UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
DEPARTMENT OF TELEMATICS ENGINEERING

Information Retrieval and Evaluation of the Privacy Risk on Twitter

MASTER THESIS

LLP/ERASMUS Education Programme

Master in Digital Systems Security, University of Piraeus
Master in Telematics Engineering, Universitat Politècnica de Catalunya

Author: Christina Kapi
Supervisor: Jordi Forné, Associate Professor,
Universitat Politècnica de Catalunya, Barcelona, Spain
Co-Supervisor: Constantinos Lambrinoudakis, Assistant Professor,
University of Piraeus, Athens, Greece

Academic Year 2013-2014

UNIVERSITAT POLITÈCNICA DE CATALUNYA
DEPARTMENT OF TELEMATICS ENGINEERING

UNIVERSITY OF PIRAEUS
DEPARTMENT OF DIGITAL SYSTEMS

Information Retrieval and Evaluation of the Privacy Risk on Twitter

MASTER THESIS

Master in Digital Systems Security, University of Piraeus
Master in Telematics Engineering, Universitat Politècnica de Catalunya

This master thesis was developed in the context of the LLP/Erasmus Education Programme between University of Piraeus and Universitat Politècnica de Catalunya. It was supervised by Jordi Forné, Associate Professor at Universitat Politècnica de Catalunya, and Constantinos Lambrinoudakis, Assistant Professor at University of Piraeus.

Disclaimer

This document reflects the results of a study that has been prepared on behalf of the Master Degree Programme “Digital Systems Security” at University of Piraeus, Greece. It was also developed in the context of the Master Degree Programme “Telematics Engineering” at Universitat Politècnica de Catalunya, Spain. The information and conclusions contained in this thesis express the author’s personal opinion and arguments, and therefore should not be interpreted that they represent the official concepts of University of Piraeus or Universitat Politècnica de Catalunya.

To my family,
for always being there for me.

To Spyros,
because he makes me smile.

Abstract

In recent times, a growing number of information retrieval applications are disposable, aiming to trace users' online behavior and activities. One of the most popular social networks, which can be considered as a valuable source of information to this kind of applications, is Twitter. Aggregated data that derive from Twitter can show great power in delivering information related to users' interests and preferences.

The process of correlating information can result in the construction of comprehensive user profiles that may disclose detailed personal information and raise challenges to users' privacy as well. Extracted behavioral patterns of users can be substantial to the development of personalization services, however, inevitably at the expense of users' privacy. Although there are a number of privacy-enhancing technologies, which strive to mitigate many of these concerns, significant gaps remain regarding the privacy protection of users.

In addition, it is essential to provide a comprehensive view on metrics which consist in quantifying privacy. Most of the efforts devoted to devise privacy metrics are quite limited, as they apply to concrete systems. The lack of suitable metrics is deterrent to the proper privacy evaluation. Therefore, even though proposed approaches have made meaningful contributions to the challenging privacy landscape, there still exists a certain ambiguity about their effectiveness and adjustment to different contexts.

In this work, we present an effort towards the construction of user profiles, through the development process of an information retrieval application. We also tackle the privacy issues related to user profiling, as personal information contained in user profiles is disclosed. The last part of this thesis approaches the theme of quantifying user privacy by applying information-theoretic notions as measures of the privacy of user profiles.

Keywords: Information Retrieval, Social Networking Services, User Profiling, Privacy-Enhancing Technologies, Measurement of User Privacy, Shannon's Entropy, Kullback–Leibler divergence.

Acknowledgements

I would like to thank several people for their determinant contribution to this master thesis; Professor Jordi Forné, my research supervisor, for his valuable support, patient guidance and useful remarks during the development of this project; Professor Constantinos Lambrinoudakis, my research co-supervisor, for his assistance and feedback, especially in the initial stages of my work, as well as his consultance throughout my postgraduate studies.

My grateful thanks are extended to David Rebollo-Monedero, and particularly to Silvia Puglisi, for their contribution and useful recommendations. I would also like to express my gratitude to Professor Sokratis Katsikas for his support and encouragement during my studies.

Finally, I wish to thank my family for the love, enthusiasm and belief in me, as well as my friends, and in particular Spyros, for their patience, love and help.

Learning is the only thing the mind never exhausts, never fears, and never regrets.

Leonardo da Vinci, 1452 –1519,
Italian Renaissance polymath

Contents

List of Figures

1 Introduction

1.1 Background and Motivation

The immense advances in information and communications technologies have significantly raised the acceptance rate of social networking applications and services[1]. Volumes of digital data and information are available instantaneously and often transmitted widely or posted on websites publicly available[2]. However, by allowing users to access micro-blogging services, such as Twitter, the universe of ineligible people that may attempt to violate the users' rights to privacy is dramatically expanded. Technologies to help users maintain their privacy online are as important today as ever before-if not more so[3].

Therefore, the widespread use of information technology has created unprecedented challenges in maintaining security, trust and privacy. Several technologies and privacy policies have been developed to protect user content on social networks, however they cannot satisfy the requirements set for data *unlinkability*. Digital traces that users leave in the micro-blogging sphere provide possibilities for generating high-quality user profiles and delivering personalized services[4], since much of the information involved in the process discloses users' interests and preferences. A lot of current applications use information extracted by social networks in order to personalize advertising, search results and relevant content. Moreover, personalization techniques have been steadily improving, providing new possibilities and making behavioral profiling more accurate.

Javier Parra Arnau[5] argues that personalization allows users to deal with the overwhelming overabundance of information, however, inevitably at the expense of their privacy. The ability of online applications to profile users based on the digital

[1] Stefanos Gritzalis, *Enhancing Web privacy and anonymity in the digital era*, Information and Communication Systems Security Laboratory, Department of Information and Communications Systems Engineering, University of the Aegean, Samos, Greece.

[2] Lucy L. Thomson, Human Rights Electronic Evidence Study, *Admissibility of electronic documentation as evidence in U. S. courts*, December 1, 2011.

[3] Ian Goldberg, *Privacy Enhancing Technologies for the Internet III: Ten Years Later*, David R. Cheriton School of Computer Science, University of Waterloo.

[4] Fabian Abel, Qi Gao, Geert-Jan Houben, Ke Tao, *Twitter-Based User Modeling for News Recommendations*, Web Information Systems, Delft University of Technology.

[5] Javier Parra Arnau, *Privacy protection of user profiles in personalized information systems*, Dissertation for the Degree of Doctor of Philosophy, Dept. of Telematics Engineering, Universitat Politècnica de Catalunya, September 2013.

evidence and trace they may leave in the online world is what enables such desired personalized service, but in the meantime, poses privacy and security risks. Thus, according to Gross and Acquisti[6], while privacy may be at risk in social networks, information is willingly provided. Most users also have fundamental gaps in understanding the way applications handle their personal information, even after reading privacy policies and terms of use. Therefore, users often reveal their online activity and permit potential adversaries access to their personal information.

Personalization based on aggregation procedures is a way to infer a user's sensitive attributes through social networks. In information retrieval, a user profile is generated through the constant monitoring of the user's online interaction and activity, and represents the user's preferences. For instance, the social graph API of Twitter is an example for semantic data aggregation of social network information, which can result in behavioral profiling. The results of the collected information are tailored to this particular user's context, thus the expected outcome is relative to the specific user providing a better user experience, but at the very same time, unique challenges in the field of privacy. In a nutshell, personalization has a great impact on technologies that allow users to retrieve information from the social web, as well as on users who employ a number of services that social networking sites offer.

However, since users exhibit a lack of awareness of the risks to which they are exposed during their online activities, there is a large unmet need for a tool which performs the adversary's analysis, but also informs the user about the privacy threats he deals with. An application which supports both user and adversary's model is suggested, able to be adapted to different needs. In greater detail, the mechanism applies advanced techniques, from an attacker's perspective, in order to develop a behavioural model derived from the analysis of information-retrieval patterns. The implementation of privacy measurements allows the user to evaluate his privacy risk level and carefully adjust his online presence. Hence, the consideration of the way personal information is expected to flow in social networking sites helps the user to think of the privacy risks and proceed accordingly.

The application relies on mathematical reflections and the corresponding interpretation for each module is also provided. In particular, the privacy risk levels,

[6] R. Gross and A. Acquisti, *Information Revelation and Privacy in Online Social Networks*, Workshop on Privacy in the Electronic Society (WPES), 2005.

as well as the related measurements that have been developed in this work, are based on mature concepts of information theory.

1.2 Aim and Objectives

The objective of this thesis is twofold. Firstly, we aim to retrieve and collect useful information through Twitter, a social networking and micro-blogging service, as well as to develop a structured user-profile. Secondly, inspired by the work of J. Parra-Arnau et. al[7], we are going to apply two information-theoretic quantities in order to display and evaluate user profile privacy.

The scientific and technical objectives of this project may be more precisely depicted as follows:

- Process of retrieving and gathering information over Twitter.
- Use of aggregated information into a format suitable for applying specific data classification and correlation techniques.
- User-profile deployment and graphical presentation.
- Assessment of privacy implications.
- Measurement of the privacy of user profiles.
- Design and implementation of a graphical user interface for the presentation of privacy measurement results.
- Guidelines for the protection of user's privacy through the risk level determination.

1.3 Summary of Contributions

Recapitulating, this thesis makes the following major contributions: At first, we devise a framework based on information retrieval and correlation techniques that enables users to measure the privacy risk level of their profiles. We also implement a comprehensive application which corresponds to the theoretical background we have analyzed and employ two information-theoretic quantities as measures of the privacy of user profiles. We consider an adversary model, on which we rely our privacy

[7] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, *Measuring the Privacy of User Profiles in Personalized Information Systems*, Future Generation Computer Systems (FGCS), 2013.

metrics and propose policies for preserving privacy through the evaluation process of privacy risk. The arguments presented in our work derive from the fields of information theory, as well as Bayes decision theory.

1.4 Structure of Thesis

This master thesis is organized as follows: Chapter 1 is considered as an introductory chapter, which refers to the objectives, the basic concepts and the organizational structure of the project. Chapter 2 describes some distinctive features of social networking services and web sites, giving emphasis to Twitter and its basic functions. Chapter 3 presents basic principles related to information privacy and analyzes further the issues related to it. In addition, it explores several technologies aimed at protecting the privacy of users. Chapter 4 tackles the privacy issues of social networking sites, which also result in user profiling. Chapter 5 focuses on measuring user privacy by applying two information-theoretic quantities, Shannon's Entropy and Kullback–Leibler divergence. Moreover, it provides an analysis of the user-profile and adversary model. Building upon this adversary model, privacy metrics are interpreted. Chapter 6 addresses the work of implementing an application that corresponds to the theoretical background of the previous chapters. It illustrates the design considerations and the architecture of the tool, as well as the methodology based on which the application was developed. In Chapter 7, implementation details, including the structure of basic application components, are also depicted. Chapter 8 provides some concluding remarks concerning the project and presents several future recommendations and extensions of this work. Lastly, Chapter 9 cites the references of the relevant literature, based on which a large part of the thesis was developed.

2 Social Media and Networking

2.1 Introduction

Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and allow the creation and exchange of user-generated content[8]. A social network is a social structure consisted of individuals or organizations, known as nodes and related with one or more types of interdependency, such as common interest, friendship, relationships of beliefs, trust, knowledge, etc. In its simplest form, a social network is actually a map of specified ties, such as friendship, between the nodes being studied. The nodes to which an individual is, therefore, connected are the social contacts of that individual. The network can also be used to measure the value that an individual gets from the social network. These concepts are possible to be displayed in a social network diagram, where nodes represent the points as well as ties represent the lines[9].

The concepts of social networks have been a subject of great interest to researchers of social and behavioral sciences which explore the activities, relationships and interactions among people in the natural world. Thus, a long and complex history, relying on perceptions in many different research disciplines is what actually defines social network analysis, a broad and growing field[10]. It is considered a structural approach that focuses on the analysis, measurement and representation of relational data. A range of methods are used to define the relationships among social entities as well as their impact on other social phenomena.

Computational and mathematical models are often used to identify patterns of linkages between entities. The basic argument is the lack of an adequate description of structural concepts through the use of natural language. Hence, the use of specialized jargon and notation is often required. Much of this is obtained from graph theory, the branch of mathematics which is related with discrete relational structures[11], as

[8] Andreas M. Kaplan, Michael Haenlein, *Users of the world, unite! The challenges and opportunities of Social Media*, Kelley School of Business, Indiana University, Business Horizons (2010) 53, 59-68, available online at www.sciencedirect.com

[9] *Social networks analysis, theory and applications*, PDF generated at Mon, 03 Jan 2011.

[10] Stanley Wasserman, Katherine Faust, *Social Network Analysis: Methods and Applications, Structural analysis in the social sciences*, ENG and New York: Cambridge University Press, 1994.

[11] Carter T. Butts, *Social network analysis: A methodological introduction*, Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California, Irvine, California, USA, Asian Journal of Social Psychology (2008), 11, 13–4.

mathematical disciplines provide a formal basis for the relevant scientific empirical results. The graph theory can be applied to the description of a social network, i.e. for the description of the actors and the relationships among them which are able to denote patterns of ties.

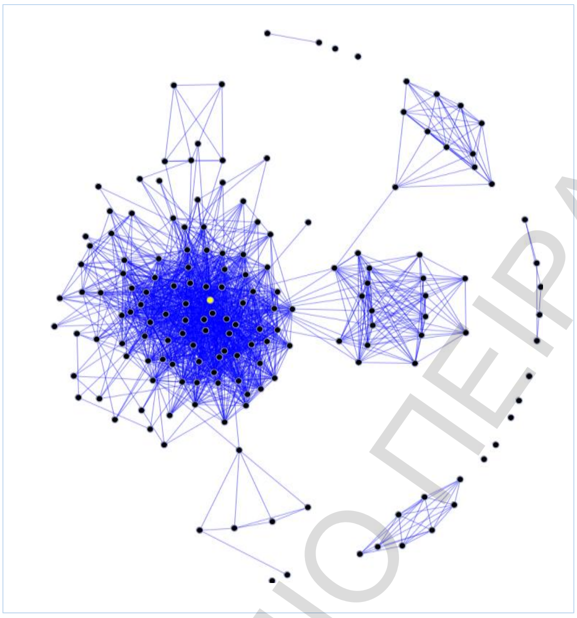


Figure 1: An example of a social network diagram[12].

Consequently, a social network is bounded by the set of entities on which it is defined[13]. Within this general definition of social networking, which was previously mentioned, there are various types of social media posing us important questions, such as the level of impact a user has, depending on his social presence. Social media are of such a high popularity that they have opened up communications across the boundaries of the world, which can actually result in opinion formation coming from multiple users connected to a social network. Moreover, the notion of influence can be more comprehensible in a case where the adversary model is present. By targeting specific users, a potential attacker is able to focus on the aggregation of personal user data and exploit his social relationship status. Therefore, new challenges arise regarding the user's privacy and protection of his personal details.

[12] *Social networks analysis, theory and applications*, PDF generated at Mon, 03 Jan 2011.

[13] Carter T. Butts, *Social network analysis: A methodological introduction*, Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California, Irvine, California, USA, Asian Journal of Social Psychology (2008), 11, 13–4.

One of the most popular sites of its kind is Twitter, whose number of users has grown fast over the last years as well as the amount of data which are stored and also available for those users. Much effort has been made to evaluate the extent of privacy threats, as the consequences of a massive data aggregation can be severe and cause negative implications for retaining privacy. Thus, the emergence of social networks, such as Twitter, has contributed in the empowerment of the attacker model and the highlighting of the privacy flaws of the information storing and gathering.

2.2 Twitter Outline

Twitter is an online social networking and microblogging service that enables users to send and read "tweets", which are text messages limited to 140 characters. It was created in March 2006 by Jack Dorsey and by July 2006, the site was launched[14]. The main concept of Twitter is the ability of making information public, if desired, through a message, text or image. Therefore, it is a real-time information network that connects users to the latest thoughts, ideas, opinions and news about what they find interesting or useful.

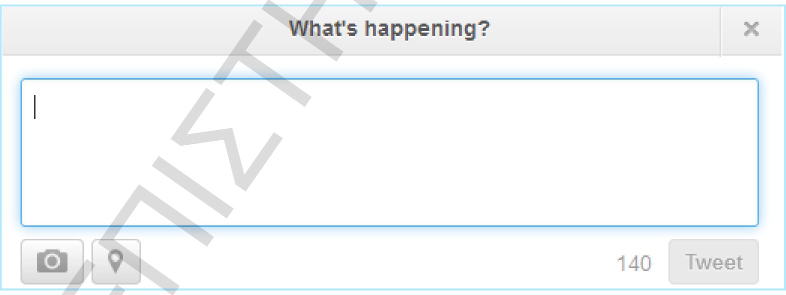


Figure 2: Composing a new tweet[15].

