such as using word-of-mouth marketing techniques and promoting brands and campaigns [11, 16].

One of the most prevalent strategies employed by influencers to entice followers towards heightened forms of emotional engagement is sexualized labour [5]. Posting sexualized images in social media is a popular form of self-presentation for young adults [1, 3, 14, 15], and it is outlined as the core tactic to attract followers for a particular type of influencers, which are categorized as “*Performers*” in [5].

This category of influencers includes adult performers/entertainers, sex workers and models. In all cases, after building an audience in mainstream social media, Performers redirect their followers to external outlets for purchasing exclusive content, often

pornographic in nature. Notable examples of such outlets are platforms like OnlyFans¹ (effectively an ‘adult’ version of Instagram), and “premium” Snapchat accounts, offering a lucrative income stream for Performers looking to monetize their online presence [2]. For social media platforms like Snapchat, the community guidelines² explicitly *prohibit accounts that promote or distribute pornographic content*. Nonetheless, it has been shown that community guidelines cannot be effectively enforced to ban adult content in social media [12]. As such, Performers who systematically violate community guidelines by posting overtly sexual content, have to use external means for managing transactions with their client base, as well as maintaining their digital presence in multiple social outlets, in case their accounts get suspended.

In this paper, we analyze data collected from *FanCentro*³, a platform where Performers can monetize their fan base via selling subscriptions to their private social media accounts. Additionally, FanCentro enables Performers to directly sell private content through a media feed, as well as chatting functionality between Performers and their subscribers. As a requirement for opening an account in FanCentro, Performers have to provide a digital copy of government-issued ID for age verification purposes. After this verification step, FanCentro, for a fraction of the paid subscriptions, handles all of the necessary transactions and administrative activities.

There are two main reasons we chose FanCentro over other similar platforms such as OnlyFans, which have gained wide mainstream media attention [4]. First, its primary focus is selling access to “premium” accounts in social platforms which, strictly, are not content marketplaces (i.e. Snapchat and Instagram). Second, FanCentro website provides a complete listing⁴ of Performer profiles, enabling us to collect data without having to employ sampling techniques which could potentially bias our findings. Our work aims to shed light on the mechanics of the semi-illicit industry of premium social media subscriptions and services offered by Performers, in the context of adult content marketplaces such as FanCentro.

2 Data Collection

We constructed a *complete* dataset with the profiles of Performers registered in FanCentro as of April 5th, 2020. In Fig. 1, we provide an illustrative example of a Performer’s profile page. We note that only Performers have public profiles and can post content, while regular users/subscribers can only interact with Performers (i.e. follow, message, like/comment to their posts) and not other users. In total, we collected the profile attributes, published content metadata, and offered products for 16,488 users. For this, we created a crawler which consumes the API used by FanCentro’s website, enabling us to collect the relevant data. Despite the “public” nature of collected information, we follow Zimmer’s approach [17]. In this regard, the data remains anonymized during all the steps of our analysis, and we report only aggregate findings.

¹ <https://onlyfans.com>.

² <https://www.snap.com/en-US/community-guidelines>.

³ <https://fancentro.com>.

⁴ In contrast, OnlyFans platform does not have such functionality.

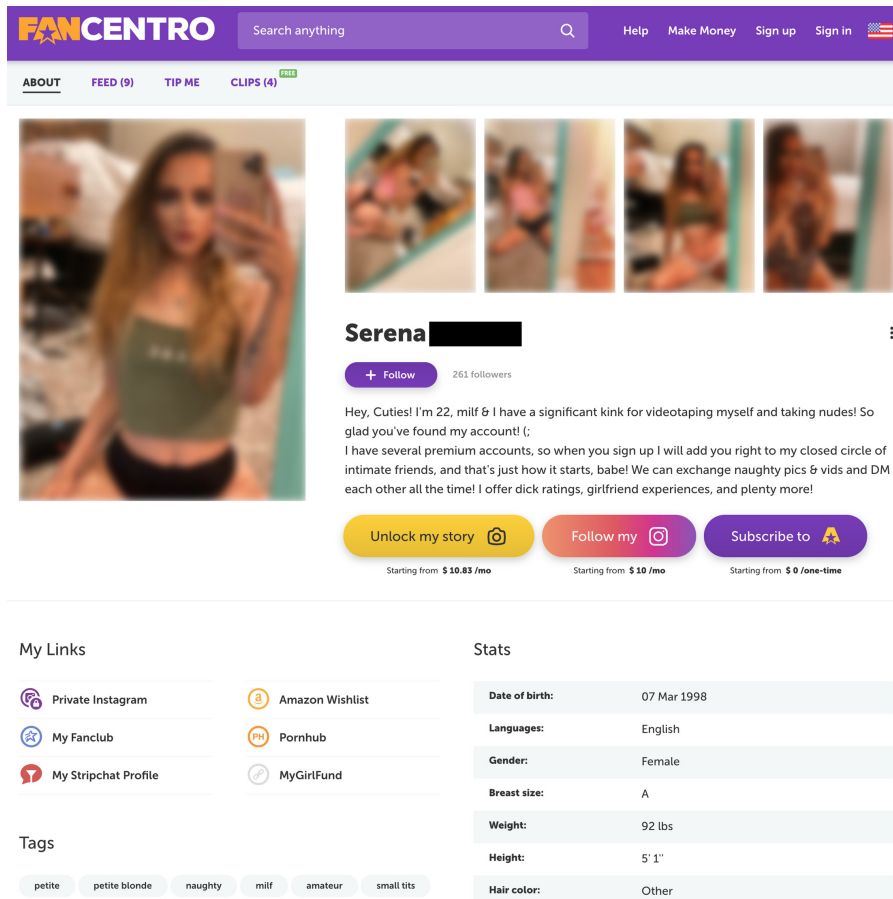


Fig. 1. An example Performer’s profile page in FanCentro.

In order to measure the activity in FanCentro in terms of new registrations of Performers, in Fig. 2, we plot the number of accounts created each week since the launch of the platform. From January of 2017 (FanCentro launch), the weekly registrations show an increasing trend until a peak was reached in November of 2018. Since then the registration rate has been generally sustained, until we observe a spike in registrations the last week of March 2020, followed by the first week of April 2020, with 196 and 161 new users, respectively. This sharp increase in new users towards the end of March 2020 is also reflected in other similar sites, and it can be linked to the coronavirus pandemic, the consequent lockdowns, and its implications for sex work [6, 10].

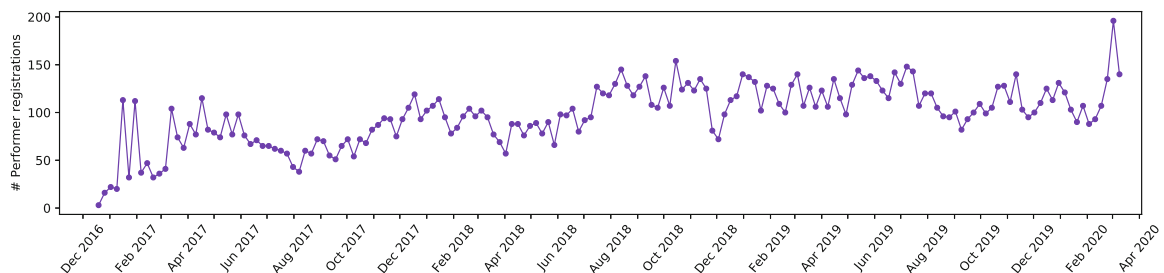


Fig. 2. Weekly registrations

3 Results and Discussion

3.1 Characterizing Performers

Table 1. Sexual identity and orientation

Sexual orientation	Sexual identity			
	Female	Male	Trans	Total
Bisexual	2486	52	41	2579
Gay	202	34	5	241
Straight	3544	141	27	3712
Trans	18	4	44	66
Total	6250	231	117	6598

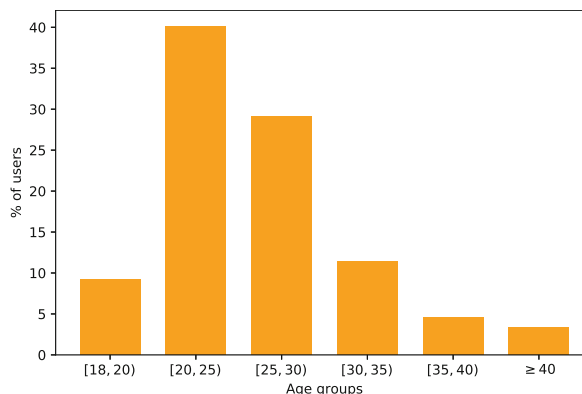


Fig. 3. Age distribution

In this section, we study the collected profiles in terms of characterizing attributes. This includes self-reported demographic information (i.e. sexual identity and orientation, age), descriptive tags, and external links to other sites, as provided by Performers. In Table 1, we report the number of profiles per sexual identity and orientation. Notably, 9,879 profiles did not include this information. Nevertheless, after analyzing the rest of the profiles, we can conclude that the majority of Performers identify as straight females. In Fig. 3, we depicted the age distribution for the profiles containing the birth-date attribute (4,526 profiles). We observe that the most common age group is 20–25 years (1,857 profiles), followed by 25–30 (1,347 profiles). The latter means that the 70% of Performers who reported their birthday are within the age bracket of 20 to 30 years. The next step of our analysis focused on the tags used by the Performers. In this regard, Fig. 4a shows a WordCloud representation of the most frequent tags used by Performers (found in 4,558 profiles). We observe that they mostly include pornographic terms, with “sexy” and “ass” being the most popular (1,472 and 928 occurrences, respectively). The outcomes of the analysis of the external links are depicted in Fig. 4b. We can observe that the most common external links from the profiles collected in our dataset are Instagram and Twitter, closely followed by public Snapchat accounts. This indicates that Performers orchestrate their online presence across multiple social outlets, enabling them to reach and engage a diverse audience. Moreover, Amazon wish lists, webcam modelling (“*camming*”) platforms [8] and porn sites have a relevant representation.

3.2 Exploring the Supply and Demand

In order to get an insight into the activities performed in FanCentro, we analyze the metadata information related to the collected profiles, including the amount of funds

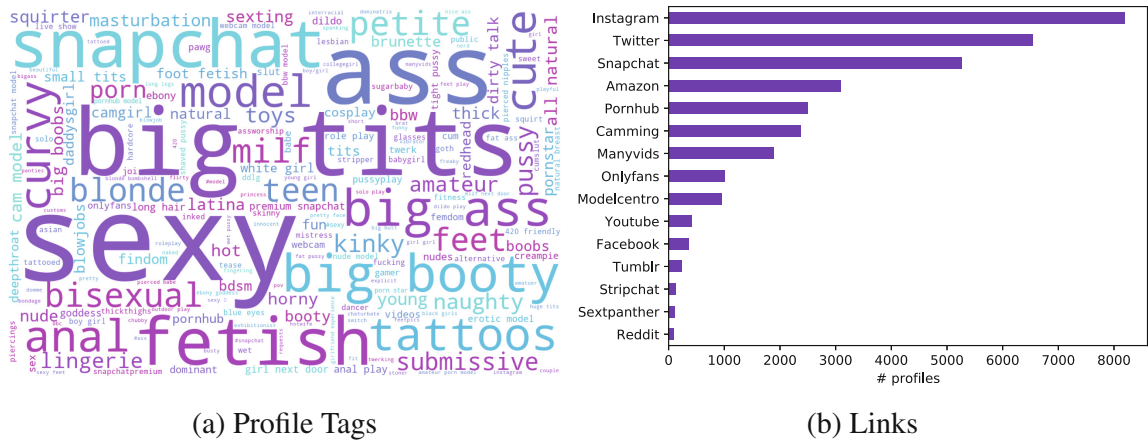


Fig. 4. Descriptive characteristics of Performers profiles: tags and links to external sites.

payable to the Performer in the next payout (revenue), the followers and the content posted by Performers.⁵

The *revenue* reflects the monetary sum of recurring sales (i.e. subscriptions) at crawl time, plus any income from one-off payments (including gratuity/tips, video clip sales and ‘lifetime access’ services) that are on hold by FanCentro until the next payout to the content creator.⁶ Provided the dynamic nature of subscriptions and content produced by Performers, revenue is a quantity that fluctuates due to a variety of reasons, including cancellation of subscriptions, chargebacks, external factors governing Performers’ popularity, etc. To assess the extent to which the revenue fluctuates over time, we use a snapshot of FanCentro profiles that we collected on March 2nd, 2020. To this end, a two-tailed Kolmogorov-Smirnov test was used, revealing no significant differences in Performers’ revenues between two consecutive months ($p = 0.44$). We found that the revenue distribution is extremely skewed, with the overwhelming majority of the Performers (96.4%) generating zero revenue within the aforementioned period. In Fig. 5a, we plot the revenue cumulative distribution function (CDF) for the 602 revenue-earning Performers (3.6% of profiles). We observe that 80% are below the minimum payout threshold of 100 USD⁷, meaning that only a negligible fraction of the performers in our dataset (0.8% approx.) would be certain to receive income by FanCentro during the next payout. Nonetheless, the revenues for the period between 23 March - 5 April 2020 period reach up to 12, 615 USD. In total, the gross earnings of Performers amount to 73, 607 USD for the payout period captured in our dataset.

Next, in Fig. 5b, we show the CDF of the number of followers. Contrary to the revenue, 78.5% of the profiles have followers. However, the revenue-generating Performers have up to two orders of magnitude more followers than the rest. The statistical significance of this difference was also confirmed by a two-tailed Kolmogorov-Smirnov test

⁵ Revenue is personal in nature and is normally visible only via the dashboard of each Performer. We have contacted FanCentro regarding this matter, and it has been removed from the data delivered via the public API.

⁶ FanCentro pays Influencers once a week after two weeks of the revenue generation date according to the license agreement.

⁷ <https://centroprofits.com/faq>.

($p < 0.01$). In terms of posts, Performers in total have uploaded 73, 233 photos, 43, 860 videos and 4, 867 clips, with the first two being part of their media feeds, while the clips are sold separately. Figure 5c shows the CDF of the total number of posts. We observe that Performers earning income have clearly more posts than the ones who do not, however, the majority of Performers have less than ten postings (61% and 93% for the revenue and non-revenue generating ones, respectively). Again, a two-tailed Kolmogorov-Smirnov test confirms that the difference between the distributions of the number of posts for revenue and non-revenue earning Performers is significant ($p < 0.01$). The low number of posts indicates that Performers, generally prefer to share their content in outlets different than FanCentro.

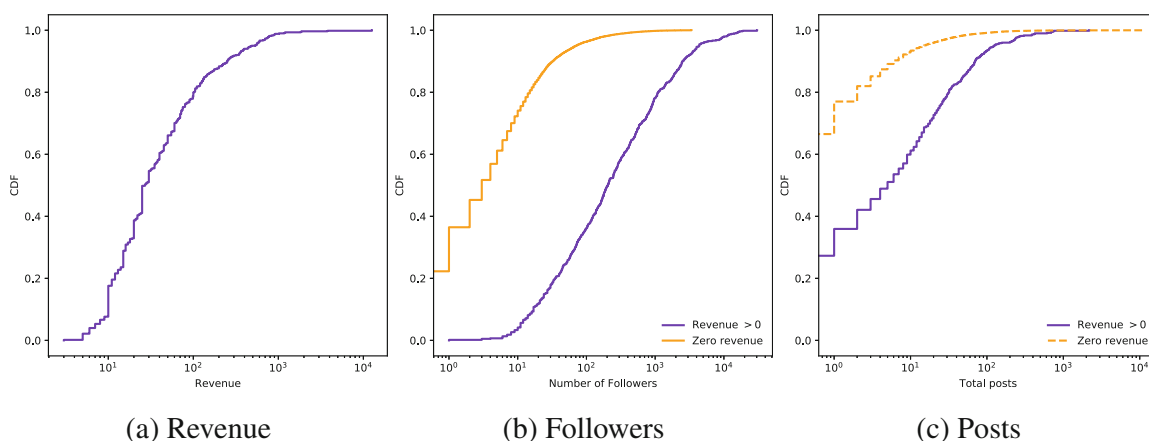


Fig. 5. Cumulative distribution functions (CDFs) of (non-zero) revenue, number of followers and posts.

3.3 FanCentro Content

To get a better understanding of the content Performers upload in FanCentro, we analyze their media feeds which, in terms of access, can contain two kinds of posts: *private* (only accessible by paying subscribers to their media feed) and *public* (freely accessible). In our dataset, the majority (89%) of posts are private (104, 737 posts), while the rest are public (12, 356 posts).

In Fig. 6, we depict the number of posts per month. We observe a consistently increasing trend in the number of posts, with a spike of 21, 300 posts in December 2018, followed by March 2020 (7, 325 posts), which is the second most active month in terms of posting activity. Next, we examine the characteristics of Performers' posts in terms of text content (titles) and user reactions, which results are depicted in Fig. 7. In our dataset, user reactions to Performers' posted content are relatively scarce, with 79% and 92% of the posts receiving zero likes and comments, respectively. This behaviour can be observed in Fig. 7a, which shows the CDFs of the reactions per post.

Notably, the majority of these posts received just one reaction, while the most popular post in our dataset has 316 likes and 55 comments. The low number of reactions

3.4 Premium Social Media Accounts

We conclude our analysis by examining the different payment models for accessing the different channels used by Performers to distribute their private content. In FanCentro, the purchasable services include access to “premium” Snapchat and Instagram accounts and the platform’s private media feed⁸. In the collected data we identified three separate payment models for accessing these services: *one-time*, *recurring* and *free trial*. The first two refer to one-off and recurring payments to access new content, respectively, while the “free trial” model allows customers to have a month of free access to the specific service, before reverting to recurring subscription payment. In Table 2, we present the distribution of the different payment models for the offered services. Private Snapchat is by far the most popular premium service, and the majority of Performers prefer offering their services as subscriptions.

Table 2. Premium services

Premium service	Payment model			
	Recurring	One-time	Free Trial	Total
Snapchat	11635	1153	41	12829
FanCentro	4716	0	5	4721
Instagram	1741	191	0	1932
Total	18092	1344	46	19482

The mean price of the Performers selling their services under one-off payments is 30 USD for Snapchat and 32 USD for Instagram. To get a deeper insight into the recurring payment model adopted by the majority of Performers, in Fig. 8 we present the distribution of subscription offerings, and in Fig. 9 we show the monthly subscription price distribution per service and total subscription duration. For simplicity, we only consider the subscription periods with more than 100 occurrences in our dataset. We note that Performers can offer their services at discounted rates as a means of promotion (similar to free trial access), which comprise a small fraction of the total offerings (2,004 in total).

In Fig. 8, we observe that the most popular service is the yearly Snapchat subscription, offered by 5,892 performers, followed by monthly Snapchat subscription (3,828 offerings) and yearly access to FanCentro feed (2,921 offerings). While three-month and half-year subscriptions exist, they are not common, accounting only for 25% of total offerings. The subscription fee is calculated on the total subscription period. As such, the monthly price generally decreases as the subscription duration increases. The monthly subscription to Performers’ premium accounts, which is the pricier option in

⁸ Recently FanCentro has introduced a purchasable direct messaging service enabling direct communication between users and Performers. Nonetheless, we excluded it from our analysis due to the low number of observations in our dataset.

all cases, on average costs 21.7 USD and 58 USD for Snapchat (Fig. 9a) and Instagram (Fig. 9c) accounts, respectively. Notably, in the first case, the price can go up to 5,000 USD, and in the second case up to 8,000 USD. In this regard, the lowest priced service is access to FanCentro media feed ($\mu = 17.3$ USD), which can cost up to 500 USD monthly (Fig. 9b). Nevertheless, the most common subscription duration is one year, priced on average 10 USD/month for Snapchat and FanCentro feed, and 14 USD/month for Instagram. Additionally, discounted rates show an average decrease of 6 USD/month for Snapchat, 14 USD/month for Instagram and 4 USD/month for FanCentro feed, when compared to the normal prices of each service, respectively.

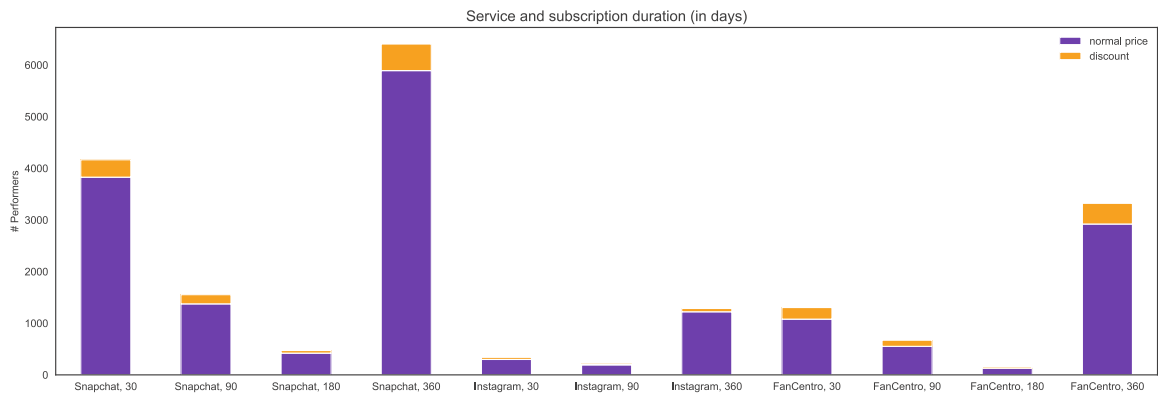


Fig. 8. Number of offerings per service and subscription duration.

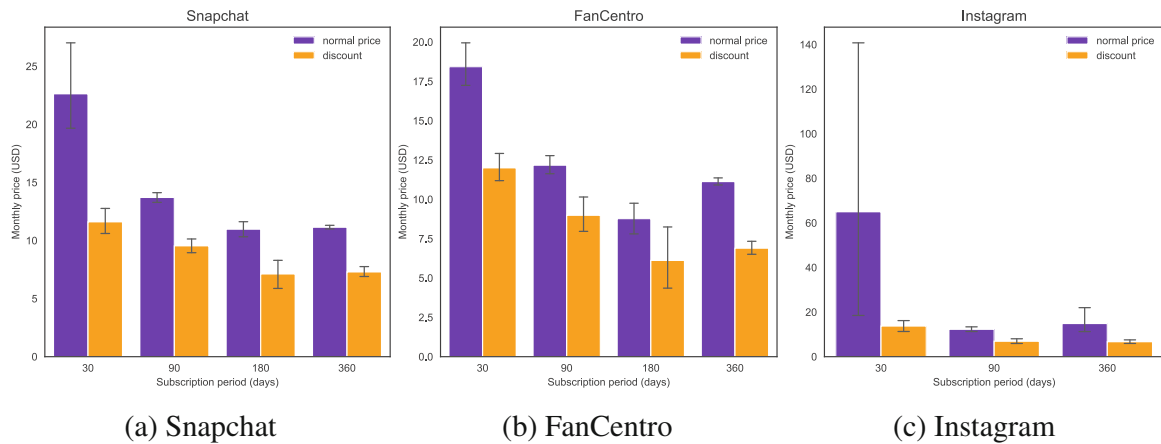


Fig. 9. Bar plots of monthly subscription price (normal and discounted) per service and subscription duration.

4 Conclusions

In this work, we performed the first quantitative analysis of the semi-illicit adult content market layered on the top of popular social media platforms like Snapchat and Instagram. To this end, we studied the demographics and activity of the selling users in FanCentro, as well as some descriptive characteristics of the content they upload. The

existence of sites like FanCentro where Performers can openly sell and promote premium social media accounts indicate that the industry built on the inefficacy of social media platforms to enforce community guidelines for effectively banning adult content is here to stay. This inefficacy is exploited and monetized in large scale, exacerbated by the fact the explicit content is staying “hidden” in private accounts, access to which is sold through the different models studied.

Moreover, our findings indicate that the coronavirus-induced lockdowns have accelerated the growth of this marketplace. This phenomenon is also reflected by the rise of other influencer-centric adult content markets, such as OnlyFans, which observed a major increase in traffic during the coronavirus pandemic [6, 10]. In part, this is due to the fact that a large number of sex workers lost their original revenue streams because of the virus; in addition, an increasing number of influencers transition to online sex work as a means to adapt to the economic downturn which caused companies to reduce marketing budgets, that would have been otherwise used for sponsored content [4]. The strong online presence of Performers across multiple popular social media sites where they openly promote their paid content signals the shift of online adult content industry towards an increasingly mainstream, gig economy. Nonetheless, the proliferation of adult content flowing unobstructed through social media, diffused and being promoted via users with large followings, might pose a serious threat to the safety of mainstream online communities, especially for the younger users.

Acknowledgements. This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the project *LOCARD* (<https://locard.eu>) (Grant Agreement no. 832735).

The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions.

References

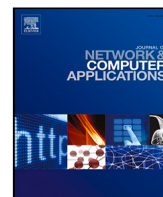
1. Baumgartner, S.E., Sumter, S.R., Peter, J., Valkenburg, P.M.: Sexual self-presentation on social network sites: who does it and how is it perceived? *Comput. Hum. Behav.* **50**, 91–100 (2015)
2. Clarke, L.: The x-rated world of premium Snapchat has spawned an illicit underground industry (2019). <https://www.wired.co.uk/article/premium-snapchat-adult-models>. Accessed 04 May 2020
3. Daniels, E.A.: Sexiness on social media: the social costs of using a sexy profile photo. *Sex. Media, Soc.* **2**(4), 2374623816683522 (2016)
4. Downs, C.: OnlyFans, Influencers, and the Politics of Selling Nudes During a Pandemic (2020). <https://www.elle.com/culture/a32459935/onlyfans-sex-work-influencers/>. Accessed 14 May 2020
5. Drenten, J., Gurrieri, L., Tyler, M.: Sexualized labour in digital culture: Instagram influencers, porn chic and the monetization of attention. *Gender Work Org.* **27**(1), 41–66 (2020)
6. Drolet, G.: Sex Work Comes Home (2020). <https://www.nytimes.com/2020/04/10/style/camsoda-onlyfans-streaming-sex-coronavirus.html>. Accessed 04 May 2020

7. Gómez, A.R.: Digital fame and fortune in the age of social media: A classification of social media influencers. *aDRResearch: Revista Internacional de Investigación en Comunicación* (19), 8–29 (2019)
8. Henry, M.V., Farvid, P.: “Always hot, always live”: computer-mediated sex work in the era of camming. *Women’s Stud. J.* **31**(2), 113–18 (2017)
9. Khamis, S., Ang, L., Welling, R.: Self-branding, “micro-celebrity” and the rise of social media influencers. *Celebr. Stud.* **8**(2), 191–208 (2017)
10. Lee, A.: Coronavirus is bad news for Big Porn but great news for OnlyFans (2020). <https://www.wired.co.uk/article/coronavirus-porn-industry-onlyfans>. Accessed 04 May 2020
11. Lou, C., Yuan, S.: Influencer marketing: how message value and credibility affect consumer trust of branded content on social media. *J. Interact. Advert.* **19**(1), 58–73 (2019)
12. Lykousas, N., Gómez, V., Patsakis, C.: Adult content in social live streaming services: characterizing deviant users and relationships. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 375–382. IEEE (2018)
13. Nandagiri, V., Philip, L.: Impact of influencers from Instagram and Youtube on their followers. *Int. J. Multidisc. Res. Modern Educ.* **4**(1), 61–65 (2018)
14. van Oosten, J.M., Peter, J., Boot, I.: Exploring associations between exposure to sexy online self-presentations and adolescents’ sexual attitudes and behavior. *J. Youth Adolesc.* **44**(5), 1078–1091 (2015)
15. van Oosten, J.M., Vandenbosch, L.: Sexy online self-presentation on social network sites and the willingness to engage in sexting: a comparison of gender and age. *J. Adolesc.* **54**, 42–50 (2017)
16. Terranova, T.: Attention, economy and the brain. *Cult. Mach.* **13** (2012)
17. Zimmer, M.: “But the data is already public”: on the ethics of research in Facebook. *Ethics Inf. Technol.* **12**(4), 313–325 (2010)



Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Intercepting Hail Hydra: Real-time detection of Algorithmically Generated Domains

Fran Casino^{a,b}, Nikolaos Lykousas^a, Ivan Homoliak^c, Constantinos Patsakis^{a,b,*},
Julio Hernandez-Castro^d

^a Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece

^b Information Management Systems Institute, Athena Research Center, Artemidos 6, Marousi 15125, Greece

^c Faculty of Information Technology, Brno University of Technology, Czech Republic

^d School of Computing, University of Kent, United Kingdom

ARTICLE INFO

Keywords:

Malware
Domain Generation Algorithms
Botnets
DNS
Algorithmically Generated Domain

ABSTRACT

A crucial technical challenge for cybercriminals is to keep control over the potentially millions of infected devices that build up their botnets, without compromising the robustness of their attacks. A single, fixed C&C server, for example, can be trivially detected either by binary or traffic analysis and immediately sink-holed or taken-down by security researchers or law enforcement. Botnets often use Domain Generation Algorithms (DGAs), primarily to evade take-down attempts. DGAs can enlarge the lifespan of a malware campaign, thus potentially enhancing its profitability. They can also contribute to hindering attack accountability.

In this work, we introduce HYDRAS, the most comprehensive and representative dataset of Algorithmically-Generated Domains (AGD) available to date. The dataset contains more than 100 DGA families, including both real-world and adversarially designed ones. We analyse the dataset and discuss the possibility of differentiating between benign requests (to real domains) and malicious ones (to AGDs) in real-time. The simultaneous study of so many families and variants introduces several challenges; nonetheless, it alleviates biases found in previous literature employing small datasets which are frequently overfitted, exploiting characteristic features of particular families that do not generalise well. We thoroughly compare our approach with the current state-of-the-art and highlight some methodological shortcomings in the actual state of practice. The outcomes obtained show that our proposed approach significantly outperforms the current state-of-the-art in terms of both classification performance and efficiency.

1. Introduction

The continuous arms race between malware authors and security researchers has pushed modern malware to evolve into highly sophisticated software, capable of infecting millions of devices. The vast amount of sensitive information that can be extracted from compromised devices, coupled with the harnessing of their resources and processing power, provides a wide range of monetisation methods fuelling a flourishing worldwide underground economy.

While device infection is the key that paves the way in, the main objectives are generally persistence and orchestration. An orchestrating entity, the botmaster, manages infected devices (bots) which in many cases can scale to the order of millions, creating a botnet (Singh et al., 2019). The botmaster manages a Command and Control (C&C) server that communicates with the bots. This communication must preserve

some degree of unlinkability to thwart any attempts to identify the botmaster. To ensure unlinkability, and as a counter-measure against take-down operations, botnets frequently make use of domain fluxing (Perdisci et al., 2012; Zang et al., 2018) through Domain Generation Algorithms (DGAs). DGAs produce a vast amount of domain names, which bots try to communicate with iteratively to find the actual C&C server. However, only a small part of them is registered and active, creating a hydra effect (Nadji et al., 2013). The botmaster may regularly pivot control between domains, thus hampering the task of seizing control of the botnet. This is helped by the fact that an outsider cannot determine which domains will be used, nor statically block all these requests. The latter stems from the fact that there are too many domains, and the seed yielding a particular sequence of domains might

* Corresponding author at: Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece.

E-mail addresses: francasino@unipi.gr (F. Casino), nlykousas@unipi.gr (N. Lykousas), ihomoliak@fit.vutbr.cz (I. Homoliak), kpatsak@unipi.gr (C. Patsakis), J.C.Hernandez-Castro@kent.ac.uk (J. Hernandez-Castro).

<https://doi.org/10.1016/j.jnca.2021.103135>

Received 31 January 2021; Received in revised form 19 April 2021; Accepted 2 June 2021

Available online 25 June 2021

1084-8045/© 2021 Elsevier Ltd. All rights reserved.

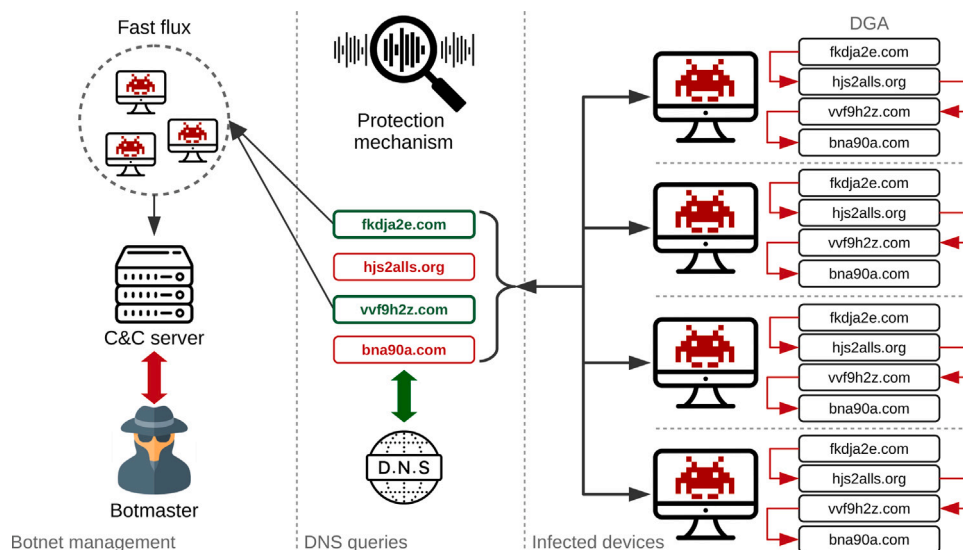


Fig. 1. The *modus operandi* of a typical DGA-powered botnet (Patsakis et al., 2020).

be unknown or change frequently. Fig. 1 illustrates the *modus operandi* of a typical DGA-powered botnet.

Currently, there are several families of DGAs employed by various malware with varying rates of requests and different characteristics. This heterogeneous landscape hinders the timely and accurate detection of an Algorithmically-Generated Domain (AGD) (Yadav and Reddy, 2012) request, which could serve as a precise indicator of compromise (IoC) of a host at the network level. Recent research tries to categorise DNS requests per DGA, often exploiting WHOIS-based features. From the perspective of an ISP, CSIRT or CERT such an approach might be beneficial. However, we argue that in terms of endpoint security that strategy cannot be considered adequate. First, the network operator does not generally know which concrete DGA is utilised by the malware that infected a given host in her network. Second, the utilisation of the WHOIS database introduces a significant time delay that in many situations cannot be tolerated.

1.1. Contributions

Motivated by the continuous evolution of DGAs we introduce a dataset collecting real-world domains called HYDRAS, which consists of more than 95 million domains belonging to 105 unique DGA families. To the best of our knowledge, this is the largest and most representative DGA dataset to date.

During the analysis of our dataset, the possibility of differentiating between benign and malicious requests in real time is discussed, as well as the identification of the malware families using them. Based on information learned from the analysis of our dataset, a novel feature set is designed and implemented, which includes lexical and statistical features over the collected DGAs, as well as English gibberish detectors. Using the proposed feature set and a Random Forest as a representative of ensemble classifiers, we perform a thorough evaluation of our dataset and show that our feature set together with the Random Forest classifier outperform the state-of-the-art approaches in terms of both classification performance and overhead.

Next, inherent biases in related works are highlighted. These biases can be attributed to the suboptimal selection of datasets and/or features, preventing their application in general, real-world scenarios. For example, employing a dataset comprising only a few families that exhibit very characteristic patterns might ease the classification task, providing accurate detection rates (e.g., the generators for cryptolocker, ramnit, geodo, locky, tempedreve, hesperbot, fobber and dircrypt provide a uniform distribution of letters, Tran

et al., 2018; Woodbridge et al., 2016), but will inevitably lead to ad-hoc solutions that are too specific and cannot be generalised. This is, unfortunately, a common practice in the existing literature.

A typical example is only considering families like bamital, CCleaner or chir in the datasets, which all produce hexadecimal values of specific length as second-level domains (SLDs). It is obvious that one can easily differentiate benign domains from such DGAs with almost 100% accuracy by merely checking whether the SLD is a hex value of a specific length. Nonetheless, not all DGAs families are so easy to detect in real scenarios.

Due to the particularities of this research field and the methodologies used by the current state-of-the-art, we also highlight some recommendations for fairer future evaluations. These are particularly relevant for comparing the results of our experiments with other approaches.

1.2. Organisation

The rest of this work is organised as follows. In Section 2, DGA-related preliminaries are briefly described, and in Section 3 a thorough review of related work is provided. Then, in Section 4, our dataset is detailed. Afterwards, we describe our approach, including methodology, feature extraction and the tools and algorithms employed. In Section 5, we describe the proposed features. We provide the results of our experiments in Section 6, Sections 7 and 8, where they are compared to the state-of-the-art. In Section 9, an analysis of the outcomes as well as a methodological analysis and comparison with the literature is provided. Finally, the paper concludes in Section 10, discussing open issues for future research.

2. Background

DGAs are one of the main pillars behind the success of botnets. They were first conceived more than a decade ago, and they have been steadily refined over the years by successive generations of malware developers. These algorithms generate a set of AGDs to communicate with C&C servers, thus eliminating the risks associated with using static IP addresses (Antonakakis et al., 2017; Nadji et al., 2017). In Patsakis and Casino (2019), the authors generalise the notion of DGAs by extending them to other protocols beyond DNS, and they propose the term Resource Identifier Generation Algorithms (RIGAs). The authors show how decentralised permanent storage (DPS) has some potential drawbacks and exploitable characteristics for armouring a botnet, a fact that has already been exploited in the real world (Anomali Labs, 2019)

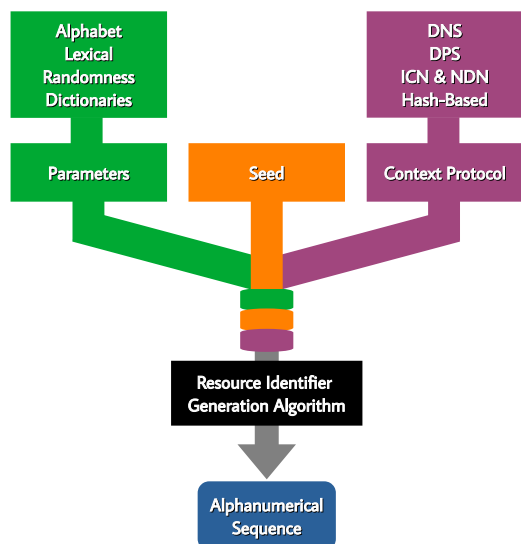


Fig. 2. RIGA generation flow, including diverse context protocols (e.g., DPS and DNS).

due to the immutability properties of DPS. Fig. 2 depicts the hierarchy of RIGAs.

In its most basic form, DGAs create a set of domain names by using a deterministic pseudo-random generator (PRNG) (Sood and Zeadally, 2016; Perdisci et al., 2012). Therefore, (infected) devices belonging to a botnet query a set of domains generated by the DGA until they are correctly resolved to a valid IP, corresponding to the C&C server. Since the location of the C&C server dynamically changes, blacklisting domains is a very inefficient protection technique. Additionally, this makes seizing the botnet much more difficult, since one would need to take (register) all domain names generated by the DGA (with a given seed) for disrupting the botmaster only for a short amount of time. This process will generally be very costly, typically involving thousands of domain names for stopping the botnet for just a day. Hence, the botmaster benefits from this asymmetry between the high ratio of generated domains to registered ones. That makes her operation cheap, as compared to the cost of defending against it, which involves registering all possible domains.

3. Related work

3.1. Traditional approaches

According to the literature, there are two main DGA families: (i) **Random-based** DGA methods, which use a PRNG to generate a set of characters that form a domain name, and (ii) **Dictionary/Wordlist-based** DGA methods, which use a dictionary to produce domains. Nevertheless, one may also consider other types of DGA families, which use more subtle approaches, i.e. valid domains that were previously hacked to hide their C&C servers (i.e. domain shadowing) (Liu et al., 2017) as well as DGAs that generate domains that are very similar to existing valid domains (Bader, 2015), further hindering the detection task. Considering the dependency of the pre-shared secret (or seed) on time, Plohmann et al. (2016) further categorise DGAs into: (i) *time-independent and deterministic*, (ii) *time-dependent and deterministic*, and (iii) *time-dependent and non-deterministic*.

In the case of random-based DGA detection, a common practice is to analyse some features of the domain names and their lexical characteristics to determine whether a DGA has generated them (Aviv and Haeberlen, 2011; Yadav et al., 2012). Moreover, auxiliary information such as WHOIS and DNS traffic (e.g. frequent NXDomain responses) is often used to detect abnormal behaviours (Zhou et al., 2013; Jiang

et al., 2010; Antonakakis et al., 2012). Other approaches use machine learning-based techniques and combine the previous information to identify Random-based DGA such as in Jiang et al. (2010), Yadav and Reddy (2012), Manadhata et al. (2014) and Zhao et al. (2015).

Due to their inherent construction, wordlist-based DGA detection represents a challenging task for classifiers. In this regard, the common approach is to use machine learning approaches (e.g., feature-based classification and deep learning) to distinguish between benign and malicious DGAs. The use of random forest classifiers (RF) based on a set of features such as word correlation, frequency, and part-of-speech tags was first proposed in Yang et al. (2018). Similarly, Selvi et al. suggested the use of RF with masked n-grams as a feature, achieving a remarkable accuracy in the binary classification task (Selvi et al., 2019). Berman (2019) put forward a methodology based on Capsule Networks (CapsNet) to detect AGDs; the author compared his method with well-known approaches such as RNNs and CNNs, and the outcomes showed similar accuracy yet better computational cost. Xu et al. (2019) suggested the combination of n-gram and a deep CNNs to create an n-gram combined with character-based domain classification (n-CBDC) model that does not require domain feature extraction. Vinayakumar et al. (2019) implemented a set of deep learning architectures with Keras and classical machine learning algorithms to classify DGA families. Their best configuration uses RNNs and SVMs with a radial basis function (SVM-RBF). Yang et al. (2020) present a heterogeneous deep neural network framework, which extracts the local features of a domain name as well as a self-attention based Bi-LSTM to extract further global features. Their outcomes showed higher accuracy than traditional DGA classifiers. Finally, a recent approach based on the probabilistic nature of wordlist-based DGAs was proposed in Patsakis and Casino (2021). In their work, Patsakis et al. proposed the combination of feature-based extraction with a probabilistic-based threshold to fully capture wordlist-based AGDs. Moreover, their method was capable of detecting both real-world and custom DGAs created to fool traditional detectors.

3.2. Adversarial and anti-forensic approaches

Recently the exciting development of deploying anti-forensic techniques in DGAs has become popular. This aims to create hard-to-detect DGA families and to fight against high performing classifiers. Anderson et al. (2016) proposed a generative adversarial network (GAN), which can learn from and bypass classical detectors. Afterwards, they improved the performance of AGD detectors after training them with the data generated by the GAN. Alaeiyan et al. (2020) proposed a DGA family created with a genetic algorithm considering lexical features such as pronounceability. Their experiments showed that such a DGA family was hard to detect by classical approaches. In a similar vein, Yun et al. (2020) used n-gram distribution and the pronounceability/readability of domains as a basis to create a novel DGA based on neural language models and the Wasserstein GAN (WGAN), which reduced detection rates in traditional DGA techniques.

Spooren et al. (2019) showed that their deep learning RNN performs significantly better than classical machine learning approaches. Besides, the authors stressed that one of the issues of manual feature engineering is that an adversary may adapt her strategy if she knows which features were used in the detection. Fu et al. (2017) proposed two DGAs using hidden Markov models (HMMs) and probabilistic context-free grammars, which were tested on state-of-the-art detection systems. Their results revealed their DGAs hindered the detection rate known detectors.

Finally, and due to the widespread use of covert or encrypted communication channels in DNS (e.g., DNSCurve, DNS over HTTPS and DNS over TLS) and in C&C connections in general (Zander et al., 2007; Homoliak et al., 2014), malware creators have an additional tool to hide their activity, rendering many traditional DGA detection useless.

Nevertheless, as shown by Patsakis et al. (2020), NXDomain detection can still be carried out in such a scenario. This also applies to feature extraction, so DGA families can still be classified with high performance.

4. The HYDRAS dataset

In this section, the HYDRAS dataset is introduced, which consists of a collection of benign and AGD domains, both real-world as well as adversarial. The name of the dataset originates from the insightful parallelism suggested by [Nadji et al. \(2013\)](#) between DGA-powered botnets and the mythical ancient Greek monster.

Benign domains are sampled from the Alexa 1M dataset. But since the Alexa dataset contains sites, not domains, it had to be preprocessed. First, all top-level domain names (e.g., `.com`, `.org`) are removed from each entry and only the SLD are kept. Then, the duplicates were pruned since some web pages have multiple entries in the dataset (e.g., `google.com` and `google.co.in`) or been subdomains of identical services (e.g., various blogs of `blogspot.com`). Finally, all internationalised domain names were removed, since they are encoded using Punycode¹ representation. After preprocessing the 1M Alexa dataset, the final dataset contains 915,994 unique domains.

The use of small and unrepresentative datasets, unfortunately very frequent in the literature, leads to several biases and other issues that can easily lead towards wrong analysis and misleading conclusions. For instance, the public feed of DGAs provided by the Network Security Research Lab at 360² as well as the DGArchive ([Plohmann et al., 2016](#)) provide real-world datasets with millions of samples from many DGA families. Nonetheless, despite the numerous samples in both these datasets, many malware families are significantly underrepresented. A demonstrative example is the `xshellghost` family in the 360 dataset, which contains only a single sample at the time of writing. Thankfully, the researchers at 360 have reversed the code of this DGA.³

Since the provision of many samples is required to perform an adequate evaluation of any detection technique, we utilised the available code of poorly represented DGAs to enlarge our dataset. The dataset was initialised with several public DGA repositories, e.g., J. Bader's ([Bader, 2020](#)), A. Abakumov's ([Abakumov, 2020](#)), and P. Chaignon's ([Chaignon, 2020](#)). As explained above, we additionally used DGA code available at these and other repositories to generate additional samples for underrepresented DGA families. In these cases, a few random seeds and/or an extended date range to obtain new samples was used.

Since we used the code of the DGAs, the added domains have identical characteristics to original ones and might occur in the real-world. Thus, these AGDs could have been collected in a real setting. Moreover, the SLDs of three adversarially designed DGAs, namely `deception`, `deception2` ([Spooren et al., 2019](#)) and `khaos` ([Yun et al., 2020](#)) were added.

In summary, our dataset consists of 95,325,598 AGDs belonging to a total of 109 families, from which 105 are unique.⁴ The families included, along with their corresponding number of collected samples, are reported in [Table 1](#). The dataset is available for download at <https://zenodo.org/record/3965397> ([Casino et al., 2020](#)).

5. Proposed features

We thoroughly analysed the AGDs in our dataset, as well as the ideas behind existing AGD detection approaches in the literature. We found out that the basic strategy for detecting non-wordlist-based DGAs is to take advantage of the fact that they, in general, make little effort to be human-memorable, as they typically are randomly generated. Moreover, even if they show a high correlation with readable words in terms of vowel/consonant usage, etc., the generated domains are expected to contain zero to only a few words having a short length.

¹ <https://tools.ietf.org/html/rfc3492>

² <https://data.netlab.360.com/dga/>

³ <https://github.com/360netlab/DGA/blob/master/code/xshellghost/dga.py>

⁴ A few DGAs are used by multiple families.

5.1. Approach to feature extraction

A general description of our approach is as follows: On receiving a domain name, we first cache it to see correlations with previous ones. Then, we try to determine whether the SLD matches some specific patterns, e.g., whether it is a hex value, its combination of vowels/consonants, length, etc. Later, after removing all digits, we try to break the remaining characters into words. Within these words, the short ones (e.g., stop words, articles) are pruned and study the remaining to determine whether they are real words or just gibberish. Moreover, the entropy of the domain is computed and a subset of the patterns created during the correlation process. All the above provide us with several features that can be efficiently used to determine whether a domain name is benign or not, without the need for external information (e.g., WHOIS) or waiting for the DNS resolution revealing whether it is an NXDomain. In this way, a significant number of requests are pruned, regardless of their outcome.

Using the insights from our analysis of DGA families in the dataset, several features were engineered, defined in [Table 2](#). The first set of parameters is computed when trying to identify valid n-grams and words. For the former, we train our n-gram model with Alexa n-grams and lengths three, four and five. For the latter, the `wordninja`⁵ word splitter was used, which probabilistically analyses its input using NLP based on the unigram frequencies of the English Wikipedia. Hence, the domain is split into meaningful words, according to a minimum word-length w . Therefore, only terms which contain at least w characters are considered as significant. Then, we compute the percentage of the domain characters which are meaningful, by calculating the ratio γ between characters belonging to words and the domain's total length. Next, two more sets of features are computed according to statistical attributes as well as ratios using the previously calculated features.

Gibberish Detection. In addition, a Gibberish detection layer is used, which consists of two methods. The first one is a 2-character Markov chain Gibberish detector,⁶ which is trained with English text to determine how often characters appear next to each other. Therefore, a text string is considered valid if it obtains a value above a certain threshold for each pair of characters. The second is a Gibberish classifier.⁷ In this case, the method checks mainly three features of the text: whether (i) the amount of unique characters is within a typical range, (ii) the number of vowels is within a standard range and (iii), the word to char ratio is in a healthy range. Finally, the entropy of a subset of the alphanumeric sequences is computed, to enrich the feature set.

An exemplified overview of the feature extraction process is illustrated in [Fig. 3](#).

A Comparison with the State-of-the-Art. Despite the fact that n-grams and some of the ratio features used in this paper are well-known and have been previously used in the literature, the combination presented in this work is novel. Moreover, we propose the use of two different Gibberish detectors, the vowel distribution of the specific n-grams computed from the Alexa domains, the statistical features computed over the different length-based words extracted by `wordninja`, and the entropy used in a subset of this novel features.

6. Classification experiments

We assess the power of our proposed features (see [Section 5](#)) to differentiate between malicious and benign domains (i.e., binary classification), as well as between several families of DGAs (i.e., multiclass classification). Since both empirical and theoretical results have shown that a combination of models (in an ensemble) can increase classification performance ([Dietterich, 2000](#); [Valentini and Masulli,](#)

⁵ <https://github.com/keredson/wordninja>

⁶ <https://github.com/rrenaud/Gibberish-Detector>

⁷ <https://github.com/ProgramFOX/GibberishClassifier-Python>

Table 1

Distribution of records per DGA in our dataset. DGAs in green denote those which were frequently underrepresented, so they were run to create more samples, while purple indicates adversarial ones.

Class	Support	Class	Support	Class	Support	Class	Support
bamital	86892	feodo	247	omexo	41	sisron	2580
banjori	439423	fobber	2000	padcrypt	246096	sphinx	174726
bedep	7814	fobber_v1	298	pandabanker	32484	suppobox	98304
beebone	72	fobber_v2	299	pitou	74314	sutra	3295
bigviktor	999	gameover	22723000	pizd	16384	symmi	65
blackhole	732	geodo	576	post	66000	szribi	20661
bobax/kraken/oderoor	30459	gozi	163529	proslifean	218399	tempedreve	13323
ccleaner	12000	goznym	364	pushdo	380427	tinba	72719
chinad	750312	gspy	100	pushdotid	6000	tinynuke	52832
chir	100	hesperbot	16512	pykspa	1996763	tofsee	2100
conficker	2082010	infy	5220	pykspa_v1	44688	torpig	18716
corebot	20931	khaos	10000	pykspa_v2_fake	798	tsifiri	59
cryptolocker	368196	kingminer	252	pykspa_v2_real	198	ud2	491
cryptowall	56624	locky	994381	pykspa2	1248	ud3	20
darkshell	49	madmax	4850	pykspa2s	9960	ud4	70
deception	149854	makloader	256	qadars	630127	vawtrak	17807
deception2	149908	matsnu	40050	qakbot	4579999	vidro	62567
diamondfox	279	mirai	2716	qghost	23	vidrotid	101
dircrypt	11210	modpack	107	qsnatch	1246482	virut	23669176
dmsniff	70	monerodownloader	2995	ramdo	6000	volatilecedar	498
dnschanger	1499578	monerominer	364271	ramnit	150662	wd	32172
dromedan	10000	murofet	13824213	ranbyus	578080	xshellghost	12001
dyre	2046998	murofetweekly	600000	redyms	91	xxhex	1900
ebury	2000	mydoom	2599	rovnix	207996	zloader	29992
ekforward	3649	necurs	12751075	shifu	2554		
emotet	431048	nymaim	700102	shiotob/urlzone/bebloh	37031	Total	95,325,598
enviserv	500	nymaim2	110511	simda	24345		

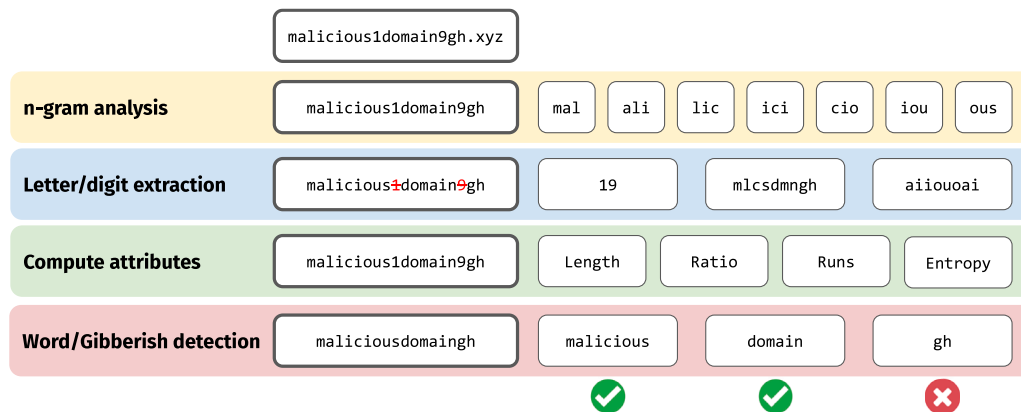


Fig. 3. Exemplified overview of the feature extraction process.

2002; Barandela et al., 2003; Kuncheva, 2014) even in the case of imbalanced datasets (Barandela et al., 2003), we opt for a Random Forest, which is a non-parametric ensemble classifier. Random Forest has previously achieved outstanding performance results in DGA classification tasks (Alaeiyan et al., 2020; Anderson et al., 2016; Selvi et al., 2019).

The hyperparameters of the Random Forest algorithm were tuned with grid search, to maximise classification performance in the task of distinguishing between benign and malicious domains over a subset of our dataset. We found that best performance is achieved using an ensemble of 100 decision trees with unlimited depth and bootstrap aggregation (i.e., bagging), where each new tree is fitted from a bootstrap sample of the training data (Breiman, 1996).

All our experiments were performed on a system equipped with an NVIDIA TITAN Xp PG611-c00 to speed-up the computations, utilising the scikit-learn⁸ library. The performance of the trained classifiers is evaluated using the standard classifications metrics of Precision, Recall, F_1 score and the area under the curve (AUC).

In all experiments, the same feature set⁹ were used and employed standard 10-fold cross-validation to avoid overfitting and get a roughly unbiased estimate of the performance of the trained models. Although the same feature sets for all experiments were used, we optimised weights of the features per DGA family, and thus targeting a binary classification (i.e., per DGA detection). In a binary classification, this is a justified setting since feature weights for a particular DGA family are not expected to vary in time. However, this is opposed to multi-class classification (requiring frequent feature/weight tuning), which we argue is not convenient for DGA detection since it deals with the more challenging classification problem.¹⁰ Also, it should be noted that the multiclass classification is not the focus of this work, and those experiments carried out only for the sake of comparison with state-of-the-art approaches.

⁹ We note that, on occasion, a few peculiarities might eliminate one feature due to reasons further detailed.

¹⁰ I.e., intuitively, it is more difficult to find accurate separating hyperplanes among multiple classes than between two classes.

⁸ <https://scikit-learn.org>

Table 2
Features used in our approach and their corresponding description.

Feature set	Notation	Description
Alphanumeric sequences	<i>Dom</i>	Domain without TLD
	<i>Dom - D</i>	<i>Dom</i> without digits
	<i>Dom - 3G</i>	Set of 3-grams of <i>Dom</i>
	<i>Dom - 4G</i>	Set of 4-grams of <i>Dom</i>
	<i>Dom - 5G</i>	Set of 5-grams of <i>Dom</i>
	<i>Dom - W</i>	Domain concatenated words
	<i>Dom - WS</i>	Domain concatenated words with spaces
	<i>Dom - WD</i>	<i>Dom - D</i> concatenated words
	<i>Dom - WDS</i>	<i>Dom - D</i> concatenated words with spaces
	<i>Dom - W2</i>	Domain concatenated words of length > 2
	<i>Dom - W3</i>	Domain concatenated words of length > 3
Statistical attributes	<i>L - HEX</i>	The domain name is represented with hexadecimal characters
	<i>L - LEN</i>	The length of <i>Dom</i>
	<i>L - DIG</i>	The number of digits in <i>Dom</i>
	<i>L - DOT</i>	The number of dots in the raw domain
	<i>L - CON - MAX</i>	The maximum number of consecutive consonants <i>Dom</i>
	<i>L - VOW - MAX</i>	The maximum number of consecutive vowels <i>Dom</i>
	<i>L - W2</i>	Number of words with more than 2 characters in <i>Dom</i>
	<i>L - W3</i>	Number of words with more than 3 characters in <i>Dom</i>
Ratios	<i>R - CON - VOW</i>	Ratio of consonants and vowels of <i>Dom</i>
	<i>R - Dom - 3G</i>	Ratio of benign grams in <i>Dom - 3G</i>
	<i>R - Dom - 4G</i>	Ratio of benign grams in <i>Dom - 4G</i>
	<i>R - Dom - 5G</i>	Ratio of benign grams in <i>Dom - 5G</i>
	<i>R - VOW - 3G</i>	Ratio of grams that contain a vowel in <i>Dom - 3G</i>
	<i>R - VOW - 4G</i>	Ratio of grams that contain a vowel in <i>Dom - 4G</i>
	<i>R - VOW - 5G</i>	Ratio of grams that contain a vowel in <i>Dom - 5G</i>
	<i>R - WS - LEN</i>	<i>Dom - WS</i> divided by <i>L - LEN</i>
	<i>R - WD - LEN</i>	<i>Dom - WD</i> divided by <i>L - LEN</i>
	<i>R - WDS - LEN</i>	<i>Dom - WDS</i> divided by <i>L - LEN</i>
	<i>R - W2 - LEN</i>	<i>Dom - W2</i> divided by <i>L - LEN</i>
	<i>R - W2 - LEN - D</i>	<i>Dom - W2</i> divided by <i>Dom - D</i>
	<i>R - W3 - LEN</i>	<i>Dom - W3</i> divided by <i>L - LEN</i>
	<i>R - W3 - LEN - D</i>	<i>Dom - W3</i> divided by <i>Dom - D</i>
Gibberish probabilities	<i>GIB - 1 - Dom</i>	Gibberish detector 1 applied to <i>Dom</i>
	<i>GIB - 1 - Dom - WS</i>	Gibberish detector 1 applied to <i>Dom - WS</i>
	<i>GIB - 1 - Dom - D</i>	Gibberish detector 1 applied to <i>Dom - D</i>
	<i>GIB - 1 - Dom - WDS</i>	Gibberish detector 1 applied to <i>Dom - WDS</i>
	<i>GIB - 1 - Dom - W2</i>	Gibberish detector 1 applied to <i>Dom - W2</i>
	<i>GIB - 1 - Dom - W3</i>	Gibberish detector 1 applied to <i>Dom - W3</i>
	<i>GIB - 2 - Dom</i>	Gibberish detector 2 applied to <i>Dom</i>
	<i>GIB - 2 - Dom - WS</i>	Gibberish detector 2 applied to <i>Dom - WS</i>
	<i>GIB - 2 - Dom - D</i>	Gibberish detector 2 applied to <i>Dom - D</i>
	<i>GIB - 2 - Dom - WDS</i>	Gibberish detector 2 applied to <i>Dom - WDS</i>
	<i>GIB - 2 - Dom - W2</i>	Gibberish detector 2 applied to <i>Dom - W2</i>
<i>GIB - 2 - Dom - W3</i>	Gibberish detector 2 applied to <i>Dom - W3</i>	
Entropy	<i>E - Dom</i>	Entropy of <i>Dom</i>
	<i>E - Dom - WS</i>	Entropy of <i>Dom - WS</i>
	<i>E - Dom - D</i>	Entropy of <i>Dom - D</i>
	<i>E - Dom - WDS</i>	Entropy of <i>Dom - WDS</i>
	<i>E - Dom - W2</i>	Entropy of <i>Dom - W2</i>
	<i>E - Dom - W3</i>	Entropy of <i>Dom - W3</i>

The set of experiments ran in this work aims to provide a solid proof of performance and accuracy for our approach. First, the detailed outcomes when applied to the HYDRAS dataset are provided. Next, we select two well-known state-of-the-art proposals using a method similar to ours, also based on Random Forest and implementing their own set of features. We compared the performance of such solutions to that of our method by applying them to the HYDRAS dataset.

6.1. Binary classification using the HYDRAS dataset

In the current experiment, several binary Random Forest classifiers that correspond to DGA family detectors were cross-validated — each detector is represented by a single such classifier. To build an input sub-dataset of each DGA family detector, random sampling was employed without replacement on AGDs from the corresponding family (or benign samples) to fit a 1:1 ratio with the benign domains, resulting in a balanced sub-dataset. To ensure the statistical significance of the results, each cross-validation execution was repeated 100 times (with

different randomly selected sample subsets). In detail, for each DGA family in our dataset we ensured 1:1 ratio with benign domains, with the dataset size per each DGA detector of family f equal to:

$$2 \min(|H[f]|, |A|), \quad (1)$$

where A represents samples of Alexa dataset and $H[f]$ represents samples of a particular family f in the HYDRAS dataset H . Hence, in the case that the number of samples in $H[f]$ is greater than in A , we employ random sampling without replacement (across repeated experiments) on $H[f]$ to reduce its size to the size of A . In the opposite case, when the size of A is greater than the size of $H[f]$, the same random sampling is employed to reduce the size of A , ensuring 1:1 ratio. Note that for each of the 100 runs of cross-validation, different sample sets were randomly selected from the Alexa dataset as benign class representatives.¹¹ Also, note that due to the particular format in

¹¹ Nevertheless, for thorough evaluation, different ratios to reproduce the actual binary classification experiment will be later used (see Section 9).

Table 3
Performance measures for binary classification (in percentage).

Class	Prec.	Recall	F_1	σ_{F_1}	AUC	Class	Prec.	Recall	F_1	σ_{F_1}	AUC	Class	Prec.	Recall	F_1	σ_{F_1}	AUC
bamital	100	100	100	0	100	gspy	99.67	100	99.83	0.29	100	qakbot	100	99.97	99.98	0.01	99.99
banjori	99.99	99.74	99.87	0.01	99.87	hesperbot	100	99.85	99.92	0.01	99.93	qghost	100	100	100	0	100
bedep	100	100	100	0	100	infy	100	99.97	99.98	0.01	99.98	qsnatch	99.77	99.86	99.81	0.02	99.81
beebone	100	99.07	99.53	0.40	99.54	khaos	99.47	96.47	97.95	0.10	97.98	ramdo	100	100	100	0	100
bigviktor	91.07	76.44	83.11	0.32	87.85	kingminer	100	97.62	98.8	<0.01	98.81	ramnit	100	99.95	99.97	0.02	99.98
blackhole	100	99.86	99.93	<0.01	99.93	locky	100	99.79	99.9	0.04	99.9	ranbyus	100	99.99	100	0	100
bobax/ /kraken /oderoor	100	99.62	99.81	0.01	99.81	madmax	100	99.98	99.99	<0.01	99.99	redyms	100	99.63	99.82	0.32	99.82
ccleaner	100	100	100	0	100	makloader	100	100	100	0	100	rovnix	100	100	100	0	100
chinad	100	100	100	0	100	matsnu	95.73	97.74	96.72	0.2	96.69	shifu	100	99.63	99.82	0.01	99.82
chir	100	100	100	0	100	mirai	100	99.96	99.98	<0.01	99.98	shiotob/ /urlzone /bebloh	100	99.99	99.99	<0.01	99.99
conficker	99.99	99.64	99.81	0.02	99.82	modpack	100	100	100	0	100	simda	99.99	99.66	99.83	0.01	99.83
corebot	100	99.98	99.99	<0.01	99.99	monerodownloader	100	100	100	0	100	sisron	100	100	100	0	100
cryptolocker	100	99.98	99.99	<0.01	99.99	monerominer	100	100	100	0	100	sphinx	100	100	100	0	100
cryptowall	100	99.87	99.93	0.02	99.94	murofet	100	99.99	100	<0.01	100	suppobox	96.84	98.3	97.57	0.02	97.55
darkshell	100	97.5	98.73	0	98.75	murofetweekly	100	100	100	0	100	sutra	100	99.97	99.98	<0.01	99.98
deception	99.03	97.00	98.00	0.07	98.02	mydoom	100	99.6	99.8	0.01	99.8	symmi	100	93.85	96.83	<0.01	96.92
deception2	98.25	96.15	97.19	0.1	97.22	neurus	100	99.89	99.95	0.02	99.95	szribi	99.98	99.68	99.83	0.03	99.83
diamondfox	100	97.13	98.55	<0.01	98.57	nymaim	100	99.6	99.8	0.03	99.8	tempedreve	100	99.56	99.78	0.02	99.78
dircrypt	100	99.94	99.97	<0.01	99.97	nymaim2	98.35	97.99	98.17	0.07	98.17	tinba	100	99.92	99.96	0.02	99.96
dmsniff	100	95.71	97.81	<0.01	97.86	omexo	100	100	100	0	100	tinynuke	100	100	100	0	100
dnschanger	100	99.93	99.96	0.02	99.97	padcrypt	100	100	100	0	100	tofsee	99.94	99.92	99.93	0.02	99.95
dromedan	100	100	100	0	100	pandabanker	100	100	100	0	100	torpig	100	99.79	99.89	0.01	99.9
dyre	100	100	100	0	100	pitou	100	99.89	99.94	0.02	99.95	tsifiri	100	100	100	0	100
ebury	100	100	100	0	100	pizd	99.43	99.62	99.52	0.09	99.52	ud2	100	100	100	0	100
ekforward	100	100	100	0	100	post	100	100	100	0	100	ud3	100	100	100	0	100
emotet	100	100	100	0	100	proslifeban	100	99.63	99.81	0.04	99.82	ud4	100	96.19	98.06	0.43	98.1
enviserv	100	100	100	0	100	pushdo	99.94	98.99	99.46	0.02	99.46	vawtrak	99.92	99.44	99.68	0.01	99.68
feodo	100	100	100	0	100	pushdotid	100	99.62	99.81	0	99.81	vidro	100	99.78	99.89	0.03	99.89
fobber	100	99.85	99.92	<0.01	99.93	pykspa	100	99.7	99.85	0.01	99.85	vidrotid	100	96.04	97.98	<0.01	98.02
fobber_v1	100	100	100	0	100	pykspa_v1	100	99.28	99.64	0	99.64	virut	99.97	99.99	99.98	<0.01	99.98
fobber_v2	100	99.78	99.89	0.1	99.89	pykspa_v2_fake	100	99.77	99.89	0.01	99.89	volatilecedar	99.93	100	99.97	0.06	100
gameover	100	100	100	0	100	pykspa_v2_real	100	99.77	99.88	0.01	99.88	wd	100	100	100	0	100
geodo	100	100	100	0	100	pykspa2	100	99.12	99.56	0.06	99.56	xshellghost	100	99.93	99.97	0.01	99.97
gozi	95.28	95.93	95.6	0.11	95.59	pykspa2s	100	97.47	98.72	<0.01	98.74	xxhex	100	99.96	99.98	0.02	99.98
gozonym	100	99.27	99.63	0.16	99.63	qadars	100	99.98	99.99	0.01	99.99	zloader	100	100	100	0	100

which several families are available in their reverse-engineered form, as well as in the AGD repositories used to initialise our dataset, the $L - DOT$ feature could not be used homogeneously, and thus, was excluded from this experiment.

The averaged outcomes of the binary classification can be seen in Table 3. We can observe that most of the DGA families are classified almost without errors, obtaining precision and recall metrics of above 99.9%. Even in the case of families with small representation (e.g., darkshell, omexo, qghost, and ud3), the classifier can discern between benign and malicious domains in almost 100% of cases, with only a few exceptions. Moreover, the standard deviation of the F_1 (i.e., σ_{F_1}) achieves values $\sim 1\%$ (in most cases only $\leq 0.01\%$), which showcases the robustness of both the classifier and the proposed feature set.

The lowest accuracy was obtained by bigviktor (with a precision of 91.07% and a recall of 76.44%) followed by suppobox, gozi, matsnu, khaos and symmi with F_1 scores ranging between 95% and 98%. This is due to the fact that most of these families use a composition of English dictionary words to create AGDs, so the extracted lexical features are not always able to properly differentiate them from our benign dataset or are adversarial. In the case of such dictionary-based families, a further enhancement based on probabilistic-based methodologies¹² is an interesting but challenging future research direction. In the case of adversarially designed DGAs, the accuracy obtained is close to the one reported for the dictionary-based DGAs, which showcases the difficulty of capturing such families. A more detailed comparison and

¹² Note that these methodologies have demonstrated remarkable detection performance (Patsakis and Casino, 2021).

analysis of adversarially designed DGAs is later presented in Section 7. Overall, the outstanding detection rates showed in Table 3 by using the same feature set across such a big dataset proves the robustness and adaptability of our approach. Note that the more divergent families (and samples) are, the more difficult is to select a common set of features which can capture them accurately.

Feature Weighting. For the sake of clarity, the average feature weight is depicted in the binary classification task in A and Table 12. The features exhibiting a high influence on the binary classification are $R - Dom - 4G$, $R - Dom - 5G$, $R - WS - LEN$, $R - WDS - LEN$, $GIB - 1 - Dom - D$, and $L - LEN$. Therefore, the n-grams, as well as the length of the domain and the valid words it contains, seem to be the most relevant features. Besides that, the relevance of each feature varies according to the family. Feature weights provide some insights into how to try to further enhance the performance of our method, e.g., by employing various feature selection methods to reduce the number of features, which might be convenient in power-constrained devices such as IoT device or smartphones.

6.2. Binary classification — comparison with state-of-the-art approaches

In this experiment, we compare the quality of our features and classification approach with two well-known approaches from the literature, namely the one proposed by Choudhary et al. (2018) and the method leveraged by Woodbridge et al. (2016). Therefore, each feature set was implemented according to the corresponding specifications, and applied them to the HYDRAS dataset. Thereafter, the same setup was used than in our binary classification experiment and computed the F_1 -score for each method. The outcome of this experiment is reported in Table 4.

Table 4
A comparison with state-of-the art (binary classification).