Each user owns a personal page and is able to receive tweets from people he follows. Thus, a user does not need permission to have access to other users' tweets, which can be constantly updated. A Twitter user is able to choose who he wants to follow, which allows him to have total control of what tweets he receives on his homepage. He can also have an unlimited amount of followers, i.e. people who follow him, but only follow and stay in touch with people he wants and cares about.

[14] Wikipedia, The free encyclopedia, *Twitter*, http://en.wikipedia.org/wiki/Twitter.

[15] *Twitter*, https://twitter.com/i/connect.

A great feature of Twitter is hashtags. A hashtag is a word or a phrase prefixed with the symbol "#". It can be used to mark individual messages as relevant to a particular group or belonging to a particular topic. In this way, a means of grouping messages is provided, since a user can search for a hashtag and get the set of messages that are related to it[16].

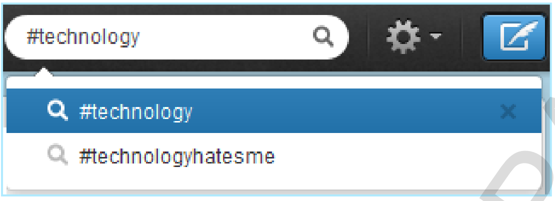


Figure 3: Searching in Twitter for recent posts containing the hashtag "#technology".

Users can also communicate with other users through tweets and create a kind of discussion. To accomplish this, the special symbol "@" is used in combination to the screen name of the user, to whom the tweet is addressed. Moreover, a user is able to share a tweet written by another user with his followers. The process of re-posting someone else's tweet is called retweet.

[16] Wikipedia, The free encyclopedia, *Hashtag*, http://en.wikipedia.org/wiki/Hashtag.

3 Internet Privacy

3.1 Basic Principles and Concepts

Privacy itself is a multifaceted concept. It was defined by Alen Westin as: *"the claim of individuals, groups, or institutions to determine for themselves when, how and to what extent information about them is communicated to others"*[17]. It is actually a fundamental human right and according to S. Warren and L. Brandeis stands for *"the right of an individual to be let alone"*[18]. It premises the protection of personal data and comes in several forms, relating to what one wishes to keep private.

The term of personal data refers to data relating to a living individual who is or can be identified either from the data or from the data in conjunction with other information that is in, or is likely to come into, the possession of the data controller[19]. Personally identifiable information (PII), as used in information security and privacy law, is a similar to personal data term and refers to information which can be used to distinguish or an individual's identity. In addition, data protection indicates the protection of personal details and enhances the rights of individuals to privacy.

Laws for the protection of privacy have been extensively adopted and tend to converge around the principle that individuals should have control over their personal information[20]. However, law regulations and amendments cannot always protect efficiently informational privacy, since it cannot be easily defined as it is vague and ambiguous. While the law supports the existence of the right to privacy with various interpretations of constitutional reforms, privacy experts attempt to further determine the right to privacy by enumerating what it allows and where it applies[21].

Information systems which collect personal information should also prevent the privacy violation. Therefore, the need for incorporating privacy requirements into the system design methodologies is immense. To meet this goal, the notion of privacy needs to be formed into a technical requirement.

[17] Wikipedia, the free encyclopedia, *Privacy*, http://en.wikipedia.org/wiki/Privacy.

[18] Samuel D. Warren and Louis D. Brandeis, *The Right to Privacy*, Harvard Law Review, Vol. 4, No. 5 (Dec. 15, 1890), pp. 193-220

[19] The Office of the Data Protection Commissioner, *Law on data protection-Data Protection Acts*, http://dataprotection.ie/ViewDoc.asp?fn=%2Fdocuments%2Flegal%2FLawOnDP.htm&CatID=7&m=l

[20] Data protection and privacy laws, https://www.privacyinternational.org/issues/data-protection-and-privacy-laws.

[21] Brian Reed, *The "Mysteries of Human Life": Dealing with an Ambiguous Right to Privacy*, Planned Parenthood v. Casey 505 U.S. 833, 851 (1992).

Review of academic research places emphasis on protection of users' privacy throughout the process of system design, in terms of eight privacy requirements namely identification, authentication, authorization, data protection, anonymity, pseudonymity, unlinkability and unobservability[22]. The first three requirements are mainly security requirements; however they are included due to their key role in the privacy protection and illegal disclosure of personal information. By addressing these requirements technical experts aim to minimize or eliminate the collection of user identifiable data[23]. Andreas Pfitzmann and Marit Hansen[24] propose in their paper a precise terminology regarding the core privacy requirements, which need to be considered while implementing Internet-based applications:

- *Anonymity* ensures that a user cannot be identified nor be tracked online within a set of users, the anonymity set.
- *Pseudonymity* refers to the use of pseudonyms[25] as identifiers. It is used in cases where anonymity cannot be implemented and the user must be accountable for his actions.
- *Unlinkability* does not allow a data controller to sufficiently distinguish two interaction steps of the same user.
- *Unobservability* prevents a data controller from determining whether an operation is being performed and ensures anonymity of the user(s) involved.

Regarding the data protection, Pinsent Masons[26] states that legal procedures apply whenever a data controller processes personally identifiable information. In order to comply with the Data Protection Act[27], a data controller must conform to the following eight principles:

[22] Christos Kalloniatis, Evangelia Kavakli, and Stefanos Gritzalis, *Addressing privacy requirements in system design: the PriS method*, September 2008, Volume 13, Issue 3, pp 241-255.

[23] Christos Kalloniatis, Evangelia Kavakli, and Stefanos Gritzalis, *Addressing privacy requirements in system design: the PriS method*, September 2008, Volume 13, Issue 3, pp 241-255.

[24] Andreas Pfitzmann, Marit Hansen, *Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management – A consolidated proposal for terminology*, 2008, http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

[25] *"Pseudonym comes from Greek "pseudonumon" meaning "falsely named" (pseudo: false; onuma: name). Thus, it means a name other than the "real name". To avoid the connotation of "pseudo" = false, some authors call pseudonyms simply nyms".* Source: Andreas Pfitzmann, Marit Hansen, *Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management – A consolidated proposal for terminology*, 2008.

[26] Pinsent Masons, International law firm, http://www.out-law.com/page-413.

[27] This information is based on UK law.

- Data should be processed fairly and lawfully and may not be processed unless the data controller can satisfy one of the conditions for processing set out in the Act.
- Data should be obtained only for specified and lawful purposes.
- Data should be adequate, relevant and not excessive.
- Data should be accurate and, where necessary, kept up to date.
- Data should not be kept longer than is necessary for the purposes for which it is processed.
- Data should be processed in accordance with the rights of the data subject under the Act.
- Appropriate technical and organizational measures should be taken against unauthorized or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.
- Data should not be transferred to a country or territory outside the European Economic Area, unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to the processing of personal data.

Hence, privacy experts wish to incorporate comprehensive policies and procedures for handling personally identifiable information, during the design of the system architecture. Several types of policies include foundational privacy principles and rules of behavior, as well as guidelines which and system-level policies. Current research, regarding the design approach, highlights the path for protecting users' privacy by developing a correct and uncorrupted environment with secure configuration. Privacy policies and associated procedures must be considered for PII incident response and data breach notification, privacy in the system development life cycle process, limitation of collection, disclosure, sharing, and use of PII, as well as the consequences of failure to follow privacy rules.

By limiting PII collections to the least amount necessary to conduct its mission, potential negative consequences in case of data privacy breaches involving PII can be prevented[28]. Therefore, protecting privacy means protecting individuals'

[28] Erika McCallister, Tim Grance, Karen Scarfone, Recommendations of the National Institute of Standards and Technology, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* NIST Special Publication 800-122, April 2010.

rights to control how personal information is collected and promulgated. Both privacy and security must be considered fundamental design goals for any system and must be factored into the specification of the system's policies, processes, architectures, and technologies. The use of privacy technologies is a way of strengthening the ability of system to protect individual privacy and secure personal information[29].

3.2 Challenges of Information Privacy

Concerns over information privacy are widespread, since people are sharing more personal information online than ever. Yun Shen and Siani Pearson[30] present in their study Solove's privacy taxonomy to identify privacy issues. The taxonomy displays a specific structure and proves that there are connections between different harms and problems:

- Information Collection

  Harms: Surveillance, Interrogation
- Information Processing

  Harms: Aggregation, Identification, Insecurity, Secondary Use, Exclusion
- Information Dissemination

  Breach of Confidentiality, Disclosure, Exposure, Increased Accessibility, Blackmail, Appropriation, Distortion
- Invasion

  Harms: Intrusion, Decisional Interference

An obvious factor in risk perception is the amount of sensitive information that is being shared online. If more personal details are being provided, perceptions of risk are likely to increase. Information collection creates potential privacy violations based on the process of data gathering and in many instances, the user is not aware of the harms incurred by such processes. Hence, information that may be especially sensitive could be exposed. Also, privacy issues relating to information dissemination arise from the revelation of personal data or the threat of spreading information.

[29] Smart Card Alliance, *Privacy and Secure Identification Systems: The role of smart cards as a privacy-enabling technology*, ID-03001, February 2003.

[30] Yun Shen, Siani Pearson, *Privacy Enhancing Technologies: A Review*, HP Laboratories, HPL-2011-113.

According to G.W. van Blarkom et al.[31], the security concern is about the protection of the self, and the related ones, in the physical, mental, self-esteem, social, reputation and other senses against threats from the outside world. This concern includes misuse and abuse of personal data in the widest sense: not just static information, but also information of a dynamic nature could be revealed. This security issue not only embodies the abuse of personal data, but also covers the reception of threatening and unwanted information.

The person involved must be informed in case of data collections and further processing, as information is only allowed to be collected for specific, explicit and legitimate purposes. However, more precautions need to be taken to ensure data protection and users' privacy. Controlled data disclosure, such as identity management or privacy enabling techniques can contribute to the limitation of privacy violation issues, which are caused by aggregation[32].

Also, another challenging task is quantifying privacy risk. Identifying privacy risks and providing integrated solutions to reduce or deter privacy loss is one of the fundamental issues of this field. Therefore, there is still breeding ground for measuring privacy associated with users' online activities and providing a set of recommendations to mitigate privacy threats, as well.

3.3 Privacy Protection Techniques

3.3.1 Privacy by Design

The privacy issues triggered by each technology can be identified early in the system's life cycle and those same issues can be properly aligned across multiple projects to ensure that the application, as a whole, uses technology in a way that consistently complies with privacy protection requirements[33]. Considering the person who uses the application as the initiator of the interaction with the system

[31] G.W. van Blarkom, J.J. Borking, J.G.E. Olk, *Handbook of Privacy and Privacy-Enhancing Technologies - The case of Intelligent Software Agents*, Privacy Incorporated Software Agent (PISA) Consortium, the Hague, the Netherlands, 2003.

[32] Yun Shen, Siani Pearson, Privacy Enhancing Technologies: A Review, HP Laboratories, HPL-2011-113.

[33] Privacy Office, U.S. Department of Homeland Security Washington, *Privacy Technology Implementation Guide*, August 16, 2007.

architecture, of great importance is to convince users that the application will not infringe their privacy and that accurate measures of data protection are in place.

Privacy enabling procedures need to be incorporated on the process of system design, while important decisions related to privacy and data protection need to be consciously taken by the user[34]. Hence, privacy requirements must be employed as a part of the design cycle for successfully addressing privacy vulnerabilities, as well as techniques that conduce to eliminating the identified risks. The outcome is that privacy becomes a crucial component of the core functionality. Thus, it is integral to the system, without diminishing functionality[35].

3.3.2 Privacy Enhancing Technologies

Echoing the need for privacy, specific mechanisms can be implemented focusing on the privacy of the users and the protection of their personal data. These technological solutions are extensively known under the name *Privacy Enhancing Technologies* (PETs). There is no commonly accepted definition of Privacy enhancing technologies, although G.W. van Blarkom et al.[36] describe them as «a system of ICT measures protecting informational privacy by eliminating or minimizing personal data thereby preventing unnecessary or unwanted processing of personal data, without the loss of the functionality of the information system»[37].

PETs are sometimes thought of as substitutes for other instruments of privacy protection, such as laws and regulatory bodies that enforce and implement legislation. However, PETs are better thought of as complementary to other instruments with which they must cooperate in order to provide a robust form of privacy protection. Relative to privacy issues legislation is, first and foremost, the instrument to which PETs must relate, incorporating legal principles into technical specifications.

[34] Andreas Krisch, *RFID Privacy Issues*, Contribution to the RFID Expert Group Meeting, 10 July 2007.

[35] Ann Cavoukian , *Privacy by Design, The 7 Foundational Principles, Implementation and Mapping of Fair Information Practices*, Information & Privacy Commissioner, Ontario, Canada.

[36] G.W. van Blarkom, J.J. Borking, J.G.E. Olk, *Handbook of Privacy and Privacy-Enhancing Technologies - The case of Intelligent Software Agents*, Privacy Incorporated Software Agent (PISA) Consortium, the Hague, the Netherlands, 2003.

[37] G.W. van Blarkom, J.J. Borking, J.G.E. Olk, *Handbook of Privacy and Privacy-Enhancing Technologies - The case of Intelligent Software Agents*, Privacy Incorporated Software Agent (PISA) Consortium, the Hague, the Netherlands, 2003.

PETs can be classified into the following five categories: basic anti-tracking technologies, cryptography-based methods from private information retrieval (PIR), TTP-based approaches, collaborative mechanisms and data-perturbative techniques[38]. A lot of these technologies may in fact be combined to result in having the highest possible level of privacy.

Basic Anti-Tracking Technologies

Tracking technologies are able to identify users, in the sense of accurately measuring the location and orientation as users move and interact across different sessions or multiple web domains. Tracking mechanisms are fundamental components of personalized services, as they enable systems to tail after users and thus, facilitate user profiling.

The Internet allows tracing online activities of users, even though this operation may raise few concerns. Tracking is not a single mechanism, but rather a combination of one or more individual approaches[39] and can be implemented at different levels of communication. For instance, an Internet Service Provider is able to uniquely identify a user through the source IP address, which can be used as a reference for the user's geographical location, whereas a Personalized Service Provider may use a cookie to associate a user with any previous online activity. Thus, user’s anonymity can be endangered as websites can process personal data or facilitate the identification of a user’s behavior.

However, there are several techniques to prevent an adversary from tracking a user, such as covering or even blocking the parameters employed to identify online activities of users. Enabling dynamic IP addresses and rejecting hypertext transfer protocol (HTTP) cookies are two basic methods to avoid tracking. The identification of users through IP addresses actually fails when a large number of users share a single IP address, while rejecting HTTP cookies may be an alternative policy to preserve privacy. The problem of the second approach is that it can disable other web services and might reduce the effectiveness of personalized ones.

[38] Javier Parra Arnau, *Privacy protection of user profiles in personalized information systems*, Dissertation for the Degree of Doctor of Philosophy, Dept. of Telematics Engineering, Universitat Politècnica de Catalunya, September 2013.

[39] MozillaZine, *User Tracking*, http://kb.mozillazine.org/User_tracking.

Some additional mechanisms, which increase the level of protection against cookies regarding the anonymity of users, are Crowds, Onion Routing, Tor, Hordes and Freedom. These PET entities provide the user the option to enable or disable cookies, and therefore offer protection against this potential threat.

Private Informational Retrieval

Private Information Retrieval (PIR) refers to cryptography-based protocols that enable a user to privately retrieve a data item of a database, in such a way that the database administrator is not aware of which particular item was retrieved. There are two primary classes of PIR schemes: information-theoretic PIR, and computational PIR. In the case of the first one, an attacker is unable to determine any information related to the user's query, even if he owns unlimited computing power. In the latter, the privacy of the query is preserved only against adversaries restricted to polynomial-time computations[40].

A naive approach would be to allow users to obtain a copy of the entire database and retrieve the desired data item locally. This solution, known as *trivial* PIR scheme, is impractical and involves a tremendous communication overhead. However, it provides users with the maximum level of privacy in the information theoretic sense[41].

Most developed and deployed privacy-enhancing technologies, such as Onion Routing and Mix networks, are limited to keeping private the identity of users through anonymization[42]. PIR protocol, on the other hand, is able to provide protection at important application domains, by keeping retrieval information private. It actually aims at transferring less data, while still preserving user privacy. In the context of Web search, it allows a user to seek out information in an online database without informing the database provider regarding the search query or response.