Class	Choudhary et al. (2018)		Woodbridge et al. (2016)		Our method		Class	Choudhary et al. (2018)		Woodbridge et al. (2016)		Our method		Class	Choudhary et al. (2018)		Woodbridge et al. (2016)		Our method	
	F_1	σ_{F_1}	F_1	σ_{F_1}	F_1	σ_{F_1}		F_1	σ_{F_1}	F_1	σ_{F_1}	F_1	σ_{F_1}		F_1	σ_{F_1}	F_1	σ_{F_1}	F_1	σ_{F_1}
bamital	99.99	<0.01	99.99	<0.01	100	0	gspy	99.82	0.32	100	0	99.83	0.29	qakbot	97.93	0.17	99.98	0.02	99.98	0.01
banjori	96.53	0.04	99.75	0.01	99.87	0.01	hesperbot	97.18	0.09	99.91	0.02	99.92	0.01	ghost	98.61	2.41	97.19	1.18	100	0
beedep	99.15	0.03	100	0	100	0	infy	99.65	0.11	99.98	<0.01	99.98	0.01	qsnatch	95.15	0.14	96.38	0.09	99.81	0.02
beebone	97.05	4.52	98.86	0.78	99.53	0.4	khaos	59.68	0.56	85.78	0.18	97.95	0.1	ramdo	99.84	0.02	99.99	<0.01	100	0
bigviktor	80.82	0.93	85.47	0.79	83.11	0.32	kingminer	99.40	0.2	99.93	0.11	98.8	<0.01	ramnit	97.25	0.18	99.95	0.02	99.97	0.02
blackhole	99.82	0.14	99.93	<0.01	99.93	<0.01	locky	95.18	0.14	99.84	0.01	99.99	0.04	ranbyus	99.54	0.02	99.99	<0.01	100	0
bobax/ /kraken /oderoor	95.18	0.09	99.74	0.03	99.81	0.01	madmax	99.00	0.05	99.99	<0.01	99.99	<0.01	redyms	98.37	0.54	99.46	0.94	99.82	0.32
ccleaner	99.87	0.01	100	0	100	0	makloader	99.94	0.11	99.94	0.11	100	0	rovnix	99.97	0.01	99.99	<0.01	100	0
chinad	99.95	<0.01	99.99	<0.01	100	0	matsnu	85.77	0.37	85.74	0.26	96.72	0.2	shifu	97.15	0.20	99.52	0.05	99.82	0.01
chir	100	0	100	0	100	0	mirai	99.09	0.09	100	0	99.98	<0.01	shiotob/ /urlzone /bebloh	99.10	0.02	99.99	0.01	99.99	<0.01
conficker	93.02	0.16	99.02	0.08	99.81	0.02	modpack	96.49	1.47	99.92	0.13	100	0	simda	95.22	0.19	95.04	0.28	99.83	0.01
corebot	99.78	0.05	99.99	0.01	99.99	<0.01	monero- downloader	99.91	0.02	100	0	100	0	siron	99.86	0.04	100	0	100	0
cryptolocker	98.87	0.02	99.99	0.01	99.99	<0.01	monerominer	99.95	0.01	100	0	100	0	sphinx	99.73	0.02	99.99	0.01	100	0
cryptowall	97.13	0.17	99.96	0.01	99.93	0.02	murofet	99.3	0.06	100	0	100	<0.01	suppobox	79.64	0.42	80.32	0.38	97.57	0.02
darkshell	99.18	0.71	99.59	0.71	98.73	0	murofetweeky	99.99	<0.01	100	0	100	0	sutra	99.73	0.07	99.98	<0.01	99.98	<0.01
deception	64.96	0.52	86.31	0.19	98.00	0.07	mydoom	98.61	0.15	99.80	0.03	99.80	0.01	symmi	90.46	1.05	97.42	0.43	96.83	<0.01
deception2	57.80	0.27	82.14	0.49	97.19	0.10	necurs	96.84	0.09	99.94	0.01	99.95	0.02	szribi	97.64	0.12	99.58	0.04	99.83	0.03
diamondfox	98.60	0.39	99.39	0.21	98.55	<0.01	nymaim	93.17	0.16	98.99	0.12	99.80	0.03	tempedreve	94.68	0.06	99.54	0.05	99.78	0.02
dircrypt	97.65	0.02	99.97	<0.01	99.97	<0.01	nymaim2	77.62	0.45	80.01	0.52	98.17	0.07	tinba	99.23	0.05	99.97	0.01	99.96	0.02
dmsniff	96.24	1.15	100	0	97.81	<0.01	omexo	100	0	99.60	0.70	100	0	tinynuke	99.99	0.01	100	0	100	0
dnschanger	98.64	0.03	99.97	0.01	99.96	0.02	padcrypt	99.82	0.03	99.99	<0.01	100	0	tofsee	99.92	0.03	99.88	0.12	99.93	0.02
dromedan	98.53	0.04	100	0	100	0	pandabanker	99.95	0.01	100	0	100	0	torpig	95.80	0.14	99.08	0.09	99.89	0.01
dyre	99.99	<0.01	100	0	100	0	pitou	99.22	0.01	99.59	0.14	99.94	0.02	tsifiri	100	0	100	0	100	0
ebury	99.83	0.09	100	0	100	0	pizd	88.29	0.10	88.27	0.26	99.52	0.09	ud2	99.97	0.06	100	0	100	0
ekforward	99.28	0.09	99.96	0.01	100	0	post	99.99	0.01	100	0	100	0	ud3	100	0	98.37	1.41	100	0
emotet	99.77	0.04	99.99	0	100	0	proslifean	94.13	0.14	99.48	0.06	99.81	0.04	ud4	96.40	0.35	100	0	98.06	0.43
enviserv	99.23	0.13	99.93	0.12	100	0	pushdo	92.00	0.14	97.52	0.06	99.46	0.02	vawtrak	90.53	0.24	95.56	0.18	99.68	0.01
feodo	99.46	0.12	100	0	100	0	pushdotid	98.15	0.09	99.80	0.04	99.81	0	vidro	95.77	0.19	99.82	0.03	99.89	0.03
fobber	98.64	0.17	99.96	<0.01	99.92	<0.01	pykspa	94.12	0.24	99.51	0.03	99.85	0.01	vidrotd	93.03	0.97	97.06	<0.01	97.98	<0.01
fobber_v1	99.92	0.08	100	0	100	0	pykspa_v1	93.80	0.51	99.55	0.04	99.64	0	virut	97.02	0.16	97.68	0.03	99.98	<0.01
fobber_v2	98.20	0.54	100	0	99.89	0.10	pykspa_v2_fake	96.41	0.14	99.63	0.02	99.89	0.01	volatilecedar	99.37	0.21	99.83	0.15	99.97	0.06
gameover	99.97	0.02	100	0	100	0	pykspa_v2_real	95.60	0.19	99.46	0.03	99.88	0.01	wd	99.99	<0.01	100	0	100	0
geodo	99.79	0.09	99.96	0.03	100	0	pykspa2	92.63	0.87	99.18	0.03	99.56	0.06	xshellghost	98.63	0.08	99.96	0.01	99.97	0.01
gozi	88.65	0.41	92.45	0.27	95.60	0.11	pykspa2s	92.21	0.29	99.07	0.29	98.72	<0.01	xxhex	99.60	0.03	100	0	99.98	0.02
goznym	92.89	0.77	99.68	0.08	99.63	0.16	qadars	99.70	0.08	99.99	<0.01	99.99	0.01	zloader	99.93	0.01	100	0	100	0

As it can be observed, all proposals succeed in capturing most of the families. The statistical features and n-gram-based features used in the different methods can recognise the patterns in the AGDs which belong to classical DGA families (i.e. the old ones which exhibit random-based generation patterns). In the case of more sophisticated families such as *rovnix*, *volatilecedar*, *beebone*, *banjori*, *locky*, *pushdo*, *proslikefan*, *symmi*, *goznm* and *pykspa*, all methods succeed to differentiate them from benign domains. Nevertheless, the method of Choudhary et al. reported notably worse accuracies for *pushdo*, *proslikefan*, *banjori*, *beebone*, *symmi*, *pykspa* and the variants of *pyskpa* than the method of Woodbridge et al. and our approach.

In the case of the rest of dictionary-based families, the adversarially-generated DGAs, and other novel DGA families, our method clearly outperforms the other approaches in most cases, with exception of *bigviktor*, in which the outcomes obtained by all methods are similar. For instance, in the case of *simda*, *vawtrak*, *gozi* and *qsnatch*, our method obtains a F_1 3% higher than the rest of methods.

In the case of more sophisticated dictionary-based families such as *matsnu*, *nyaim* and *suppobox*, our method outperforms the rest of approaches by approximately 10% in the case of *matsnu*, and close to 18% in the case of *nyaim* and *suppobox*. Finally, the highest difference was observed in the comparison with the adversarially generated DGAs.

In this regard, for the *deception* DGA, our approach outperforms Choudhary et al. by a 33% and Woodbridge et al. by a 12%. The *deception2* family exhibits similar outcomes, yet this time our approach outperforms Choudhary et al. by a 40% and Woodbridge et al. by a 15%. Finally, for the *khaos* DGA, the differences are close to 38% when comparing to Choudhary et al. and approximately 12% in the case of Woodbridge et al.

The average F_1 measure per family, as well as the standard deviation of the total average, are depicted in Table 5. As it can be observed, since a significant amount of DGA families are detected with high performance by all approaches, the average F_1 outcomes are high in all methods. Nevertheless, the robustness of our methodology is highlighted by the σ value, since it translates into a very high detection rate across all families, outperforming the rest of approaches.

The aforementioned comparison supports our idea to use a novel and upgradeable dataset such as HYDRAS, since most of the families present in the datasets used in the state-of-the-art approaches belong either to the random-based DGA category or to the thoroughly analysed set of dictionary-based families with specific patterns. In both cases, such families can be captured with high classification performance by using well-known feature sets. The latter implies that researchers should evaluate their methods with novel more complex families since approaches that test their accuracy with old datasets are no longer proving their validity versus the current DGA landscape.

6.3. Binary classification using other datasets

To compare the quality of our approach when applied to other datasets, two datasets from the recent literature are selected. First, we selected the dataset presented in Anand et al. (2020), which contains 50,600 samples that are split to benign and malicious AGDs in 50:50 ratio (i.e., 25,300 benign samples and 25,300 malicious ones). The authors used several machine learning methods and reported accuracy between 94.9% and 97.0% for the C5.0 algorithm — the one that achieved the highest accuracy. In our case, using the same dataset, we achieved the accuracy of 98.9%, which is between 1.9% and 4% higher than any of the proposed methods in Anand et al. (2020).

The second comparison was done using the dataset created in Selvi et al. (2019). In this case, the authors tested their approach with a dataset consisting of 64,000 samples — similar to the previous case, the samples were split in 50:50 ratio for the benign and malicious classes. The authors of Selvi et al. (2019) used a Random Forest classifier

Table 5
Average outcomes per DGA class in the binary classification comparison.

	Choudhary et al. (2018)	Woodbridge et al. (2016)	Our method
Average F_1	96.019	98.323	99.454
σ	7.424	4.289	1.810

and achieved an accuracy of 98.9% by using their 2-gram setup with 34 features. In our case, the accuracy achieved was 96.7%, that is only 2% below the outcome achieved by the authors. Nevertheless, the main drawback of their approach is the computational time required to compute such n-grams, which grows exponentially. In fact, the authors of Selvi et al. (2019) needed 1.21 h for execution of a complete n-fold experiment with three repetitions. In our case, the same experiment took approximately one minute. The latter emphasises that the trade-off between accuracy and computational time is also an important aspect (see Section 8 for the details).

6.4. Multiclass classification using other datasets

In this experiment, we aim at predicting the DGA families, given a set of AGDs, in a multiclass classification setting. We argue that this experiment has low practical utility in contrast to binary classification, and is provided only due to a fair comparison with related work.

A fair comparison of the multiclass classification's performance cannot be made if different approaches use different datasets. Although the best option to compare a performance of related work vs. our approach would be to use the HYDRAS dataset, several problems arise: (1) many implementations of feature extractors are unavailable, (2) classifiers need to be fine-tuned and details of parameters are often not presented in papers. Intuitively, due to a large number of DGA families contained in the HYDRAS dataset, the multiclass classification using HYDRAS dataset yields worse outcomes than in small datasets. This is the common problem of multiclass classification, especially when the classes are not well separated (Silva-Palacios et al., 2017).

Therefore, this experiment is based on another dataset introduced in Bader (2020), which was already evaluated by several papers. For example, this dataset was evaluated by Alaeiyan et al. (2020). We further extend the comparison with the other two approaches reported in the literature, namely DeepDGA (Anderson et al., 2016) and Phoenix (Schivoni et al., 2014). We performed a multiclass classification using our Random Forest classifier and our proposed feature set, this time slightly changing its configuration from the one used for the binary classification (i.e., an ensemble of 200 trees with unlimited depth). The repository of this dataset no longer contained samples of the *RunForestrun* family, so it was not included in the comparison. Moreover, the *Tinba* family had repeated SLD entries, which would lead to a biased and possibly unrepresentative classification (i.e., the same samples could easily end up both in the training and testing sets, thus reporting a 100% detection in most of the validations partly due to this fact). Therefore, a subset of the *Tinba* samples that are presented in our dataset was used. It is worth to note that this issue was ignored or not reported by the rest of approaches using this dataset.

The outcomes of the multiclass classification are depicted in Table 6. Even though in some cases Phoenix and DeepDGA showed better performance (e.g., for *Padcrypt*, *Qadars*, and *Symmi*), our method outperformed the rest, both in accuracy and recall, followed by the method proposed in Alaeiyan et al. (2020). Furthermore, we observed that there were relatively small differences for most families, and for the most part all methods reported high accuracy for similar families. In some cases, the reported performance metrics were equal to zero, which is related to the lack of samples. The latter means that the classifier could not be properly trained.

We can observe in Table 12 of Appendix that feature relevance in the multiclass setting differs substantially from the reported in the binary classification. In other words, the relevance of the features will

Table 6

Multiclass classification outcomes in percentages, using different performance metrics. The averages are weighted according to the number of samples in each family.

Class	Phoenix (Schiavoni et al., 2014)			DeepDGA (Anderson et al., 2016)			Alaeiyan et al. (2020)			Our method		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
banjori	100	100	100	100	100	100	100	100	100	99.80	100	99.90
chinad	74.2	84.1	78.84	98.8	96.3	97.53	92.3	19.5	32.20	87.99	97.27	92.39
corebot	88.4	97.4	92.68	100	100	100	100	97.4	98.68	100	72.50	84.06
dircrypt	0	0	0	0	0	0	0	0	0	0	0	0
downloader	100	100	100	100	100	100	100	100	100	100	100	100
dnschanger	0	0	0	0	0	0	0	0	0	0	0	0
fobber	3.2	2	2.46	3.1	7.6	4.40	0	0	0	37.62	12.67	18.95
gozi	7.1	50.0	12.43	3.0	3.0	3.00	0	0	0	0	0	0
javascrip	2.0	3.4	2.52	0	0	0	0	0	0	0	0	0
locky	0	0	0	0	0	0	0	0	0	0	0	0
murofet	99.9	98.0	98.94	100	98.0	98.99	99.7	99.4	99.55	99.91	99.79	99.85
necurs	22.2	15.8	18.46	28.0	32.7	30.17	30.2	5.1	8.73	32.24	8.15	13.02
newgoz	97.2	94.1	95.62	100	99.6	99.80	98	89.9	93.78	99.40	100	99.70
kraken	88.1	52.3	65.64	90.7	60.3	72.44	98.5	90	94.06	99.30	99.93	99.61
padcrypt	79.3	100	88.46	88.5	100	93.90	100	65.2	78.93	87.50	29.17	43.75
proslifean	3.0	19.4	5.20	4.0	23.5	6.84	0	0	0.00	25.00	2.00	3.70
pykspa	85.2	61.4	71.37	89.1	80.3	84.47	84.7	99.4	91.46	83.50	86.86	85.15
qadars	60.4	81.1	69.24	85.2	84.7	84.95	96.9	16.3	27.91	69.70	34.50	46.15
qakbot	53.5	55.0	54.24	57.9	56.6	57.24	54.5	82.1	65.51	65.69	76.70	70.77
ramnit	1.2	3.3	1.76	0	0	0	0	0	0	0	0	0
ranbyus	2.1	6.5	3.17	0	0	0	0	0	0	0	0	0
shiotob	96.7	81.6	88.51	98.3	89.8	93.86	84.7	91.4	87.92	92.26	90.00	91.12
simda	63.0	99.0	77.00	98.3	89.8	93.86	89.6	100	94.51	91.33	99.00	95.01
sisron	100	100	100	100	100	100	100	100	100	100	100	100
suppobox	32.4	79.3	46.00	67.4	74.6	70.82	97.3	68.5	80.40	90.61	98.43	94.36
symmi	98.3	96.6	97.44	98.3	100	99.14	98.3	100	99.14	92.31	56.25	69.90
tempedreve	27.6	67.1	39.11	43.8	96.3	60.21	57.2	75.1	64.94	59.98	71.94	65.42
tinba	25.3	64.6	36.36	49.9	98.2	66.17	100	99.7	99.85	99.52	99.94	99.73
vavtrak	30.9	9.7	14.77	68.3	87.5	76.72	100	8.3	15.33	69.47	66.00	67.69
Total	93.25	90.46	91.40	94.64	92.49	93.28	94.49	95.20	94.41	95.39	95.49	95.25

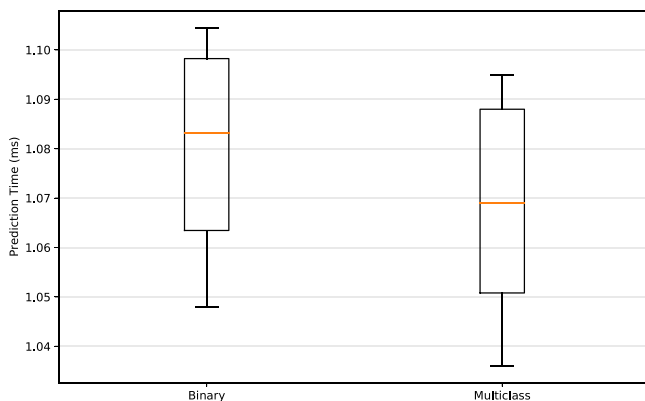


Fig. 4. Prediction time per SLD, in milliseconds.

be different according to the particularities of the collected samples, highlighting the importance of using the same dataset to avoid biased comparisons across different approaches. In the particular case of the dataset selected to perform the multiclass classification, the features $L - LEN$, $E - Dom$, $L - DIG$, $R - WD - LEN$, and $R - WDS - LEN$ were the most relevant. The latter means that the length (i.e., specific families create only AGDs of fixed length), the number of digits, the meaningful words in the SLD, and the entropy of the SLD enabled to predict which specific family created a given AGD, with a high precision.

7. Classification of adversarially designed AGDs

To further assess the quality of our selected features, we opted to use it against three especially “hard to detect” DGAs. These DGAs, deception, deception2 (Spooren et al., 2019), and khaos (Yun et al., 2020) are specially crafted, using machine learning methods, to

evade detection (see Section 2). While our features are generic and not targeted towards identifying any particular set of these families, we also manage to detect adversarially designed AGDs with significantly better performance than in previous works (see Table 8). In detail, the precision achieved by our approach is by 15% to 30% better. Similarly, the recall and F_1 score are by more than 10% better in almost all cases. It may also be observed that in some cases, the detection rates slightly vary if the ratio of malicious to benign samples is increased. This fact will be discussed more thoroughly in Section 9. Nonetheless, the F_1 score is at least 92.48%, which indicates that our method is very effective even when confronted against specially crafted DGAs — a challenge that is very close to represent the most challenging scenario.

7.1. Invalid domains

It should be noted that, during our study, many invalid domains generated by these DGAs were identified. To the best of our understanding, the researchers simply left the neural networks to generate AGDs that bypassed the filters, without double-checking their validity in real scenarios. As a result, the neural networks identified that the use of the hyphen character managed to bypass some filters, priming them to overuse it. Therefore, one can observe thousands of domains which either start or finish with a hyphen, which are unfortunately invalid according to RFC 1123 (Braden, 1989). Similarly, many of them do not conform to RFC 5891 for Internationalised Domain Names (IDNs) (Klensin, 2010), by having hyphens as third and fourth characters, but not starting with an “xn”, so they are rejected by ICANN.¹³ The issue is particularly relevant in the DGAs generated by Spooren et al. (2019) spanning across 1.64% of the samples. In the khaos family, the dataset contains 19 IDNs. Note that in all following experiments, only DGAs which do not produce IDNs are considered.

¹³ <https://www.icann.org/en/system/files/files/idn-guidelines-10may18-en.pdf>

Table 7

A detection of unknown DGA families represented by adversarially designed DGAs (leave-one-out experiment).

DGA	Precision	Recall	F1
khaos	100	85.40	92.12
deception	99.99	84.71	91.72
deception2	99.99	73.08	84.44

7.2. Detection of unknown families

We performed another experiment to test the capability of our approach to detect previously unknown DGA families. In this regard, a leave-one-out experiment were leveraged with the same configuration as in the binary classification experiments (i.e., 10-fold cross validation) but in contract to it the target family was completely hidden to the training phase. In other words, we tried to predict whether a set of AGDs is benign or malicious without previous knowledge of the DGA family generating them. The outcome of this experiment is reported in Table 7. As it can be observed in the table, our approach is able to correctly classify most of the samples, achieving a slightly lower F_1 -score than the one reported in our binary classification (see Table 8), due to a general decrease of the recall values. Note that the most affected family is `deception2`, yet our approach still outperforms the original works in which these families were proposed, thus showcasing the robustness of our features once again.

8. Overhead of our approach

We measured several statistics during our experiments with the intention to demonstrate the practical aspects of our approach. We measured the duration of the features computation and the prediction time in both the binary and multiclass setting with our dataset. The average time required to compute all the features for an SLD is 1.48 ms, while the prediction times from both classification experiments are depicted in Fig. 4. Note that the figure does not include the training time, which scales linearly with the size of the dataset. However, training is an action that is very rare (e.g., repeated after weeks or months) and it can be performed offline. Hence, it does not incur performance degradation to the operation of AGD detector. Finally, both the features computation time and the prediction times are measured without any parallelisation to enable a fair comparison with related work.

8.1. Binary classification

In terms of performance comparison for the binary classification, the work proposed in Anand et al. (2020) did not report any computational cost nor performance metrics. In the case of Selvi et al. (2019), the authors reported a total experiment time of 1.21h in their 2-gram setup with 34 features, which is the one that yielded the highest accuracy. In our case, the time required for the same experiment, including the n-fold validation, is between 10–20 s, and close to one minute including the feature computation of all the SLDs of the dataset, without considering parallelisation. Therefore, the trade-off between accuracy (which in our case is just 2% lower) and computational time (i.e. two orders of magnitude faster) of our method is clearly outperforming the method presented in Selvi et al. (2019).

8.2. Multiclass classification

Considering related works that have studied the multiclass classification (see Table 6), a fine-grained comparison of processing performance was not possible since the achieved performance was not systematically reported and in some cases not stated at all. Moreover, such performance highly varies depending on the exact hardware configuration, and thus, the exact replication of each of the environments

used is challenging. For instance, the authors of Schiavoni et al. (2014) report only that experiments required time in the order of minutes. In the case of Anderson et al. (2016), authors stated that their model required expensive training periods of 14 h (considering 300 epochs and Alexa subsampling), as well as a classification time of 7 min for approximately 13k samples (i.e. 0.03 s per sample) in a GPU-based setup. In work presented in Alaeiyan et al. (2020), the authors reported a feature computation time of around 217 min for 252,757 samples, which implies approximately 0.05 s per sample. Moreover, the authors also specified a classification time of around 60 min, which translates into 0.014 s on average to classify a sample.

In sum, our approach outperformed related work in the multiclass experiment in time requirements by one order of magnitude for both the feature computation as well as in the prediction. This means that our method is suitable for real-time AGD detection and classification, even in environments with very high traffic volumes.

9. Discussion

In this section, we focus on a quantitative comparison of the most relevant DGA detectors of the related work, and we analyse the existing limitations in the literature.

9.1. Quantitative comparison

The DGA research field has several open challenges; one of them is the continuous appearance of new families. In contrast to other research fields that rely on standardised benchmarks (e.g., computer vision), DGA-based datasets need to be updated frequently to be able to prove the performance of the detection methods is of relevance in real scenarios. Recently, Zago et al. (2020) created a balanced and structured dataset containing 38 families. Nevertheless, although their approach is sound, it does not keep pace with the recent evolution of malware campaigns. Solving this issue is hard, and although there are research efforts in adversarial classification (see Section 2), we argue that further research should focus on upgradeable versions of datasets and include version tracking. In our dataset, we introduce a collection of 105 families, which better reflects the complexity and challenge of the current landscape. Moreover, it is worth to note that obtaining enough samples of specific families can be a cumbersome and difficult task due to a number of reasons, such as the inner algorithmic structure of the DGA¹⁴ or the fact that it has not been reverse engineered yet. Therefore, we argue that obtaining a perfectly balanced dataset that contains all possible DGA families is extremely difficult, borderline impossible. In this regard, it should be stressed that due to the heterogeneity and non-replicability of some datasets as well as the range of different techniques applied in related work, a perfect and fair comparison is unfeasible.

As it can be observed in Table 9, the results of our work were achieved using the biggest dataset in terms of the number of DGA families (and samples). This compares favourably and leads to a much more challenging classification task when compared to related work. Due to the continuous evolution of malware, a high number of supported DGA families (including the most recent ones) is a critical capability required of any successful DGA detector that can be deployed in real scenarios. Note that data sources of related works in the table share some common repositories such as Bambenek, DGArchive, and NetLab 360. However, only a few authors used reverse-engineered DGAs to populate their datasets, as seen in Table 9 and described in Section 4. With regard to the detection method used, LSTM is the most prevalent, followed by ML classifiers, from which RF is known to report the best classification

¹⁴ E.g., it can create only a small set of samples due to the use of specific dictionaries.

Table 8

Binary classification against adversarially designed DGAs. First row of each family denotes the reported results in the original work.

DGA	Method	Precision	Recall	F1
khaos	Yun et al. (2020)	68.00	98.00	80.30
	Our approach — ratio 1:1	99.47	96.47	97.95
	Our approach — ratio 1:10	96.08	90.73	93.32
	Our approach — ratio 1:100	96.55	89.63	92.96
deception	Spooren et al. (2019)	84.40	87.10	85.72
	Our approach — ratio 1:1	99.03	97.00	98.00
	Our approach — ratio 1:10	96.21	93.86	95.02
	Our approach — ratio 1:100	96.12	93.29	94.68
deception2	Spooren et al. (2019)	77.50	81.50	79.45
	Our approach — ratio 1:1	98.25	96.15	97.19
	Our approach — ratio 1:10	94.44	91.29	92.84
	Our approach — ratio 1:100	94.56	90.50	92.48

Table 9

A quantitative comparison of our work with the most relevant state-of-the-art approaches.

Ref.	Features	# DGAs	AGD samples	Method	Dataset source
Anderson et al. (2016)	Lexical, entropy	10	110,000	GAN, LSTM, RF	Manually crafted
Schiavoni et al. (2014)	Lexical	5	1,153,516	DBSCAN	SIE framework, Exposure blacklist and other public implementations
Curtin et al. (2019)	WHOIS, lexical, smashword score	41	1,280,000	RNN	DGArchive and Several GitHub repositories
Koh and Rhodes (2018)	Lexical	4	4,000	ELMo	Several GitHub repositories
Yu et al. (2017)	Time-based, query response, domain name	19	4,739,563	LSTM/CNN	Farsight security/DGArchive
Tran et al. (2018)	Domain name	37	169,831	LSTM	Bambenek
Lison and Mavroeidis (2017)	Domain name	58	2,900,000	RNN	DGArchive, Bambenek
Yu et al. (2018)	Lexical	N/A	1,000,000	CCN/RNN	Bambenek
Mac et al. (2017)	Entropy, lexical	37	81,490	Several ML methods	Bambenek
Choudhary et al. (2018)	Lexical	19	34,264,306	Random Forest and DNN	Bambenek, DGArchive
Li et al. (2019)	Lexical, query response	5	160,000	Several ML methods	Bambenek
Jyothsna et al. (2018)	Lexical	19	245,872	DNN	Bambenek, Netlab 360
Chen et al. (2018)	Domain name	60	1,687,806	LSTM	Bambenek, Netlab 360
Sivaguru et al. (2018)	Time-based and domain name	15	551,086	Several Binary Classifiers	Real traffic, Bambenek
Attardi and Sartiano (2018)	Domain name	19	135,056	LSTM, BLSTM	Bambenek and Netlab 360
Zago et al. (2019)	Entropy, lexical, similarity	17	16,000	Several ML methods	Netlab 360, DGArchive, DNS-BH
Bharathi and Bhuvana (2019)	Domain name	19	245,872	LSTM, BLSTM	Bambenek, Netlab 360
Khehra and Sofat (2018)	Entropy, lexical	5	272,209	CNN/RNN	Stratosphere dataset (Stratosphere Labs, 2020)
Zago et al. (2020)	Lexical	38	30,799,449	Several ML methods	UMUDGA
Anand et al. (2020)	Lexical	19	25,300	Several ML methods	Netlab 360
Yang et al. (2020)	Domain name	20	100,000	BLSTM, HDNN	Fu et al. (2017)
Alaeiyan et al. (2020)	Lexical, pronounceability	30	252,757	Genetic algorithm and RF	Bader (2020)
Almashhadani et al. (2020)	Entropy, randomness, lexical	20	208,190	Several ML methods	DGArchive, Bambenek
Selvi et al. (2019)	Entropy, lexical	26	252,757	RF	Bader (2020)
Our approach	Entropy, lexical, gibberish	105	95,325,598	RF	HYDRA Dataset (Casino et al., 2020)

performance. Regarding the features, both lexical (e.g., ratios of letters, n-grams, words) and entropy-based ones seem to occur most widely.

The methods that use side-information (e.g., WHOIS, timing, etc.) cannot prevent compromised hosts from contacting the C&C server, and thus bring an additional cost in terms of time which makes them prohibitively slow for real-time detection and incident response. Undeniably, caching and whitelisting can significantly reduce such a cost; however, this is expected to occur every time the host has to connect with a new domain or a DGA has a new seed, which is unrealistic in most scenarios.

9.2. Fair comparison and evaluation of reproducibility

As previously reported in Section 6, the comparison of any two approaches should be made under the same contextual settings (i.e., benchmarks and performance metrics); otherwise, the interpretation of the results might be biased and unduly favour an approach with less challenging settings. To analyse the quality and the methodologies used in related research, we reviewed the works from [Table 9](#) in terms of reproducibility and presentation of the outcomes. In detail,

we verified whether the authors explicitly reported their evaluation methodology, their dataset collection procedure (for reproducibility purposes), and sufficient details about their outcomes (for a fair comparison). The results of this effort are shown in [Table 10](#), where we observe that there are some serious methodological issues mainly due to the lack of experimental setup description, biased performance measures (e.g not reporting widely used metrics to enable fair comparison) being used and/or extremely imbalanced datasets (e.g. not enough samples for unbiased training), and the aggregation of classification results by averaging, while not reporting information about some poorly performing families. Note that [Table 10](#) is not intended to criticise the related works, on the contrary, it aims to establishing a common ground to improve the transparency and contributions of the literature.

Moreover, when the reported results are aggregated (i.e. the outcomes are not reported per class but as an overall aggregate), the unbalanced nature of the dataset, as well as the fact of hiding the classification performance per family hinders the objective interpretation of the results (e.g., a very small sample set could not reflect the characteristics of a family, and specific sampling ratios of benign to

Table 10
Methodological limitations of related work.

Methodological limitations	References
Not reported samples, extremely imbalanced benchmarks or lack of robust performance measures	Yu et al. (2017), Curtin et al. (2019), Tran et al. (2018), Lison and Mavroeidis (2017), Yu et al. (2018), Mac et al. (2017), Chen et al. (2018), Anand et al. (2020), Chen et al. (2018), Zago et al. (2019), Khehra and Sofat (2018)
Aggregated classification outcomes	Tran et al. (2018), Lison and Mavroeidis (2017), Yu et al. (2018), Choudhary et al. (2018), Jyothsna et al. (2018), Chen et al. (2018,?), Sivaguru et al. (2018), Anand et al. (2020), Attardi and Sartiano (2018), Zago et al. (2019), Bharathi and Bhuvana (2019), Khehra and Sofat (2018), Zago et al. (2020), Alaeiyan et al. (2020), Almashhadani et al. (2020), Yang et al. (2020), Selvi et al. (2019)

malicious domains might result in statistically biased outcomes). This, in turn, translates into approaches that might obtain highly accurate results for some families, while they are unable to detect other families; however, the occurrence of this phenomenon cannot be discerned from the reported results, and thus a fair comparison cannot be made.

The same benchmark settings in the case of multiclass classification are even more critical to allow for a fair comparison since the more families used the more difficult is to classify them, especially taking into account their random nature. Moreover, since several DGAs can create the same pattern, the more samples collected the more possibilities of overlapping the domain names (Alaeiyan et al., 2020; Zago et al., 2020; Patsakis and Casino, 2021; Mac et al., 2017), which increases chances for misclassification thus making the problem more challenging. Nevertheless, in the case of binary classification with a statistically sound methodology, highly accurate detection of underrepresented families indicates the robustness of the selected features, thus showcasing the high performance of the detection method even in extreme cases.

9.3. Sound evaluation methodology

We argue that an objective comparison of the results among different approaches is possible only through a sound evaluation methodology. To do so, it is imperative to include the reporting of the performance over data with the same ratio of AGD samples to benign domains. Moreover, such experiments should be repeated several times with different sample sets (e.g., 100 times in our case and using 10-fold cross-validation in each iteration), since methodologies using a single run of n-fold cross-validation only shuffles samples within the selected sample set. For instance, one could perform a 1:1 ratio classification between Alexa and a malicious family with 50 samples. In this setup, which is frequently adopted by ML practitioners in this field, only 50 samples of Alexa would be selected and shuffled in the cross-validation. Therefore, the rest of the Alexa samples will not be used unless the experiments are repeated with different samples to produce a statistically sound outcome. In this regard, repeating the experiments and selecting a different set of samples in each iteration provides a better representation of the characteristics of each family and hence a much more realistic accuracy. In this setup, low values of the standard deviation in the results indicate the desirable stability and robustness of the method. Unfortunately, this methodology is rarely adopted in the field.

9.4. Ratios of malicious to benign samples

The well-known imbalance problem (Liu et al., 2009) argues that there are much less malicious events than benign ones when performing, e.g., traffic analysis and intrusion detection in real scenarios (Shab-tai et al., 2012; Wang et al., 2020; Liu et al., 2018). Nevertheless, in the

area of DGA analysis, there is no wide consensus on the common ratio of benign to malicious domains that one can commonly find in real-world settings. This is due to some DGAs generating only a few domains per day, while others might create them in the hundreds or thousands. Therefore, we made a conscious effort to evaluate the performance of our approach under malicious to benign ratios different than 1:1, and hence we explored ratios of 1:10 and 1:100 as well. During our experiments, if a malicious family had more samples than the size of the benign dataset, these were randomly under-sampled, to obtain the desired ratios. Table 11 shows the results obtained by using the F_1 measure and its standard deviation across 100 repetitions of the 10-fold cross-validation, selecting different samples in each iteration.

When we compare the outcomes obtained across all ratios (i.e. 1:1, 1:10, and 1:100), we can observe that dictionary-based and adversarial families obtain slightly worse accuracy when the malicious to benign ratio is increased. The latter occurs because these DGAs create domains that have structural similarities with Alexa domains, which increases the difficulty of the classification task due to the overlapping features. Nevertheless, as it can be observed in the rest of cases, the variance of the outcomes according to each sampling ratios is minimal, which denotes stable results. This means that our approach and its features represent homogeneously benign domains, and thus, they are able to accurately distinguish them from malicious ones, regardless of the sample ratio. This showcases the quality of the feature selection as well as the statistical confidence of the classification.

A proper methodology should also be considered when using automated approaches for machine learning, such as H2O,¹⁵ auto-sklearn,¹⁶ AutoKeras¹⁷ etc. Such libraries may hyper-optimize parameters for many methods and generate a model which maximises, for instance, the F_1 score. We argue that this *unique win* should not be considered as the best method since, as discussed above, this solution has to be weighed along with the efficiency of the rest of the models over the same family, and considering several repetitions, only in this way providing statistical soundness.

10. Conclusion

Nowadays, modern malware has evolved into highly sophisticated software, which can be used to infect millions of devices. This enables hard-to-detect and resilient malware campaigns, which have turned cybercrime into a profitable “business”. To enable faster and more accurate botnet detection, and to speed-up take-down operations, a new DGA detection method using machine learning is presented. In essence, our method stands out from the rest in terms of accuracy and performance because we use more comprehensive features and a broader and more representative dataset. We only identified a case in which our outcomes were slightly below these obtained by other methods, yet the time required was between one and two orders of magnitude lower in our case. The relevance of our features is manifested in three ways. First, it achieves an almost optimal detection rate in the binary classification problem for the broadest possible set of DGA families. Second, our features allow us to outperform the current state-of-the-art also in multiclass classification, using the same datasets presented in other works. Finally, our approach was able to detect adversarially designed DGAs, including the experiments in which our system was not trained to detect such families (i.e. assuming no previous knowledge).

Additionally, our methodology is more rigorous than most seen in the field to date, avoiding common pitfalls in the literature that focus on DGAs with many non-obvious constraints. Setting aside feature extraction, our work highlights the inherent biases of datasets and

¹⁵ <https://www.h2o.ai/>

¹⁶ <https://github.com/automl/auto-sklearn>

¹⁷ <https://github.com/keras-team/autokeras?spm=a2c65.11461447.0.0.68b37903yEmaw3>

Table 11

Binary classification outcomes with different ratios of malicious to benign domains (we always assume a higher number of benign domains in the ratios).

Class	Ratio 1:10		Ratio 1:100		Class	Ratio 1:10		Ratio 1:100		Class	Ratio 1:10		Ratio 1:100	
	F_1	σ	F_1	σ		F_1	σ	F_1	σ		F_1	σ	F_1	σ
bamital	99.98	0.03	99.97	0.03	gspy	100	0	100	0	qakbot	99.88	0.03	99.87	0.03
banjori	99.60	0.13	99.51	0.18	hesperbot	99.90	0.05	99.90	0.05	qhost	100	0	100	0
bedep	99.93	0.03	99.93	0.03	infy	99.92	0.03	99.93	0.08	qsnatch	99.07	0.36	98.82	0.14
beebone	99.53	0.40	99.53	0.40	khaos	93.33	0.69	92.96	0.83	ramdo	99.98	0.03	99.98	0.03
bigviktor	93.56	0.58	93.26	0.28	kingminer	98.80	<0.01	98.80	<0.01	ramnit	99.90	0.05	99.93	0.03
blackhole	100	0	99.98	0.04	locky	99.63	0.06	99.77	0.12	ranbyus	99.93	0.06	99.95	0.05
bobax/ /kraken /oderoor	99.82	0.03	99.68	0.12	madmax	99.90	0.05	99.93	0.06	redyms	100	0	100	0
ccleaner	100	0	99.98	0.03	makloader	99.94	0.11	100	0	rovnix	100	0	99.98	0.03
chinad	99.97	0.03	100	0	matsnu	91.86	0.67	91.58	0.46	shifu	99.68	0.10	99.70	0.05
chir	100	0	100	0	mirai	99.95	0.05	99.90	0.05	shiotob/ /urlzone /bebloh	99.95	0.05	99.98	0.03
conficker	99.45	0.05	99.23	0.13	modpack	100	0	100	0	simda	99.03	0.10	99.09	0.14
corebot	99.97	0.03	99.97	0.03	monerodown- loader	100	0	100	0	sisron	100	0	100	0
cryptolocker	99.95	0.05	99.95	0.05	monerominer	100	0	100	0	sphinx	100	0	99.97	0.06
cryptowall	99.87	0.06	99.95	0.05	murofet	100	0	99.95	<0.01	suppobox	92.39	0.37	91.92	0.59
darkshell	100	0	99.58	0.73	murofetweekly	100	0	100	0	сутra	99.95	0.05	99.97	0.06
deception	95.02	0.29	94.69	0.44	mydoom	99.73	0.03	99.75	0.05	symmi	96.55	0.48	96.55	0.48
deception2	92.85	0.86	92.49	0.46	neccurs	99.85	0.09	99.87	0.03	szribi	99.70	0.09	99.60	0.17
diamondfox	98.61	0.10	98.67	0.10	nymaim	99.46	0.12	99.38	0.10	tempedreve	99.62	0.06	99.7	0.10
dircrypt	99.85	<0.01	99.92	0.03	nymaim2	94.53	0.58	94.37	0.41	tinba	99.92	0.03	99.95	0.05
dmsniff	98.06	0.43	98.32	0.39	omexo	99.55	0.69	99.60	0.70	tinynuke	100	0	100	0
dnschanger	99.95	0.05	99.88	0.03	padcrypt	99.98	0.03	100	0	tofsee	99.90	0.05	99.80	0
dromedan	99.9	0.10	99.93	0.03	pandabanker	99.98	0.03	99.97	0.06	torpig	99.48	0.13	99.62	0.03
dyre	100	0	100	0	pitou	99.87	0.03	99.80	0.13	tsifiri	99.16	0.83	99.72	0.49
ebury	99.98	0.03	100	0	pizd	96.90	0.09	96.61	0.25	ud2	100	0	99.97	0.06
ekforward	99.92	0.03	99.93	0.03	post	100	0	100	0	ud3	100	0	100	0
emotet	99.98	0.03	99.98	0.03	proslifean	99.75	0.09	99.51	0.15	ud4	98.30	0.43	98.30	0.43
enviserv	99.97	0.06	100	0	pushdo	98.46	0.06	98.28	0.16	vawtrak	98.77	0.19	98.54	0.36
feodo	99.93	0.12	99.93	0.12	pushdotid	99.72	0.08	99.73	0.06	vidro	99.83	0.08	99.88	0.03
fobber	99.93	0.03	99.88	0.03	pykspa	99.67	0.08	99.67	0.18	vidrotid	97.82	0.28	97.98	<0.01
fobber_v1	100	0	100	0	pykspa_v1	99.58	0.03	99.63	0.08	virut	99.80	0.09	99.88	0.03
fobber_v2	100	0	100	0	pykspa_v2_fake	99.65	0.05	99.51	0.08	volatilecedar	99.90	0.10	99.93	0.12
gameover	100	0	99.98	0.03	pykspa_v2_real	99.55	0.09	99.58	0.03	wd	100	0	100	0
geodo	100	0	100	0	pykspa2	99.52	0.04	99.56	<0.01	xshellghost	99.93	0.03	99.85	0.05
gozi	93.04	0.19	92.68	0.52	pykspa2s	98.29	0.15	98.20	<0.01	xxhex	100	0	99.95	0.09
goznm	99.54	0.08	99.49	0.08	qadars	99.93	0.03	99.93	0.03	zloader	100	0	100	0

methodologies in previous literature that report many close to perfect results; however, these results may be true for only a very limited and unrepresentative number of DGA families. Notably, we stress the methodological errors in the use of machine learning with, e.g., the use of very few samples and in some cases aggregated classification outcomes preventing a clear comparison. In this regard, a dataset with more than 95 million AGDs is constructed and shared, providing the extracted features to the community. While this facilitates the reproducibility of our results, we also allow fellow researchers to use a significantly richer baseline dataset, both in terms of number of families and samples.

In future work, we aim to enhance our semantic classification by using other training sources in order to increase the accuracy of both English and non-English domain names. Moreover, we will explore wordlist-based DGA detection in more depth by using probabilistic approaches based on word repetition and similar features. Finally, we will study the impact of dimensionality reduction techniques in our dataset.

CRediT authorship contribution statement

Fran Casino: Validation, Investigation, Data curation, Writing - original draft, Writing review & editing, Project administration. **Nikolaos Lykousas:** Software, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Ivan Homoliak:** Software, Data curation, Writing - original draft, Writing - review & editing. **Constantinos Patsakis:** Conceptualization, Investigation, Writing - original draft,

Writing - review & editing, Supervision, Funding acquisition. **Julio Hernandez-Castro:** Methodology, Writing - original draft, Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the projects CyberSec4Europe (<https://www.cybersec4europe.eu>) (Grant Agreement no. 830929), LOCARD (<https://locard.eu>) (Grant Agreement no. 832735) and YAKSHA, (<https://project-yaksha.eu/project/>) (Grant Agreement no. 780498). Also, this work was supported by the H2020 ECSEL project VALU3S (876852) and the internal project of Brno University of Technology (FIT-S-20-6427). The Titan Xp used for this research was generously donated by NVIDIA Corporation.

The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the author.

Table 12

The average weight denotes the relevance (in percentage) of certain features in a corresponding classification setting. Since the alphanumeric sequences are used as the input to compute the rest of the features, they do not have any weights.

Feature set	Notation	Avg. weight (Binary)	Avg. weight (Multiclass)
Statistical attributes	<i>L - HEX</i>	1.38	0
	<i>L - LEN</i>	5.32	19.76
	<i>L - DIG</i>	3.93	11.84
	<i>L - DOT</i>	N/A	2.81
	<i>L - CON - MAX</i>	1.36	0.69
	<i>L - VOW - MAX</i>	0.32	0.18
	<i>L - W2</i>	0.45	1.01
	<i>L - W3</i>	1.59	0.07
Ratios	<i>R - CON - VOW</i>	1.47	0.61
	<i>R - Dom - 3G</i>	1.57	1.53
	<i>R - Dom - 4G</i>	15.23	2.18
	<i>R - Dom - 5G</i>	11.70	0.43
	<i>R - VOW - 3G</i>	0.75	0.36
	<i>R - VOW - 4G</i>	0.65	0.33
	<i>R - VOW - 5G</i>	0.62	0.26
	<i>R - WS - LEN</i>	7.95	2.42
	<i>R - WD - LEN</i>	2.61	9.35
	<i>R - WDS - LEN</i>	6.20	7.40
	<i>R - W2 - LEN</i>	1.95	0.84
	<i>R - W2 - LEN - D</i>	1.28	0.33
	<i>R - W3 - LEN</i>	2.61	0.16
	<i>R - W3 - LEN - D</i>	2.54	0.16
Gibberish probabilities	<i>GIB - 1 - Dom</i>	5.12	0.48
	<i>GIB - 1 - Dom - WS</i>	1.67	0.52
	<i>GIB - 1 - Dom - D</i>	5.76	0.48
	<i>GIB - 1 - Dom - WDS</i>	0.82	0.48
	<i>GIB - 1 - Dom - W2</i>	0.46	0.46
	<i>GIB - 1 - Dom - W3</i>	0.97	0.07
	<i>GIB - 2 - Dom</i>	0.92	0.75
	<i>GIB - 2 - Dom - WS</i>	0.87	0.80
	<i>GIB - 2 - Dom - D</i>	1.09	0.69
	<i>GIB - 2 - Dom - WDS</i>	0.89	1.26
	<i>GIB - 2 - Dom - W2</i>	0.19	0.24
	<i>GIB - 2 - Dom - W3</i>	1.00	0.11
Entropy	<i>E - Dom</i>	2.10	12.66
	<i>E - Dom - WS</i>	1.44	8.19
	<i>E - Dom - D</i>	2.27	4.76
	<i>E - Dom - WDS</i>	1.51	3.39
	<i>E - Dom - W2</i>	0.63	1.62
	<i>E - Dom - W3</i>	0.81	0.16

Appendix A. Feature relevance

To showcase the relevance of the features used in our approach, the specific values of weights are provided in Table 12 in the case of our binary classification using the HYDRAS dataset (see Section 6) and the multiclass classification performed with the dataset introduced in Bader (2020). We computed the weights in different setups to highlight the fact that the importance of several features may vary according to the families analysed. The latter, which seems straightforward, is nevertheless worth showcasing given a specific subset of families, so that further insights can be discovered. Further discussion about such outcomes is presented in Section 6.

References

Abakumov, A., 2020. DGA repository. <https://github.com/andrewaeva/DGA>.
 Alaeiyan, M., Parsa, S., P., V., Conti, M., 2020. Detection of algorithmically-generated domains: An adversarial machine learning approach. *Comput. Commun.*
 Almashhadani, A.O., Kaiiali, M., Carlin, D., Sezer, S., 2020. Maldomdetector: A system for detecting algorithmically generated domain names with machine learning. *Comput. Secur.* 93, 101787.
 Anand, P.M., Kumar, T.G., Charan, P.S., 2020. An ensemble approach for algorithmically generated domain name detection using statistical and lexical analysis. *Procedia Comput. Sci.* 171, 1129–1136, Third International Conference on Computing and Network Communications (CoCoNet'19).
 Anderson, H.S., Woodbridge, J., Filar, B., 2016. DeepDGA: Adversarially-tuned domain generation and detection. In: *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security. AISec '16*, ACM, New York, NY, USA, pp. 13–21.

Anomali Labs, 2019. Interplanetary storm. <https://www.anomali.com/blog/the-interplanetary-storm-new-malware-in-wild-using-interplanetary-file-systems-ipfs-p2p-network>.
 Antonakakis, M., et al., 2012. From throw-away traffic to bots: Detecting the rise of DGA-based malware. In: *Proceedings of the 21st USENIX Conference on Security Symposium*. USENIX Association, p. 24.
 Antonakakis, M., et al., 2017. Understanding the mirai botnet. In: *26th USENIX Security Symposium (USENIX Security 17)*. USENIX Association, Vancouver, BC, pp. 1093–1110.
 Attardi, G., Sartiano, D., 2018. Bidirectional LSTM models for DGA classification. In: *International Symposium on Security in Computing and Communication*. Springer, pp. 687–694.
 Aviv, A.J., Haerleren, A., 2011. Challenges in experimenting with botnet detection systems. In: *Proceedings of the 4th Conference on Cyber Security Experimentation and Test. CSET'11*, USENIX Association, Berkeley, CA, USA, p. 6.
 Bader, J., 2015. The DGA of pykspa “you skype version is old”. <https://www.johannesbader.ch/2015/03/the-dga-of-pykspa/>.
 Bader, J., 2020. Domain generation algorithms (DGAs) of malware reimplemented in Python. https://github.com/baderj/domain_generation_algorithms.
 Barandela, R., Valdovinos, R.M., Sánchez, J.S., 2003. New applications of ensembles of classifiers. *Pattern Anal. Appl.* 6 (3), 245–256.
 Berman, D.S., 2019. DGA capsnet: 1D application of capsule networks to DGA detection. *Information 10* (5), 157.
 Bharathi, B., Bhuvana, J., 2019. Domain name detection and classification using deep neural networks. *Commun. Comput. Inf. Sci.* 969, 678–686.
 Braden, R., 1989. RFC1123: Requirements for Internet Hosts-Application and Support.
 Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24 (2), 123–140.
 Casino, F., Lykousas, N., Homoliak, I., Patsakis, C., Hernandez-Castro, J., 2020. HYDRA dataset. Zenodo, <http://dx.doi.org/10.5281/zenodo.3965397>.
 Chaignon, P., 2020. DGA collection. <https://github.com/pchaigno/dga-collection>.

- Chen, Y., Zhang, S., Liu, J., Li, B., 2018. Towards a deep learning approach for detecting malicious domains. In: 2018 IEEE International Conference on Smart Cloud, SmartCloud. IEEE, pp. 190–195.
- Choudhary, C., et al., 2018. Algorithmically generated domain detection and malware family classification. In: International Symposium on Security in Computing and Communication. Springer, pp. 640–655.
- Curtin, R.R., Gardner, A.B., Grzonkowski, S., Kleymenov, A., Mosquera, A., 2019. Detecting DGA domains with recurrent neural networks and side information. In: Proceedings of the 14th International Conference on Availability, Reliability and Security. pp. 1–10.
- Dieterich, T.G., 2000. Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. Springer, pp. 1–15.
- Fu, Y., Yu, L., Hambolu, O., Ozcelik, I., Husain, B., Sun, J., Sapra, K., Du, D., Beasley, C.T., Brooks, R.R., 2017. Stealthy domain generation algorithms. *IEEE Trans. Inf. Forensics Secur.* 12 (6), 1430–1443. <http://dx.doi.org/10.1109/TIFS.2017.2668361>.
- Homoliak, I., Ovsonka, D., Greg, M., Hanacek, P., 2014. NBA of obfuscated network vulnerabilities' exploitation hidden into HTTPS traffic. In: The 9th International Conference for Internet Technology and Secured Transactions, ICITST-2014. IEEE, pp. 310–317.
- Jiang, N., Cao, J., Jin, Y., Li, L.E., Zhang, Z., 2010. Identifying suspicious activities through DNS failure graph analysis. In: The 18th IEEE International Conference on Network Protocols. pp. 144–153.
- Jyothsna, P., Prabha, G., Shahina, K., Vazhayil, A., 2018. Detecting DGA using deep neural networks (DNNs). In: International Symposium on Security in Computing and Communication. Springer, pp. 695–706.
- Khehra, G., Sofat, S., 2018. Botscoop: Scalable detection of DGA based botnets using DNS traffic. In: 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT. IEEE, pp. 1–6.
- Klensin, J., 2010. Internationalized Domain Names in Applications (IDNA): Protocol. Tech. rep., RFC 5891.
- Koh, J.J., Rhodes, B., 2018. Inline detection of domain generation algorithms with context-sensitive word embeddings. In: 2018 IEEE International Conference on Big Data, Big Data. IEEE, pp. 2966–2971.
- Kuncheva, L.L., 2014. Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons.
- Li, Y., Xiong, K., Chin, T., Hu, C., 2019. A machine learning framework for domain generation algorithm (DGA)-based malware detection. *IEEE Access*.
- Lison, P., Mavroeidis, V., 2017. Automatic detection of malware-generated domains with recurrent neural models. *arXiv preprint arXiv:1709.07102*.
- Liu, D., Li, Z., Du, K., Wang, H., Liu, B., Duan, H., 2017. Don't let one rotten apple spoil the whole barrel: Towards automated detection of shadowed domains. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. CCS '17, ACM, New York, NY, USA, pp. 537–552.
- Liu, X., Wu, J., Zhou, Z., 2009. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern.* B 39 (2), 539–550.
- Liu, Z., Zeng, Y., Zhang, P., Xue, J., Zhang, J., Liu, J., 2018. An imbalanced malicious domains detection method based on passive DNS traffic analysis. *Secur. Commun. Netw.* 2018.
- Mac, H., Tran, D., Tong, V., Nguyen, L.G., Tran, H.A., 2017. DGA botnet detection using supervised learning methods. In: Proceedings of the Eighth International Symposium on Information and Communication Technology. ACM, pp. 211–218.
- Manadhata, P.K., Yadav, S., Rao, P., Horne, W., 2014. Detecting malicious domains via graph inference. In: Kutylowski, M., Vaidya, J. (Eds.), *Computer Security - ESORICS 2014*. Springer International Publishing, Cham, pp. 1–18.
- Nadji, Y., Antonakakis, M., Perdisci, R., Dagon, D., Lee, W., 2013. Beheading hydras: Performing effective botnet takedowns. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. pp. 121–132.
- Nadji, Y., Perdisci, R., Antonakakis, M., 2017. Still beheading hydras: Botnet takedowns then and now. *IEEE Trans. Dependable Secure Comput.* 14 (5), 535–549.
- Patsakis, C., Casino, F., 2019. Hydras and IPFS: A decentralized playground for malware. *Int. J. Inf. Secur.*
- Patsakis, C., Casino, F., 2021. Exploiting statistical and structural features for the detection of domain generation algorithms. *J. Inform. Secur. Appl.* 58, 102725.
- Patsakis, C., Casino, F., Katos, V., 2020. Encrypted and covert DNS queries for botnets: Challenges and countermeasures. *Comput. Secur.* 88, 101614.
- Perdisci, R., Corona, I., Giacinto, G., 2012. Early detection of malicious flux networks via large-scale passive DNS traffic analysis. *IEEE Trans. Dependable Secure Comput.* 9 (5), 714–726.
- Plohm, D., Yakdan, K., Klatt, M., Bader, J., Gerhards-Padilla, E., 2016. A comprehensive measurement study of domain generating malware. In: 25th USENIX Security Symposium, USENIX Security 16. USENIX Association, Austin, TX, pp. 263–278.
- Schiavoni, S., Maggi, F., Cavallaro, L., Zanero, S., 2014. Phoenix: DGA-based botnet tracking and intelligence. In: Dietrich, S. (Ed.), *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer International Publishing, Cham, pp. 192–211.
- Selvi, J., Rodríguez, R.J., Soria-Olivas, E., 2019. Detection of algorithmically generated malicious domain names using masked N-grams. *Expert Syst. Appl.* 124, 156–163.
- Shabtai, A., Moskovitch, R., Feher, C., Dolev, S., Elovici, Y., 2012. Detecting unknown malicious code by applying classification techniques on opcode patterns. *Secur. Inform.* 1 (1), 1.
- Silva-Palacios, D., Ferri, C., Ramírez-Quintana, M.J., 2017. Improving performance of multiclass classification by inducing class hierarchies. *Procedia Comput. Sci.* 108, 1692–1701.
- Singh, M., Singh, M., Kaur, S., 2019. Issues and challenges in DNS based botnet detection: A survey. *Comput. Secur.* 86, 28–52.
- Sivaguru, R., et al., 2018. An evaluation of DGA classifiers. In: 2018 IEEE International Conference on Big Data, Big Data. IEEE, pp. 5058–5067.
- Sood, A.K., Zeadally, S., 2016. A taxonomy of domain-generation algorithms. *IEEE Secur. Priv.* 14 (4), 46–53.
- Spooren, J., et al., 2019. Detection of algorithmically generated domain names used by botnets: A dual arms race. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. SAC '19, ACM, New York, NY, USA, pp. 1916–1923.
- Stratosphere Labs, 2020. Stratosphere labs datasets overview. <https://www.stratosphereips.org/datasets-overview>.
- Tran, D., Mac, H., Tong, V., Tran, H.A., Nguyen, L.G., 2018. A LSTM based framework for handling multiclass imbalance in DGA botnet detection. *Neurocomputing* 275, 2401–2413.
- Valentini, G., Masulli, F., 2002. Ensembles of learning machines. In: Italian Workshop on Neural Nets. Springer, pp. 3–20.
- Vinayakumar, R., Soman, K.P., Poornachandran, P., Akarsh, S., Elhoseny, M., 2019. Improved DGA domain names detection and categorization using deep learning architectures with classical machine learning algorithms. In: Hassanien, A.E., Elhoseny, M. (Eds.), *Cybersecurity and Secure Information Systems: Challenges and Solutions in Smart Environments*. Springer International Publishing, Cham, pp. 161–192.
- Wang, Q., et al., 2020. Malicious domain detection based on K-means and SMOTE. In: *Computational Science – ICCS 2020*. Springer International Publishing, Cham, pp. 468–481.
- Woodbridge, J., Anderson, H.S., Ahuja, A., Grant, D., 2016. Predicting domain generation algorithms with long short-term memory networks. *arXiv preprint arXiv:1611.00791*.
- Xu, C., Shen, J., Du, X., 2019. Detection method of domain names generated by DGAs based on semantic representation and deep neural network. *Comput. Secur.* 85, 77–88.
- Yadav, S., Reddy, A.L.N., 2012. Winning with DNS failures: Strategies for faster botnet detection. In: Rajarajan, M., Piper, F., Wang, H., Kesidis, G. (Eds.), *Security and Privacy in Communication Networks*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 446–459.
- Yadav, S., Reddy, A.K.K., Reddy, A.L.N., Ranjan, S., 2012. Detecting algorithmically generated domain-flux attacks with DNS traffic analysis. *IEEE/ACM Trans. Netw.* 20 (5), 1663–1677.
- Yang, L., Liu, G., Dai, Y., Wang, J., Zhai, J., 2020. Detecting stealthy domain generation algorithms using heterogeneous deep neural network framework. *IEEE Access* 8, 82876–82889.
- Yang, L., et al., 2018. A novel detection method for word-based DGA. In: International Conference on Cloud Computing and Security. Springer, pp. 472–483.
- Yu, B., Gray, D.L., Pan, J., De Cock, M., Nascimento, A.C., 2017. Inline DGA detection with deep networks. In: 2017 IEEE International Conference on Data Mining Workshops, ICDMW. IEEE, pp. 683–692.
- Yu, B., Pan, J., Hu, J., Nascimento, A., De Cock, M., 2018. Character level based detection of DGA domain names. In: 2018 International Joint Conference on Neural Networks, IJCNN. IEEE, pp. 1–8.
- Yun, X., Huang, J., Wang, Y., Zang, T., Zhou, Y., Zhang, Y., 2020. Khaos: An adversarial neural network DGA with high anti-detection ability. *IEEE Trans. Inf. Forensics Secur.* 15, 2225–2240.
- Zago, M., Gil Pérez, M., Martínez Pérez, G., 2019. Scalable detection of botnets based on DGA: Efficient feature discovery process in machine learning techniques. *Soft Comput.*
- Zago, M., Perez, M.G., Perez, G.M., 2020. UMUDGA: A dataset for profiling DGA-based botnet. *Comput. Secur.* 92, 101719.
- Zander, S., Armitage, G., Branch, P., 2007. A survey of covert channels and countermeasures in computer network protocols. *IEEE Commun. Surv. Tutor.* 9 (3), 44–57.
- Zang, X., Gong, J., Mo, S., Jakalan, A., Ding, D., 2018. Identifying fast-flux botnet with AGD names at the upper DNS hierarchy. *IEEE Access* 6, 69713–69727.

Zhao, G., Xu, K., Xu, L., Wu, B., 2015. Detecting APT malware infections based on malicious DNS and traffic analysis. *IEEE Access* 3, 1132–1142.

Zhou, Y., Li, Q.-S., Miao, Q., Yim, K., 2013. DGA-based botnet detection using DNS traffic. *J. Internet Serv. Inf. Secur.* 3, 116–123.

Fran Casino is a postdoctoral researcher in the Department of Informatics at Piraeus University (Piraeus, Greece). He obtained his B.Sc. degree in Computer Science in 2010 and his M.Sc. degree in Computer Security and Intelligent Systems in 2013, both from Rovira i Virgili University in Tarragona, Catalonia, Spain. He received a Ph.D. in Computer Science from the Rovira i Virgili University in 2017 with honours (A cum laude) as well as the best dissertation award. He was visiting researcher in ISCTE-IUL (Lisbon-2016). He has participated in several European-, Spanish- and Catalan-funded research projects and he has authored more than 50 publications in peer-reviewed international conferences and journals. His research focuses on pattern recognition, and data management applied to different fields such as privacy and security protection, recommender systems, smart health, supply chain, and blockchain.

Nikolaos Lykousas is a Ph.D. student at the University of Piraeus studying deviant behaviour in modern Social Networks. He received his Master's degree in Intelligent interactive Systems from Universitat Pompeu Fabra in 2017, where he was awarded with an academic excellence scholarship for his achievements, after receiving his B.S. degree from the Department of Informatics at the University of Piraeus, in 2016. Since his undergraduate studies, he has participated as a research engineer in several EC funded projects and has gained considerable experience in the fields of Big Data Analytics, Cybersecurity, Digital Privacy and Cloud computing.

Ivan Homoliak was born in Rimavska Sobota, Slovakia, in 1987. He received the B.S. degree in information technology, the M.S. degree in intrusion detection and supervised machine learning, and the Ph.D. degree in adversarial intrusion detection in network traffic from the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Czech Republic, in 2007, 2012, and 2016, respectively. From

2016 to 2020, he was a Postdoctoral Research Fellow with the Singapore University of Technology and Design (SUTD), and he worked in the project aimed at distributed computing and blockchains. Prior to that, Ivan worked on the project focusing on insider threat detection in SUTD, where he pursued the application of ML in this area and analysis of security issues. Currently, he is the research scientist at FIT BUT, and he focuses on security aspects of blockchains and their applications.

Constantinos Patsakis holds a B.Sc. in Mathematics from the University of Athens, Greece and an M.Sc. in Information Security from Royal Holloway, University of London. He obtained his Ph.D. in Cryptography and Malware from the Department of Informatics of the University of Piraeus. His main areas of research include cryptography, security, privacy, data anonymisation and cybercrime.

He has authored numerous publications in peer-reviewed international conferences and journals, and he has been teaching computer science courses in European universities for more than a decade. Dr Patsakis has been working in the industry as a freelance developer and security consultant. He has participated in several national (Greek, Spanish, Catalan and Irish) and European R&D projects. Additionally, he has worked as a researcher at the UNESCO Chair in Data Privacy at the Rovira i Virgili University (URV) in Tarragona, Catalonia, Spain and as a research fellow at Trinity College, Dublin Ireland. Currently, he is an Associate Professor at the University of Piraeus and adjunct researcher of Athena Research and Innovation Center.

Julio Hernandez-Castro is a Professor of Cybersecurity at the University of Kent, in the UK. He received a degree in Mathematics in 1995, an M.Sc. in Coding Theory and Network Security in 1999 and a Ph.D. in Computer Science in 2003. His interests are primarily in Computer Security, especially Steganography and Steganalysis, but also in Machine Learning applications to Cybersecurity, RFID/NFC Security, and in advancing the study of Randomness Generation and Testing, particularly on hardware and constrained IoT devices. He additionally works in studying ransomware and other types of malware. He is a keen chess player and Python enthusiast.

Received April 14, 2020, accepted June 18, 2020, date of publication June 24, 2020, date of current version July 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004727

Unravelling Ariadne's Thread: Exploring the Threats of Decentralised DNS

CONSTANTINOS PATSAKIS^{1,2}, (Member, IEEE), FRAN CASINO¹, (Member, IEEE),
NIKOLAOS LYKOUSAS¹, AND VASILIOS KATOS³, (Member, IEEE)

¹Department of Informatics, University of Piraeus, 185 34 Piraeus, Greece

²Information Management Systems Institute, Athena Research Center, 151 25 Marousi, Greece

³Department of Computing and Informatics, Bournemouth University, Poole BH12 5BB, U.K.

Corresponding author: Constantinos Patsakis (kpatsak@unipi.gr)

This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the projects CyberSec4Europe (Grant Agreement no. 830929), LOCARD (Grant Agreement no. 832735) and ECHO (Grant Agreement no 830943). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

ABSTRACT The current landscape of the core Internet technologies shows considerable centralisation with the big tech companies controlling the vast majority of traffic and services. This situation has sparked a wide range of decentralisation initiatives with blockchain technology being among the most prominent and successful innovations. At the same time, over the past years there have been considerable attempts to address the security and privacy issues affecting the Domain Name System (DNS). To this end, it is claimed that Blockchain-based DNS may solve many of the limitations of traditional DNS. However, such an alternative comes with its own security concerns and issues, as any introduction and adoption of a new technology typically does - let alone a disruptive one. In this work we present the emerging threat landscape of blockchain-based DNS and we empirically validate the threats with real-world data. Specifically, we explore a part of the blockchain DNS ecosystem in terms of the browser extensions using such technologies, the chain itself (Namecoin and Emercoin), the domains, and users who have been registered in these platforms. Our findings reveal several potential domain extortion attempts and possible phishing schemes. Finally, we suggest countermeasures to address the identified threats, and we identify emerging research themes.

INDEX TERMS Blockchain, blockchain forensics, cybercrime, DNS, malware.

I. INTRODUCTION

One could argue that there is a periodic paradigm bounce between centralisation and decentralisation in computer science. A representative example is the transition from mainframes with dummy terminals to personal computers or the shift from centralised local storage to the cloud. Although the Internet was in principle designed to be distributed and decentralised by nature, in reality, the control is placed onto a relatively limited number of stakeholders and the quest for further decentralisation is becoming an imminent need. Such requirement manifests in many ways, see for example the case of net neutrality, or the concept of crowdsourcing which attempts to address efficiency and sustainability issues. As such, in recent years, we are witnessing an increasing demand and creation of decentralised services.

The associate editor coordinating the review of this manuscript and approving it for publication was Kuo-Hui Yeh¹.

A noteworthy example of decentralisation is the blockchain technology, which is being widely deployed in various and diverse fields [1]. In different forms, the decentralisation wave is gradually reaching traditional centralised services, such as DNS. DNS is a distributed database with a centralised data governance model, primarily controlled by The Internet Corporation for Assigned Names and Numbers (ICANN). In this regard, ICANN manages the top-level domains (TLDs) and the operation of root name servers. In practice, in order for a client to contact a host of a particular domain name, it first issues a query to a DNS server to obtain the host's IP address. For efficiency the DNS server may maintain a copy of this information in its cache, depending on how often this domain is requested. In the case where the DNS server does not hold the information requested, the query is propagated to the root name server. Next, the root name server will find the servers for the corresponding TLD and then forward the query to the corresponding authoritative name server, which would return the requested IP.

While DNS is currently one of the oldest yet critical Internet application-level protocols, it has several drawbacks that mandate its replacement. For instance, DNS does not support cryptographic primitives by default. Although DNS supports security extensions through DNSSEC, these are not widely used. As a result, any query and response can be intercepted by anyone on the same network, exposing this service to numerous threats, primarily through man-in-the-middle type of attacks. Indicatively, there can be confidentiality and privacy violations through passive eavesdropping, as well as integrity breaches since anyone on the same network, may inject a response of an intercepted query. Moreover, totalitarian regimes can exploit DNS to censor unwanted web pages and services. Furthermore, in the past years the DNS servers have been both attack targets - see for example DNS poisoning attacks - as well as components of an attack vector, as they have been used in amplification denial of service attacks.

Motivation: The issues above have driven the research community to seek alternative solutions to DNS. Some initiatives include DNS over HTTPS [2] and DNS over TLS [3] and [4] while others are looking into solutions to provide alternatives to ICANN's centralisation paradigm. One of the most promising decentralised solutions is blockchain DNS which has already been adopted by several chains such as Ethereum, Namecoin and Emercoin. Despite being in their infancy, blockchain domains have attracted the interest of several prominent stakeholders. A notable example is Alibaba, who recently filed a patent for a blockchain-based domain name management system.¹ A brief overview of blockchain DNS together with some degree of scepticism is presented in [5]. To date, blockchain DNS is already being exploited by cybercriminals.² Therefore, we argue that there is a need to explore threat models related to novel blockchain solutions,³ as well as decentralised file storage systems [6]. The decentralisation of services may undoubtedly provide a plethora of possibilities in terms of privacy, security and democratisation. Nevertheless, substantial changes in the backbone of well-established services and infrastructures may come at a high cost. Adversaries are expected to opportunistically take advantage of such changes by exploiting the lack of knowledge, experience and maturity of the users and deployments, as well as the inherent flaws that exist in the early stages of a new technology. At the same time, the use of encrypted and covert communications adds another layer of difficulty to detect infected systems [7], for instance, in the case of botnets. Therefore, it is imperative to raise awareness on the opportunities as well as the emerging security threats. This paper aims to fill this research gap by providing an overview of the current state of the art and practice (Section II), a detailed presentation of the emerging

threats and how they could be amplified (Section III). Further to merely speculating future threats, we perform an investigation and analysis of the currently available blockchain DNS ecosystem and illustrate the presence of risks. To this end, in Section IV, we showcase the results of an in-depth analysis of Namecoin, Emercoin and Blockchain DNS. Our findings show that there are ongoing domain extortion activities and indicate that possible phishing campaigns have already been deployed. It should be noted that the threats discussed and the conclusions drawn from the statistical analysis could be extended to other Blockchain DNS systems. Finally, some remarks and findings are further discussed, along with possible countermeasures in Section V.

To the best of our knowledge, the previous work in this field was limited to the research by Kaodner *et al.* [8] back in 2011 who analysed the Namecoin domain. The authors studied an early version of the Namecoin domain; however, they identified issues such as domain squatting which was an anticipated threat. In our work, the analysis is considerably extended by providing a detailed study of Namecoin and Emercoin data in terms of domains, addresses and their corresponding timelines. We perform an analysis and empirical evaluation of the current state of practice in real-world blockchain DNS systems. Moreover, we identify extortion schemes, pricing schemes and discuss both domain squatting and typo squatting. The recent high rate of domain registrations and the observation that particular parties registered a considerable number of domains - some in the order of thousands - indicate that blockchain DNS in its current state may not constitute a safe and secure ecosystem. As such, the broader adoption of such solutions, despite their attractive features, should be approached with scepticism.

II. BACKGROUND

A. BLOCKCHAIN-BASED DNS

Decentralised systems were in principle used to improve the robustness and availability of domain name resolution tasks as well as enabling the feature of bypassing censorship campaigns and tampering, as discussed in [9]–[14]. Some of the research initiatives in this area focused on developing specific TLDs, such as in the Dot-P2P project (with the .p2p TLD) [15]. In this regard, although the idea of using P2P networks to perform distributed domain name resolution was interesting, their performance entailed several drawbacks [16]. Nevertheless, only up until recently, the adoption of distributed DNS is progressively gaining ground [5], mainly due to the inherent features of blockchain technology, such as immutability, verifiability, and trust. These features, when introduced to registrar systems, can enable functional and real-world scale distributed DNS systems. According to Scopus, Web of Science and Google Scholar, a set of approaches, some of which are fully functional, have appeared in the literature since 2016. In what follows, we describe and analyse the main features of the most relevant and adopted solutions. The work presented by Hari *et al.* [17] is one of the first works that

¹<https://domainnamewire.com/2019/08/15/alibaba-files-blockchain-domain-name-patent-application/>

²<https://www.digitalshadows.com/blog-and-research/how-cybercriminals-are-using-blockchain-dns-from-the-market-to-the-bazaar/>

³https://en.bitcoinwiki.org/wiki/Blockchain_Projects_List

propose the use of blockchain to develop a DNS infrastructure. The authors discuss the benefits of such a system over the main threats and drawbacks of traditional models such as compromised hosts, spoofing, trust management, and its heavy dependence on PKIs. Benshoof *et al.* [18] proposed a system named D³NS, which uses a distributed hash table and a domain name ownership implementation based on the Bitcoin blockchain. They aim to replace the top-level DNS and certificate authorities, offering increased scalability, security and robustness. Liu *et al.* [19] proposed a blockchain-based decentralisation DNS resolution method with distributed data storage to mitigate single points of failure and domain name resolution data tampering. Gourley and Tewari [20] proposed the use of blockchain to enhance the certificate validation procedure to create an improved DNS security extension, providing the same benefits with DNSSEC while overcoming its main drawbacks. Similarly, in an attempt to reduce the level of trust in certificate authorities, Guan *et al.* [21] presented AuthLedger, a blockchain-based system that provides efficient and secure domain name authentication. BlockZone, of Wang *et al.* [22], uses a replicated network of nodes to offer efficient name resolution through an improved Practical Byzantine Fault Tolerance (PBFT) consensus mechanism.