TTP-Based Mechanisms

[40] Meredith L. Patterson, Len Sassaman, *Subliminal Channels in the Private Information Retrieval Protocols*. In: Proceedings of the 28th Symposium on Information Theory in the Benelux, 2007.

[41] Sergey Yekhanin, *Locally Decodable Codes and Private Information Retrieval Schemes*, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, July 2007.

[42] Femi Olumofin, Ian Goldberg, *Revisiting the Computational Practicality of Private Information Retrieval*, Cheriton School of Computer Science, University of Waterloo.

A trusted third party (TTP) is an entity which facilitates interactions between two different parties. Hence, another way to protect user privacy is to use a TTP-based technology as an intermediary between the user and the untrusted personalized information system. In the context of this scenario, the external system is not aware of the user ID and only the identity of the TTP involved in the communication is revealed. Therefore, TTP-based approaches are quite common because, in general, they offer a reasonable trade-off between efficiency, accuracy and privacy[43].

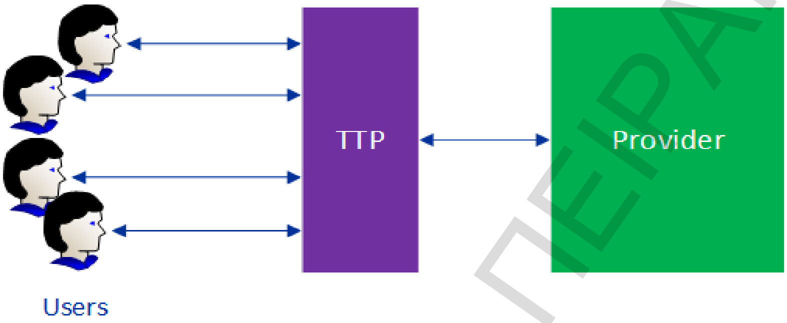


Figure 4: A TTP-based scheme[44].

One of the flaws of this technology is that personalization services are restricted, since the TTP forwards personal information of the user on his behalf. A solution to this issue could be the use of *pseudonymizers*. Pseudonymisers receive queries from users and, prior to forwarding them to the external system, they replace the real IDs of the users by fake ones, pseudonyms. In such way, pseudonymizers conceal the real user IDs, while corresponding pseudonyms forward the replies from the providers to the users.

An alternative TTP-based approach to protect users' identities from an untrusted service provider is the use of *anonymizers*. An anonymizer is an aid to use services anonymously within a network and filters all directly identifiable personal data from the data that is required to establish connections. Thus, this technique replaces the information with information that does not trace back to the user. In the current scenario, such a mechanism is implemented between the user and the service

[43] Agusti Solanas, Josep Domingo-Ferrer, Antoni Martınez-Ballest, *Location Privacy in Location-Based Services: Beyond TTP-based Schemes*, 1st International Workshop on Privacy in Location-Based Applications (PILBA 2008) within 13th European Symposium on Research in Computer Security (ESORICS 2008), Malaga, Spain, Oct 2008. ISBN: 1613-0073.

[44] Muhamed Ilyas, Dr. R. Vijayakumar, *LPM: A distributed architecture and algorithms for location privacy in LBS*, International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.

provider. It acts as intermediary, receives the requests from users requesting to use the services and filters the identification information from the request headers before forwarding to the servers of the service provider. The identities of the clients are not revealed to the service provider, but only to the Anonymizer service, which is the only component that needs to be trusted. Anonymizers aim to preserve users' anonymity and as a result their privacy.

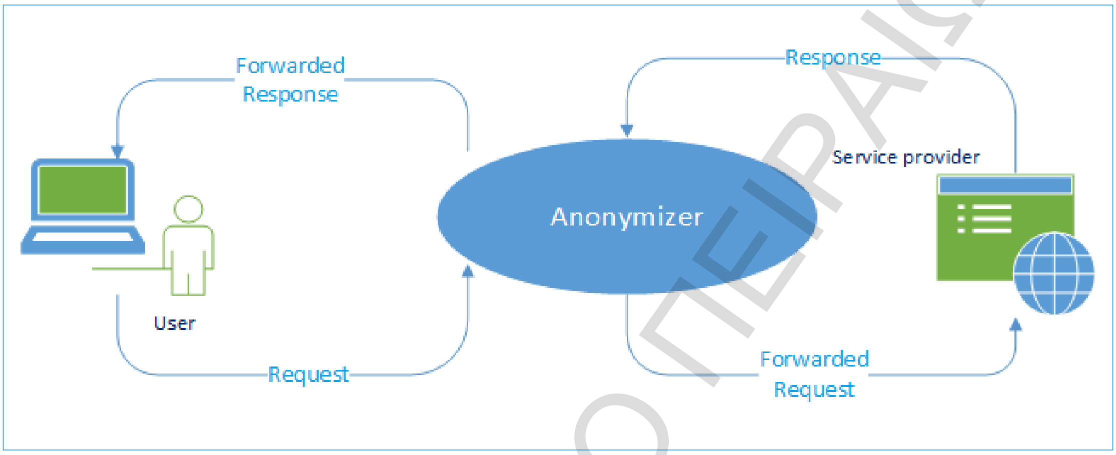


Figure 5: The Anonymizer.

Digital credentials also provide fine-grained privacy control at every step in the life-cycle of certified personal information that is being managed[45]. Granted by a TTP, they provide users sufficient privileges to accomplish a particular transaction without completely revealing their identity. Moreover, TTP is not required to be online at the time of service access.

As mentioned above, Mix nodes can be employed to offer privacy protection. A Mix node is a processor which receives a number of messages as input, modifies their appearance and timing by using some cryptographic transformation, and outputs a randomly permuted list of function evaluations of the input items. This process is accomplished without revealing the relationship between input and output elements[46].

One more technology based on TTP, which was previously mentioned, is *onion routing*. Goldschlag et al.[47] introduced onion routing as a means to establish an

[45] Stefan Brands, *Non-Intrusive Identity Management*, McGill School of Computer Science & Credentica, March 23, 2004.

[46] Leticia Fernández Franco, *A survey and comparison of anonymous communication systems: anonymity and security*, Universitat Oberta de Catalunya (UOC), June 2012.

[47] David M. Goldschlag, Michael G. Reed, Paul F. Syverson, *Hiding Routing Information*, Workshop on Information Hiding, Cambridge, UK, May, 1996.

anonymously redirected encrypted path through a network with full control of routing decisions and identity disclosure left in the hands of the sender. This approach is called onion routing because it relies upon a layered object to direct the construction of an anonymous, bidirectional, real-time virtual circuit between two communication parties, an initiator and responder, and is presented as a flexible, communication infrastructure that is resistant to both eavesdropping and traffic analysis.

To use a network of onion routers, users randomly choose a path through the network and construct a circuit - a sequence of nodes which will route traffic[48]. Because individual routing nodes in each circuit only know the identities of adjacent nodes, and because the nodes further encrypt multiplexed virtual circuits, studying traffic patterns does not yield much information about the paths of messages.

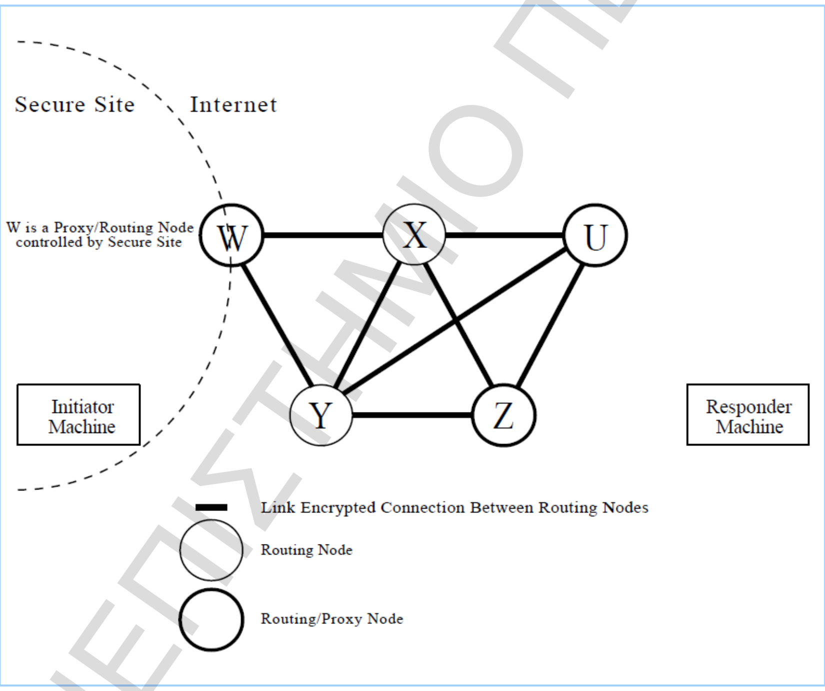

Secure Site
Internet
W is a Proxy/Routing Node controlled by Secure Site
W
X
U
Y
Z
Initiator Machine
Responder Machine
Link Encrypted Connection Between Routing Nodes
Routing Node
Routing/Proxy Node

Figure 6: Onion Routing Topology[49].

Although onion routing may be used for anonymous communication, it differs from anonymous remailers in two ways: communication is real time and bidirectional, and the anonymous connections are application independent[50].

[48] Aniket Kate, Greg M. Zaverucha, Ian Goldberg, *Pairing-Based Onion Routing with Improved Forward Secrecy*, David R. Cheriton School of Computer Science, University of Waterloo.

[49] David M. Goldschlag, Michael G. Reed, Paul F. Syverson, *Hiding Routing Information*, Workshop on Information Hiding, Cambridge, UK, May, 1996.

User Collaboration

One popular approach based on user collaboration is the *Crowds* system, which was named by the operation of grouping users into a large and geographically diverse group (crowd) that collectively issues requests on behalf of its members[51]. The functionality of Crowds system relies on the principle of enabling communication, while protecting the anonymity of the stakeholders.

Main goal of the Crowds is the anonymity, which is provided at the level of intermediate nodes and final recipient. A crucial element that the Crowds system addresses is the degree of anonymity[52] provided. Actually, anonymity as a requirement is not accurate while designing a system. The definition of the expected anonymity degree is important, as it can vary from absolute privacy to provably exposed, where the attacker can evince the identity of the receiver. By using degrees of anonymity, the anonymity properties provided by this technology are characterized against several classes of attackers.

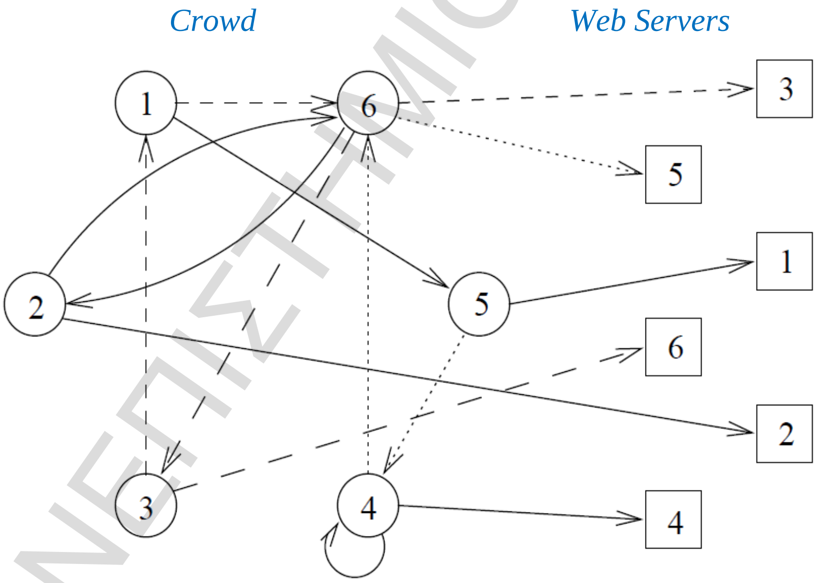


Figure 7: Communication Flow within the Crowd[53].

[50] Ashish T. Bhole, Savita H. Lambole, *Design and Implementation of Distributed Security using Onion Routing*, International Conference & Workshop on Recent Trends in Technology, (TCET) 2012.

[51] Michael K. Reiter, Aviel D. Rubin, *Crowds: Anonymity for Web Transactions*, AT&T Labs-Research.

[52] *Degrees of Anonymity: Absolute privacy, beyond suspicion, provable innocence, possible innocence, exposed, provably exposed*. Source: Michael K. Reiter, Aviel D. Rubin, *Crowds: Anonymity for Web Transactions*, AT&T Labs-Research.

[53] Michael K. Reiter, Aviel D. Rubin, *Crowds: Anonymity for Web Transactions*, AT&T Labs-Research.

There are also certain types of attacks that the Crowds system has limited effect. Common source for those types of attacks are malicious nodes which act against the crowd they belong in. Crowd protocol is also unable to protect user privacy against the collusion of all participants. Finally, another important drawback is the additional traffic intrinsic to this forwarding mechanism[54].

These deficiencies are actually present in most of the PETs that act effectively on user collaboration. Another attempt to overcome these drawbacks has been made by A. Erola et al.[55], who propose a variation of the original Crowds protocol. In particular, they present a P2P protocol, which exploits social networks in order to protect the privacy of users from profiling mechanisms. Essentially similar to Crowds, this scheme aims at grouping users with contiguous interests based on social networks.

In the scenario of personalized Web search, a methodology[56] to distort user profiles against an external observer is illustrated. The main concept lies on combining original queries with false ones in order to obfuscate profiles of interests. A shortcoming regarding query forgery is the distinction of false queries from the real ones[57].

Data perturbation

The perturbation method attempts to preserve privacy in case an adversary aims at obtaining a particular user profile. The original (private) profile is distorted to reduce the risk of user profiling. Therefore, in a social network, nodes that look structurally similar may be indistinguishable to an adversary, despite external information. A certain level of anonymity is accomplished through structural similarity.

It is assumed that an attacker can leverage semantics of profile data and background knowledge related to the published data, in order to measure the behavior

[54] Javier Parra Arnau, *Privacy protection of user profiles in personalized information systems*, Dissertation for the Degree of Doctor of Philosophy, Dept. of Telematics Engineering, Universitat Politècnica de Catalunya, September 2013

[55] A. Erola, J. Castella-Roca, A. Viejo, and J. M. Mateo-Sanz, *Exploiting social networks to provide privacy in personalized Web search*, J. Syst., Softw., vol. 84, no. 10, pp. 1734-745, 2011.

[56] D. Rebollo-Monedero, J. Forné, *Optimal query forgery for private information retrieval*, IEEE Trans. Information Theory, vol. 56, no. 9, pp. 4631-4642, 2010.

[57] D. Rebollo-Monedero, J. Forné, A. Solanas, A. Martinez-Ballesté, *Private Location-Based Information Retrieval through User Collaboration*. Computer Communications, 33 (6): 762-774, 2010.

of users and determine patterns. However, a noise model can be constructed that exports realistic-looking parameters, but also a pair of conflicting factors. Hence, the resulting perturbed profile can be aggregated with that of others. In such way, the approach preserves individual user privacy, even though allows accurate construction of statistical results that refer to the overall[58].

In the scenario of personalized Web search, query forgery is an approach of data perturbation, where forged queries are generated on behalf of the user. By receiving both genuine and false queries, the search engine which in this case is the adversary, would not be able to obtain an accurate user profile. A widely known application which employs query forgery is TrackMeNot. It is referred to a web browser extension that generates false queries and sends them to different search engines on behalf of the user.

To summarize, degrading the quality of information about a user’s queries or interests, as appropriate, has as a result the enhancement of user’s privacy. The elusive information that is introduced consists in making it difficult for an adversary to deploy malicious activities.

[58] Charu C. Aggarwal, Tarek Abdelzaher, *Integrating sensors and social networks*, Chapter 14.

4 User Privacy in Social Networks

4.1 Introduction

Online social networks have made a huge impact on human interaction and have profoundly transformed the way in which personal information is stored, processed and used. Users around the world are able to create integrated personality profiles, as well as express and share ideas, thoughts, feelings or personal details with online friends. Thus, social networks include highly sensitive information, as they enable in-person communication and encourage users to reveal more private information than they would otherwise[59]. The storage and process of personally identifiable information pose a number of risks to users' privacy, as it is possible to be used for various means of personalization.