Some work focused on IoT systems, and their communication protocols have also been proposed. For example, Duan *et al.* [23] presented DNSLedger, a hierarchical multi-chain structure in which domain name management and resolution are performed in a decentralised way. The authors claim that their system could enhance IoT-related communication technologies due to its efficiency. BlockONS, proposed by Yoon *et al.* [24], is a system that aims to overcome classical problems related to DNS resolution, namely DNS cache poisoning, spoofing, and local DNS cracking. The authors propose a robust and scalable object name service appropriate for an IoT ecosystem. ConsortiumDNS was introduced by Wang *et al.* [25] as a system based on a three-layer architecture composed by consortium blockchain, a consensus mechanism and external storage. The authors claim that their approach increases the efficiency of the overall system, compared to other well-known approaches such as Namecoin or Blockstack. Finally, a number of patented designs of Blockchain-based DNS systems is found in [26], [27].

Currently, there are several relevant and widely adopted blockchain DNS projects. Handshake⁴ is one of the most widely supported technologies, which aims to offer an alternative to existing certificate authorities. Therefore, Handshake aims to replace the root zone file and the DNS name resolution and registration services worldwide. The Ethereum name service (ENS)⁵ uses smart contracts to manage the .eth registrar by means of bids and recently added the support for .onion addresses. Namecoin⁶ is a cryptocurrency

based on Bitcoin, with additional features such as decentralised name system management, mainly for the .bit domain. It was the first project to provide an approach to address Zooko's triangle since the system is secure, decentralised and human-meaningful. Nevertheless, contrary to well-established blockchains like Bitcoin, Namecoin's main drawback is its insufficient computing power, which makes it more vulnerable to the 51% attack. Practically, if an adversary manages to get a slight majority of the computing power, they may rewrite the whole chain. Blockstack [28] is a well-known blockchain-based naming and storage system that overcomes the main drawbacks of Namecoin. Blockstack's architecture separates control and data planes, enabling seamless integration with the underlying blockchain. EmerDNS⁷ is a system for decentralised domain names supporting a full range of DNS records. EmerDNS operates under the "DNS" service abbreviation in the Emercoin NVS. Nebulis⁸ is a globally distributed directory that relies on the Ethereum ecosystem and smart contracts to store, update, and resolve domain records. Moreover, Nebulis proposes the use of off-chain storage (i.e. IPFS) as a replacement for HTTP. OpenNIC⁹ deserves a special mention since it is a hybrid approach in which a group of peers manages namespace registration, yet the name resolving task is fully decentralised. OpenNIC provides DNS namespace and resolution over a set of domains, including those maintained by blockchain solutions such as EmerDNS and New Nations.¹⁰ Moreover, OpenNIC resolvers have recently added access to domains administered by ICANN. In addition to namespace registrar, users can also create their own TLD on request. It should also be noted that OpenNIC has recently voted to drop support for .bit after rampant abuse from malware operators. It is worth mentioning that this decision was taken after a voting process by the OpenNIC members. Table 1 summarises the main features of the most relevant Blockchain-DNS systems.

TABLE 1. Technical characteristics of the most relevant DNS systems. Although Blockstack is blockchain agnostic, it is mainly used with Bitcoin blockchain.

Method	Pedigree Platform	Registrar and Resolution Management	TLD Examples
ICANN	Network of Servers and resolvers	Centralised	.com .net .org
OpenNIC	Decentralised Servers	Hybrid	.bbs .pirate .libre
ENS	Ethereum	Decentralised	.eth .onion
Handshake	Bitcoin	Decentralised	unrestricted
Blockstack	Blockchain agnostic	Decentralised	.id .podcast .helloworld
Emercoin	Bitcoin	Decentralised	.coin .bazar .emc
Namecoin	Bitcoin and Peercoin	Decentralised	.bit

⁷<https://emercoin.com/en/documentation/blockchain-services/emerdns/emerdns-introduction>

⁸<https://www.nebulis.io/>

⁹<https://www.opennic.org/>

¹⁰<http://www.new-nations.net/>

⁴<https://handshake.org/>

⁵<https://ens.domains/>

⁶<https://www.namecoin.org/>

Internet users can reach the TLDs offered by Namecoin, OpenNIC, New Nations, and EmerDNS (e.g. .bit, .coin, .emc, .lib and .bazar) through various browser extensions such as peername, blockchain-DNS and friGate [29]. The process is outlined in Figure 1.

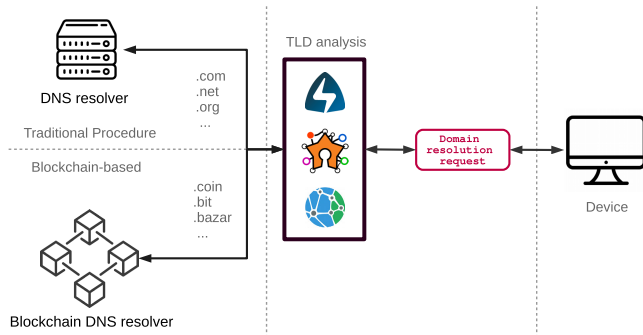


FIGURE 1. Workflow of the browser extensions procedure to enable resolution of EmerDNS, Namecoin, New Nations and OpenNIC domains. The extension inspects the TLD of the requested domain and directs the query to the corresponding DNS system.

B. DOMAIN NAME FRAUD

Apart from the protocol weaknesses DNS carries, there are also several attacks on the underlying processes. For instance, when a business registers its name on a specific TLD, an adversary may opportunistically register the same name to another TLD. This attack is known as *domain squatting*. ICANN, being the central authority of the main TLDs, has the capability to alleviate from such attacks. Another attack stems from the failure of timely renewal of a domain name once its registration has lapsed. An adversary may use automated systems to buy the domain name preemptively. This is referred to as *dropcatching*.

Another attack vector emerges from the typos that people usually make when typing. In this regard, an adversary may register a domain that looks like a known domain, but with a small typo. This is usually called *typosquatting*. A particular case of typosquatting is the exploitation of linguistic collisions. In this attack, the adversary tries to exploit the fact that a typo in a word may result in a word in another language. Therefore, since the search engines correct such errors in the search results, a malicious domain may appear as a legitimate result of a query for the target domain, poisoning the search results. In *bitsquatting*, the adversary tries to exploit the possible network errors that may introduce some noise in the response of a DNS server. In this sense, since there is only one bit of difference between the registered and the target domain, the smallest hardware error could trigger the attack. *Homograph* attacks attempt to exploit the visual resemblance of one domain with another, registering, e.g. punycodes for target domain names so that the IDN looks similar to it in the browser. In *soundsquatting*, the adversary registers domains that sound similar to the target domain. In *combosquatting* the adversary tries to trick a user into trusting a domain because it looks like the original, yet has some additional

words appended or prepended. The latter is something that many legitimate sites may do as well for publicity, so the user is accustomed to trusting them. Similarly, in *AbbrevSquatting* the adversary registers a possible abbreviation of a domain name. Since mobile devices have limited space to illustrate information, an adversary may embed a trusted domain name in the second-level domain names. This tactic is known as *levelsquatting*. In Table 2, we provide a categorisation of the related work in terms of attacks and scope (traditional DNS and distributed DNS). Table 3 illustrates most of these attacks with examples.

TABLE 2. Overview of domain attacks related works.

	DNS	Distributed DNS
[30]–[35]	Typosquatting	
[36], [37]	Bitsquatting	
[38]	Combosquatting	
[39]	Soundsquatting	
[40]	Abbrevsquatting	
[41]–[43]	Homograph	
[44]	Levelsquatting	
[45]	Dropcatching	
[46]	Linguistic-collision	
[47]	Domain squatting	
[8]		Domain squatting
This work		Domain & typo squatting

TABLE 3. Examples of types of domain fraud.

Attack	Benign	Malicious
Domain squatting	facebook.com	facebook.new
Typosquatting	facebook.com	facebok.com
Bitsquatting	facebook.com	fcebook.com
Combosquatting	facebook.com	yourfacebook.com
Soundsquatting	facebook.com	phacebook.com
Abbrevsquatting	fb.com	fbk.com
Homograph	facebook.com	facebook.com
Levelsquatting	facebook.com	facebook.com.maldom.com
Linguistic-collision	adobe.com	idobe.com

C. DISTRIBUTED PLATFORMS AND C2

Nowadays, advanced and sophisticated malware campaigns continuously emerge, and some are already employing the services offered by decentralised technologies such as blockchain and distributed file storage (DFS). In the case of botnets, the use of technologies such as DFS systems prevents the generation of *non-existent* domain errors (NXDomain responses), which is a well-known Indicator of Compromise (IoC) type for malware using domain generation algorithms. In this regard, Patsakis et al. [6] extended the definition of domain generation algorithms (i.e. a family of pseudo-random domain name generators to which an infected host can dynamically identify the location of its C2 server) into a more generic framework, namely Resource Identifier Generation Algorithms (RIGA). Moreover, the authors showed how DFS like IPFS could enhance malware campaigns due to their attractive features such as immutability, efficiency and negligible costs. Botnet C2 management through Blockchain systems is also a noteworthy threat as

proposed by Ali *et al.* [48] and used in the case of the Cerber ransomware, analysed in by Pletnick *et al.* [49]. In this case, the malware retrieves the C2 address from the transaction information of the bitcoin blockchain. A more recent threat is the use of encrypted and covert communication channels such as in the case of DNSsec, DNS over HTTPS (DoH) and DNS over TLS (DoT). Although these technologies hinder the possibility of using NXDomain information leaks to detect suspicious behaviour, Patsakis *et al.* showed that even in such case some patterns might emerge [7], which can be used to identify and classify Domain Generation Algorithm (DGA) families accurately. Regarding the recently developed Blockchain-DNS systems, there are emerging uses of these for cybercriminal activities such as the setup of illicit market places.¹¹

TABLE 4. Main characteristics of blockchain DNSs.

Property	Description
Trust	Verifiable and robust consensus mechanisms
Decentralisation	The network is completely distributed with no central entities
Availability	The availability of the network depends on multiple peers and not on a single entity.
Censorship-resistant	Access to information and domain name resolution are not subject to borders or bans
Robustness	Resilient to attacks that affect centralised DNS systems such as MiM, spoofing, cache poisoning, cracking.
Unlimited Resources	A high number of simultaneous users sharing their assets.
Namespace Freedom	Registration of new SLDs and TLDs
Automated Management	Auctions to register domain names, fast and transparent ownership control

III. THE DECENTRALISED DNS THREAT

A blockchain-based DNS solution offers the features and benefits as summarised in Table 4. In this regard, one could argue that the traditional DNS seems to be outdated, compared to the novel blockchain DNSs. In any case, the traditional DNS proved its worth in terms of reliability and scalability from the early 80s until today with modest adjustments. However, blockchain-based DNSs are not short of introducing new and emerging threats, giving opportunities for the development of novel attack vectors [50]–[53]. In the following sections, we present and analyse the most well-known threats as well as identify novel ones. We also discuss their potential impact on sociotechnical systems. Figure 2 is an overview of the emerging threats surrounding the blockchain DNS.

A. MALWARE

Malware actors are among the prime beneficiaries of blockchain-based DNS services. This enabling technology provides the capability to register a substantial number of domains with low entropy. Currently, malware authors use DGAs to generate domain names (i.e. algorithmically generated domains or AGDs); however, since most short and

¹¹<https://www.digitalshadows.com/blog-and-research/how-cybercriminals-are-using-blockchain-dns-from-the-market-to-the-bazar/>

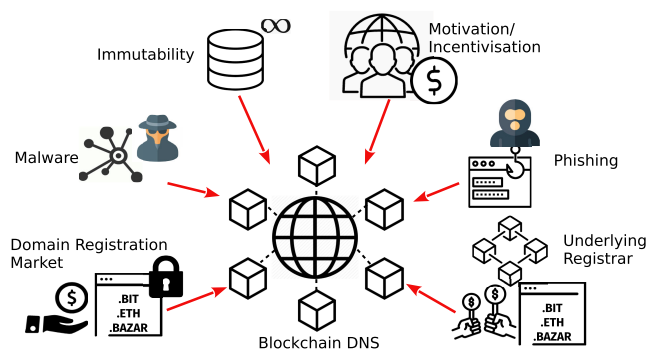


FIGURE 2. An overview of main threats of blockchain DNSs.

meaningful domain names are not available, they resorted to the use of long and random-looking domain names. Upon infecting a host, a bot that uses a DGA issues many Non-Existent Domain (NXDomain) requests to resolve the C2 server. The surge of DNS queries and corresponding NXDomain answers can be analysed, potentially providing attribution by singling out the underlying DGA.

With the use of blockchain-based DNS systems, the conventional NXDomain requests will not be issued (see next section), hence hindering the detection mechanisms. Moreover, by using domain names with lower entropy, many filtering and machine learning approaches are rendered useless.

Even more, the use of blockchain-based DNSs introduces further challenges for malware analysts. When performing static analysis on the reverse-engineered code, the analyst and the tools that they use must have knowledge on the new domains and who maintains them as the function calls can considerably differ. Traditional filters for domain names will fail to reveal calls to a `.bit` domain for instance, as the resolution mechanism, is completely different. In fact, the use of the blockchain DNS from various botnets¹² to connect to the C2 servers has reportedly created more issues in the analysis, attribution, and takedown. As reported by deteque,¹³ more than 100 domains registered in blockchain DNS registrars were used by C2 servers in 2018 implying that their use is actively being exploited by cybercriminals. In light of the above, OpenNIC has recently decided¹⁴ to drop support for `.bit` domains.

In addition, it should be noted that requests to agreeably benign domains, e.g. `google.com`, may resolve to IP addresses not owned by the domain. The same of course applies for case sensitive domains, e.g. `GoOgle.com`, or the use of spaces, e.g. `google.com`. While Handshake, for instance, may have already taken some precautionary measures for the highly visited domain names, this does not prevent the use of existing domain names with less visibility in being exploited to serve

¹²<https://blog.netlab.360.com/threat-alert-a-new-worm-fbot-cleaning-adminer-is-using-a-blockchain-based-dns-en/>
<https://www.microsoft.com/security/blog/2018/03/07/behavior-monitoring-combined-with-machine-learning-spoils-a-massive-dofail-coin-mining-campaign/>

¹³<https://www.deteque.com/news/abused-top-level-domains-2018/>

¹⁴https://wiki.opennic.org/votings/drop_namecoin

malware. Unfortunately, these DNS servers register and may resolve domains which are case sensitive, indicating another form of phishing and domain squatting that could be used in the near future. It should be noted that an adversary could still perform fast fluxing and change the IP addresses that are used whenever deemed necessary by simply performing a transaction in the chain. From a digital forensics perspective however the whole history and timeline of the fast-flux will be preserved due to the immutability feature of the blockchain. Finally, it should be noted that blockchain-based schemes often provide strong privacy guarantees, preventing law enforcement agencies from tracking the perpetrators, providing them with the perfect cover-up for their operations.

B. UNDERLYING REGISTRAR MECHANISM

The main approach for registering domains in blockchain DNSs is to perform bids or auctions, replacing the first-request, first-served concept followed in traditional, centralised DNS. However, by exploiting vulnerabilities in the underlying bid system, an attacker may obtain control of domains as recently observed with the `apple.eth` domain grab.¹⁵ Moreover, most blockchain DNS systems such as Emercoin allow the registration of case sensitive domains, which is not possible in traditional systems. The latter, if paired with some other unrestricted practices such as the use of spaces, non UTF-8 or ASCII characters, may lead to an explosion of the (alternative) domain namespace where legitimate domains may not be easily distinguishable. Such a situation is likely to raise trust issues towards the DNS service in general. Note that the attack mentioned above could be prevented and reverted in traditional DNS, but not in blockchain DNSs. As such, the registration processes and implementation of the underlying smart contracts will need to be extensively studied.

In essence, an uncontrolled and fully decentralised DNS type of service may lead to having *parallel* Internets. Note that each blockchain DNS system enables the registration of arbitrary sets of TLDs, which may overlap with existing ones. Therefore, the same domain would resolve to different IPs, depending on the blockchain DNS system used. For instance, even if not used, the domain `google.com` is registered in Emercoin in block 252362.¹⁶ This opens a whole new avenue of possibilities, in which users can have access to a myriad of contents without restriction. Yet, in many instances, they could be owned by a malicious entity. The latter problem, as discussed in the following sections, is exacerbated by other properties such as immutability.

C. DOMAIN REGISTRATION MARKET

In the least sinister scenario, we consider the case of one registering the domain name of an existing, legitimate webpage. Since the blockchain TLDs are not known to the vast

¹⁵<https://www.coindesk.com/ethereum-name-service-auction-exploited-to-grab-apple-domain-and-it-cant-be-undone>

¹⁶<https://explorer.Emercoin.com/block/252362>

majority of people, it is expected that some will rush to opportunistically buy such names requesting a good payment in exchange for the name. As presented in more detail in the empirical evaluation, this practice is already taking place. Block 160356 of Emercoin¹⁷ illustrates such requests were the fees range from \$600 to \$20,000.

The problem is an extension of domain backordering as in this instance we are not dealing with expired domains, but with new TLDs. The existence of ICANN and intermediates, e.g. registrars, allows in many cases the arbitration or even the shutdown or handing over of a domain name. However, blockchain systems do not support such remediation mechanisms. In fact, at the time of writing, one can register a name for an arbitrary amount of time in Emercoin. For instance, there are many domain names in Emercoin which are registered for thousands of years, e.g. there are domains registered up to 5014 and 12012 in blocks 200590 and 380209, respectively.¹⁸

D. PHISHING

Phishing is a fraudulent practice which targets an audience to obtain valuable personal information by using impersonation of entities, persons and more techniques. According to the *State of the Phish 2019* by Proofpoint [54], the number of compromised accounts by these attacks varied from 38% to 65% from 2017 to 2018. This type of attack leverages socially engineering methods to trick users into performing activities that will benefit the attacker in some way, usually financially [55]. Email is the most popular avenue for a phishing attack, with more than 90% of successful cyber-attacks/security breaches being initiated from a spoofed email [56]. In fact, the automated capabilities of this attack, coupled with the incapacity of users to identify a phishing attack [57] may render the threat even more effective. There are many factors which augment this threat and most relate to the human. For instance, the timing of the attack, the authoritative writing, as well as the exploitation of common practices in an organisation, may significantly encourage the user into accepting the email as legitimate. Furthermore, the use of spoofed or compromised email accounts further complicates the situation.

In the context of blockchain DNSs, the above issues can be exacerbated. The users are accustomed to visiting specific web pages and sending emails to particular accounts. If these accounts are pointing to a similar address, e.g. changing the TLD, many users are highly likely to be tricked. The use of puny codes for phishing or the use of different TLDs can become an effective ingredient of an attack vector. With the introduction of blockchain DNSs, an adversary has far more options as there is a wide range of domains that are becoming available at a minimum cost. Practically, this means that not only the phishing sites may have a similar domain name with

¹⁷<https://explorer.emercoin.com/block/160358>

¹⁸<https://explorer.emercoin.com/block/200590> and <https://explorer.emercoin.com/block/380209>

legitimate ones, but with the use of, e.g. a Let's Encrypt¹⁹ certificate, the fraudulent web pages may have valid and trusted HTTPS support. Therefore, the phishing page may have all the distinctive elements, from the UI, the HTTPS support and the valid domain name, making it very difficult for a common user to distinguish the original from the phishing page.

E. LACK OF MOTIVATION

Motivation under the blockchain DNS paradigm is clearly related to the features offered by such a system, including censorship resistance as one of the main attractions. Nevertheless, these desirable features come at a cost, since decentralised systems rely completely on their nodes and their participation [58]. Therefore, keeping the user's interest in blockchain DNSs is critical.

Unarguably, blockchain's adoption in a myriad of scenarios is a reality [1]. Nevertheless, not all blockchain-based projects succeed. In this regard, according to Deadcoins²⁰ there are approximately 1000 dead cryptocurrencies and more than 660 attempts to promote fraudulent cryptocurrencies. Interestingly enough, as of 2018, ICO scams have already raised more than 1 billion dollars.²¹ Despite the existence of some awareness campaigns such as HoweyCoin,²² the lack of a specific and interoperable framework to pursue such deviant behaviour enables the persistence of these practices. In the case of blockchain, this may hinder the creation of new projects as well as the persistence of well-known and established ones. One of the main problems that could arise is an unbalanced/unstable computational power, which could compromise the underlying consensus mechanisms and trigger, for instance, a 51% attack. Note that this attack may be applied regardless of the number of users that use a blockchain DNS solution, as the attack is targeted towards the nodes that store the blockchain which, depending on the rewards they have, their participation may decrease over time. The latter may allow an adversary to control the blockchain and compromise its integrity without having to exploit any software vulnerability of the system.

F. IMMUTABILITY

The immutability property of blockchains, although standing as one of the main beneficial features, may also be abused for malicious purposes. Well-known blockchains such as Bitcoin Satoshi Vision (BSV)²³ and Bitcoin Blockchain have suffered from serving as an illegal data storage that cannot be deleted [59], [60]. The lack of verifiable deletion mechanisms enables DFS systems such as IPFS and IndImm²⁴ to host and disseminate illegal content [6]. Therefore, neither contents nor domain names are subject to a take-down mechanism.

¹⁹<https://letsencrypt.org>

²⁰<https://deadcoins.com/>

²¹<https://www.ccn.com/ico-scams-have-raised-more-than-1-billion-report-claims/>

²²<https://www.howeycoins.com/index.html>

²³<https://bitcoinsv.io/>

²⁴<https://en.cryptonomist.ch/2019/07/29/indimm-ripple-blockchain/>

Moreover, strategies as blacklisting domains are unpractical if the number of domains is high.

From a legal perspective, the GDPR does not consider the immutable nature of blockchains and DFS. In this sense, novel decentralised technologies implement features that are not aligned with current regulations and their requirements, which prevents the possibility to apply requests such as the right to be forgotten [61], [62]. Thus, the aforementioned facts make the combination of blockchain DNS and DFS systems a fertile playground for building malicious ventures. For instance, at the time of writing, Emercoin supports I2P (Invisible Internet Project) links; well-known for their anonymity, however, given the continuous rise of IPFS and other DFS solutions, blockchain DNS systems may support IPFS in the near future. The support of a permanent and distributed storage such as IPFS, combined with blockchain DNS, can allow the creation of a permanent link that cannot be taken down. It should be noted that there are already initiatives towards such direction, e.g. Unstoppable Domains.²⁵ Evidently, the combination of both would be ideal for the distribution of infringing content that would become permanently available for everyone who has access to the link.

IV. ANALYSIS OF REAL-WORLD DATA

To assess the extent and risk of these threats, we conducted an analysis of real-world data. In the first set of experiments, we used the BDNS extension²⁶ and in the second one we used the Namecoin²⁷ and the Emercoin²⁸ blockchain platforms. We argue that the most critical domain names are the top ones as captured in the Alexa domain global ranking system²⁹ since they handle most of the user traffic. Therefore, if an adversary would like to take over a domain, a domain in the Alexa top 1,000 domains would offer them the highest impact. In addition we constructed a dataset merging the top 1 million Alexa domains³⁰ with the Cisco Umbrella 1 Million³¹ dataset.

In what follows, we will refer to *AIK* as the dataset of the second-level domain (SLD) names of the Alexa top 1,000 domains collected and as *TOP1m* as the SLDs of the merge of the Alexa and Umbrella top 1 million datasets at the time of writing. The intuition behind having two distinct datasets is that *AIK* is small and can be used for exhaustive search without abusing the service provider's resources, while the *TOP1m* allows for a more extensive analysis that can be performed offline.

A. USING THE BDNS EXTENSION

BDNS is an open-source extension for Chrome and Firefox. The goal of the extension is to resolve .bit, .lib, .emc, .coin,

²⁵<https://unstoppabledomains.com>

²⁶<https://blockchain-dns.info>

²⁷<https://www.namecoin.org/>

²⁸<https://emercoin.com/>

²⁹<https://www.alexa.com/topsites>

³⁰<http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>

³¹<https://umbrella.cisco.com/blog/cisco-umbrella-1-million>

.bazar and OpenNIC domains.³² The extension monitors the requests of the browser for domains. If the domain falls within the supported TLDs, it uses a RESTful interface to resolve the IP.

Based on this concept, we created a crawler which sends queries using this REST interface and tries to resolve A1K domains with any of these TLDs. The search showed that 464 domains out of the potential 25,000 web pages (i.e. generated from the combination of A1K with the different TLDs) were registered. These 464 web pages were mapped to 465 IPs, as one of the DNS records mapped a domain to two IPs. Interestingly, 21 of these IPs were private and 444 public. The latter were actually 55 unique addresses, one of which was used to resolve 220 of these web pages, and 81 belong to another IP address, signifying a high concentration. In terms of countries, these domains resolve to 15 countries, as illustrated in Table 5.

TABLE 5. Distribution by country.

Country	IPs	Country	IPs	Country	IPs
DE	238	CA	5	AT	1
US	146	SG	3	HK	1
CN	20	GB	2	IT	1
FR	12	NL	2	SE	1
RU	9	SC	2	TW	1

Going a step further, we browsed each of the domains. From the 464 domains, 163 did not resolve to a valid server or returned an error on the server-side and 9 to a default welcome landing page of a service, e.g. IIS Server. Then, 80 pages redirected the user to a porn web page (<https://iusr.co>) which belonged to the same IP address (192.243.100.192). Note that the latter IP served only this web page except for one page that was down. Then, many of the pages resolved to placeholder pages. Three of them resolved to the same IP (161.97.219.84) pointing to “Computer Rehab domain hosting”, 11 pointed to a parking domain of dotbit.me with the same IP (144.76.12.6). Sixty-seven domains were registered as part of the project New Nations <http://www.new-nations.net> from a single IP (178.254.31.11). The latter IP also resolved 76 more web pages that were divided into three placeholder web pages (ww1.partenka.net, ww17.cikidot.com, ww38.partenka.net) with 63 in the first one, 3 in the second and 9 in the last one. Notably, from the domains that resolved to the same one listed in A1K (34), almost half of them (16) belonged to porn web sites. The rest 18 of them belonged to 11 web pages, including Wikipedia, Instagram and mega.

B. THE NAMECOIN DATA

Namecoin was the first widely used Blockchain DNS, becoming a reference point for more recent approaches such as Emercoin and Blockstack. This blockchain manages the registrar of the .bit TLD through a straightforward procedure, in which users specify the SLD that they wish to register

(which will be later appended with a .bit), as well as the resolving IP and other secondary parameters. At the time of writing this article, Namecoin has a total of 106,659 active domains (i.e. they have been recently created or periodically renewed by their owners). Nevertheless, despite the restrictions imposed by the registrar procedure and the data structure template to be added in the blockchain as well as the deviant behaviour of some users, we found some relevant statistics that showcase the potential of Namecoin as a platform to impulse illicit activities.

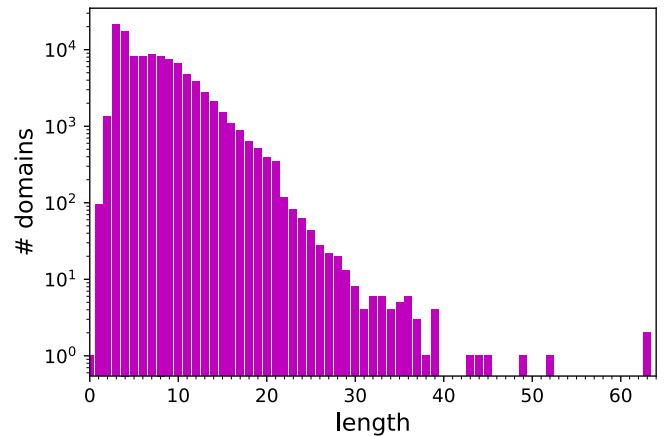


FIGURE 3. Length distribution of domain names registered in Namecoin. Note that values are represented in logarithmic scale.

As a foreseeable tendency, most users opted for registering domains of low length (from the set of domains offered by ICANN, practically all SLDs with length lower than six are already registered or reserved), as described in Figure 3. As already discussed, this hinders procedures such as AGD detection. Clearly, the fact that a domain has to be renewed every certain time at a small cost, a feature which is not implemented in Emercoin, prevents the ownership of domains for long periods if there is no revenue. Nevertheless, this does not seem a constraint for some users, as seen in Table 6. More concretely, more than 87,000 addresses registered at least one domain, yet there are users that own more than 1,000 domains, which often contain the words *sex*, *porn*, *stream*, *hack* as well as other SLDs from well-known brands and companies. Although most of them do not resolve to an IP, this may change in almost real-time with a simple update. Finally, it is worthwhile noting that the intersection of the SLDs of the TOP1m SLDs with the unique SLDs registered in Namecoin is 32,446 and if we count the naming variations (lower/capitals) 32,865, which account for 30,81% of the 106,659 registered domains. Again, using *dnstwist*, we identified 6,299 domains that belong to A1K and whose names have been registered with different typo variations.

C. THE EMERCOIN DATA

Emercoin blockchain is one of the most well-known services for domain registration. In total, the blockchain contains 54,210 records at the time of writing. Interestingly,

³²<https://www.opennic.org/>

TABLE 6. Top 10 addresses in Namecoin with most registered domains.

Namecoin address	# domains
MyZTAGS74akZBiqYPKuvD3zGCfL8tGmXpz	1900
N256bGgH4E84P8fcEcLs4m1YCXZYzB6nzAm	43
NJ6HHqGu9mmW25XgyGoj7V6hPoCSkQLnQ6	40
MwyGuUCawVzCcCSonJpWjN1Kcioq7TNM92	17
MzB1bm2QDmqpmAKearPev4QxAXTWj1kZRI	7
MwAaZiRFGiVcTfVh2bJshN5WXTectocjY	6
NAfxmnNyNoXTxCXtp3R7TZdy1SVqu885ax	4
N2pF7NKSQG73fUkgq9ZSxsjGAHnRH81P7D	4
MyFUY4gCVGYs7TfNxxuGNaf2k6hqrQEky	4
NHtkpFy3yWwYsAwbkEUSr1uwFXX57xded3	3

TABLE 7. Lexical statistics for domain names registered in Emercoin.

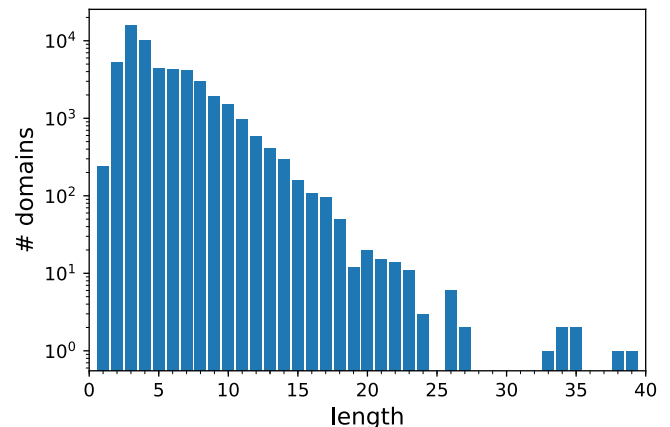
Feature	Registered domains
.com TLD	44
Punycode (xn-)	1261
Capital letters	316
Whitespace character	35

although the naming requirements of Emercoin specify that only lowercase alphanumeric ASCII characters are allowed, the chain contains case sensitive domains not only for the advertised TLDs but for standard TLDs like .com. The distribution of the domains is illustrated in Table 7. In this regard, we observed that most of the addresses registered one or two domains (i.e. more than 43,500 addresses registered at least one domain in Emercoin), while some addresses registered more than 1,000 domains, as showed in Table 8. Many of these records contained an IP, an email address, or a note advertising that the domain is for sale. More concretely, by querying the Emercoin blockchain, we found that up to 617 domains contain the words “for sale” in their *value* field, and in most cases an email to contact. Moreover, when searching for “\$” in the *value* field, the search returned more than 100 domains with a specific sale value. Finally, correlating the A1K dataset with the Emercoin chain returned 1,045 domains, which correspond to 328 unique SLDs registered with different TLD variants. The intersection of the SLDs of the TOP1m SLDs with the unique SLDs registered in Emercoin is 12,214 and if we count the naming variations (lower/capitals and different TLDs) 31,587, which is 58.27% of the 54,210 registered domains. Moreover, using `dnstwist`³³ we identified 9,634 domains that belong to A1K and whose names have been registered with different typo variations (typosquatting).

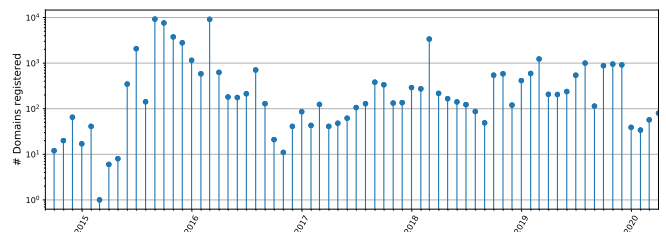
The domain name length distribution is depicted in Figure 4. Notably, most of the domains have lengths below five, with three letters being the most registered domains (as in the case of Namecoin). As previously stated, these SLD are no longer available in ICANN, since they are already registered, and are among the first to be registered once a new TLD appears. Given the high correlation with ICANN domains, it is expected that many of them, if they do not belong to the corresponding ICANN owners, are highly likely to be used for malicious activities such as phishing or cybersquatting.

³³<https://github.com/elceef/dnstwist>**TABLE 8.** Top 10 addresses in Emercoin with most registered domains.

Emercoin address	# domains
ETkxi1X1CeX2QDSWp3CDmuDj7jJZtftfNF	4255
EKzDF4RAHAt8tWdQGbvR9zm7PjrHeth7Rm	3068
EUKa9nrsqX8udF8UpfCGLcYQG8cfT98ZvT	707
EQADxQhroZwGnQAyirFtNbwwojykcifQv3	253
EYBExDLR3aqZunRj6NuyRC9TXt8NHKKXWZ	196
ENnpjY8YQr5rvKNc1TY6kkBwsDZXwmEiY2	150
EWwX61CW9TorzZ7Dy1dmnfKYPxz7dBMGxJ	137
EaQkdxCMPVzMXtTFqYaQxV7wQ1qqLy8aXF	58
ELRNsgvTbV83MyPdD5ACf1xyemLFV7Sued	53
ESCWovPDaX55KcPXC3bdkKWqbH4zBEiwnRd	46

**FIGURE 4.** Length distribution of domain names registered in Emercoin. Note that values are represented in logarithmic scale.

Finally, some statistics of the domain registering behaviour over time are depicted in Figure 5, which shows the domains registered from the beginning of the blockchain up to March 2020. Notably, we can see some peaks in its lifetime.

**FIGURE 5.** Timeline of registered domains in Emercoin. Note that values are represented in logarithmic scale.

The distribution over time of the domains registered with .com was also explored. As seen in Figure 6, such practices, although not alarmingly numerous, are still active in 2020. Therefore, the registrar system still allows anybody to register domains with TLDs different than those offered by Emercoin. This situation can enable several of the threats presented earlier, such as the vulnerabilities with the underlying registrar, which in turn may enable malware and phishing campaigns, as well as cybersquatting.

Finally, global statistics for Namecoin and Emercoin were produced. Currently, there are more than 140K domain names registered in both blockchains, but only 5,266 have an IP address associated with them in their registrar blocks.³⁴

³⁴<https://blockchain-dns.info/explorer/>

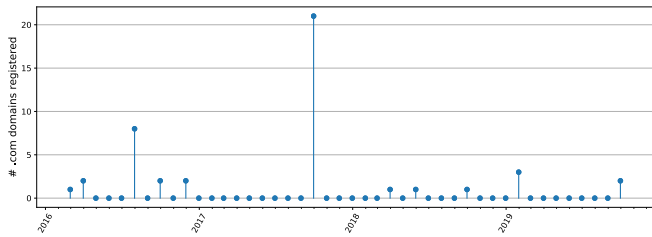


FIGURE 6. Timeline of .com domains registered in Emercoin.

Out of these 5,266, we computed the distribution of TLDs (Table 9). We can observe that most of the domains belong to .coin, .bit, .lib, .bazar, and .emc. Note that some of the other TLDs should not be “available”, considering the whitespace character. Next, we explored the distribution of IP addresses controlling these domains, according to the data contained in the blockchain. In this regard, the top 15 IPs used for that purpose are described in Table 10. We observe that 192.243.100.192 is the IP address to which most domains resolve (i.e. a total of 1957 domains).

TABLE 9. Distribution of TLDs resolving to an IP in both Emercoin and Namecoin.

TLD	Number	TLD	Number	TLD	Number
coin	1261	\$	1	net	1
bit	1045	oz	1	ln	1
lib	1017		1	in	1
bazar	998	bbs	1	9988	1
emc	861	news	1	kib	1
i2p	19	ua	1	fashion	1
neo	14	luxsocks	1	woshiwo321	1
com	8	mayun	1	name	1
onion	3	years	1	www	1
cn	3	pi	1	cion	1
coin	2	aaatttaa	1	mec	1
eth	2	io	1	su	1
enc	1	liib	1	biz	1
org	1	linux	1	1010	1

TABLE 10. Top 15 IPs to which domains resolve in both Emercoin and Namecoin.

IPs	Domains	IPs	Domains
192.243.100.192	1957	78.107.255.15	53
144.76.12.6	448	192.241.241.153	45
202.108.22.5	402	202.108.8.82	45
192.227.233.13	340	81.2.247.158	45
178.128.220.134	144	94.242.60.7	37
185.31.209.8	88	185.61.138.167	32
178.32.148.152	67	46.29.251.130	29
92.63.101.1	53		

In order to go a step further and explore whether the information contained in the blockchain is valid from a domain to IP address resolution perspective, we extracted all the domains and IP addresses from Namecoin and Emercoin and attempted to resolve them. Surprisingly, the results indicated that there were only 273 and 471 unique IPs resolving Namecoin and Emercoin domains, respectively. The latter supports the data illustrated in Table 10, where there are multiple domains hosted by only a few IPs. However, there might be

cases where domain data have not been properly registered or updated in these chains.

V. DISCUSSION AND COUNTERMEASURES

Arguably, the aforementioned threats seem to portray an obscure future. In what follows, we propose a set of mitigation strategies and mechanisms for each of the identified threats.

As identified, Emercoin registrar allows some theoretically forbidden patterns and characters, including the .com TLDs. These practices, although uncommon, are still active, as seen in Figure 6. In the case of Namecoin, the periodic renewal mechanism, as well as the fact of only controlling one TLD, allows a higher degree of control. Yet, both blockchains have similar patterns and user behaviours as analysed in Sections IV-B and IV-C. As such, more robust mechanisms have to be implemented in the future to avoid deviant behaviours. These mechanisms should cover the whole registrar procedure in an end-to-end manner, from the auction systems (e.g. with robust smart contracts and revocation mechanisms, triggered following a condition such as a majority vote) to the proper checking of the data structures stored in the blockchain so that malicious/unexpected information cannot be inserted. Other solutions and functionalities such as forks, which will be later described for the case of the immutability threat, could also be adopted.

In the case of cybersquatting, several strategies have been implemented by systems like Handshake, in which they pre-reserve the top 100k Alexa domains. Other similar policies may be implemented in future decentralised DNS systems as well as a controlled flow of domains being registered, to prevent users from registering arbitrary amounts of domains. Due to the unrestricted nature of Blockchain DNS systems, users may register the most used SLDs and append one of the multiple TLDs offered by the new blockchain DNS registrars. As previously stated, the appearance of blockchain DNS systems which aim to register and resolve all the domain spectrum (both in terms of SLDs and TLDs), may create different versions of the Internet. In this scenario, the challenge of controlling the domain name registration as well as the resolution will require unprecedented security and privacy mechanisms.

The email had always accommodated a noteworthy attack surface due to the lack of security considerations since its inception. The evolution of email security at some point called upon the DNS infrastructure in an attempt to prevent certain types of spam and phishing. Email security policies and protocols such as the Sender Policy Framework (SPF), Domain Keys Identified Mail (DKIM) and Domain Message Authentication Reporting (DMARC) which depend upon DNS can be extended and adapted to force checks on domains and prevent domain spoofing attempts. In addition, the email clients should include scanning and checking functionality to distinguish between the different emerging parallel Internets attributed to different blockchain DNS entries. The email servers (and MTAs in general) could enforce tighter

policies by requiring properly configured DMARC services. In essence, the email ecosystem could act in this instance as the gatekeeper prior to entering the blockchain DNS controlled realm.

The decentralised nature of blockchain DNS is expected to change and improve the botnets' C2 communication channels by providing more effective Rendez-vous algorithms than the current DGAs. Fewer NXDomain responses, covert channels and encrypted communications are expected. Traffic analysis, similar to the one described by [7], is expected to be less effective. This new state of play would require more proactive approaches such as hunting for synthesised IoC type of patterns in the blockchain itself, not only limited in the domain information but also all available metadata. The immutability of the blockchain would allow to continuously and reliably study the botnets' modus operandi and respond with mitigation actions.

The immutability of blockchains requires other approaches to counter malicious records. Although less popular, forks are a well-known mechanism to "delete" data from the blockchain [62]. Nevertheless, forks are used only in exceptional cases and are not considered to be an efficient solution, since they add a prohibitive overhead to the system, especially if the number of deletion requests is high. Other strategies regarding the block consolidation mechanism (the number of blocks created in front of the actual block for it to be considered safe) can also be explored, yet, again, they could hinder the efficiency of the system. In terms of blockchains, technical efforts to circumvent immutability while preserving their inherent security are steadily emerging [62].

Finally, it should be emphasised that for such initiatives to become mainstream and not a tool for cybercrime, they need to build trust in their services. At their current form, it is evident that both Namecoin and Emercoin have already a number of issues as their users face privacy and security challenges. Therefore, moderation solutions must be developed to protect the reputation of the emerging ecosystem. The moderation may prevent poisoning of the chains and removal of malicious records making the users trust the provided services.

VI. CONCLUSION

When a disruptive technology such as blockchain enters the realms of one of the core Internet services such as DNS, it is imperative that the security community invests a significant amount of effort to study and investigate the security implications. The DNS hijacking incident back in 2014 where 300K routers were compromised,³⁵ albeit having a high impact to businesses, is minuscule compared to the potential damage malicious actors can cause when the blockchain DNS becomes widely accepted. This paper attempted to tessellate the emerging threats and provide insight into the associated risks introduced by moving from a centralised to a fully

³⁵https://www.theregister.co.uk/2014/03/04/team_cymru_ids_300000_compromised_soho_gateways/

decentralised DNS. The thorough analysis and evaluation of several open TLD registries such as OpenNIC as well as two well-known blockchain-based DNS systems, namely Emercoin and Namecoin, showcased that the actual solutions are far from being adopted by the users due to several security and reliability issues. Therefore, from a forensic investigation perspective, the use of blockchain is a mixed blessing; on the one hand, some of the evidence will be stored in a forensically sound manner. On the other, the introduction of yet another technology into the Internet backbone will not only increase the complexity leading to a potentially wider attack surface but will also result in significant attribution challenges. Future work will focus on the exploration of other blockchain-based DNS systems and the elaboration of ontologies and security models to overcome the main drawbacks of such systems, with the aim to provide a reliable and sustainable decentralised DNS landscape.

REFERENCES

- [1] F. Casino, T. K. Dasaklis, and C. Patsakis, "A systematic literature review of blockchain-based applications: Current status, classification and open issues," *Telematics Inform.*, vol. 36, pp. 55–81, Mar. 2019.
- [2] P. Hoffman and P. McManus, *DNS Queries Over HTTPS (DoH)*, document RFC 8484, 2018, p. 21. [Online]. Available: <https://tools.ietf.org/html/rfc8484>
- [3] Z. Hu, L. Zhu, J. Heidemann, A. Mankin, D. Wessels, and P. Hoffman, *Specification for DNS Over Transport Layer Security (TLS)*, document RFC 7858, 2016. [Online]. Available: <https://tools.ietf.org/html/rfc7858>
- [4] S. Dickinson, D. Gillmor, and T. Reddy, *Usage Profiles for DNS over TLS and DNS over DTLS*, document RFC 7841, Internet Engineering Task Force, 2018, vol. 10.
- [5] E. Karaarslan and E. Adiguzel, "Blockchain based DNS and PKI solutions," *IEEE Commun. Standards Mag.*, vol. 2, no. 3, pp. 52–57, Sep. 2018.
- [6] C. Patsakis and F. Casino, "Hydras and IPFS: A decentralised playground for malware," *Int. J. Inf. Secur.*, vol. 18, no. 6, pp. 787–799, Dec. 2019, doi: [10.1007/s10207-019-00443-0](https://doi.org/10.1007/s10207-019-00443-0).
- [7] C. Patsakis, F. Casino, and V. Katos, "Encrypted and covert DNS queries for botnets: Challenges and countermeasures," *Comput. Secur.*, vol. 88, Jan. 2020, Art. no. 101614.
- [8] H. A. Kalodner, M. Carlsten, P. Ellenbogen, J. Bonneau, and A. Narayanan, "An empirical study of namecoin and lessons for decentralized namespace design," in *Proc. WEIS*, 2011.
- [9] Z. Liu, E. S.-J. Swildens, and R. D. Day, "Domain name resolution using a distributed DNS network," U.S. Patent 7725 602 B2, May 25, 2010.
- [10] C. Cachin and A. Samar, "Secure distributed DNS," in *Proc. Int. Conf. Dependable Syst. Netw.*, 2004, pp. 423–432.
- [11] V. Ramasubramanian and E. G. Sirer, "The design and implementation of a next generation name service for the Internet," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*. New York, NY, USA: Association for Computing Machinery, 2004, pp. 331–342.
- [12] M. Wachs, M. Schanzenbach, and C. Grothoff, "A censorship-resistant, privacy-enhancing and fully decentralized name system," in *Proc. Int. Conf. Cryptol. Netw. Secur.* Cham, Switzerland: Springer, 2014, pp. 127–142.
- [13] Z. Qiang, Z. Zheng, and Y. Shu, "P2PDNS: A free domain name system based on P2P philosophy," in *Proc. Can. Conf. Electr. Comput. Eng.*, 2006, pp. 1817–1820.
- [14] M. Abu-Amara, F. Azzedin, F. A. Abdulhameed, A. Mahmoud, and M. H. Sqalli, "Dynamic peer-to-peer (P2P) solution to counter malicious higher domain name system (DNS) nameservers," in *Proc. 24th Can. Conf. Electr. Comput. Engineering (CCECE)*, May 2011, pp. 001014–001018.
- [15] D. Storm. (2010). *P2P DNS to Take on ICANN After us Domain Seizures*. [Online]. Available: <https://www.computerworld.com/article/2469753/p2p-dns-to-take-on-icann-after-us-domain-seizures.html>
- [16] R. Sancho and R. Lopes Pereira, "Hybrid peer-to-peer DNS," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2014, pp. 977–981.

- [17] A. Hari and T. V. Lakshman, "The Internet blockchain: A distributed, tamper-resistant transaction framework for the Internet," in *Proc. 15th ACM Workshop Hot Topics Netw. (HotNets)*, 2016, pp. 204–210.
- [18] B. Benshoof, A. Rosen, A. G. Bourgeois, and R. W. Harrison, "Distributed decentralized domain name service," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, May 2016, pp. 1279–1287.
- [19] J. Liu, B. Li, L. Chen, M. Hou, F. Xiang, and P. Wang, "A data storage method based on blockchain for decentralization DNS," in *Proc. IEEE 3rd Int. Conf. Data Sci. Cyberspace (DSC)*, Jun. 2018, pp. 189–196.
- [20] S. Gourley and H. Tewari, "Blockchain backed DNSSEC," in *Proc. Int. Conf. Bus. Inf. Syst.*, in Lecture Notes in Business Information Processing, vol. 339, 2019, pp. 173–184.
- [21] Z. Guan, A. Garba, A. Li, Z. Chen, and N. Kaaniche, "AuthLedger: A novel blockchain-based domain name authentication scheme," in *Proc. 5th Int. Conf. Inf. Syst. Secur. Privacy*, 2019, pp. 345–352.
- [22] W. Wang, N. Hu, and X. Liu, "Blockzone: A blockchain-based DNS storage and retrieval scheme," in *Artificial Intelligence and Security*. Cham, Switzerland: Springer, 2019, pp. 155–166.
- [23] X. Duan, Z. Yan, G. Geng, and B. Yan, "DNSLedger: Decentralized and distributed name resolution for ubiquitous IoT," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2018, pp. 1–3.
- [24] W. Yoon, I. Choi, and D. Kim, "BlockONS: Blockchain based object name service," in *Proc. IEEE Int. Conf. Blockchain Cryptocurrency (ICBC)*, May 2019, pp. 219–226.
- [25] X. Wang, K. Li, H. Li, Y. Li, and Z. Liang, "ConsortiumDNS: A distributed domain name service based on consortium chain," in *Proc. IEEE 19th Int. Conf. High Perform. Comput. Commun., IEEE 15th Int. Conf. Smart City, IEEE 3rd Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Dec. 2017, pp. 617–620.
- [26] H. Li, H. Ma, L. Haopeng, Z. Huang, X. Yang, K. Li, and H. Wang, "Blockchain-based domain name resolution system," U.S. Patent App. 15/768 833, May 30, 2019.
- [27] H. Li, X. Wang, Z. Lin, J. Wu, X. Si, K. Li, X. Yang, and H. Wang, "Systems and methods for managing top-level domain names using consortium blockchain," U.S. Patent App. 10/178 069, Oct. 4, 2019.
- [28] M. Ali, J. Nelson, R. Shea, and M. J. Freedman, "Blockstack: A global naming and storage system secured by blockchains," in *Proc. USENIX Annu. Tech. Conf. (USENIX ATC)*, Denver, CO, USA: USENIX Association, Jun. 2016, pp. 181–194.
- [29] Emercoin. (2019). *Emercoin Links & Resources*. [Online]. Available: <https://emercoin.com/en/documentation/links-resources>
- [30] D. B. Gilwit, "The latest cybersquatting trend: Typosquatters, their changing tactics, and how to prevent public deception and trademark infringement," *Wash. UJL Pol'y*, vol. 11, p. 267, Jan. 2003.
- [31] B. Edelman, "Large-scale registration of domains with typographical errors," in *Domain Name Typosquatter Still Generating Millions*. Cambridge, MA, USA: Harvard Univ., 2003.
- [32] B. Liu, C. Lu, Z. Li, Y. Liu, H. Duan, S. Hao, and Z. Zhang, "A reexamination of internationalized domain names: The good, the bad and the ugly," in *Proc. 48th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2018, pp. 654–665.
- [33] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, "Seven Months' worth of mistakes: A longitudinal study of typosquatting abuse," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2015, pp. 1–13.
- [34] J. Spaulding, S. Upadhyaya, and A. Mohaisen, "The landscape of domain name typosquatting: Techniques and countermeasures," in *Proc. 11th Int. Conf. Availability, Rel. Secur. (ARES)*, Aug. 2016, pp. 284–289.
- [35] M. T. Khan, X. Huo, Z. Li, and C. Kanich, "Every second counts: Quantifying the negative externalities of cybercrime via typosquatting," in *Proc. IEEE Symp. Secur. Privacy*, May 2015, pp. 135–150.
- [36] A. Dinaburg, "Bitsquatting: DNS hijacking without exploitation," *Proc. BlackHat Secur.*, Jul. 2011.
- [37] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen, "Bitsquatting: Exploiting bit-flips for fun, or profit?" in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, 2013, pp. 989–998.
- [38] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis, "Hiding in plain sight: A longitudinal study of combosquatting abuse," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 569–586.
- [39] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the use of homophones in domain squatting," in *Proc. Int. Conf. Inf. Secur.* Cham, Switzerland: Springer, 2014, pp. 291–308.
- [40] P. Lv, J. Ya, T. Liu, J. Shi, B. Fang, and Z. Gu, "You have more abbreviations than you know: A study of abbreviating abuse," in *Proc. Int. Conf. Comput. Sci. Cham, Switzerland: Springer*, 2018, pp. 221–233.
- [41] V. Le Pochat, T. Van Goethem, and W. Joosen, "Funny accents: Exploring genuine interest in internationalized domain names," in *Proc. Int. Conf. Passive Act. Netw. Meas.* Cham, Switzerland: Springer, 2019, pp. 178–194.
- [42] D. Chiba, A. A. Hasegawa, T. Koide, Y. Sawabe, S. Goto, and M. Akiyama, "DomainScouter: Understanding the risks of deceptive IDNs," in *Proc. 22nd Int. Symp. Res. Attacks, Intrusions Defenses (RAID)*, 2019, pp. 413–426.
- [43] F. Quinkert, T. Lauinger, W. Robertson, E. Kirda, and T. Holz, "It's not what it looks like: Measuring attacks and defensive registrations of homograph domains," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Jun. 2019, pp. 259–267.
- [44] K. Du, H. Yang, Z. Li, H. Duan, S. Hao, B. Liu, Y. Ye, M. Liu, X. Su, G. Liu, Z. Geng, Z. Zhang, and J. Liang, "TL;DR hazard: A comprehensive study of levelsquatting scams," in *Security and Privacy in Communication Networks*, S. Chen, K.-K. R. Choo, X. Fu, W. Lou, and A. Mohaisen, Eds. Cham, Switzerland: Springer, 2019, pp. 3–25.
- [45] N. Miramirkhani, T. Barron, M. Ferdman, and N. Nikiforakis, "Panning for gold.Com: Understanding the dynamics of domain dropcatching," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 257–266.
- [46] M. Joslin, N. Li, S. Hao, M. Xue, and H. Zhu, "Measuring and analyzing search engine poisoning of linguistic collisions," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 1311–1325.
- [47] Y. Zeng, T. Zang, Y. Zhang, X. Chen, and Y. Wang, "A comprehensive measurement study of domain-squatting abuse," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.
- [48] S. T. Ali, P. McCorry, P. H.-J. Lee, and F. Hao, "ZombieCoin 2.0: Managing next-generation botnets using bitcoin," *Int. J. Inf. Secur.*, vol. 17, no. 4, pp. 411–422, Aug. 2018.
- [49] S. Pletinckx, C. Trap, and C. Doerr, "Malware coordination using the blockchain: An analysis of the cerber ransomware," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, May 2018, pp. 1–9.
- [50] R. Amado. (2018). *How Cybercriminals are Using Blockchain DNS: From the Market to the Bazar*. [Online]. Available: <https://www.digitalshadows.com/blog-and-research/how-cybercriminals-are-using-blockchain-dns-from-the-market-to-the-bazar/>
- [51] J. Sanders. (2019). *Blockchain-Based Unstoppable Domains is a Rehash of a Failed Idea*. [Online]. Available: <https://www.techrepublic.com/article/blockchain-based-unstoppable-domains-is-a-rehash-of-a-failed-idea/>
- [52] F. ul Hassan, A. Ali, S. Latif, J. Qadir, S. Kanhere, J. Singh, and J. Crowcroft, "Blockchain and the future of the Internet: A comprehensive review," 2019, *arXiv:1904.00733*. [Online]. Available: <http://arxiv.org/abs/1904.00733>
- [53] R. Rasmussen and P. Vixie, "Surveying the dns threat landscape," Internet Identity, Tech. Rep., 2013.
- [54] Proofpoint. (2019). *2019 State of the Phish Report*. [Online]. Available: <https://www.proofpoint.com/us/resources/threat-reports/state-of-phish>
- [55] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2091–2121, 4th Quart., 2013.
- [56] Phishme. (2016). *Enterprise Phishing Susceptibility and Resiliency Report*. [Online]. Available: <https://www.infosecurityeurope.com/novadocuments/351537?v=636276130024130000>
- [57] C. Iuga, J. R. C. Nurse, and A. Erola, "Baiting the hook: Factors impacting susceptibility to phishing attacks," *Human-centric Comput. Inf. Sci.*, vol. 6, no. 1, p. 8, Dec. 2016.
- [58] P. G. Lopez, A. Montresor, and A. Datta, "Please, do not decentralize the Internet with (permissionless) blockchains!" 2019, *arXiv:1904.13093*. [Online]. Available: <https://arxiv.org/abs/1904.13093>
- [59] BBC. (2019). *Child Abuse Images Hidden in Crypto-Currency Blockchain*. [Online]. Available: https://www.bbc.com/news/technology-47130268?ocid=socialflow_twitter
- [60] R. Matzutt, J. Hiller, M. Henze, J. H. Ziegeldorf, D. Müllmann, O. Hohlfeld, and K. Wehrle, "A quantitative analysis of the impact of arbitrary blockchain content on bitcoin," in *Proc. 22nd Int. Conf. Financial Cryptogr. Data Secur. (FC)*. Cham, Switzerland: Springer, 2018, pp. 420–438.

- [61] E. Politou, E. Alepis, and C. Patsakis, "Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions," *J. Cybersecurity*, vol. 4, no. 1, Jan. 2018, Art. no. tyy001.
- [62] E. Politou, F. Casino, E. Alepis, and C. Patsakis, "Blockchain mutability: Challenges and proposed solutions," *IEEE Trans. Emerg. Topics Comput.*, early access, Oct. 25, 2019, doi: [10.1109/TETC.2019.2949510](https://doi.org/10.1109/TETC.2019.2949510).



CONSTANTINOS PATSAKIS (Member, IEEE) received the B.Sc. degree in mathematics from the University of Athens, Greece, the M.Sc. degree in information security from the Royal Holloway, University of London, and the Ph.D. degree in cryptography and malware from the Department of Informatics, University of Piraeus.

He was a Researcher with the UNESCO Chair in Data Privacy, Rovira i Virgili University (URV), Tarragona, Spain, and a Research Fellow with Trinity College, Dublin, Ireland. He is currently an Assistant Professor with the University of Piraeus and an Adjunct Researcher with the Athena Research and Innovation Center. He has authored numerous publications in peer-reviewed international conferences and journals. He has been teaching computer science courses in European universities for more than a decade. He has been working in the industry as a Freelance Developer and a Security Consultant. He has participated in several national (Greek, Spanish, Catalan, and Irish) and European Research and Development projects. His main areas of research include cryptography, security, privacy, data anonymization, and data mining.



FRAN CASINO (Member, IEEE) received the B.Sc. degree in computer science and the M.Sc. degree in computer security and intelligent systems from Rovira i Virgili University, Tarragona, Spain, in 2010 and 2013, respectively, and the Ph.D. degree (*cum laude*) in computer science from Rovira i Virgili University, in 2017, and the best dissertation award. He was a Visiting Researcher with ISCTE-IUL, Lisbon, in 2016. He is currently a Postdoctoral Researcher with the

Department of Informatics, Piraeus University, Piraeus, Greece. He has participated in several European-, Spanish- and Catalan-funded research projects. He has authored more than 40 publications in peer-reviewed international conferences and journals. His research interests include pattern recognition, and data management applied to different fields, such as privacy and security protection, recommender systems, smart health, and blockchain.



NIKOLAOS LYKOUSAS received the B.S. degree from the Department of Informatics, University of Piraeus, in 2016, and the master's degree in intelligent interactive systems from Universitat Pompeu Fabra, in 2017, where he was awarded with an academic excellence scholarship for his achievements. He is currently pursuing the Ph.D. degree with the University of Piraeus studying deviant behavior in modern social networks. Since his undergraduate studies, he has participated as a Research Engineer in several EC funded projects and has gained considerable experience in the fields of big data analytics, cybersecurity, digital privacy, and cloud computing.



VASILIOS KATOS (Member, IEEE) is currently a Professor with Bournemouth University. Prior to his current post, he was a Professor of information and communications systems security at the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, and a Principal Lecturer at the School of Computing, University of Portsmouth, where he has participated in the development of the interdisciplinary Masters course M.Sc. in Forensic IT. He has worked in the industry as a Security Consultant as an expert witness in information systems security and is NIS Expert for ENISA. He has over 100 publications in reputable scientific journals and international conferences in the area of digital forensics and incident response. He has served as a reviewer on a number of venues, e.g., IEEE COMMUNICATION LETTERS. He is an Editorial board member of *Computers & Security*.

...



Unearthing malicious campaigns and actors from the blockchain DNS ecosystem



Fran Casino^{a,b}, Nikolaos Lykousas^a, Vasilios Katos^c, Constantinos Patsakis^{a,b,*}

^a Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece

^b Information Management Systems Institute, Athena Research Center, Artemidos 6, Marousi 15125, Greece

^c Bournemouth University, Poole House P323, Talbot Campus, Fern Barrow, Poole, Dorset, BH12 5BB, UK

ARTICLE INFO

Keywords:

Blockchain
Blockchain forensics
Cybercrime
DNS
Malware
Decentralised DNS

ABSTRACT

Blockchain DNS has emerged as an alternative solution to traditional DNS to address many of its inherent drawbacks. In this regard, a blockchain DNS approach is decentralised, resilient, provides high availability, and prevents censorship. Unfortunately, despite these desirable features, the major blockchain DNS solutions to date, Namecoin and Emercoin have been repeatedly reported for malicious abuse, ranging from malware distribution to phishing. In this work, we perform a longitudinal analysis of both these chains trying to identify and quantify the penetration of malicious actors in their ecosystems. To this end, we apply a haircut blacklisting policy and the intelligence collected from various engines to perform a taint analysis on the metadata existing in these blockchains, aiming to identify malicious acts through the merge of identifying information. Our analysis provides an automated validation methodology that supports the various reports about the wide-scale abuse of these solutions showing that malicious actors have already obtained an alarming and extensive share of these platforms.

1. Introduction

With the continuous digitisation of procedures, services, and products, crime has been shifting towards the same direction. Despite the continuous evolution of artificial intelligence techniques such as machine learning, pattern recognition and natural language processing, which are capable of ingesting terabytes of unstructured data to enhance response times, and expand the capacities of security operations, attackers tend to be always a step ahead. The latter is directly related to the appearance of novel technologies, industrialisation processes, the difficulty to collect data from diverse sources in orchestrated campaigns and their timely detection, and the lack of proactive security mechanisms. As a result, cybercrime is predicted to be the third-largest economy in 2021 [1].

Meanwhile, there have been systematic efforts to address the security and privacy issues of the Domain Name System (DNS). The DNS is one of the oldest yet critical Internet application-level protocols. In this regard, recommendations and approaches for security improvements such as DNSSEC, DNSCurve, and DNS over TLS/HTTPS are hindered by the lack of adoption [2], which leave DNS exposed to several threats, including man-in-the-middle attacks, passive eavesdropping and data injection. Moreover, the hierarchical design of DNS makes it prone to

particular types of attacks such as poisoning, as well as amplification type of denial of service attacks [3]. For instance, due to the lack of authentication in the traditional DNS protocol, a DNS server cannot authenticate whether a response originates from a valid DNS resolver, which is ranked higher in the DNS hierarchy. Therefore, an attacker may query a DNS server for a known website XYZ and then send a spoofed response which falsely claims that the IP of XYZ is an attacker controlled host. However, for efficiency, DNS servers store the responses from DNS resolvers in their cache. Thus, the spoofed response will be cached in the DNS server. As a result, all users who later ask for the IP of XYZ will be redirected to host controlled by the attacker. Furthermore, freedom of speech is hard to accomplish given the actual design of DNS, since, e.g. authoritative regimes can manipulate them to block traffic and censor everything that may question them.

Recently, with the exploitation of decentralised, immutable data structures such as blockchain, several industries have found a way to promote their services and enhance their features, including security, privacy, traceability, and verifiability [4,5]. Nevertheless, the inherent immutability of such systems paired with design flaws prevent illegal and undesired content from being modified or taken down [6,7]. In this context, novel decentralised applications such as decentralised DNS

* Corresponding author at: Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece.

E-mail addresses: francasino@unipi.gr (F. Casino), nlykousas@unipi.gr (N. Lykousas), vkatos@bournemouth.ac.uk (V. Katos), kpatsak@unipi.gr (C. Patsakis).

<https://doi.org/10.1016/j.comcom.2021.08.023>

Received 9 March 2021; Received in revised form 26 June 2021; Accepted 24 August 2021

Available online 3 September 2021

0140-3664/© 2021 Elsevier B.V. All rights reserved.

systems are not an exception [8,9]. Therefore, despite the potential of BDNS systems to disrupt traditional DNS models, their inherent design flaws can be used to leverage resilient malware campaigns.

Motivation and main contributions: The threat landscape has changed considerably since the introduction of DNS, urging the community to seek alternatives for this service. These alternatives are served in two main flavours: (1) Security improvements of the existing DNS using approaches like DNS over HTTPS [10] and DNS over TLS [11], and (2) Decentralisation of DNS, with blockchain as the enabling technology. In the latter case, several approaches are already functional, with Namecoin and Emercoin being the most mature and used ones. In addition, other approaches seem to perpetuate the blockchain DNS trend, such as Handshake,¹ while some registered patents by, e.g. Alibaba² and with upcoming projects (e.g., both EXIP [12] and Butterfly [13] projects launched and ICO in 2021), which aim to extend the foundational properties of BDNS, highlight the importance of a proper design of such systems. In addition, novel browsers like Brave [14] are rapidly gaining attention due to their privacy properties, as well as other potential benefits for the users. Brave already adopts several similar mechanisms like Unstoppable domains and the Ethereum name service (ENS).

Despite the research leveraged by the community towards more secure and resilient DNS systems, adversaries are expected to opportunistically take advantage of such changes by exploiting both the technology in its early stages, as well as the lack of knowledge and experience of the end-users and system administrators. For instance, well-known malicious campaigns are still exploiting such systems. For example, BazarLoader struck again in April 2021, showcasing new specific attack patterns similar to these of Trickbot [15], as also claimed in the past [16]. It is therefore imperative to raise awareness on the emerging security threats

This work extends the initial findings of [8,9] and provides a automated and comprehensive approach towards discovering illegal activities related to blockchain DNS services to the one described in [17]. In the latter, the authors captured malicious traffic originating from blockchain DNS resolved sources and conducted a binary classification approach between benign traffic (traditional) and malicious blockchain DNS traffic. Nevertheless, the size of their dataset and the fact of differentiating between disparate types of traffic (i.e. traditional and blockchain-based) requires further research to provide more extensive and statistically sound outcomes.

In this work, we analyse the corpus of domains registered in Namecoin and Emercoin and their registered IPs. Moreover, we provide evidence of the connection between a subset of such domains and illegal activities, as reported and corroborated by several individual sources. To this end, we adapt the blacklisting poison and haircut policy of Möser et al. [18] to a blockchain DNS context. This approach enables an investigator to identify strong connections among IPs and wallets that are validated by existing attack patterns, e.g. BazarLoader [19]. Moreover, we identify traces of active attacks and campaigns and several correlations on the metadata used in both chains, namely wallets, IPs, domains and emails. In addition, by analysing the malicious IPs used by several subsets of wallets and domains, we identify potentially malicious IPs that have not been reported yet. For each investigation phase we provide a detailed description of the procedures, and a comprehensive representation of the outcomes, which prove that the existing blockchain DNS systems are far from delivering the evangelised features. To the best of our knowledge, this is the first piece of research that provides detailed and documented proof of the malicious activities carried out in both Namecoin and Emercoin by automating the analysis of internal blockchain data as well as correlated data

¹ <https://www.coindesk.com/handshake-goes-live-with-an-uncensorable-internet-browser>.

² <https://domainnamewire.com/2019/08/15/alibaba-files-blockchain-domain-name-patent-application/>.

Table 1
Main characteristics of blockchain DNSs.

Property	Description
Availability	The availability of the system depends on multiple peers and not on a single entity.
Automated management	Auctions to register domain names, fast and transparent ownership control
Censorship-resistance	Domain name resolution services and information are not subject to borders or bans
Decentralisation	The network is completely distributed with no central entities
Namespace freedom	Registration of new SLDs and TLDs
Robustness	Resilient to attacks that affect centralised DNS systems such as MiM, spoofing, cache poisoning, cracking.
Trust	Through verifiable and robust consensus mechanisms
Unlimited resources	A high number of simultaneous users sharing their assets.

from external intelligence. Moreover, we provide several automated mechanisms to leverage proactive measures and detect cybercriminal campaigns orchestrated in the core of blockchain DNS systems. Finally, our methodology illustrates how blockchain forensics can be performed beyond the cryptocurrency ecosystem, where the actual evidence are not limited to the data existing in the chain itself.

The rest of the article is organised as follows. In Section 2, we provide a general background on blockchain DNS and explore the related work. In Section 3, we describe the methodology adopted in terms of data collection and analysis, and in Section 4 we provide a thorough analysis of the registered domains in Namecoin and Emercoin, as well as the identification of the illegal activities leveraged by such domains. Finally, in Section 5, we discuss the findings of our experiments and conclude the article by providing some threads for future research.

2. Related work

As studied in the current literature [9,20–27], the main features that decentralised systems can potentially provide are availability, robustness, censorship resistance, as well as other managerial improvements. Table 1 summarises the main characteristics and features of blockchain DNS systems according to the literature.

The early strategies adopted to create decentralised DNS systems focused on the development of specific TLDs such as in the case of the Dot-P2P project (with the .p2p TLD) [28]. However, the inherent performance bottlenecks contributed to adoption delays and diminished the functionality of such systems. Only recently, and due to the progressive adoption of blockchain-based distributed DNS systems [29], the idea of functional and real-world distributed DNS systems is showing clear signs of a comeback.

There exists a set of functional approaches to blockchain-based DNS according to the scientific literature. Hari et al. [30] provided a thorough discussion about the limitations of traditional practices and the benefits of using blockchain for the development of a DNS infrastructure. In [31], Benschhof et al. proposed D³NS, which integrates a distributed hash table and domain name ownership implementation based on the Bitcoin blockchain. One of their aims is to replace the top-level DNS and certificate authorities, offering increased scalability, security and robustness. Gourley and Tewari [32] proposed the use of blockchain to improve the main drawbacks of DNSSEC in the certificate validation procedure, creating an enhanced DNS security extension. With a similar aim, Guan et al. [33] presented AuthLedger, blockchain-based system that provides efficient and secure domain name authentication. Liu et al. [34] proposed a blockchain-based decentralisation DNS resolution method with distributed data storage to mitigate single points of failure and domain name resolution tampering. BlockZone, proposed by Wang et al. [35], uses a replicated network of nodes to offer efficient name resolution supported by improved Practical Byzantine Fault Tolerance (PBFT) consensus mechanism. Yu et al. [36] proposed the use of a consortium blockchain to establish

a DNS cache resources trusted sharing model, which improves the credibility of DNS resolution results by establishing a complete chain of trust.

In the IoT communications domain, some authors have developed specific blockchain-based solutions to enhance domain name resolution and management. For instance, Duan et al. [37] presented DNSLedger, a decentralised, hierarchical multi-chain structure to provide domain name resolution services. BlockONS, proposed by Yoon et al. [38], described a robust and scalable object name service appropriate for an IoT ecosystem with the aim to overcome classical problems related to DNS resolution, namely DNS cache poisoning, spoofing, and local DNS cracking. ConsortiumDNS, presented by Wang et al. [39] is a three-layer architecture composed by a consortium blockchain, a consensus mechanism and external storage. The authors claim that their approach is more efficient compared to other well-known approaches such as Namecoin or Blockstack. Finally, a set of patented designs of Blockchain-based DNS systems can be found in [40,41].

The first system to reach a certain level of maturity was Namecoin,³ which is a cryptocurrency based on Bitcoin, with additional features such as decentralised name system management, mainly for the .bit domain. Moreover, it was the first project to provide security, decentralisation and human-meaningfulness, as required to address Zooko's triangle [27]. Nevertheless, due to the lack of support and adoption, Namecoin's main drawback is its insufficient computing power, which makes it more vulnerable to the 51% attack than other similar systems. Blockstack [42] is a blockchain-based naming and storage system that separates control and data planes, enabling seamless integration with the underlying blockchain. EmerDNS,⁴ more commonly known as Emercoin, is a blockchain DNS system which supports a wide range of DNS records. EmerDNS operates under the "DNS" service abbreviation in the Emercoin NVS. Handshake⁵ is one of the most widely supported technologies, which aims to offer an alternative to existing certificate authorities. Therefore, Handshake aims to replace the root zone file and the DNS name resolution and registration services worldwide.