Several social networking sites, such as Facebook, do not encourage the virtual contact and communication with strangers. On the contrary, they aim to set up a communication with real-life friends. Therefore, the use of pseudonyms and the desire of anonymity contradict the main purpose of most social networks, which is actually the idea of self-exposure. Moreover, the user himself is able to control his online presence and account, by defining the amount of personally identifiable information he shares, as well as through privacy settings provided by the social networking sites. However, even though a lot of users remain hesitant when it comes to sharing all manner of content through social platforms, they do not always realise that careful curation of the information they post online is essential to protect their privacy. In addition, privacy settings are generally not sufficiently understood by the average users who seldom change the default configuration and this method does not actually prevent the social networking site itself from gathering the sensitive user data[60].

As a result, social networking raises concerns about the efficiency of privacy policies in force and the impact of sharing personal information online. Questions arise about the overexposure of the users and the related to information privacy implications as well. Social-based personalization is also presented as a prominent

[59] Alessandro Acquisti, Ralph Gross, R., *Imagined communities: Awareness, information sharing, and privacy on the Facebook*, PET 2006.

[60] Steven Furnell, Costas Lambrinoudakis, Javier Lopez (Eds.), *Trust, Privacy, and Security in Digital Business*, 10th International Conference, TrustBus 2013, LNCS 8058, pp. 62–73, Springer-Verlag Berlin Heidelberg 2013.

example of privacy risks[61]. Hence, the capability of social media to endow an adequate and sufficient response to the challenges, which are posed in the context of sharing and processing personal information, needs to be evaluated.

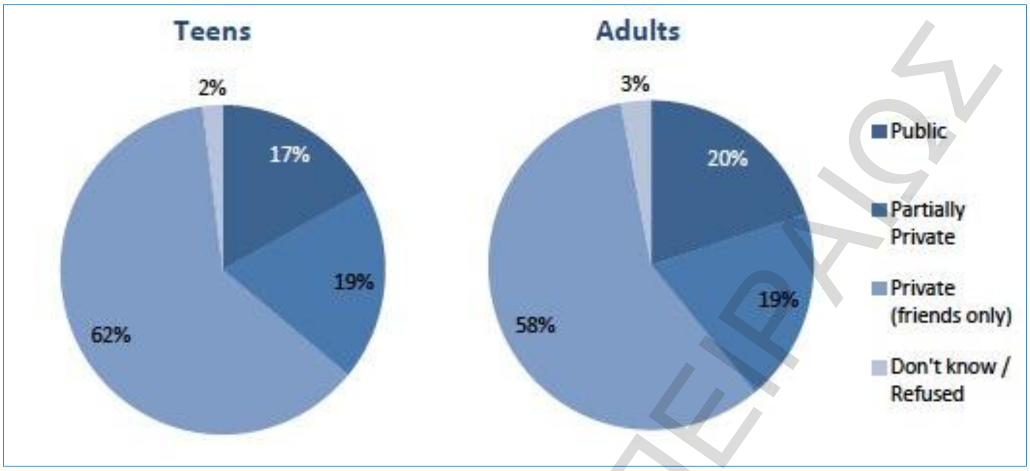


Figure 8: Teens' and adults' privacy settings on social media sites[62].

Current research and scientific articles from multiple fields, such as philosophy, sociology, psychology and data protection regulation, attempt to analyse different aspects of this phenomenon and establish new rules to better cope with privacy risks derived from social media.

4.2 Privacy Issues in Twitter

In addition to the benefits of using social network sites, such as Twitter, there might be some risks associated with the misuse of personal data obtained online. Privacy preservation concerns both user data content and relevant information, as these information items may be attractive targets for privacy invading attacks. As mentioned in previous section, Twitter is a popular micro-blogging and social network service that allows people to share messages of 140 characters in length. While Twitter permits people to share information among friends or followers, the

[61] Eran Toch, Yang Wang, Lorrie Faith Cranor, *Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems*, Published online: 10 March 2012, DOI 10.1007/s11257-011-9110-z.

[62] The Pew Research Center's Internet and American Life Project, April 26 – May 22, 2011 Spring Tracking Survey.

default privacy setting displays all messages as public. Thus, anyone who signs up for Twitter is able to see them, as they may be posted to a public timeline website[63].

The enhancing personalization using information derived from social networking sites has been increased by the introduction of the Application Programming Interfaces (API)[64]. Twitter also offers an API, which can be used to export and collect information for diverse purposes, such as personalization. Information retrieval and processing are integral part of inferring essential information related to the online behavior and activities of users. Therefore, this process aims at profiling users, in the sense of aggregating information about an individual user.

Privacy concerns may arise due to the digital storage and process of personal information, in a way that the rightful owner can be harassed or for purposes he could never have thought or approved[65]. Anyone can have access to public tweets, since there are no strict privacy policies regarding the relationships of friendship among users. Also, users may be uninformed about the possibility of changing the default settings or may be worried about the implications the various settings may have. Therefore, even though individual messages may not seem incriminating, but instead harmless, they can be easily retrieved and aggregated with other kinds of personal information. The results consists in obtaining quite rich information related to the location of users, their personality, their interests and favorite topics of discussion as well.

In a nutshell, to address privacy concerns derived from Twitter, and in general, social networks, several technological measures have been developed, such as PETs, aiming to protect published information from unauthorized audiences and raise the users' awareness when it comes to sharing personally identifiable information. As mentioned in the previous chapter, PETs include a wide range of mechanisms that consist in protecting users' privacy; a complex issue involving several stakeholders, such as users, PETs industry and developers, social networking providers, regulatory

[63] Lee Humphreys, Phillipa Gill, Balachander Krishnamurthy, *Privacy on twitter-How much is too much? Privacy issues on Twitter.*

[64] Eran Toch, Yang Wang, Lorrie Faith Cranor, *Personalization and privacy: A survey of privacy risks and remedies in personalization-based systems*, Published online: 10 March 2012, DOI 10.1007/s11257-011-9110-z.

[65] Lee Humphreys, Phillipa Gill, Balachander Krishnamurthy, *Privacy on twitter-How much is too much? Privacy issues on Twitter.*

bodies and third parties. Therefore, there is a constant need for combined solutions, in which various technologies and targeted regulatory guidelines conspire to create a system with adequate enforcement and control powers[66].

[66] Steven Furnell, Costas Lambrinoudakis, Javier Lopez (Eds.), *Trust, Privacy, and Security in Digital Business*, 10th International Conference, TrustBus 2013, LNCS 8058, pp. 74–84, Springer-Verlag Berlin Heidelberg 2013.

5 Measuring the Privacy of User Profiles

5.1 User - Profile Model

In the motivating scenario, users publish short messages on Twitter to share their thoughts and various events of their daily life. The digital traces they leave can be used to model user preferences and deliver personalized services. Therefore, the information revealed, such as specific topics of interest, implies the extraction of a profile of interests or *user profile*.

While conducting a security analysis, the profile of the user, which can be carefully examined by an external attacker, has to be taken into consideration. Several parameters can be estimated to prevent the adversary from acting maliciously, such as detecting temporal patterns in the profiles or violating users' rights to privacy, since behavioural tracking relies heavily on users' personal information.

In the current use case scenario, a user is able to utilize a web-based application on the condition he owns a Twitter account. The application focuses on Twitter users and analyzes their tweets in order to generate statistical results, through a range of classification techniques, according to main topics of discussion. Hence, the user's degree of interest on each topic is measured depending on the number and content of tweets he posts with relation to these topics. The user's profile is presented through a graphical format, as it is depicted in Figure 9.

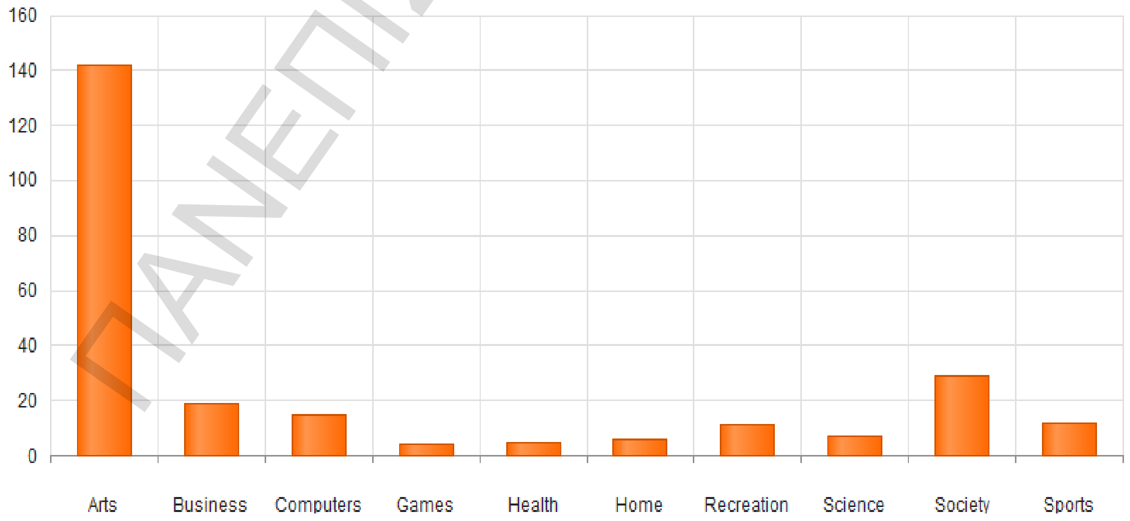


Figure 9: Modeling a User Profile with *Privacy Measurement and Analysis* application.

In particular, the user model is displayed as a bar chart, where the user's preferences are plotted vertically on the horizontal axis. The other axis represents a discrete value, the frequency of each topic. The size of each bar is proportional to the popularity of each category in the user profile.

The classification of user data is an element to be considered in the definition of the user-profile model and, hence, in the adversary model. Privacy metrics shall be analyzed in the next sections in accordance with the user profile model, which poses a considerable threat to individual privacy.

5.2 Adversary Model

A security case study requires the analysis of adversary behavior in order to incorporate sufficient protection to the threatened system. For developing relevant outcomes, potential adversarial behavior is modeled. An attacker is considered as any entity capable of having access to user's tweets, conducing to obtain his profile model. However, it is crucial to note that the adversary typically has to operate within the constraints of the environment provided by the target environment[67].

Adversary modeling is very important, as understanding the goals of the attacker contributes to provide a set of security and privacy considerations. Also, privacy is quantified according to this entity. An effort must be made to develop new methods based on understanding adversary behavior, procedures as well as techniques. However, depending on the adversary, protection guidelines could vary. All threat scenarios is very difficult to be covered, thus, the used metrics may significantly diverge, depending on the assumptions made about the attacker.

In the context of the proposed case study, the intention of adversary is to track users over time and develop profiles of their interests, characteristics, and main topics of discussion. Tracking is also used for diverse types of aggregate measurements, such as website traffic statistics or effective exposure of advertising[68]. In current use case scenario, users' online activities are recorded and correlated. In addition, the information conveyed allows the attacker to extract a profile of actual interests for a particular user and compute the corresponding graph chart.

[67] John Lowry, Rico Valdez, Brad Wood, *Adversary Modeling to Develop Forensic Observables*.

[68] Claude Castelluccia, Arvind Narayanan, *Privacy considerations of online behavioural tracking*, European Network and Information Security Agency (ENISA), October 2012.

A user profile can sometimes be extremely detailed. Thus, the obfuscation of users' interests can also be considered as an indirect way to achieve a certain level of privacy protection. Bearing in mind the model of user profile, which was defined in the previous section, a user could deface the attacker's concept by modifying the content of his tweets, i.e. his topics of discussion. Hence, this action would result in the perturbation of the user profile and his bar chart of interests as well. Thereby, the adversary could not obtain much valuable information, since the genuine profile of the user would have been altered. The attacker believes that the observed behavior characterizes the actual user's profile; however, the graph chart does not reflect the actual preferences of the user. In the relevant literature, the perturbed profile is referred as the user's apparent profile and it differs from the actual user profile.

The adversary model also depends on what capabilities or intentions the attacker is presumed to have[69]. In accordance with our case scenario, the main goal of the privacy attacker is the *identification or individuation* of the user. In the context of the scenario considered, the adversary attempts to *identify* a user, in the sense of distinguishing him from the rest of the population, by detecting deviations between the user's interests with respect to the average profile of the population. Users' actions are being tracked and any entity capable of profiling users based on the information they disclose is regarded as adversary.

5.3 Privacy Metrics

5.3.1 Entropy and Divergence as Measures of Privacy

In this project, two fundamental information-theoretic quantities are used for the measurement of privacy. According to J. Parra-Arnau et al.[70], the Shannon entropy and Kullback–Leibler (KL) divergence reflect the intuition that an adversary is able to compromise user privacy as long as the apparent user profile diverges from the uniform profile. The interpretation of both metrics is performed to provide a set of arguments about their usage as privacy level parameters. The symbol H will denote entropy and D relative entropy or Kullback-Leibler (KL) divergence.

[69] Wikipedia, the free encyclopedia, *Adversary*, http://en.wikipedia.org/wiki/Adversary_(cryptography)

[70] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, *Measuring the Privacy of User Profiles in Personalized Information Systems*, Future Generation Computer Systems (FGCS), 2013.

User privacy is measured as the Shannon entropy of the apparent user distribution. The *entropy* H(p) of a discrete random variable X with probability distribution p is a measure of its uncertainty, defined as

$$\mathrm{H}(X) = \text{-E} \log p(X) = -\sum_{x} p(x) \log p(x) .$$

Given two probability distributions p(x) and q(x) over the same alphabet, the KL *divergence* D(p||q) is defined as

$$\mathrm{D}(\mathrm{p}||\mathrm{q}) = \mathrm{E}_p \log \frac{p(X)}{q(X)} = \sum_{x} p(x) \log \frac{p(x)}{q(x)} .$$

In information theory, the KL divergence is often referred to as *relative entropy*, as it may be regarded as a generalization of entropy of a distribution, relative to another. Therefore, entropy is a special case of KL divergence, as for a uniform distribution u on a finite alphabet of cardinality n,

$$\mathrm{D}(p||u) = \log \mathrm{n} - \mathrm{H(p)}.$$

The KL divergence provides a measure of discrepancy between probability distributions, assuring that D(p||q) $\geq$ 0, with equality if, and only if, $p = q$. Hence, relation D$(p||u)$ = log n - H(p), between entropy and KL divergence, implies that H(p) $\leq$ log n, with equality if, and only if, $p = u$. Consequently, as stated in the research of D. Rebollo-Monedero et al., *entropy maximization* is a special case of *divergence minimization*, ideally achieved when the distribution obtained as optimization variable is identical to the reference one.

5.3.2 Privacy Metrics against Identification

Starting point of the identification issue is to assume that the adversary aims to individuate a user in the sense of attempting to distinguish him from the population of users. Sound arguments regarding the use of entropy and divergence as measures of privacy are expounded by Edwin T. Jaynes' rationale about entropy maximization methods[71].

[71] Edwin T. Jaynes, *On the rationale of maximum-entropy methods*, Proc. IEEE, vol. 70, no. 9, pp. 939{952, Sep. 1982.

According to J. Parra Arnau[72], the key idea under Jaynes' rationale is that the entropy of an apparent user profile, modeled by a diagram of frequencies across categories of interests, may be regarded as a measure of privacy, or more accurately, anonymity. As stated, the main concept is that the method of types from information theory establishes an approximate monotonic relationship between the likelihood of a PMF[73] in a stochastic system and its entropy. In a nutshell, the higher the entropy of a profile, the higher is its probability, and therefore greater is the number of the users who behave according to it. Consequently, the entropy is appropriately considered as a measure of anonymity; however, not in the sense that the user's identity remains unknown, but only in the sense that the user's profile is more private, since it is assumed as more common and less interesting to an adversary whose objective is to target particular users.

In his work, Javier Parra Arnau also states that KL divergence is a measure of discrepancy between probability distributions, which includes Shannon's entropy as the special case when the reference distribution is uniform. If the distribution of the population's average profile is known, the divergence between the user and the population's profile constitutes a measure of privacy, implying that the lower the divergence is, the more private the profile can be considered. Thus, KL divergence is similarly regarded as a measure of anonymity, in the sense that a profile of interests which matches the population's, does not require perturbation.

To summarize, in the context of the identification problem, the modification of the user's profile, in a way that approaches the population's average profile, results in the minimization of KL divergence, and therefore in preserving a certain level of user anonymity. Under this interpretation, more effort is needed, from the attacker's perspective, to distinguish a given user from the population of users. A lower divergence also implies a lower privacy risk, or more precisely, a lower anonymity loss. Maximizing the Shannon's entropy also allows the user to be unnoticed. Thus, KL divergence and Shannon's entropy are interpreted as privacy metrics under the assumption that the adversary aims at targeting users who diverge from the average profile.

[72] Javier Parra Arnau, *Privacy protection of user profiles in personalized information systems*, Dissertation for the Degree of Doctor of Philosophy, Dept. of Telematics Engineering, Universitat Politècnica de Catalunya, September 2013.
[73] PMF stands for Probability Mass Function.

6 Twitter Profiling and Privacy Measurement Tool

6.1 Introduction

The purpose of this work is the implementation of an application, which profiles Twitter users and measures their privacy taking into consideration two fundamental information-theoretic quantities, Shannon's entropy and KL divergence. The application is capable of displaying results about the privacy risk level of the users and aims to address privacy challenges from a technical perspective. In particular, the proposed tool supports the following functionalities:

- Search of results for a particular Twitter user.
- Retrieval of relevant tweets.
- Categorization of a user's tweets with respect to his interests and database development for storing information.
- Development of a diagram of interests regarding the user's profile.
- Measurement of a user's privacy according to information theory parameters.

It presents a comprehensive view regarding privacy concerns derived from social networks, such as Twitter. Of great importance is to illustrate an effective privacy indicator in terms of measuring privacy risk or gain. Online profiling is perceived as a threat to privacy, as it relies heavily on users' personal interests. Linking together every piece of information that can be retrieved about an individual poses serious issues, since the user himself is often not aware of the danger that lies at the digital traces he leaves. The current application presents a way of user profiling and displays results about a user's privacy, inducing him to take into consideration the severity of risks he deals with and protect himself appropriately.

6.2 Design Considerations

The application was developed under the assumption that the adversary has complete access to the functionalities of the tool, which were stated in the previous section, and also owns a Twitter account. He is capable of obtaining information that is relevant to a user's interests and measuring the privacy level of this user. Therefore, this application is available to everyone, such as simple users who aim to audit and manage their profile, and also measure its privacy, as well as potential adversaries who take advantage of the tool in order to target users who vary from the population of users.

Therefore, the profiling activities of users on Twitter are recorded and correlated at any time, due to the great volume of user information to which an attacker has access. The information retrieved allows him to develop user patterns. However, a simple user is familiar with the capabilities of the adversary, who is limited to specific computational methods and techniques, and is able to protect himself. He is able to realize the risks he deals with and modify his actual profile. Moreover, population's average profiles are also accessible, in the sense that the discrepancy between these profiles and the user profile can be easily estimated. Hence, the crucial difference between the two categories of users is their pure intentions while using the application.

6.3 Methodology and Architectural Design

In this section, the main components structuring the application are displayed in great detail. The architectural design and behavior of the application is depicted, focusing on the way the components interact with each other. The analysis is performed with respect to the basic elements involved, as well as the objectives raised at the beginning of the project.

6.3.1 Information Retrieval in Twitter

Users share tweets of different content and size. The amount of tweets published can be enormous; however obtaining a specific number of tweets is not straightforward. Twitter applies limitations on the way tweets are retrieved, such as retrieving up to 3,200 recent tweets for a particular user, including retweets from other users. Moreover, twitter API applies a rate limit policy, which allows 180 calls per 15 minutes for each authenticated user. Hence, not all tweets that are publicly available can be retrieved, whereas the complexity of an API call is limited as well.

The application uses the *Original REST API*, which allows developers to access core Twitter data. In the context of this work, the sample consists of two hundred (200) users, while the number of tweets for each user is two hundred and fifty (250). Thus, the total number of tweets that have been aggregated is fifty thousand (50,000). This number of tweets was used to define the population on which we applied our experimental analysis.

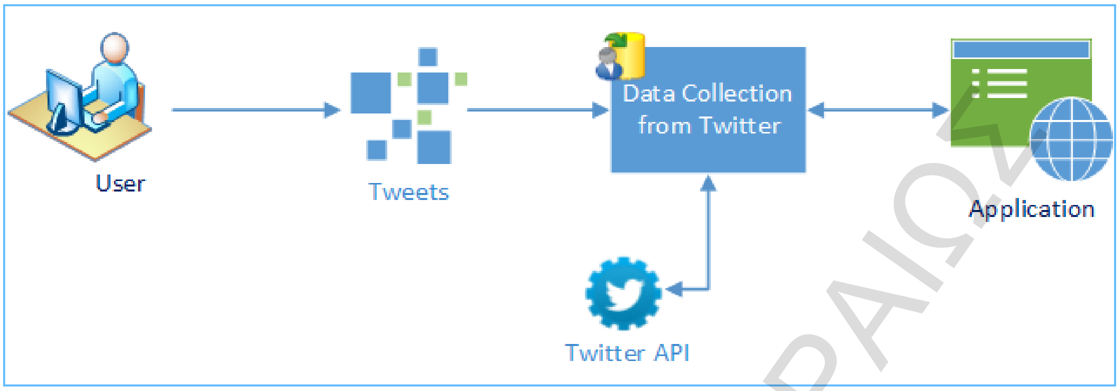


Figure 10: Retrieving Information from Twitter.

6.3.2 Categorizing and Storing the Tweets

This section provides an overview of the tweets classification and storage process. This procedure includes text processing and two different classification techniques which lead to the correlation of the extracted information.

As stated in the work of Fabrizio Sebastiani[74], *text categorization* (also known as *text classification* or *topic spotting*) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set. In the context of this project, we are able to categorize tweets using the following techniques: *Naïve Bayes Classification* and *TextWise*[75].

Naïve Bayes Classification

A naive Bayes classifier is a simple probabilistic classifier, which relies on the application of Bayes' theorem with strong (naive) independence assumptions[76].

A PHP library was used to perform the classification of tweets into specific categories based on Bayesian decision theory. Also, a training set was defined, consisted of fourteen categories, which are depicted below:

- *Arts*
- *Computers and Technology*

[74] Fabrizio Sebastiani, *Text Categorization*, Universita di Padova, Padova, Italy.
[75] http://www.textwise.com/.
[76] Wikipedia, the free encyclopedia, Naïve Bayes, http://en.wikipedia.org/wiki/Naive_Bayes_classifier.

- *Games*
- *Health*
- *Home*
- *Holidays and Celebrations*
- *News*
- *Business*
- *Science*
- *Sentiment and Feelings*
- *Society*
- *Sports*
- *Beauty, Fashion and Style*
- *Social Networks and Online Communities*

A simple technique, which calculates the probability of word frequency in each category, was used.

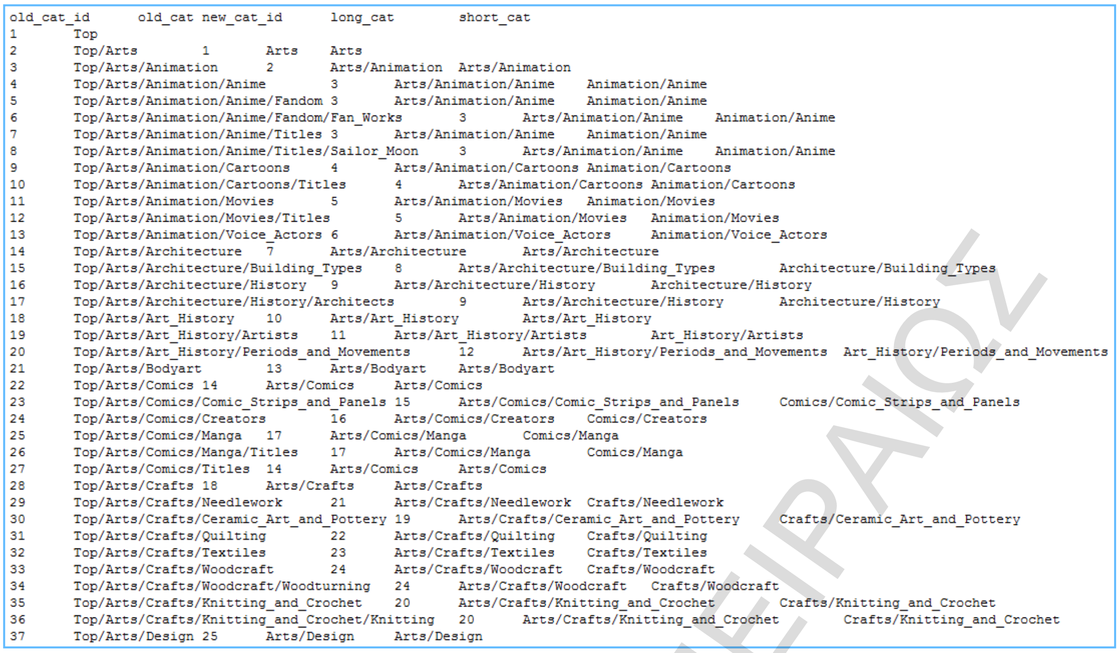
TextWise Classification

TextWise semantic categorization produces a weighted analysis of tags and subjects based on any kind of text[77]. It provides an API that gives developers access, among other services, to category ones. This service identifies the main topical categories for a query's text and returns the categories ordered by weight[78]. TextWise API uses a hierarchical schema based on ODP (Open Directory Project) which is consisted of 770 categories.

Through the TextWise API, we managed to classify the users' tweets with respect to the predefined categories. This categorization strategy also allowed us to develop user profiles and be able to evaluate the privacy risk level. Last but not least, the TextWise categorization was preferred over the Bayesian during the experimental analysis, since it analyzes text using patented semantic technology and a more integrated and effective training data set. In the case of Bayesian classification, the training set was more imperfect, as it was developed only in the context of this project; however, categorization process was not the main focus of our work, as well as the development of a flawless classification tool.

[77] http://www.textwise.com/categorization.

[78] http://textwise.com/api-services.



```
old_cat_id      old_cat new_cat_id      long_cat        short_cat
1       Top
2       Top/Arts        1       Arts    Arts
3       Top/Arts/Animation      2       Arts/Animation  Arts/Animation
4       Top/Arts/Animation/Anime        3       Arts/Animation/Anime    Animation/Anime
5       Top/Arts/Animation/Anime/Fandom 3       Arts/Animation/Anime    Animation/Anime
6       Top/Arts/Animation/Anime/Fandom/Fan_Works       3       Arts/Animation/Anime    Animation/Anime
7       Top/Arts/Animation/Anime/Titles 3       Arts/Animation/Anime    Animation/Anime
8       Top/Arts/Animation/Anime/Titles/Sailor_Moon     3       Arts/Animation/Anime    Animation/Anime
9       Top/Arts/Animation/Cartoons     4       Arts/Animation/Cartoons Animation/Cartoons
10      Top/Arts/Animation/Cartoons/Titles      4       Arts/Animation/Cartoons Animation/Cartoons
11      Top/Arts/Animation/Movies       5       Arts/Animation/Movies   Animation/Movies
12      Top/Arts/Animation/Movies/Titles        5       Arts/Animation/Movies   Animation/Movies
13      Top/Arts/Animation/Voice_Actors 6       Arts/Animation/Voice_Actors     Animation/Voice_Actors
14      Top/Arts/Architecture   7       Arts/Architecture       Arts/Architecture
15      Top/Arts/Architecture/Building_Types    8       Arts/Architecture/Building_Types        Architecture/Building_Types
16      Top/Arts/Architecture/History   9       Arts/Architecture/History       Architecture/History
17      Top/Arts/Architecture/History/Architects        9       Arts/Architecture/History       Architecture/History
18      Top/Arts/Art_History    10      Arts/Art_History        Arts/Art_History
19      Top/Arts/Art_History/Artists    11      Arts/Art_History/Artists        Art_History/Artists
20      Top/Arts/Art_History/Periods_and_Movements      12      Arts/Art_History/Periods_and_Movements  Art_History/Periods_and_Movements
21      Top/Arts/Bodyart        13      Arts/Bodyart    Arts/Bodyart
22      Top/Arts/Comics 14      Arts/Comics     Arts/Comics
23      Top/Arts/Comics/Comic_Strips_and_Panels 15      Arts/Comics/Comic_Strips_and_Panels     Comics/Comic_Strips_and_Panels
24      Top/Arts/Comics/Creators        16      Arts/Comics/Creators    Comics/Creators
25      Top/Arts/Comics/Manga   17      Arts/Comics/Manga       Comics/Manga
26      Top/Arts/Comics/Manga/Titles    17      Arts/Comics/Manga       Comics/Manga
27      Top/Arts/Comics/Titles  14      Arts/Comics     Arts/Comics
28      Top/Arts/Crafts 18      Arts/Crafts     Arts/Crafts
29      Top/Arts/Crafts/Needlework      21      Arts/Crafts/Needlework  Crafts/Needlework
30      Top/Arts/Crafts/Ceramic_Art_and_Pottery 19      Arts/Crafts/Ceramic_Art_and_Pottery     Crafts/Ceramic_Art_and_Pottery
31      Top/Arts/Crafts/Quilting        22      Arts/Crafts/Quilting    Crafts/Quilting
32      Top/Arts/Crafts/Textiles        23      Arts/Crafts/Textiles    Crafts/Textiles
33      Top/Arts/Crafts/Woodcraft       24      Arts/Crafts/Woodcraft   Crafts/Woodcraft
34      Top/Arts/Crafts/Woodcraft/Woodturning   24      Arts/Crafts/Woodcraft   Crafts/Woodcraft
35      Top/Arts/Crafts/Knitting_and_Crochet    20      Arts/Crafts/Knitting_and_Crochet        Crafts/Knitting_and_Crochet
36      Top/Arts/Crafts/Knitting_and_Crochet/Knitting   20      Arts/Crafts/Knitting_and_Crochet        Crafts/Knitting_and_Crochet
37      Top/Arts/Design 25      Arts/Design     Arts/Design
```

Figure 11: Textwise Categorization Mapping[79].

After the categorization step, tweets were stored into a database. As previously mentioned, TextWise was finally used as classification technique during the experimental analysis procedure. Two hundred and fifty (250) tweets were stored for each user. The process of storing tweets shall be analyzed further in the next chapter.

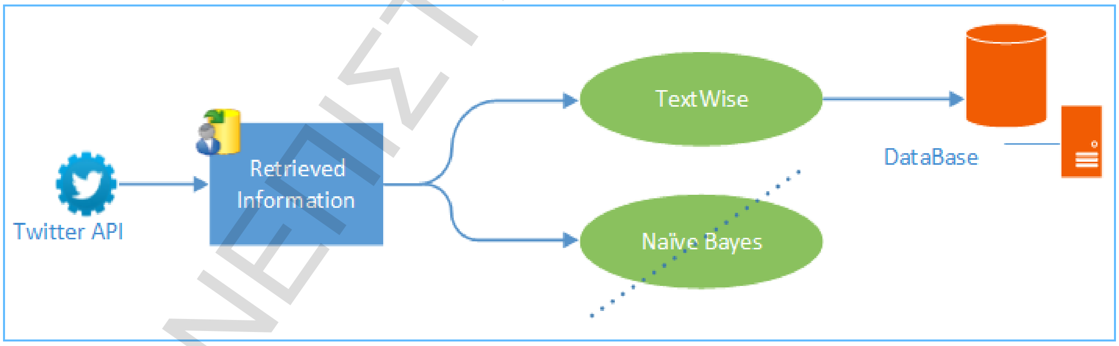


Figure 12: Categorizing and Storing Tweets.

6.3.3 User Profiling

User profiling implies collecting information about a particular user. The information, which was retrieved through Twitter API, after the classification process, includes specific attributes related to the users' interests. TextWise technique consists

[79] http://www.textwise.com/api_docs/labels/2010-ODP-Topic-Category-Mapping.txt.

in constructing comprehensive user profiles. Correlation of the information was actually accomplished through user profiling, which as discussed in previous chapters, raises a significant threat to user privacy. User profiles can be used for various purposes, most of which result in the disclosure of personal information. Therefore, a user's individual characteristics allows to tail after his online behavior.

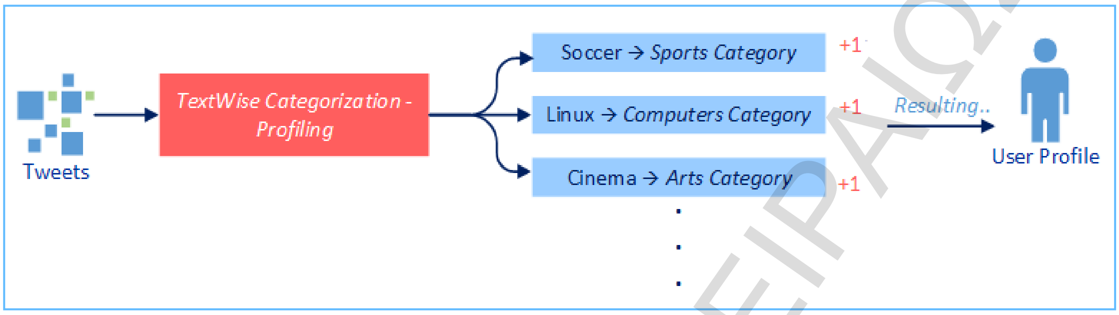


Figure 13: Process of Profiling Users.