In addition to the above systems, there are two approaches that are based on the Ethereum blockchain, the Ethereum name service⁶ (ENS), and Nebulis.⁷ The former uses smart contracts to manage the .eth registrar through bids. Moreover, ENS recently added the support for .onion addresses. The latter is a globally distributed directory that relies on the Ethereum ecosystem and smart contracts to store, update and resolve domain records. Moreover, Nebulis uses decentralised storage technologies such as IPFS as a replacement for HTTP. Table 2 summarises the main features of the discussed DNS approaches.

Finally, OpenNIC⁸ is a unique case, since it is a hybrid approach in which a group of peers manages namespace registration, yet the name resolving task is fully decentralised. OpenNIC provides DNS namespace and resolution for an extensive set of domains, including those managed by EmerDNS, and New Nations.⁹ In addition, OpenNIC resolvers have recently added access to domains administered by ICANN. Notably, OpenNIC has dropped the support for .bit domains due to malware abuse.¹⁰ As stated in the corresponding voting:

“Over the past year .bit domains have started being used as malware hubs due to their anonymous nature. Since there is no way to contact the owner of those domains, it creates a backscatter effect, and a number of people running public T2 servers have seen domains blacklisted, emails blocked, and shutdown notices from their providers.”

³ <https://www.namecoin.org/>.

⁴ <https://emercoin.com/en/documentation/blockchain-services/emerdns/emerdns-introduction>.

⁵ <https://handshake.org/>.

⁶ <https://ens.domains/>.

⁷ <https://www.nebulis.io/>.

⁸ <https://www.opennic.org/>.

⁹ <http://www.new-nations.net/>.

¹⁰ https://wiki.opennic.org/votings/drop_namecoin.

Currently, several malicious campaigns are exploiting the features of the blockchain DNS ecosystem. Setting aside the massive cybersquatting attacks [9] and hosting of malicious marketplaces, e.g. Joker's Stash [43,44], the blockchain DNS approach has been exploited by many malware families as it provides *bulletproof hosting* [45]. The latter cannot be considered a recent development as reports about the abuse of .bit domains date back to 2013 [46]. From that point onward a number of regular reports emerged on specific malware families exploiting the blockchain DNS ecosystem. For instance, Fbot botnet used domains resolved by Emercoin to communicate with its *command and control* (C2) servers [47] and the same approach was used by Cerber [48]. In general, as reported by FireEye [49], blockchain DNS domain have been used for hosting C2 servers of many malware families, including but not limited to Necurs, AZORult, Emotet [50], Terdot, Gandcrab [51], SmokeLoader [52], and very recently Trickbot [19].

Table 2 summarises the main features of the most relevant Blockchain-DNS systems.

Internet users can reach the TLDs offered by Namecoin, OpenNIC, New Nations, and EmerDNS (e.g. .coin, .emc, .lib and .bazar) through various browser extensions such as peername, blockchain-DNS and friGate [53]. The domain name resolution procedure is outlined in Fig. 1.

Finally, despite the theoretical and desired features previously described, blockchain DNS systems have several drawbacks, which can be exploited by malicious actors [9,54,55]. Patsakis et al. [9] explored the main blockchain DNS systems and identified a set of challenges and threats related to their underlying registrar mechanisms, malware and phishing campaigns, and the immutability of data residing in such systems. Similarly, Xia et al. [56] performed a qualitative analysis of the Ethereum Name Service and discussed their challenges. Recently, Huang et al. [17] explored the traffic generated by sites resolved by blockchain DNS systems and analysed its patterns. Despite the fact that their dataset contains few benign samples, their outcomes showed that they could differentiate between traditional domains and blockchain DNS domains that were known to leverage malicious activities, according to VirusTotal.

Following an analysis of the literature, the main drawbacks identified by researchers to detect malicious activities in blockchain DNS systems are (i) the lack of automated tools to pair the activities performed in the blockchain with external intelligence tools, (ii) the difficulty to extract interoperable metrics (e.g., behavioural indicators) to identify malicious behaviours, and (iii) the unstructured nature of data, which prevents the application of policies extendable to other frameworks. To the best of our knowledge, our work is the first to propose a fully automated pipeline leveraging a structured data analysis and feature collection, which is used to correlate blockchain data with external intelligence sources and apply proactive policies to effectively detect malicious behaviours as well as cybercrime campaigns.

3. Methodology

As already discussed, in a blockchain DNS system, one registers a domain by paying through the corresponding cryptocurrency, e.g. Namecoin, Emercoin, etc. Setting aside the monetary transactions which may hinder money-laundering acts, the maliciousness stems from the content that such a domain has. Currently, we are well aware that blockchain DNS systems have been exploited by malicious actors for several malware campaigns or black marketplaces, as discussed in Section 2. One may ponder about the extent of this exploitation, as it is infeasible to collect all the content, and even if it were possible, it would be impossible to collect the content that existed and was flagged malicious.

To alleviate this challenge and create a ground truth, we base our analysis on the domains and IPs that are registered in these blockchains. To this end, we initially perform a dump of these blockchains to collect all the domain names and the IPs that have been used by them.

Table 2
Main characteristics of the most relevant DNS systems. Although Blockstack is blockchain agnostic, it is mainly used with Bitcoin blockchain.

Method	Pedigree platform	Registrar & resolution management	TLD examples
ICANN	Network of servers and resolvers	Centralised	.com .net .org
Namecoin	Bitcoin and Peercoin	Decentralised	.bit
Emercoin	Bitcoin	Decentralised	.coin .bazar .emc
ENS	Ethereum	Decentralised	.eth .onion
Handshake	Bitcoin	Decentralised	Unrestricted
Blockstack	Blockchain agnostic	Decentralised	.id .podcast .helloworld
OpenNIC	Decentralised servers	Hybrid	.bbs .pirate .libre

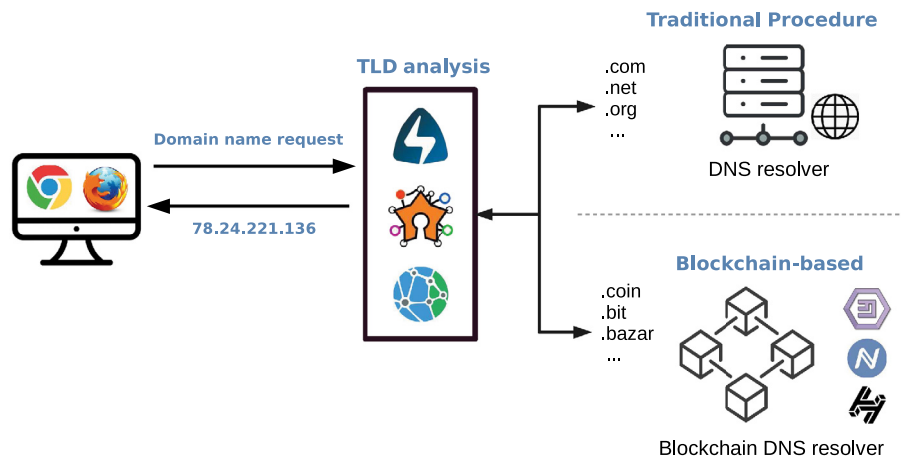


Fig. 1. Workflow of the domain name resolution procedure. The extension analyses the TLD of the requested domain and directs the query to the corresponding DNS system.

Contrary to traditional DNS systems, in blockchain-based DNS all the history of a domain, including the IPs that were used to provide the content is recorded and publicly accessible.

Additionally, we aim to establish a baseline approach to perform blacklisting and use it to measure the number of malicious wallets and domains. A straightforward process is to use an intelligence engine to query these domains. However, taking into consideration only the domains is not very effective as most of these domains are not indexed, and only a few intelligence engines collect data about them. Moreover, it is highly possible that the logs that they have may not refer to the domains per se, but their IP addresses. This can be attributed to the fact that the DNS query is performed to a non-standard TLD and the engine drops it. Nonetheless, the connection to the IP is recorded. Therefore, one has to consider whether the IP has been used for other malicious activities, e.g. spamming, phishing etc.

We argue that the blacklisting policies of Möser et al. [18] that were applied in Bitcoin to trace money laundering can be adopted in the blockchain DNS chains to identify malicious activity. To this end, we adapt the poison and haircut policies as follows. Let us assume that wallet W_1 has registered a domain D_1 which is mapped to IP_1 . If IP_1 is flagged as malicious, then the wallet is flagged as malicious. Similarly, if wallet W_2 has registered a domain D_2 which is also mapped to IP_1 , then wallet W_2 is also flagged as malicious. In essence, a malicious IP “poisons” all the wallets that are attached to it. Nonetheless, once we have a malicious IP in a wallet, it taints the rest of the IPs of the wallet. Using the haircut policy of Möser et al. we consider the rest of the IPs as *suspicious*. Therefore, poisoning is applied to domains and wallets, while haircut is applied to IPs. The two policies are illustrated in Fig. 2.

Based on the above, we first need to look for the domains and then extract intelligence about the IPs that are used. Using the above, we attempt to identify any emerging patterns and whether the tainting approach provides any insight regarding upcoming threats.

4. Experimental setup

To investigate malicious activities related to the use of blockchain DNS platforms, we analysed the contents of both Namecoin¹¹ and the Emercoin¹² blockchains. Namecoin was the first widely used Blockchain DNS, becoming a reference point for more recent approaches such as Emercoin and Blockstack. This blockchain manages the registrar of the .bit TLD through a straightforward procedure, in which a registrant specifies the SLD that they wish to register (which is subsequently appended with the .bit TLD), as well as the resolving IP and other secondary parameters. The Emercoin blockchain is one of the most well-known services for domain registration. Surprisingly enough, although the naming requirements of Emercoin specify that only lowercase alphanumeric ASCII characters are allowed, the chain contains case sensitive domains not only for the advertised TLDs but for traditional TLDs like .com. In the following sections, we describe the details of each phase of our approach, which are detailed in Fig. 3.

4.1. Data collection and dataset structure

For the purposes of this research, we downloaded all the data from the two most widely used chains supporting blockchain DNS, which at the time of writing are Emercoin and Namecoin, in the form of JSON files. From these files, we extracted a subset of relevant information, namely domain names, IPs and emails (by using the *value* field), and the wallets associated to each domain, to create a curated dataset. Based on this, our dataset consists of a set of unique 5985 IP addresses. Note that the set of IP addresses consists of the public IPs as there were many occurrences of private IPs. Most likely, the private IP addresses are acting as placeholders for future record updates. We also noted invalid IPs or containing typos, for instance, one of the four integers of an

¹¹ <https://www.namecoin.org/>.

¹² <https://emercoin.com/>.

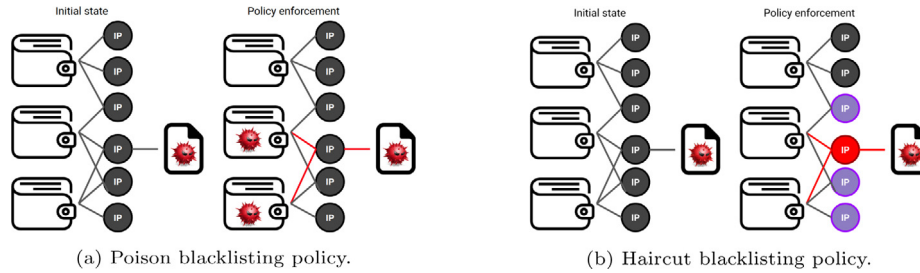


Fig. 2. Wallet and IP blacklisting with the poison (a) and haircut (b) policies.

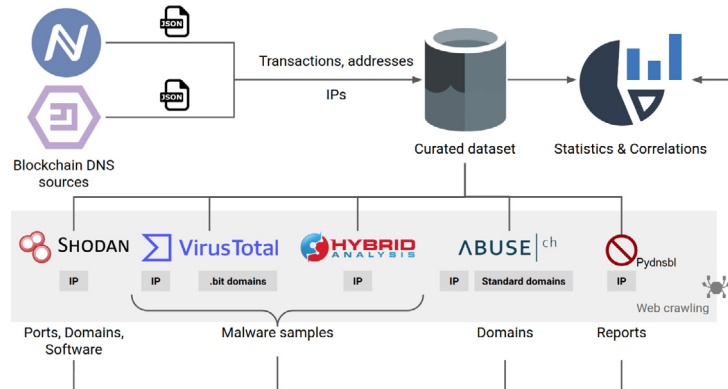


Fig. 3. Outline of the methodology for analysing blockchain DNS data.

IPv4 address contained a number greater than 255. These IP addresses were pruned as they provided no tangible value from an investigation perspective. Therefore, we ended up with 5130 public IPs being used in Namecoin, 919 in Emercoin, and 55 IPs are in both chains.

In addition, the dataset contains 2469 Emercoin wallets and 61 357 Namecoin wallets, which are related to these IPs in distinct ways. Finally, the number of domains related to these IPs are 4452 in the case of Emercoin, and 27 403 in the case of Namecoin. Nonetheless, not all of them are valid domain names. There are multiple domains which do not conform to the DNS format, e.g. they contain non allowed characters, have registered the same domain with combinations of upper and lower case characters etc. As a result, the resulting numbers of domains are 2675 for Emercoin and 27 261 for Namecoin.

The first step in our intelligence collection was to query the registered domains in the available engines. Due to the fact that these TLDs are not widely available, only a few engines provide actual information. In our research, we used VirusTotal, which at the time of writing supports only queries for .bit domains. From the 27 261 domains that were queried, only 661 were recorded in VirusTotal, 195 of which were reported malicious. Notably, these malicious domains were associated with 576 unique public IP addresses, implying that almost all of them have been updated several times. The fluxing rate of these IP addresses will be discussed in Section 4.2.2. Based on our poison blacklisting policy, since these domains are reported as malicious, the associated wallets that have registered them, and the IPs that have hosted them are poisoned, hence flagged as malicious.

Next, we submitted all the extracted IPs from Namecoin and Emercoin to VirusTotal, Hybrid Analysis, and Shodan, and collected the information that each platform has about them. We queried the 5985 unique IPs to which domains have been mapped in VirusTotal and Hybrid Analysis to determine how many of them are linked with malware samples that they have analysed. Notably, 1550 (25.9% of the total) IPs are reported malicious in the two platforms as they are correlated with 32 340 unique samples. Moreover, using intelligence from the different sources provided by Abuse,¹³ we identified some more IPs

Table 3

Identified malware distributed by IPs where Emercoin and Namecoin map their domains.

Type	Families
Banking malware	Ursnif, Chthonic, Dridex, Panda, BankBot, ClipBanker, Cerberus, Feodo, Geodo, heodo, Gozi, Vawtrak, Qbot
Ransomware	Buran, GlobImposter, GermanWiper, GandCrab, Hermes, Phobos, Paradise, Trolldesh, Sigma, maze, locky, zerber
Loader	Hancitor
Trojan	Bifrost, emotet, DanaBot, PsiXBot
Stealer	AZORULT, Valak
Miners	xmrig, minergate, acruminer
Botnet	Gafgyt, Mirai, Ramnit
RAT	agent tesla, quasar, ghost, imminent monitor rat

being malicious, reaching to 26.18% of the total. Merging the latter with the reports of VirusTotal for the .bit domains we have 1926 malicious IP addresses. Finally, we queried VirusTotal for the rest of IPs for other malicious activity, e.g. spamming, phishing etc. Of the remaining 4062, 131 were flagged as malicious, raising the total to 2057 IPs. Practically, more than a third (34.32%) of the IPs to which domains backed by blockchain DNS are redirecting are known to be malicious.

Notably, these IPs are linked with several malware families including, but not limited to, Emotet, AZORULT, Feodo, Cerberus, GermanWiper, and GandCrab. A more comprehensive list is presented in Table 3.

Moreover, we used Pydnbsl,¹⁴ an aggregator of blacklists of IPs to determine how many of the IPs have been blacklisted. In total, 1629 of the IPs in our dataset are blacklisted. Purging the duplicate reports of the IPs, the malicious reported IPs are 3039, representing the 50.78% of the total.

Next, we correlated these IP addresses with information from Shodan. While only 2493 of the IP addresses had been monitored and

¹³ <https://abuse.ch/>.

¹⁴ <https://github.com/dmippolitov/pydnbsl/>.

indexed by this tool, we nevertheless can extract valuable intelligence. In Table 4a and b we report the ten most common ports these devices are using and the ten most common identified products by Shodan, respectively. The results indicate that most of the servers are providing web hosting, file sharing, DNS, and mail services, with a preference to Linux-powered servers, implied by the use of SSH.

4.2. Blockchain DNS analysis and correlation

In what follows, we provide a detailed analysis of both Emercoin and Namecoin blockchains. First, we provide an exploratory analysis to highlight the most active IPs and wallets of each system and their ties with malicious activities, as reported by external intelligence sources. Second, we provide a geographical coverage of the IPs of each system. Next, we focus on the potential threats of such systems and apply our blacklisting policy, namely a hop-based approach, to analyse the links between IPs, wallets, domains, and e-mails and categorise their threat level. Finally, we analyse the user's behaviour according to some features to discover patterns that could indicate potential harm, and provide a statistical analysis by correlating them with maliciously reported IPs.

4.2.1. Emercoin

In the case of Emercoin, we created several data structures to establish associations between wallets, IP addresses and domains. First, we collected some statistics regarding the IP addresses found in Emercoin, and how different wallets used them to update the *value* field of one or several domain names. In this regard, Fig. 4a provides an overview of the top 20 Emercoin IPs in terms of the number of wallets using them. As it can be observed, a small subset of IPs are associated with more than 100 wallets, yet the vast majority of IPs have only one wallet associated with them, as it can be understood by observing the decreasing pace of the values. For instance, looking at the top five, the most used IP (202.108.22.5) has been reported as malicious. In the case of the runner up 192.243.100.192, although it has not been reported as malicious, it directs to a “boutique”¹⁵ for selling Emercoin domains. The IP 192.227.233.13 is found in many expired domains and was reported as malicious, yet it is not resolving to any site at the time of writing. The IP address 178.128.220.134 is resolving to emerAPI, an Emercoin related software, which includes links to the official site, yet there is no proof of its authenticity. Finally, 185.31.209.8 is an IP announced in several Eastern Europe sites [57] to be used when registering Emercoin domains. In the latter case, several users have used it as a default option. It is worth noting that, although there are only two IPs reported as malicious in this top five, our hop-based association approach, later described in this section, flagged IPs 192.243.100.192 and 185.31.209.8 as suspicious. The latter means that, (a) the intelligence available for these sites is insufficient, (b) that such IPs are not being used with malicious intentions yet, or (c) that malicious users, like benign ones, initially used them when setting up their wallets or (d) as a means to temporarily hide their activity and redirect incoming traffic.

Next, we computed the same statistics this time considering each wallet. Fig. 4b shows the amount of IPs used by the top 20 Emercoin wallets in their registered domain(s). We can observe that several wallets contain more than 50 IPs related to them. In this regard, a clear example of the extent to which Emercoin is being used for malicious purposes is given by observing, e.g. the top three wallets, since these are associated with several malicious IPs. Moreover, the wallet ETQERUknhW2A5cBmfHN4VBqL7VGiFnKQRh has been related with the DGA of BazarLoader [19] (also known as BazarBackdoor).

In addition, we depicted in Fig. 5 the geographical coverage of the Emercoin IPs, and we compare it with the reported malicious activities

¹⁵ <https://www.ecwid.com/store/cantdoevil/Existing-Invincible-EmerDNS-Domains-Contact-p155967426>.

Algorithm 1 Hop-based Association

```

1: function COMPUTESUSPICIOUSIPs( Dict ip_to_wallet, Dict wallet_to_mail,
   Dict ip_to_domain, List malicious_ips)
2:   Dict status_ips = { };
3:   while (ip in malicious_ips) do
4:     status_ips[ip] = malicious           ▷ Store {key,value} pair.
5:     wallet_list = GetWallets (ip_to_wallet, ip)           ▷ Wallets
   associated with malicious IP.
6:     domain_list = GetDomains (ip_to_domain, ip)           ▷ Domains
   associated with malicious IP.
7:     associated_ips = GetIPs (wallet_to_mail, wallet_list,
   domain_list)           ▷ Get IPs of associated wallets and domains
8:     status_ips = UpdateDict (associated_ips) ▷ Update benign IPs
   with suspicious value
9:   end while
10: return status_ips           ▷ Dict with classified IPs
11: end function

```

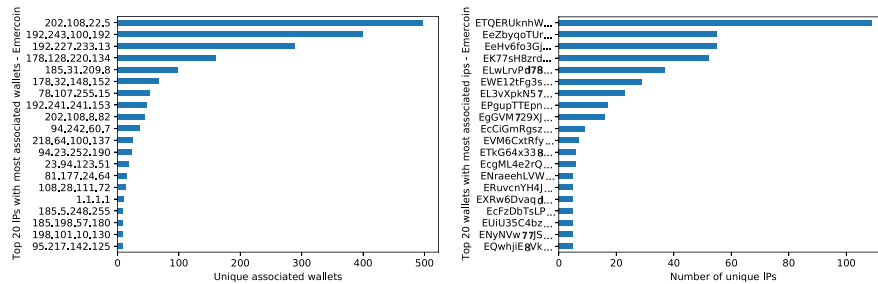
collected in Section 4.1. As identified in the maps, there is a direct correlation between the number of hosts and the malicious IPs reported. It is worth to mention that, in proportion to the amount of hosted IPs, there are less malicious IPs located in Russia and China than in other areas such as North America and Australia, according to the intelligence reports.

The next phase of the analysis focused on the identification of possible relationships between the different objects existing in these blockchain systems. More precisely, we analysed the correlations between wallets, as reported in the previous experiments, the set of “apparently” benign IPs, and the domains used in Emercoin. For this purpose, following the methodology described in Section 3, we developed a hop-based association approach, as described in Algorithm 1. More concretely, if a wallet or a domain contains a malicious IP, we tag the rest of the IPs associated with such wallet or domain as suspicious. Moreover, we use additional information from the *value* field of the curated dataset to find further relationships between such domains and wallets (e.g. wallets using the same email). In this case, we add the IP addresses of the additional wallets to the suspicious list. Following our methodology, we assume that if a wallet has used an IP reported in a malicious campaign, the rest of the associated IPs can potentially be used for similar purposes. Note that a suspicious state can only be updated by a malicious one if a specific IP is found to be malicious according to our ground truth, and that suspicious IPs do not spread their status further.

Concerning the detailed procedures and computational cost of our hop-based approach, the first step is to collect a snapshot of the whole blockchain and parse it into a structured JSON file, which is updated at regular intervals. Since this activity is performed offline, we consider this cost negligible. Next, the hop-based approach is applied to both Namecoin and Emercoin data in the order of seconds, even without parallelisation. More concretely, the cost of exploring all the IPs of a given blockchain system and, in the case they are reported as malicious, marking as suspicious the rest of IPs of the wallets containing it, is upper bounded by $O(n^2)/2$ in the case of a fully connected undirected graph. Given n nodes, the number of edges in a fully connected undirected graph is $n(n-1)/2$. As previously seen in Section 4.2, the connectivity of both Namecoin and Emercoin is far from a fully connected undirected graph, and thus the cost in such cases is much lower than $O(n^2)/2$. Moreover, note that the computational cost is also tied to the amount of dangerous IPs of the network. In other words, we only explore the wallets associated with an IP if the latter is marked as dangerous. Finally, the cost of identifying whether an IP is malicious is linear and is proportional to the time it takes to query a threat intelligence engine like VirusTotal that we used in this work.

Table 4
Statistics from Shodan.

(a) Used ports			(b) Identified software	
Port	Count	Common service	Software	Installations
80	1690	Web server	OpenSSH	1123
443	1411	Web server over SSL/TLS	Apache httpd	729
22	1068	SSH	nginx	681
53	888	DNS	Exim smtpd	276
21	386	FTP	MySQL	200
25	381	SMTP	Postfix smtpd	178
993	380	IMAP over TLS/SSL	Pure-FTPd	141
587	342	SMTP	MS IIS httpd	54
143	334	IMAP	ProFTPD	53
995	320	POP3 over SSL/TLS	Microsoft HTTPAPI httpd	28



(a) Top 20 most used IPs in Emercoin and the number of wallets using them. (b) Top 20 Emercoin wallets and the corresponding number of IPs found in their domains.

Fig. 4. Statistics about the most used IPs and biggest wallets of Emercoin.

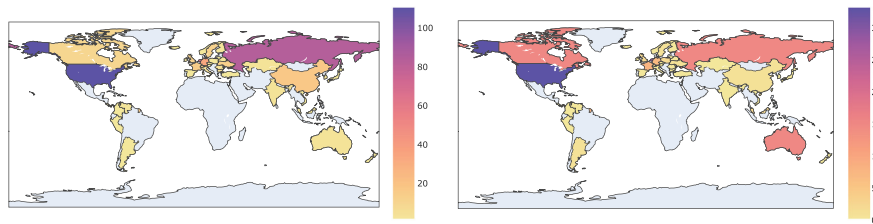


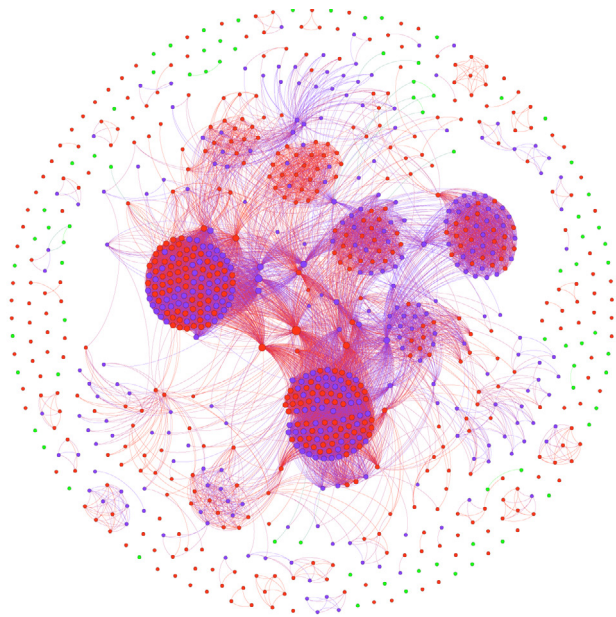
Fig. 5. Geographical coverage heatmap of IPs mapped in Emercoin (left) and the corresponding malicious reports (right).

In the case of Emercoin, our hop-based association found 280 new potentially malicious IPs, in addition to the 502 malicious IPs confirmed by the intelligence collected. Therefore, by revising our initial statistics, 74 IP addresses were found to be benign (only 8% of the IPs did not present any connection with malicious activities).

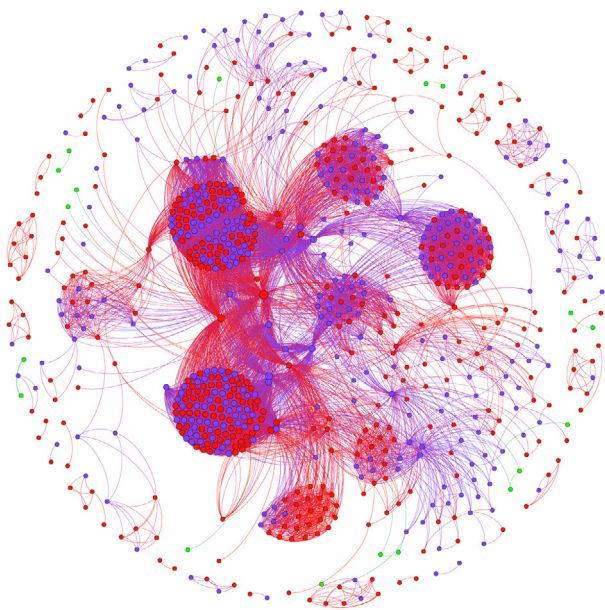
Further analysis was conducted on the intelligence collected in Section 4.1. In this regard, we use the list of IP addresses and the classification (i.e. benign, malicious and suspicious) provided by the hop-based approach. Thus, we deploy a graph-based visualisation of Emercoin (see Fig. 6a), in which nodes represent IPs, and the edge connecting two IPs represents a commonly shared interrelation in the form of, e.g. a wallet, an email, a domain or a combination of them. In the case of benign IPs, we can observe that they are mostly isolated (cf. Fig. 6a), since they have a very small representation in Fig. 6b. In the case of malicious and suspicious clusters, we can clearly identify their connections and all the associations, showcasing the relevance of the hop-based procedure to find new, potentially malicious groups of IPs. The average clustering coefficient of the network represented in Fig. 6a is 0.701 and in the case of Fig. 6b (discarding the isolated nodes) is 0.831. These numbers denote the high degree of connectivity between the nodes when they belong to a cluster, exhibiting highly interconnected communities. Fig. 7a shows the Complementary Cumulative Distribution Function (CCDF) of Emercoin. It can be clearly observed, by merging the data represented in Fig. 7a with the visual information of Fig. 6a, that there are specific peaks corresponding to high degree clusters. The number of clusters appears to be similar regardless of their

degree, for clusters with more than 10^2 elements. The latter denotes specific malicious behaviours (note that high degree clusters exist only in a malicious context as seen in Fig. 6a), which can be understood as outliers (they do not follow the initial data distribution, in which the higher the degree, the lower the amount of clusters). This malicious clusters can be potentially related to a specific campaign, orchestrated by one or several users using a closed set of IPs, wallets, emails and domains. As an additional outcome, we depicted the distribution of Eigenvector centrality in Fig. 7b. It can be observed that we have a cluster of nodes close to zero (corresponding to isolated nodes with few or none connections with highly connected nodes), and another cluster with a value above 0.08. The latter means that the nodes of the malicious clusters are highly interconnected between them and, in some cases, to other clusters. Therefore, in some occasions, the same assets (i.e. wallets, emails, IPs, or domains) have been used in more than one campaign, probably triggered by the same entities.

Finally, to identify additional relevant features, we explored the amount of updates that each domain had. In the analysed blockchain DNS systems, a domain can be updated by several reasons, such as renewing its time to live, assigning a new IP to it, or changing the *value* field to add extra options or information [58]. Our hypothesis was that highly active domains could be associated with malicious activities. In this regard, Fig. 8 shows the top 20 most active domains in terms of updates. For instance, the most updated domains are `everypony.emc` and `mymonero.coin` and in both cases these domains are associated with malicious IPs. Nevertheless, since the vast amount of Emercoin



(a) Emercoin representation including all isolated nodes, where each node represents an IP, and their size is weighted according to their connectivity. The edges represent commonly shared data between nodes, such as wallets, emails or domains.



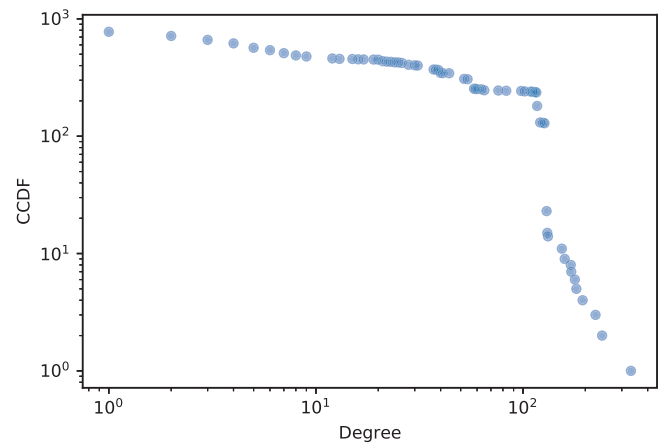
(b) Emercoin graph representation excluding isolated nodes. It can be observed that only a reduced number of benign nodes are present.

Fig. 6. Graph-based representation of the Emercoin ecosystem.

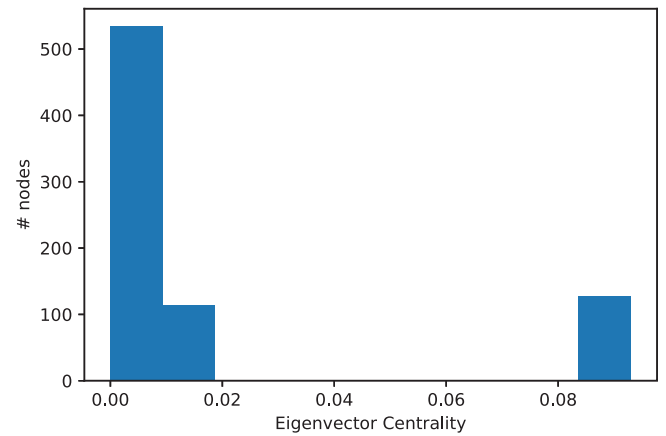
domains only contain one interaction (corresponding to their creation operation) we went a step further and explored if the combination of updates and the number of different IPs associated to each domain over time, could be used to indicate the goodness of a domain name. Therefore, we computed a ratio considering the number of IPs and the number of updates for each domain as described in (1).

$$Ratio_{IPs,updates} = \frac{\text{Number of unique IPs}}{\text{Number of updates}} \quad (1)$$

Next, we selected a range from 1 to 10 to represent the number of updates and, for each value, we computed the average $Ratio_{IPs,updates}$ for the set of benign domains and the set of malicious ones (i.e. domains



(a) CCDF of Emercoin.



(b) Eigenvector centrality distribution of Emercoin

Fig. 7. CCDF and eigenvector centrality values of the Emercoin ecosystem.

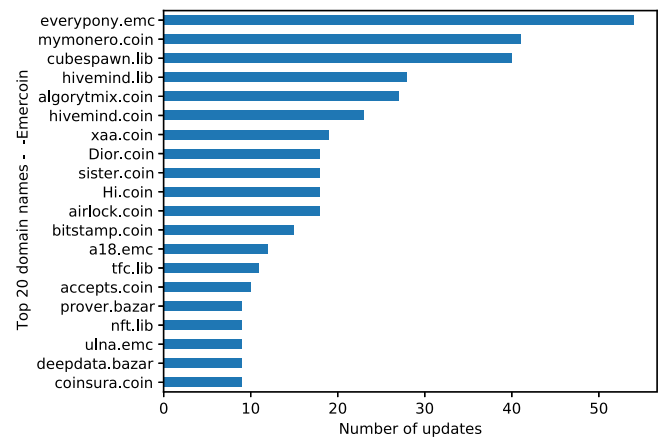


Fig. 8. Top 20 most updated domains in Emercoin.

were tagged as malicious if they contain a malicious IP in their records). Note that we considered values equal or above a specific number of updates to compute each average. The values, as well as the associated t -test outcomes, are shown in Table 5.

As observed from the t -test outcome, the IP address updates are significantly higher for the domains engaging in malicious activity than the benign ($p = 0.0002$), where a malicious domain is expected to have twice as many IP updates as a benign one. This can be used

Table 5
Different $Ratio_{IPs,updates}$ average values considering a range of update values, and the corresponding t -test outcomes. Note that the column “ ≥ 1 ” considers all the domains existing in the blockchain.

Domain type	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 6	≥ 7	≥ 8	≥ 9	≥ 10
Benign	0.928	0.492	0.409	0.298	0.257	0.234	0.234	0.228	0.212	0.200
Malicious	0.864	0.761	0.773	0.737	0.72	0.722	0.723	0.729	0.730	0.748

t -test values: statistic = -5.5507 — p -value = 0.0002

as a composite indicator of compromise and tactics, techniques and procedures.

Note that the fact that most of Emercoin domains do not have more than one update hinders the classification for domains if we consider only such case. Nevertheless, the more updates, the more evident is the difference in the behaviour between benign and malicious domains.

The latter means that malicious domains use more IPs per update than benign ones, on average. Note that the insight provided by the t -test can be complemented with the total number of IPs registered in a domain.

4.2.2. Namecoin

In the case of Namecoin, we computed the same set of data structures as with Emercoin, to analyse the different relations between IP addresses and wallets. Fig. 9a shows the correlation between the number of unique wallets and the top 20 IPs existing in Namecoin. It is noteworthy that, from the top five IPs, four were malicious except 91.250.85.116, which was found to be suspicious by our hop-based association approach.

The next data structure, graphically depicted in Fig. 9b, reports the correlation between the top wallets and the number of IPs related to each of them. It can be observed that the amount of IPs associated with these wallets is far lower than the numbers seen in Emercoin. Nevertheless, the latter is not related to a decrease in the number of malicious wallets. This is supported by the fact that, e.g. the top five wallets depicted in Fig. 9b used malicious IP addresses in their domains.