User's interests and preferences are identified and formed into categories depending on the content of tweets. Collecting rich and accurate information about a particular user is clearly the main purpose, while developing a user profile. In the context of our work, the diverse attributes of the user include a large amount of his interests, preferences and opinions. As depicted below, in the context of our project, an interactive visualization representing a user profile is provided.

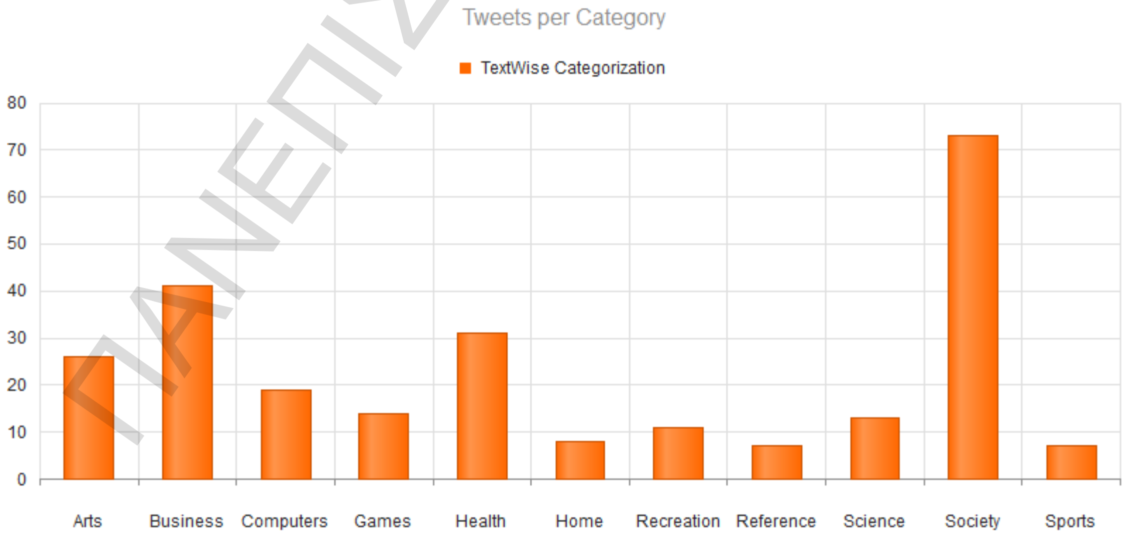


Figure 14: Performing Profiling Results for *Bill Gates* according to TextWise Categorization.

Profiling process aims at identifying individual interests of users. Hence, the possibility of providing the user a personalized experience is high.

6.3.4 Measuring User Privacy

The next step is related to the measurement of the privacy of users. Privacy metrics used in this thesis have been already introduced in previous chapter, as well as the analysis of adversary behavior. In a nutshell, the entropy of a user profile is the first approach of conducting an analysis regarding users' privacy. The use of divergence as additional privacy metric, with respect to the profile of a predefined population group, is also essential.

In this work, after the profiling process, both parameters were used to measure the privacy risk level:

- *Entropy* of user profiles.
- *Divergence* of user profiles in accordance with the average population profile, obtained during the process of classifying and storing tweets.

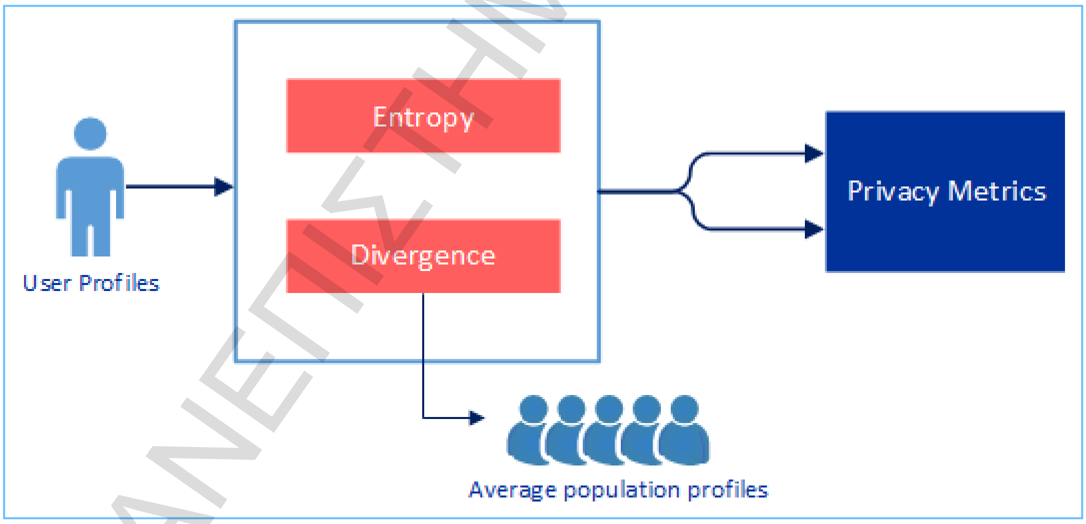


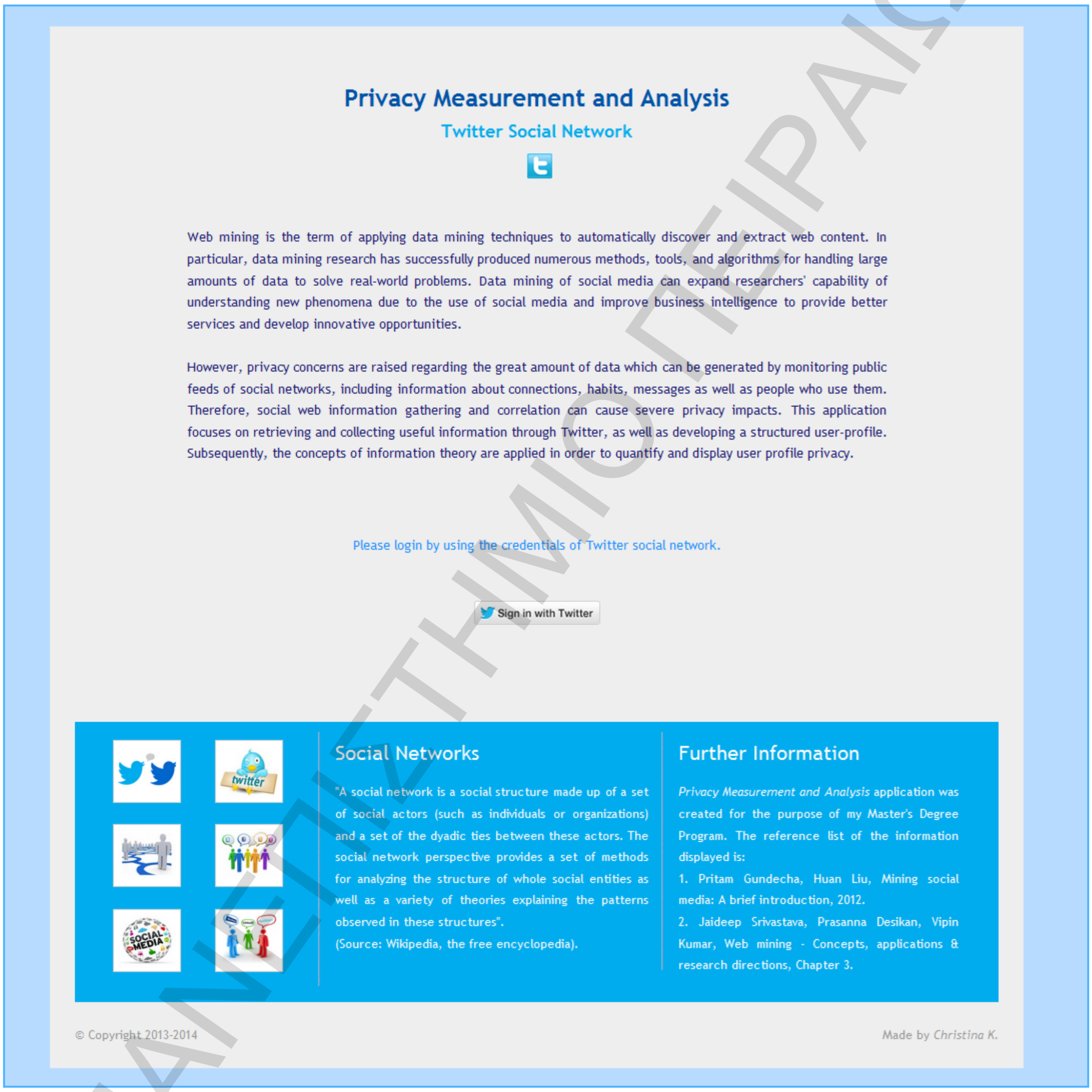
Figure 15: Displaying Privacy Metrics based on Entropy and KL Divergence.

An increasing value of entropy stands for a higher user privacy level. Displaying a reduced divergence value of a user's profile with respect to the average population profile, a *privacy gain* is also obtained.

6.4 Presenting the Privacy Measurement and Analysis Application

In this section, a scenario of use is expounded to present in detail the functionalities of the tool. The scenario specifies the way users perform tasks in the context of the environment that was developed for the purposes of this project.

The *home page* is the first page the user has access to, as entering the application.



Privacy Measurement and Analysis

Twitter Social Network

Web mining is the term of applying data mining techniques to automatically discover and extract web content. In particular, data mining research has successfully produced numerous methods, tools, and algorithms for handling large amounts of data to solve real-world problems. Data mining of social media can expand researchers' capability of understanding new phenomena due to the use of social media and improve business intelligence to provide better services and develop innovative opportunities.

However, privacy concerns are raised regarding the great amount of data which can be generated by monitoring public feeds of social networks, including information about connections, habits, messages as well as people who use them. Therefore, social web information gathering and correlation can cause severe privacy impacts. This application focuses on retrieving and collecting useful information through Twitter, as well as developing a structured user-profile. Subsequently, the concepts of information theory are applied in order to quantify and display user profile privacy.

Please login by using the credentials of Twitter social network.

Social Networks

"A social network is a social structure made up of a set of social actors (such as individuals or organizations) and a set of the dyadic ties between these actors. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures".

(Source: Wikipedia, the free encyclopedia).

Further Information

Privacy Measurement and Analysis application was created for the purpose of my Master's Degree Program. The reference list of the information displayed is:

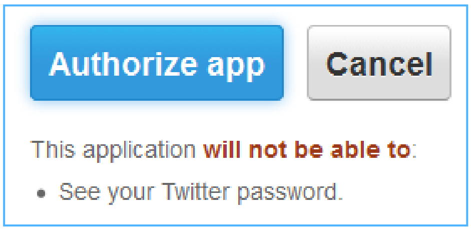
1. Pritam Gundecha, Huan Liu, Mining social media: A brief introduction, 2012.
2. Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, Web mining - Concepts, applications & research directions, Chapter 3.

© Copyright 2013-2014

Made by *Christina K.*

Figure 16: Home Page of *Privacy Measurement and Analysis* Application.

To enter and use the components the application provides, the user has to login by using the credentials of his Twitter account. The user is authenticated and provides the application the permission to have access to his personal account. The authentication is performed through the usual process of identifying an individual in Twitter, based on the username and password of the user.



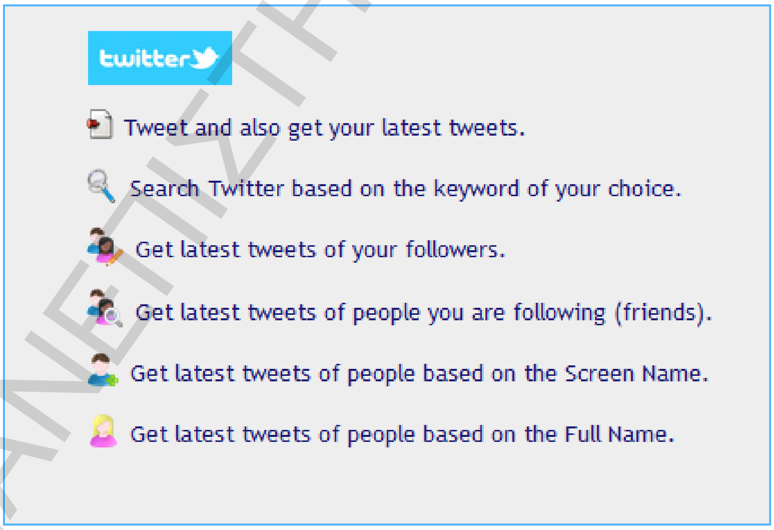
This application **will not be able to**:

- See your Twitter password.

Figure 17: Provide the application permissions for accessing the Twitter account.

The application directs the user to the *twitter functions* page. This page provides the following functionalities, which are also depicted in Figure 18:

- Tweeting and obtaining the latest tweets of the user.
- Searching Twitter based on keywords.
- Getting the latest tweets of user’s followers.
- Getting the latest tweets of user’s friends.
- Getting the latest tweets of people regarding their Full Name.
- Getting the latest tweets of people regarding their Screen Name.



Tweet and also get your latest tweets.

Search Twitter based on the keyword of your choice.

Get latest tweets of your followers.

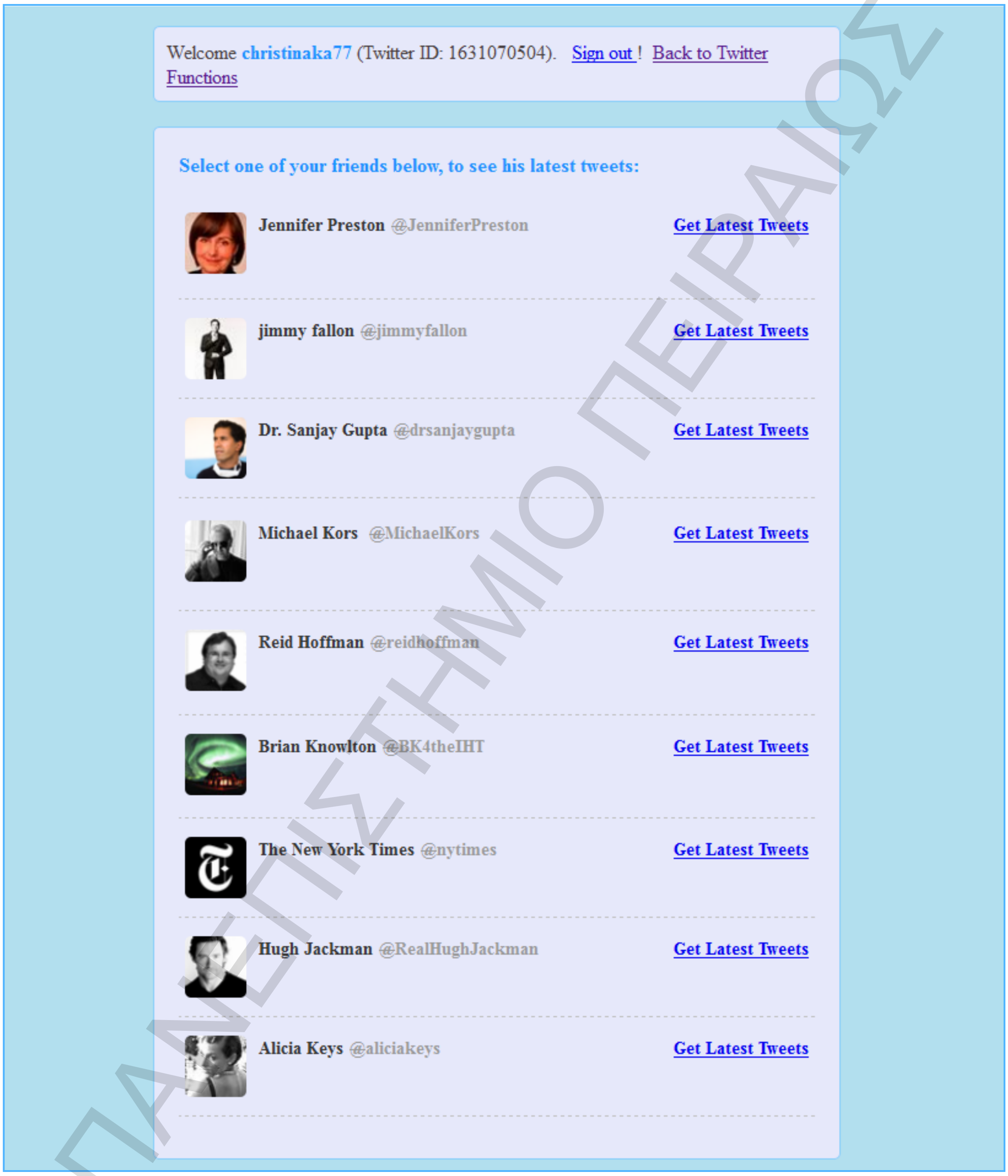
Get latest tweets of people you are following (friends).

Get latest tweets of people based on the Screen Name.

Get latest tweets of people based on the Full Name.

Figure 18: Twitter Functions.

Each function provides the user different functionalities regarding the process of information retrieval. However, for the purposes of our project we are using the function of *«getting the latest tweets of user’s friends»* exclusively. Even though all functions consist in retrieving information, the function which refers to people the user is following, allows us to collect a greater number of users; a crucial factor, the importance of which is established during the process of evaluating the privacy risk level of users.



Welcome christinaka77 (Twitter ID: 1631070504). Sign out ! Back to Twitter
Functions

Select one of your friends below, to see his latest tweets:

Jennifer Preston @JenniferPreston — Get Latest Tweets

jimmy fallon @jimmyfallon — Get Latest Tweets

Dr. Sanjay Gupta @drsanjaygupta — Get Latest Tweets

Michael Kors @MichaelKors — Get Latest Tweets

Reid Hoffman @reidhoffman — Get Latest Tweets

Brian Knowlton @BK4theIHT — Get Latest Tweets

The New York Times @nytimes — Get Latest Tweets

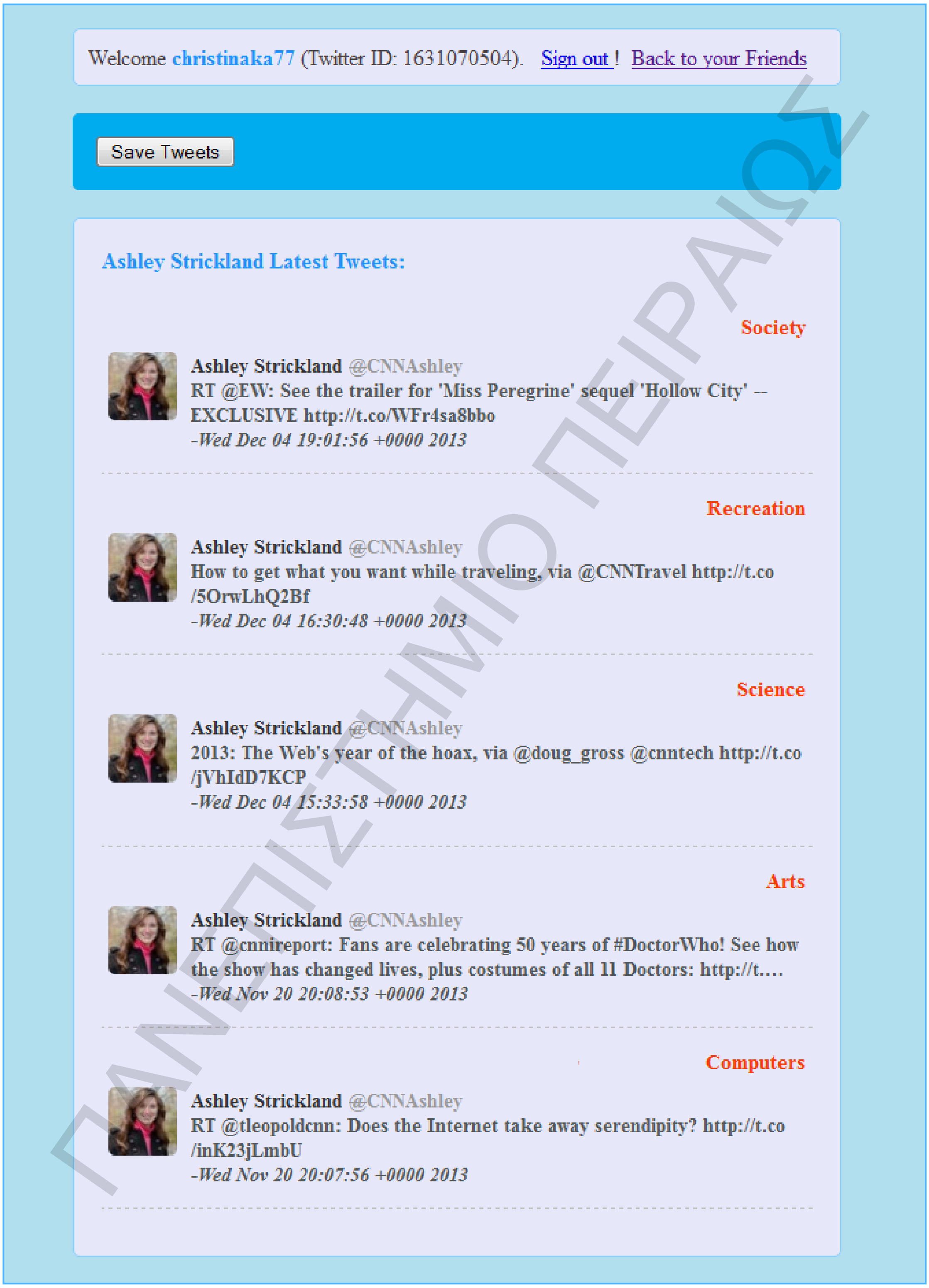
Hugh Jackman @RealHughJackman — Get Latest Tweets

Alicia Keys @aliciakeys — Get Latest Tweets

Figure 19: Function of Getting a User's Friends.

In the context of our scenario, we are retrieving the tweets of the user *Ashley Strickland*. The tweets are automatically classified into predefined categories through

TextWise categorization tool. The results of the classification process are presented on the right corner of each tweet, in orange letters.



Welcome christinaka77 (Twitter ID: 1631070504). Sign out ! Back to your Friends

Ashley Strickland Latest Tweets:

Society

Ashley Strickland @CNNAshley
RT @EW: See the trailer for 'Miss Peregrine' sequel 'Hollow City' -- EXCLUSIVE http://t.co/WFr4sa8bbo
-Wed Dec 04 19:01:56 +0000 2013

Recreation

Ashley Strickland @CNNAshley
How to get what you want while traveling, via @CNNTravel http://t.co/5OrwLhQ2Bf
-Wed Dec 04 16:30:48 +0000 2013

Science

Ashley Strickland @CNNAshley
2013: The Web's year of the hoax, via @doug_gross @cnntech http://t.co/jVhIdD7KCP
-Wed Dec 04 15:33:58 +0000 2013

Arts

Ashley Strickland @CNNAshley
RT @cnnireport: Fans are celebrating 50 years of #DoctorWho! See how the show has changed lives, plus costumes of all 11 Doctors: http://t....
-Wed Nov 20 20:08:53 +0000 2013

Computers

Ashley Strickland @CNNAshley
RT @tleopoldcnn: Does the Internet take away serendipity? http://t.co/inK23jLmbU
-Wed Nov 20 20:07:56 +0000 2013

Figure 20: Retrieving Tweets for *Ashley Strickland*.

In addition, the results are automatically depicted in the following graph chart (Figure 21).

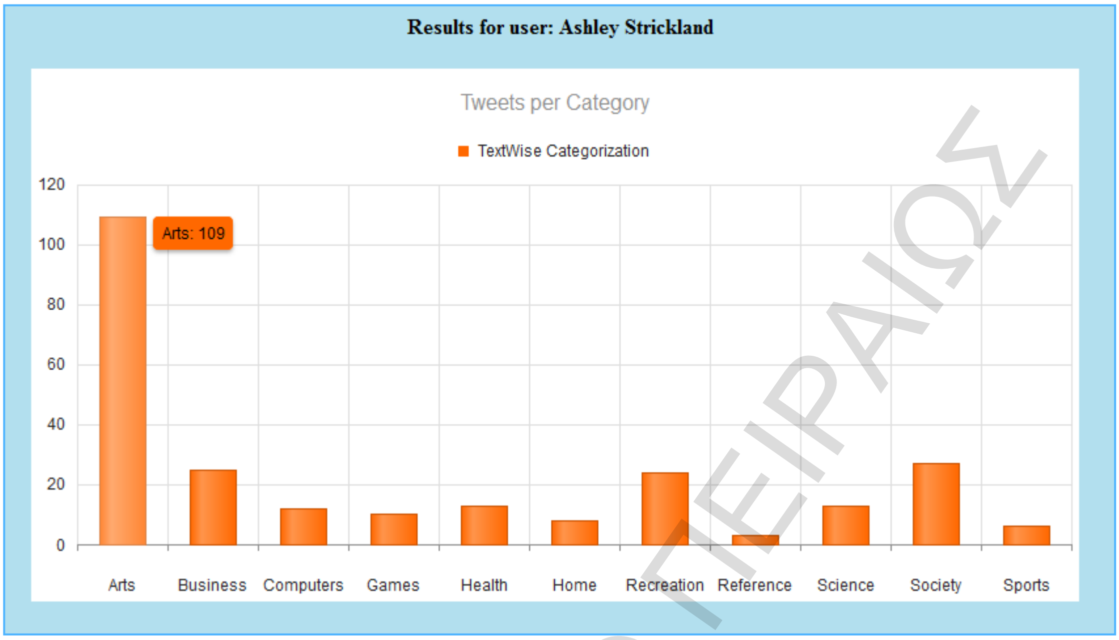


Figure 21: Classification Results for *Ashley Strickland* depicted in a Graph Chart.

The graph chart that represents the user profile provides the user a first insight into the level of his privacy (Figure 21). However, the need for a more comprehensive interpretation of the risk the user deals with, led to a more detailed approach. Consequently, the application was designed in such a way that the values of *entropy* and *divergence* are also calculated automatically; two fundamental information-theoretic quantities that are of great importance for the measurement of privacy. Along with the calculation of entropy and divergence, one more value is calculated, *maximum entropy*. Entropy and maximum entropy values actually consist in measuring the privacy risk level.

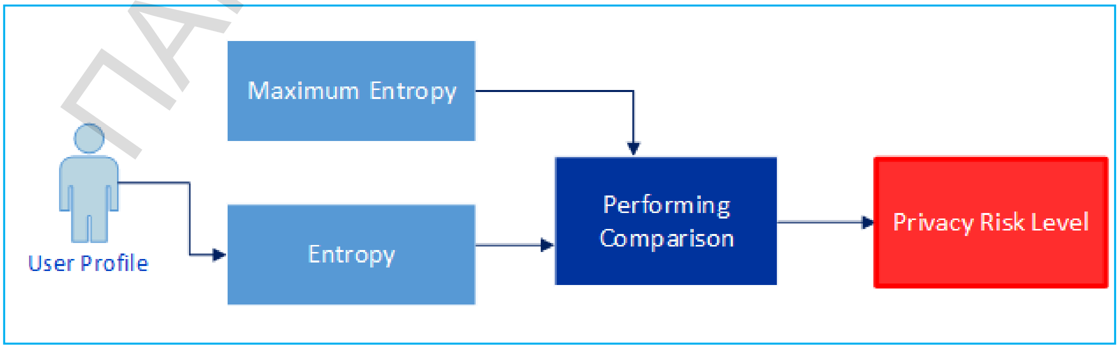
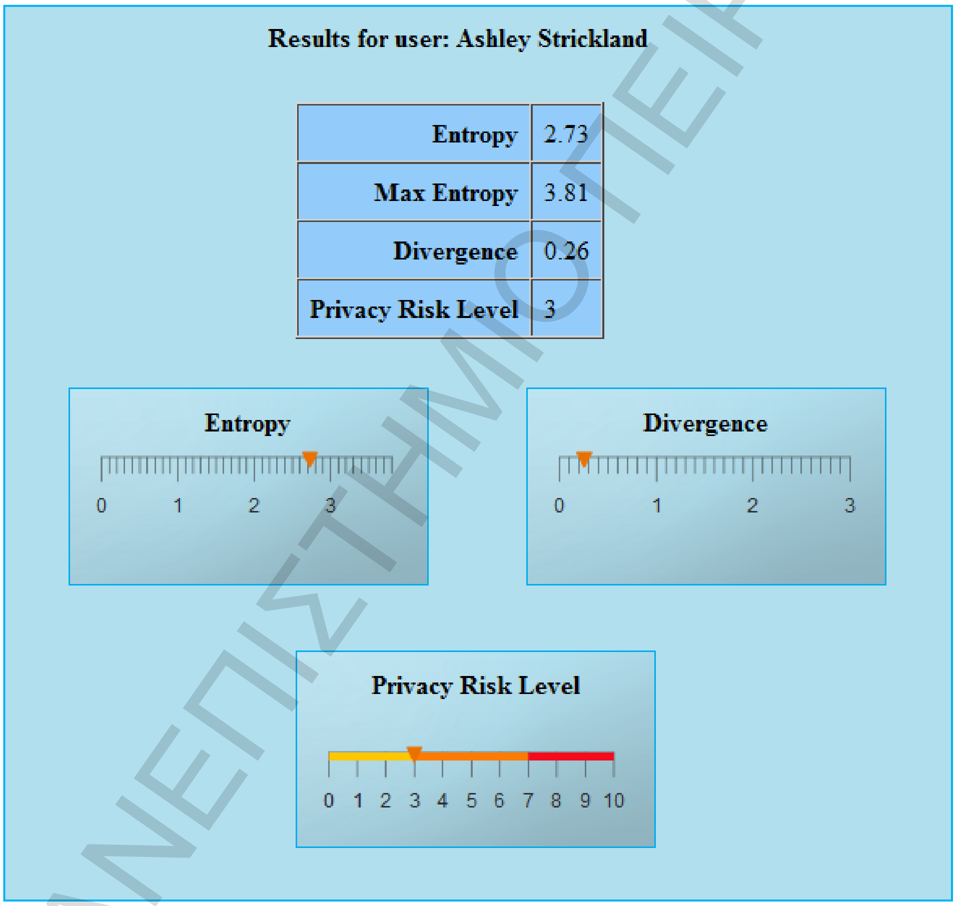


Figure 22: Process of Measuring Privacy Risk.

Therefore, the particular component contains the following elements:

- The user's profile presented as a chart of categories (Figure 21).
- Information related to privacy against an identification attack.
    - Entropy of the user's profile.
    - Maximum entropy value.
    - Divergence in correspondence with the average population's profile
    - Privacy risk level value.

Privacy information related to a user's profile, in particular to Ashley Strickland's profile, is clearly depicted in Figure 23.



Results for user: Ashley Strickland

Entropy	2.73
Max Entropy	3.81
Divergence	0.26
Privacy Risk Level	3

Figure 23: Results of Measuring the Privacy of *Ashley Strickland*'s Profile.

The higher the entropy of Ashley Strickland's profile, the greater is the number of the users who behave according to it. Moreover, given the distribution of the population's average profile, the divergence between Ashley's and the population's profile implies that the lower the divergence is, the more private the profile can be considered. Consequently, an increasing value of divergence entails that the user's profile is more identifiable, in relation to the rest of the population. In the context of our scenario, the privacy risk level of the profile is evaluated as *Level 3* (Figure 24).