Similar to Emercoin, we depicted in Fig. 10 the geographical coverage of the Namecoin IP addresses and the malicious activities collected in Section 4.1. In this case, we observe a stronger correlation between the amount of IP addresses hosted and the reported malicious activities than in the case of Emercoin.

Next, by using Algorithm 1, we computed the set of suspicious IP addresses contained in Namecoin. In this case, in addition to the 2577 malicious reported IPs, we classified 1118 as suspicious ones, leaving 1431 as benign ones (i.e. only a 28% of the IPs were not connected to maliciously reported IPs). After computing such statistics, we depicted the graph representation of the Namecoin ecosystem in Fig. 11a. As in the case of Emercoin, nodes represent the IPs, and edges represent a common value (e.g. wallet, email, domain) shared between them. If we compare the representations depicted in Figs. 11a and 11b, we can observe a substantially reduced number of benign nodes in the latter, since most of them appear to be isolated. In the case of suspicious nodes, they are correlated with malicious ones, exhibiting clearly identifiable clusters. Moreover, there are different sizes of malicious clusters, yet well represented due to the high connectivity between malicious IPs. In addition, we computed the CCDF and the eigenvector distribution and depicted them in Figs. 12a and 12b, respectively. In the former case, we can observe a similar behaviour than the one discussed in Emercoin Section. That is, a set of malicious (according to the visual analysis of Fig. 11a) high degree clusters is represented, breaking the data distribution into two identifiable subsets (i.e. the data follows a completely different distribution below and above 10^2). In addition, Fig. 12b shows the eigenvector distribution of Namecoin. Again, there are two identifiable types of nodes in terms of centrality relevance, being the ones close to 0.05 the ones which denote higher connectivity, linking different malicious clusters. The average clustering coefficient of the network represented in Fig. 11a is 0.446 and in the case of Fig. 11b (discarding the isolated nodes) is

0.694. These numbers are lower than in the case of Emercoin due to the high amount of isolated nodes existing in Namecoin. Nevertheless, we can observe a rapid growth when we discard these isolated nodes. The latter means that, despite having some clusters which are not fully interconnected (especially small-sized ones), the average connectivity of the nodes when they belong to a cluster is high.

Next, we extracted the most updated domains in Namecoin and depicted them in Fig. 13. It is worth to note that, for instance, in the case of the two most updated domains, the users always used a private IP (127.0.0.1). In this regard, the behaviour of apparently benign users is not always expected by the network in terms of information updates. Since in both cases the owner updated the domain with the same information that it previously had (i.e. without the need to do it nor any other justifiable reason). Next, we used Eq. (1) with the benign and malicious subsets of Namecoin domains to compute the values for the same range than the one used in Emercoin, and depicted the results in Table 6.

The values obtained in Namecoin denote the same behaviour than the ones observed in Emercoin, yet this time with lower average values. The latter is a consequence of the Namecoin renewal requirement, which translates into a higher number of updates per domain to overcome their expiration time. Therefore, malicious domains tend to have more IPs per update, provably to keep malicious campaigns alive during longer periods and avoid security measures such as blacklisting.

In addition to the previous experiments, we extracted the common public IPs in both Emercoin and Namecoin and found that a total of 55 IPs are shared between such systems (we did not consider public nor IPs used in well-known services or traditional DNS servers), from which 32 are malicious. The latter exhibits the possibility that the same actors are perpetrating malicious activities in both blockchains.

4.3. Use case example

To showcase some of the functionalities of the proposed correlation analysis approach, we extracted a set of malicious domains reported back in 2018 by FireEye in several campaigns, namely Gandcrab ransomware, CHESSYLITE, Neutrino and other samples [59]. First, we computed some basic statistics for each domain by querying our curated dataset. In this regard, several of the domains did not resolve to any IP (bleepingcomputer.bit, nomoreransom.bit, esetnod32.bit, emsisoft.bit, and gandcrab.bit), and some others (brownsloboz.bazar, brownsloboz.lib, and brownsloboz.emc) only contained private IPs so were not considered further. The rest of the domains were studied and their main statistics are described in Table 7. The Namecoin domains reported exhibit specific behaviours that are aligned with the outcomes reported in Section 4.2.2. With the exception of flashupd.bit and cyber7.bit, the domains used a set of different IPs which were associated with a large number of different wallets (i.e. several wallets were managing such IPs and used them in another domains as well, as reflected in Table 7, column ‘Related Wallets’). Moreover, the $Ratio_{IPs,updates}$ value of such domains (i.e. considering the total number of IPs and the number of updates), is aligned with the malicious behaviour observed in Namecoin. Next, we analysed which of these subset of domains were related in terms of IPs, wallets, or emails, and we observed that leomoon.bit lookstat.bit sysmonitor.bit volstat.bit and xoonday.bit shared common information. Moreover, we extended our search to find other domains that were correlated with these ones

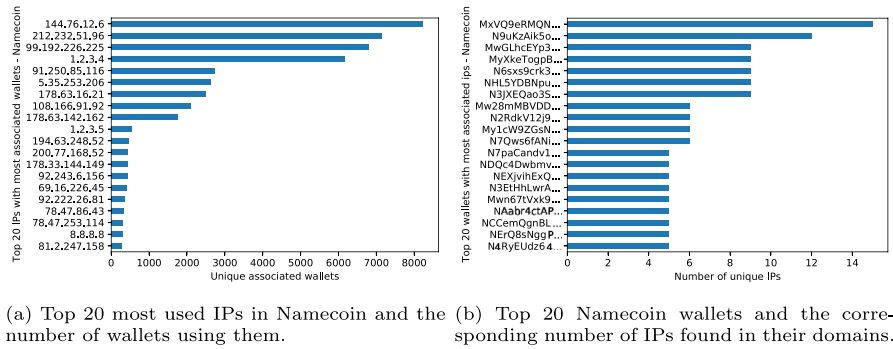


Fig. 9. Statistics about the most used IPs and biggest wallets of Namecoin.

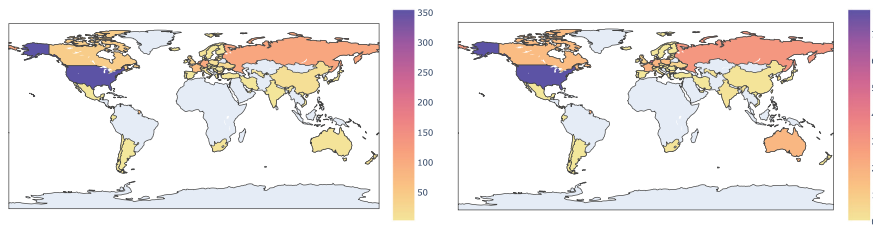


Fig. 10. Geographical coverage heatmap of IPs mapped in Namecoin (left) and the corresponding malicious reports (right).

Table 6

Different $Ratio_{IPs,updates}$ average values considering a range of update values, and the corresponding t -test outcomes. Note that the column “ ≥ 1 ” considers all the domains existing in the blockchain.

Domain type	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 6	≥ 7	≥ 8	≥ 9	≥ 10
Benign	0.396	0.304	0.278	0.249	0.225	0.216	0.194	0.146	0.137	0.121
Malicious	0.504	0.410	0.380	0.354	0.345	0.328	0.344	0.353	0.339	0.343

t -test values: statistic = -4.5437 — p -value = 0.0003

Table 7

Statistics and IP classification of the studied Namecoin domains. Domains coloured in red denote a malicious clustered group.

Domain	Updates	Related wallets	IP classification breakdown			
			Benign	Malicious	Suspicious	Total
leomoon.bit	17	71	0	9	3	12
lookstat.bit	11	35	0	3	4	7
sysmonitor.bit	15	52	0	6	5	11
volstat.bit	16	48	0	7	3	10
xoonday.bit	15	76	0	10	0	10
flashupd.bit	1	2	0	1	0	1
cyber7.bit	1	1	0	1	0	1
brownsloboz.bit	6	14	0	4	1	5

and we found the following list of domains: typeme.bit, brow-baseis.bit, silikat.bit, vedixme.bit, testikname.bit, delix.bit, cash-money-analitica.bit, foaming.bit, firststat.bit, skildexin.bit, glesifax.bit, stamexis .bit, flexz.bit, checkxod.bit, money-cash-analitica .bit. Finally, we extended the list of suspicious IP addresses by using our hop-based association approach.

4.4. Evaluation of the hop-based policy

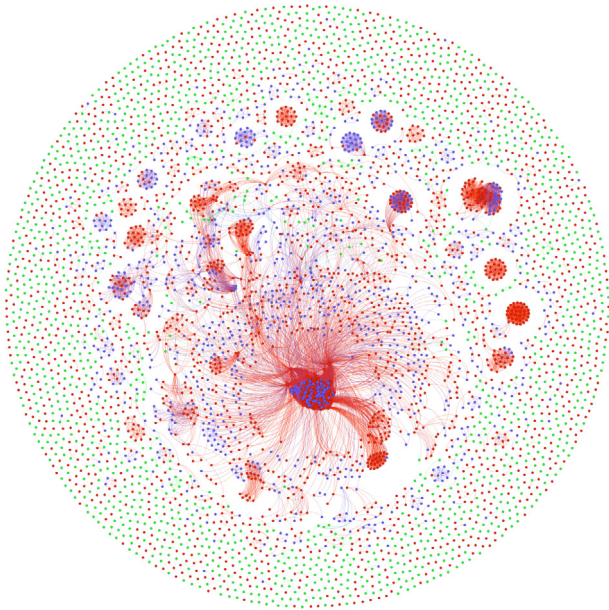
Further to our initial experiments, we also evaluated the efficacy of our hop-based policy. To achieve this one would have to determine whether IPs that were classified as suspicious from our algorithm would be later identified by threat intelligence platforms. Note that platforms such as VirusTotal do not report the first time that an IP was classified as malicious but only the last analysis result and its date.

Leaving a timeframe of approximately six months, we queried Virus-Total for the IPs that our hop-based approach had classified as suspicious. The returned results proved our hypothesis as 47 of these IPs are now reported as malicious, as seen in Table 8. It should be noted, that our approach identifies sources from which an adversary may launch an upcoming attack. Therefore, our approach correctly identified such IPs in a predictive security manner.

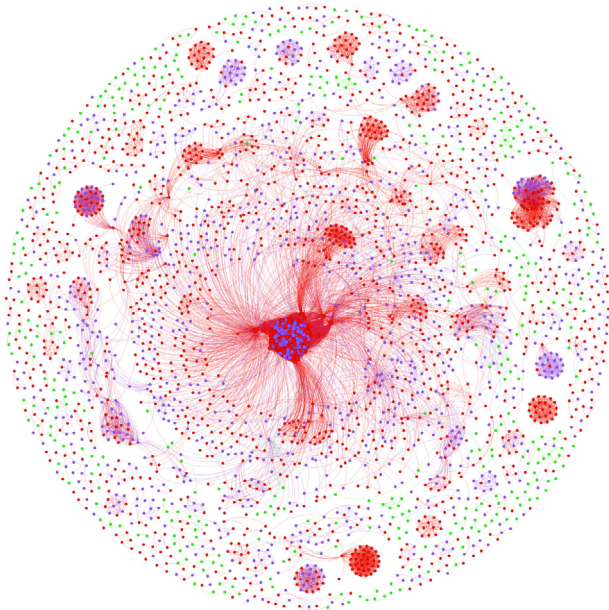
5. Discussion

One of the conclusions that can be extracted from the outcomes discussed in the previous sections is that Namecoin and Emercoin are currently primarily used for malicious purposes since a huge share of the IPs registered in Emercoin and Namecoin are directly associated with malicious activities. Such statistics hinder the adoption of blockchain DNS systems and the trust of the community towards them. Therefore, the emergence of novel solutions overcoming the main drawbacks of blockchain DNS is required. After exploring the state-of-the-art and analysing the actual status of Emercoin and Namecoin, we identified different subsets of challenges applicable to these and other blockchain DNS systems. These challenges can be mainly classified into (i) the registration procedure and users behaviour, (ii) the extraction of information flows and their links with external threat analysis systems, and (iii) the security of the underlying blockchain platform and proactive measures.

There is an urgent need to improve the robustness and security of the registration procedures in blockchain DNS systems. One clear example relies on Emercoin registrar, which allows the use of case sensitive, non UTF-8, and other forbidden patterns and characters, as well as invalid domains according to RFC 1123 [60]. Furthermore, strategies to avoid, e.g. cybersquatting, are required, such as the one



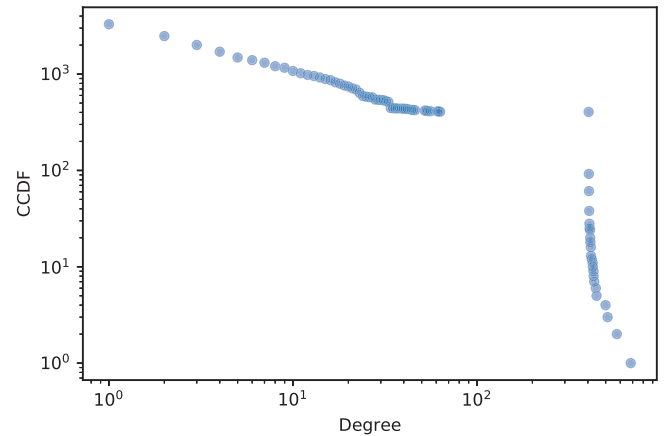
(a) Namecoin representation including all isolated nodes, where each node represents an IP, and their size is weighted according to their connectivity. The edges represent commonly shared data between nodes, such as wallets, emails or domains.



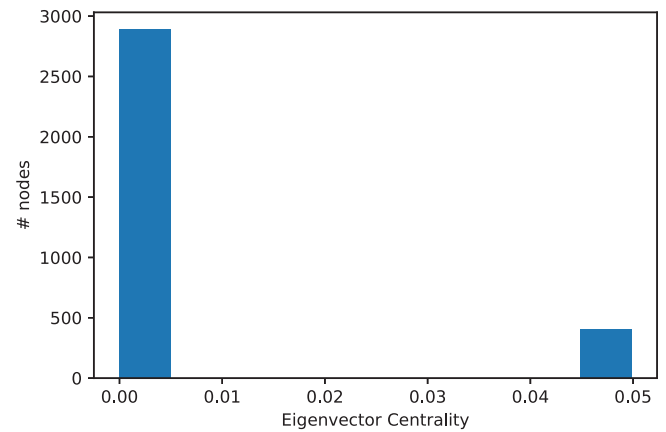
(b) Namecoin graph representation excluding isolated nodes. It can be observed that the amount of benign nodes is substantially reduced.

Fig. 11. Graph-based representation of the Namecoin ecosystem.

implemented by Handshake, which reserved the top 100k Alexa domains. In terms of user behaviours, specific control of the amount and speed of domains registered could help in detecting and reducing several campaigns. In this regard, we studied the behaviour of users and their strategies to avoid being linked or related to other activities in both Emercoin and Namecoin. While there exist several wallets containing a vast number of IPs in both systems, most malicious users follow the strategy of one-wallet one-IP. That is, to avoid being tracked, users often use different wallets with a low time-to-live (e.g. only for one IP update). The latter hinders the task of identifying malicious wallet-to-IP connections, especially since most of the interactions in



(a) CCDF of Namecoin.



(b) Eigenvector centrality distribution of Namecoin.

Fig. 12. CCDF and eigenvector centrality values of the Namecoin ecosystem.

the blockchain are of this nature. Nevertheless, our methodology is able to unveil these internal relationships by exploring the correlations in different dimensions, namely wallets, IPs, domains, and further information stored in the *value* field. For instance, we can leverage proactive security in blockchains, with, e.g., active checks focusing on the behaviour of the users, as well as the information associated with each wallet. As observed in the studied BDNS systems and due the possibility of having other potential indicators, we believe that exploring and assessing the different data managed by such systems is crucial to design the proper mitigation strategies. For example, parameters such as the amount of suspicious domains registered (e.g., domain squatting [56], or artificially generated domains [61]), the number of wallet updates, the IPs and domains registered, and the connectivity of the nodes are features that can be used to identify potentially harmful user behaviours. The latter can be augmented by our hop-based approach as well as similar methods following blacklisting policies, enhancing the reliability and trust of blockchain DNS while reducing the impact of malicious campaigns. Therefore, it is imperative to establish a holistic end-to-end approach, possibly through integrating smart contracts with revocation mechanisms [62,63], to manage the registration procedure as well as to protect blockchain DNS system from misuse. Moreover, while we have to support security and privacy initiatives, the accountability perspective, especially when it comes to critical Internet infrastructures such as DNS must also be taken into consideration.

Another issue that we encountered during our investigation is that the bulk of threat intelligence sources lack information regarding blockchain DNS systems. Moreover, the intelligence collected from the

Table 8
Originally classified suspicious IPs for which VT reports malicious activity.

185.117.119.190	192.241.241.153	54.37.229.180	89.223.88.183	185.86.148.137
91.235.129.241	210.16.101.109	108.167.140.18	185.222.202.206	193.106.31.146
51.89.177.5	192.3.12.121	5.34.180.226	185.101.105.232	111.90.149.240
45.141.84.190	5.252.176.7	45.153.184.158	185.14.187.128	209.141.36.7
23.239.84.135	31.220.23.1	192.99.178.153	95.217.74.220	172.82.152.132
45.32.236.82	185.147.14.237	145.239.47.64	185.13.36.121	64.44.51.117
195.123.237.156	93.115.28.9	185.107.94.36	5.83.163.2	51.81.112.135
194.5.249.247	138.68.149.171	185.82.202.123	109.201.133.111	104.203.229.17
23.92.93.233	107.174.86.134	108.170.40.59	173.249.5.248	5.182.210.180
109.234.35.166	104.161.32.111			

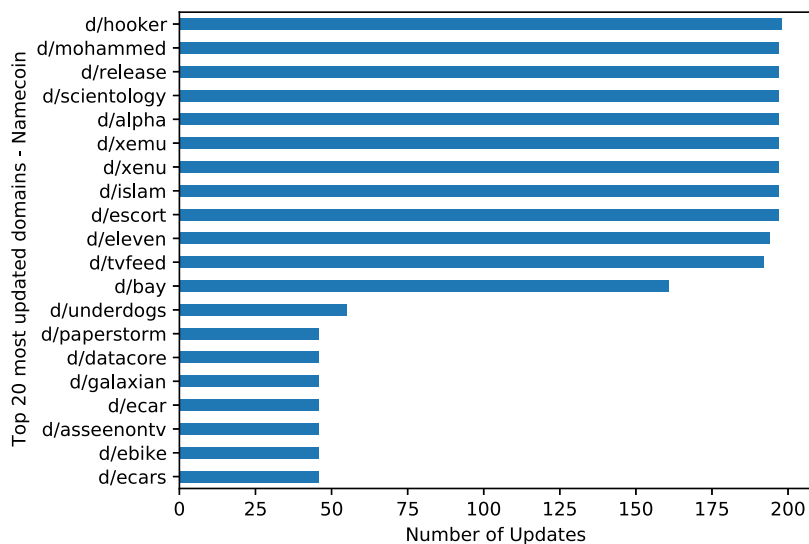


Fig. 13. Top 20 most updated domains in Namecoin.

sources used in this article is disparate and not homogeneous. For instance, only VirusTotal keeps track of requests to .bit domain but not to .coin, bazar, .lib and .emc domains. Hybrid Analysis does not keep track of any such requests. Notably, other platforms do not keep track of these domains, nor of their updates but monitor each connected IP individually. With the continuous rise of such schemes, the quest for information about such domains and their interconnections becomes even more necessary.

The timely collection of quality intelligence is crucial to detect cybercriminal campaigns and may lead to their prevention since methodologies like the one proposed in this article rely on such information to establish ground truth. Therefore, more efforts should be devoted to the active monitoring of the blockchain DNS ecosystem, including both their domains and IPs, in an automated way.

In the case of blockchain features, they are often recalled in their beneficial form, yet some of them can leverage malicious opportunities. The clearest example of this is immutability. In this regard, the impossibility of deleting records guarantees traceability and auditability of malicious campaigns, their modus operandi, and enables mitigation actions. For instance, we can leverage proactive security in blockchains, with, e.g. active checks focusing on the behaviour of the users, as well as the information associated with each wallet. The latter can be used to detect future campaigns by using, e.g. our hop-based approach as well as similar methods following haircut blacklisting policies, enhancing the reliability and trust of blockchain DNS while reducing the impact of malicious campaigns. Nevertheless, the impossibility of deleting, e.g. malicious records or illegal information, is a clear disadvantage. In this regard, there is still much work ahead to enable efficient blockchain deletion mechanisms [7,64], since actual practices mainly rely on forks, and long block consolidation mechanisms, which add prohibitive overhead to blockchain systems. Aligned with the idea of forks, well-known systems such as Bitcoin and Ethereum have opted

for forks as a solution to security issues or required protocol changes to enable further functionalities [65,66]. Therefore, fork-based strategies, including novel and robust functionalities, could help in recovering the trust in Namecoin and Emercoin.

In principle, blockchains are considered to provide some form of privacy. While there is no transaction privacy, users through the use of multiple wallet addresses may enjoy some privacy guarantees. Hence, blockchain DNS approaches, beyond decentralisation, immutability, and resilience may provide some privacy guarantees to the owner of the domains, through, e.g. pseudoanonymisation. Notably, in our research we observe that even though both chains have several thousands of wallet addresses, users have opted to share self-identifying information such as emails allowing the linking of their wallets, defying the very scope of using different wallets for registering their domains. In fact, as discussed in the previous section, this behaviour is frequent, indicating the lack of understanding of how blockchains work from the users' perspective.

6. Conclusions

In this article, we provided a thorough analysis of the most mature blockchain DNS systems, namely Namecoin and Emercoin. In addition to reviewing the actual state-of-the-art of blockchain DNS systems, we proposed a sound and automated methodology to retrieve, process, and analyse the data stored in such systems. Thereafter, we recalled a set of blacklisting policies, namely blacklisting and haircut, and used the latter in our investigation to provide an insight into how Namecoin and Emercoin are used. The outcomes of our analysis, which includes internal correlations and external intelligence linked to several campaigns, concluded that the actual blockchain DNS ecosystems are being used for malicious purposes since more than 50.7% of the IPs used by the domains registered has been reported as malicious. Moreover,

we developed a predictive association method to identify suspicious IPs (more than 24% of all IPs were tagged as suspicious), enabling proactive measures.

Finally, we identified and discussed the main challenges and proposed several ways to overcome them, according to the knowledge extracted from our analysis and the well-known flaws of blockchain DNS systems. Future work will focus on exploring other blockchain DNS systems and studying further proactive strategies to prevent malicious activities in blockchain ecosystems. Moreover, we will explore other strategies to identify malicious behaviour considering e.g., time-based thresholds, to capture potential active threats.

CRedit authorship contribution statement

Fran Casino: Conceptualisation, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualisation. **Nikolaos Lykousas:** Conceptualisation, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualisation. **Vasilios Katos:** Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Supervision, Project administration. **Constantinos Patsakis:** Conceptualisation, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the projects *CyberSec4Europe* (Grant Agreement no. 830929) and *LOCARD* (Grant Agreement no. 832735).

The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

References

- [1] World Economic Forum, The global risks report 2020, 2020, http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf.
- [2] C. Shar, State of DNSSEC Deployment 2016, Internet Society Report, 2016.
- [3] M. Anagnostopoulos, G. Kambourakis, P. Kopanos, G. Louloudakis, S. Grizalis, Dns amplification attack revisited, *Comput. Secur.* 39 (2013) 475–485.
- [4] F. Casino, T.K. Dasaklis, C. Patsakis, A systematic literature review of blockchain-based applications: Current status, classification and open issues, *Telemat. Inform.* 36 (2019) 55–81.
- [5] M.S. Ali, M. Vecchio, M. Pincheira, K. Dolui, F. Antonelli, M.H. Rehmani, Applications of blockchains in the internet of things: A comprehensive survey, *IEEE Commun. Surv. Tutor.* 21 (2) (2018) 1676–1717.
- [6] G. Ateniese, B. Magri, D. Venturi, E. Andrade, Redactable blockchain-or-rewriting history in bitcoin and friends, in: 2017 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2017, pp. 111–126.
- [7] E. Politou, F. Casino, E. Alepis, C. Patsakis, Blockchain mutability: Challenges and proposed solutions, *IEEE Trans. Emerg. Top. Comput.* (2019) 1.
- [8] H.A. Kalodner, M. Carlsten, P. Ellenbogen, J. Bonneau, A. Narayanan, An empirical study of namecoin and lessons for decentralized namespace design, in: WEIS, 2011.
- [9] C. Patsakis, F. Casino, N. Lykousas, V. Katos, Unravelling ariadne's thread: Exploring the threats of decentralised dns, *IEEE Access* 8 (2020) 118559–118571.
- [10] P. Hoffman, P. McManus, DNS queries over HTTPS (DoH), 2018, <https://tools.ietf.org/html/rfc8484>.
- [11] Z. Hu, L. Zhu, J. Heidemann, A. Mankin, D. Wessels, P. Hoffman, Specification for dns over transport layer security (tls), 2016, <https://tools.ietf.org/html/rfc858>.
- [12] The exip project, 2021, <https://exip.live/>.
- [13] The butterfly protocol, 2021, <https://www.butterflyprotocol.io/>.
- [14] The brave browser, 2021, <https://brave.com/>.
- [15] A. Brandt, Bazarloader deploys a pair of novel spam vectors, 2021, <https://news.sophos.com/en-us/2021/04/15/bazarloader/>.
- [16] Cyware, Links discovered between bazar and TrickBot, 2020, <https://cyware.com/news/links-discovered-between-bazar-and-trickbot-2909546d>.
- [17] Z. Huang, J. Huang, T. Zang, Leopard: Understanding the threat of blockchain domain name based malware, in: A. Sperotto, A. Dainotti, B. Stiller (Eds.), *Passive and Active Measurement*, Springer International Publishing, Cham, 2020, pp. 55–70.
- [18] M. Möser, R. Böhme, D. Breuker, Towards risk scoring of bitcoin transactions, in: *International Conference on Financial Cryptography and Data Security*, Springer, 2014, pp. 16–32.
- [19] J. Bader, The domain generation algorithm of bazarloader, 2020, <https://johannesbader.ch/blog/the-dga-of-bazarbackdoor/>.
- [20] E.S.-J. Swildens, R.D. Day, et al., Domain name resolution using a distributed dns network, 2010, US Patent 7, 725, 602.
- [21] C. Cachin, A. Samar, Secure distributed dns, in: *International Conference on Dependable Systems and Networks*, 2004, 2004, pp. 423–432.
- [22] V. Ramasubramanian, E.G. Sizer, The design and implementation of a next generation name service for the internet, in: *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, in: SIGCOMM '04, Association for Computing Machinery, New York, NY, USA, 2004, pp. 331–342.
- [23] M. Wachs, M. Schanzenbach, C. Grothoff, A censorship-resistant, privacy-enhancing and fully decentralized name system, in: *International Conference on Cryptology and Network Security*, Springer, 2014, pp. 127–142.
- [24] Z. Qiang, Z. Zheng, Y. Shu, P2pdns: A free domain name system based on p2p philosophy, in: 2006 Canadian Conference on Electrical and Computer Engineering, 2006, pp. 1817–1820.
- [25] M. Abu-Amara, F. Azzedin, F.A. Abdulhameed, A. Mahmoud, M.H. Sqalli, Dynamic peer-to-peer (p2p) solution to counter malicious higher domain name system (dns) nameservers, in: 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2011, pp. 001014–001018.
- [26] F. Casino, E. Politou, E. Alepis, C. Patsakis, Immutability and decentralized storage: An analysis of emerging threats, *IEEE Access* 8 (2020) 4737–4744.
- [27] S. Al-Mashhadi, S. Manickam, A brief review of blockchain-based dns systems, *Int. J. Internet Technol. Secur. Trans.* 10 (4) (2020) 420–432.
- [28] D. Storm, P2P DNS to take on ICANN after US domain seizures, 2010, <https://www.computerworld.com/article/2469753/p2p-dns-to-take-on-icann-after-us-domain-seizures.html>.
- [29] E. Karaarslan, E. Adiguzel, Blockchain based dns and pki solutions, *IEEE Commun. Stand. Mag.* 2 (3) (2018) 52–57.
- [30] A. Hari, T. Lakshman, The internet blockchain: A distributed, tamper-resistant transaction framework for the internet, in: *HotNets 2016 - Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, 2016, pp. 204–210.
- [31] B. Benshoof, A. Rosen, A. Bourgeois, R. Harrison, Distributed decentralized domain name service, in: *Proceedings - 2016 IEEE 30th International Parallel and Distributed Processing Symposium, IPDPS 2016*, 2016, pp. 1279–1287.
- [32] S. Gourley, H. Tewari, Blockchain backed dnssec, *Lect. Notes Bus. Inf. Process.* 339 (2019) 173–184.
- [33] Z. Guan, A. Garba, A. Li, Z. Chen, N. Kaaniche, Authledger: A novel blockchain-based domain name authentication scheme, in: *ICISSP 2019 - Proceedings of the 5th International Conference on Information Systems Security and Privacy*, 2019, pp. 345–352.
- [34] J. Liu, B. Li, L. Chen, M. Hou, F. Xiang, P. Wang, A data storage method based on blockchain for decentralization dns, in: *Proceedings - 2018 IEEE 3rd International Conference on Data Science in CyberSpace, DSC 2018*, 2018, pp. 189–196.
- [35] W. Wang, N. Hu, X. Liu, Blockzone: A blockchain-based dns storage and retrieval scheme, in: *Artificial Intelligence and Security*, Springer International Publishing, Cham, 2019, pp. 155–166.
- [36] Z. Yu, D. Xue, J. Fan, C. Guo, Dnstm: dns cache resources trusted sharing model based on consortium blockchain, *IEEE Access* 8 (2020) 13640–13650.
- [37] X. Duan, Z. Yan, G. Geng, B. Yan, Dnsledger: Decentralized and distributed name resolution for ubiquitous iot, in: 2018 IEEE International Conference on Consumer Electronics, ICCE 2018, Vol. 2018-January, 2018, pp. 1–3.
- [38] W. Yoon, I. Choi, D. Kim, BlockONS: Blockchain based object name service, in: *ICBC 2019 - IEEE International Conference on Blockchain and Cryptocurrency*, 2019, pp. 219–226.
- [39] X. Wang, K. Li, H. Li, Y. Li, Z. Liang, ConsortiumDNS: A distributed domain name service based on consortium chain, in: *Proceedings - 2017 IEEE 19th Intl Conference on High Performance Computing and Communications, HPCC 2017, 2017 IEEE 15th Intl Conference on Smart City, SmartCity 2017 and 2017 IEEE 3rd Intl Conference on Data Science and Systems, DSS 2017*, Vol. 2018-January, 2018, pp. 617–620.
- [40] H. Li, H. Ma, L. Haopeng, Z. Huang, X. Yang, K. Li, H. Wang, Blockchain-based domain name resolution system, 2019, US Patent App. 15/768, 833.
- [41] H. Li, X. Wang, Z. Lin, J. Wu, X. Si, K. Li, X. Yang, H. Wang, Systems and methods for managing top-level domain names using consortium blockchain, 2019, US Patent App. 10/178, 069.

- [42] M. Ali, J. Nelson, R. Shea, M.J. Freedman, Blockstack: A global naming and storage system secured by blockchains, in: 2016 USENIX Annual Technical Conference (USENIX ATC 16), USENIX Association, Denver, CO, 2016, pp. 181–194.
- [43] Seize and desist? The state of cybercrime in the post-alphabay and hansa age, 2017.
- [44] KrebsOnSecurity, Carders park piles of cash at joker's stash, 2016, <https://krebsonsecurity.com/2016/03/carders-park-piles-of-cash-at-jokers-stash>.
- [45] Abuse.ch, Bit - the next generation of bulletproof hosting, 2017, <https://abuse.ch/blog/dot-bit-the-next-generation-of-bulletproof-hosting/>.
- [46] Trend Micro, .bit domain used to deliver malware and other threats, 2013, <https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/bit-domain-deliver-malware-and-other-threats>.
- [47] H. Wang, Fbot, a satori related botnet using block-chain dns system, 2018, <https://blog.netlab.360.com/threat-alert-a-new-worm-fbot-cleaning-adbminer-is-using-a-blockchain-based-dns-en/>.
- [48] S. Pletinckx, C. Trap, C. Doerr, Malware coordination using the blockchain: An analysis of the cerber ransomware, in: 2018 IEEE Conference on Communications and Network Security, CNS 2018, 2018, pp. 1–9.
- [49] K.G. Randi Eitzman, J. Valdez, How the rise of cryptocurrencies is shaping the cyber crime landscape: Blockchain infrastructure use, 2018, <https://www.fireeye.com/blog/threat-research/2018/04/cryptocurrencies-cyber-crime-blockchain-infrastructure-use.html>.
- [50] C. Patsakis, A. Chrysanthou, Analysing the fall 2020 emotet campaign, 2020, [arXiv:2011.06479](https://arxiv.org/abs/2011.06479).
- [51] L. Abrams, Gandcrab ransomware distributed by exploit kits, appends gdcB extension, 2018, <https://www.bleepingcomputer.com/news/security/gandcrab-ransomware-distributed-by-exploit-kits-appends-gdcB-extension/>.
- [52] Microsoft, Behavior monitoring combined with machine learning spoils a massive dofoil coin mining campaign, 2018, <https://www.microsoft.com/security/blog/2018/03/07/behavior-monitoring-combined-with-machine-learning-spoils-a-massive-dofail-coin-mining-campaign/>.
- [53] Emercoin, Emercoin links & resources, 2019, <https://emercoin.com/en/documentation/links-resources>.
- [54] L. Böck, N. Alexopoulos, E. Saracoglu, M. Mühlhäuser, E. Vasilomanolakis, Assessing the threat of blockchain-based botnets, in: 2019 APWG Symposium on Electronic Crime Research (ECrime), IEEE, 2019, pp. 1–11.
- [55] K. Perlow, Mapping out decentralized namecoin and emergoin infrastructure, in: Black Hat USA, Las Vegas, 2018.
- [56] P. Xia, H. Wang, Z. Yu, X. Liu, X. Luo, G. Xu, Ethereum name service: the good, the bad, and the ugly, 2021, [arXiv preprint arXiv:2104.05185](https://arxiv.org/abs/2104.05185).
- [57] S. Null, In response to the ukrainian (and not only) prohibitions: decentralized emerDNS system against site blocking, 2019, <https://sudonull.com/post/70519-In-response-to-the-Ukrainian-and-not-only-prohibitions-decentralized-EmerDNS-system-against-site-blo>.
- [58] Emercoin, The emergoin nvs, 2021, <https://emercoin.com/en/documentation/blockchain-services/emernvs>.
- [59] J.V. Randi Eitzman, How the rise of cryptocurrencies is shaping the cyber crime landscape: Blockchain infrastructure use, 2018, <https://www.fireeye.com/blog/threat-research/2018/04/cryptocurrencies-cyber-crime-blockchain-infrastructure-use.html>.
- [60] R. Braden, Rfc1123: Requirements for internet hosts-application and support, 1989.
- [61] F. Casino, N. Lykousas, I. Homoliak, C. Patsakis, J. Hernandez-Castro, Intercepting hail hydra: Real-time detection of algorithmically generated domains, 2020, [arXiv preprint arXiv:2008.02507](https://arxiv.org/abs/2008.02507).
- [62] Y. Yu, Y. Zhao, Y. Li, X. Du, L. Wang, M. Guizani, Blockchain-based anonymous authentication with selective revocation for smart industrial applications, *IEEE Trans. Ind. Inf.* 16 (5) (2020) 3290–3300.
- [63] J. Cruz, Y. Kaji, N. Yanai, Rbac-sc: Role-based access control using smart contract, *IEEE Access* 6 (2018) 12240–12251.
- [64] K. Maeda, M. Ohtani, Y. Oishi, C. Yasumoto, J. Zhu, Deletion of blocks in a blockchain, 2020, US Patent 10, 739, 997.
- [65] F. Schär, Blockchain forks: A formal classification framework and persistency analysis, *Munich Pers. RePEc Arch.* (2020).
- [66] T. Neudecker, H. Hartenstein, An empirical analysis of blockchain forks in bitcoin, in: International Conference on Financial Cryptography and Data Security, Springer, 2019, pp. 84–92.