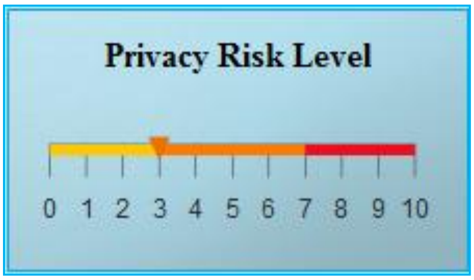
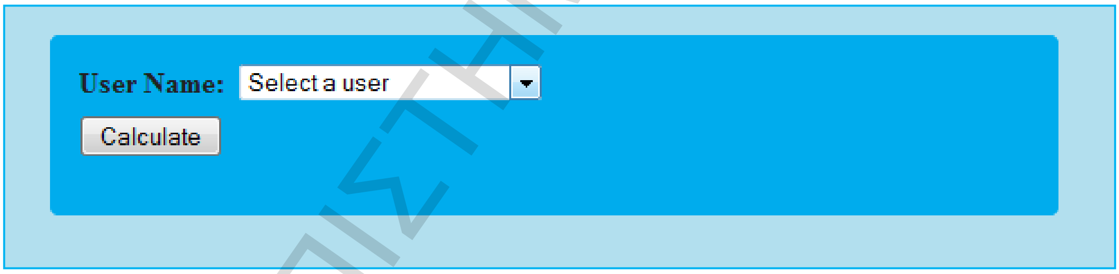


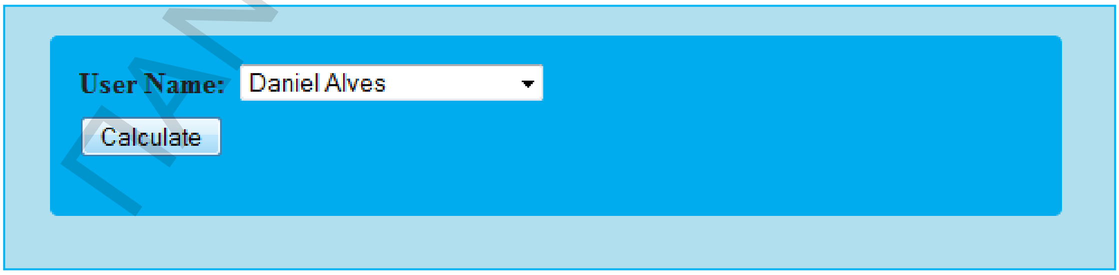
Figure 24: Privacy Risk Level of *Ashley Strickland*'s Profile.

An additional implemented functionality of the application is the calculation of the values of entropy, divergence and privacy risk for a particular twitter user through another graphical user interface. This component allows us to choose the user, in alphabetical order, and obtain at once the relevant results. The process is illustrated in the following figures:



User Name: Select a user

Figure 25: Measuring the Privacy of User Profiles (a).



User Name: Daniel Alves

Figure 26: Measuring the Privacy of User Profiles (b).

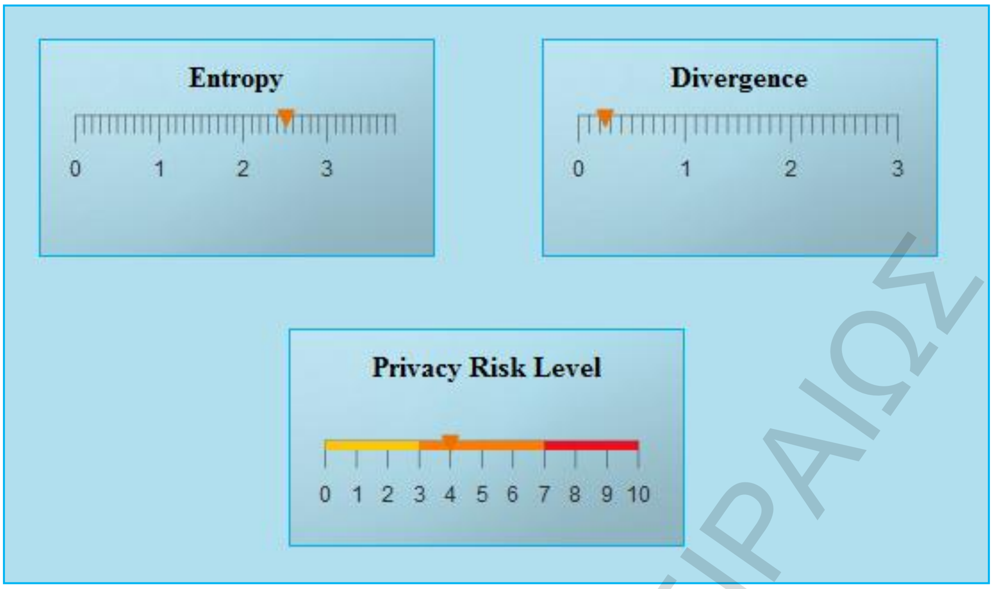


Figure 27: Results of Measuring the Privacy of *Daniel Alves'* Profile.

As mentioned above, the first element employed to measure the privacy risk level is the user's entropy. Along with the calculation of entropy and divergence, two graphical representations of the distribution of information are also generated. The histograms present the number of users for the different calculated value of entropy and divergence respectively.

We should point out that the developed functionality of the application generates both histograms.

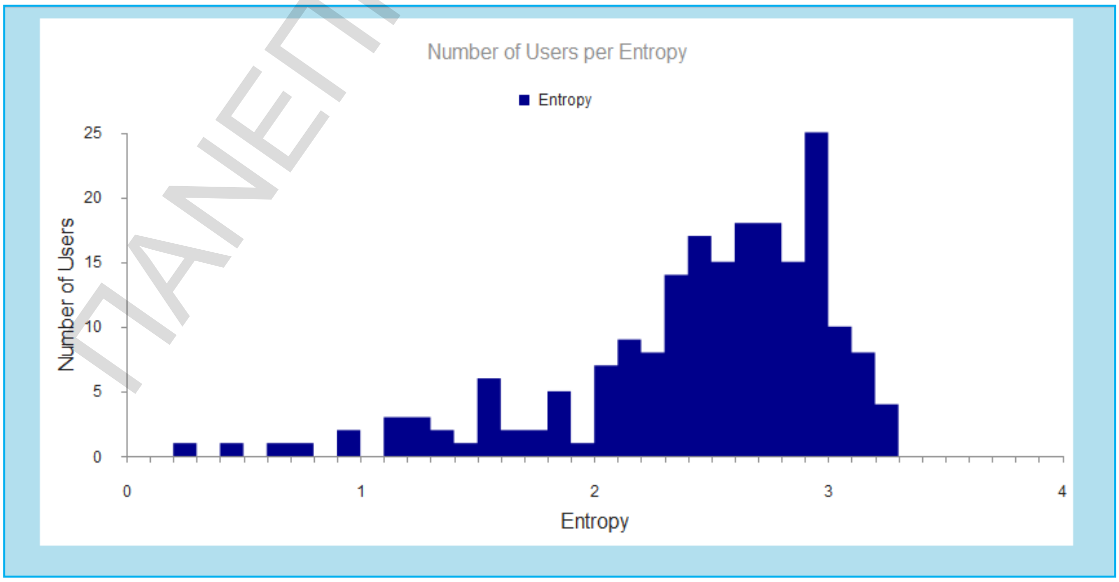


Figure 28: Histogram of Users' Entropy Values.

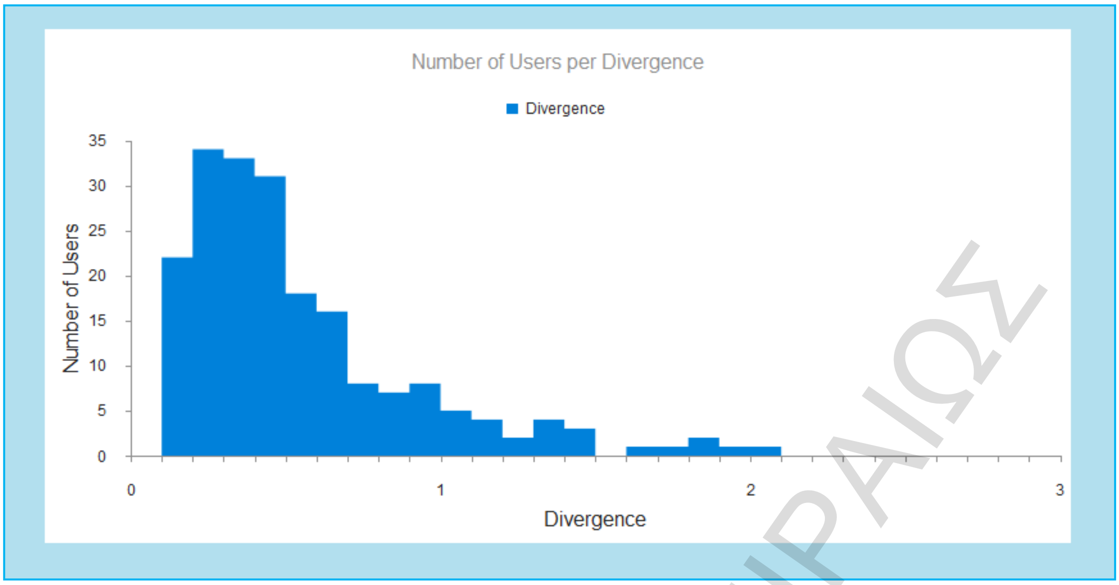


Figure 29: Histogram of Users' Divergence Values.

In addition, apart from the development of user profiles for a particular user, the application also generates a graph chart that depicts the total number of tweets per category of interests.

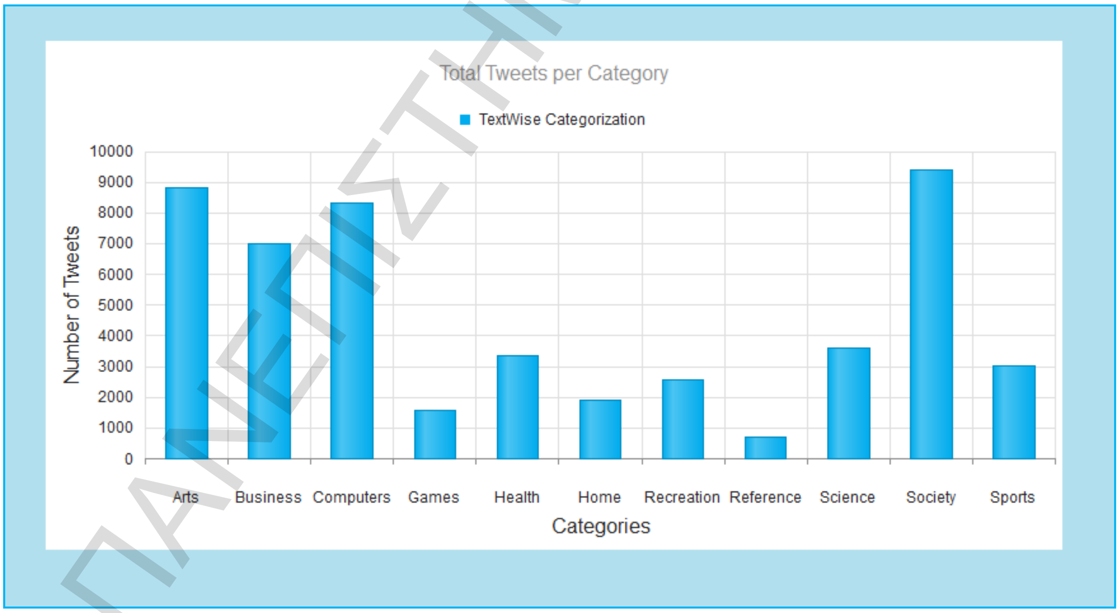


Figure 30: Chart of Total Tweets per Category.

In a nutshell, the main purpose of the component is to measure the privacy of the user, and hence evaluate the actions that need to be performed to protect the user against an external adversary. The architectural design of the application is depicted in the following illustration:

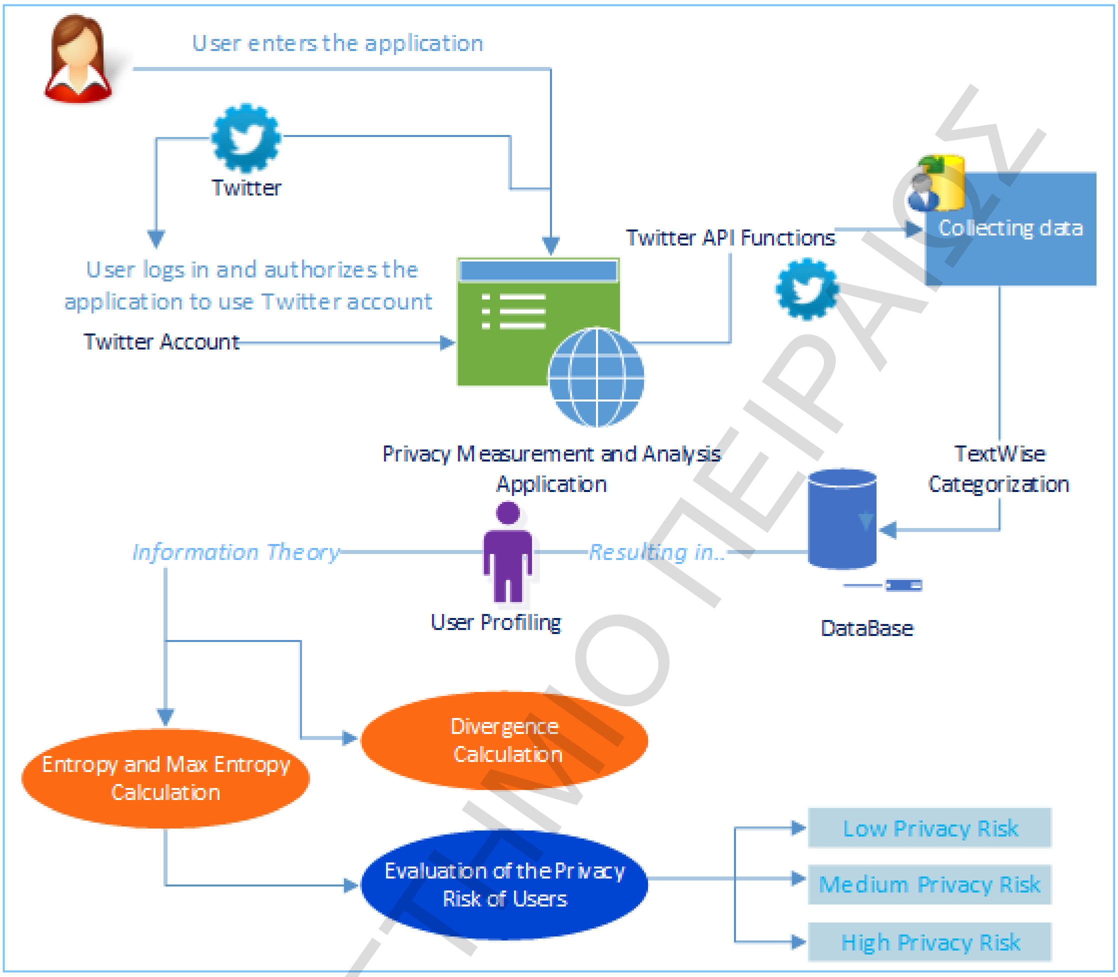

User enters the application
Twitter
User logs in and authorizes the application to use Twitter account
Twitter Account
Twitter API Functions
Collecting data
Privacy Measurement and Analysis Application
TextWise Categorization
Information Theory
Resulting in..
User Profiling
DataBase
Divergence Calculation
Entropy and Max Entropy Calculation
Evaluation of the Privacy Risk of Users
Low Privacy Risk
Medium Privacy Risk
High Privacy Risk

Figure 31: Application's Flow Chart.

7 Implementation Details

7.1 Introduction

The application proposed in this thesis has been developed from the ground up by employing open source software. In particular, three well known tools were used: the language PHP, for the development of a dynamic Web-based application; MySql, for the development and management of a database; and the server Apache, which has the ability to display dynamic pages written in PHP, while communicating with the database.

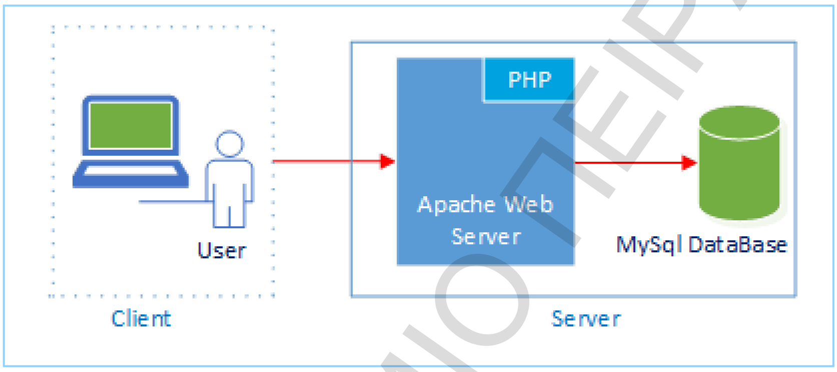


Figure 32: Open Source technologies on which the application was based.

The application aimed at retrieving information from twitter, deploying user profiles and measuring the risk level regarding user's privacy. Risk level was calculated according to the mathematical notation and assumptions made in Chapter 5. The application also provides the user the possibility of evaluating the results of the privacy measurement, by interpreting this information and adopting the corresponding privacy protection policies. Thus, the user's profile is the input of the identification module, which allows the generation of useful indicators or even alarms to the user regarding his privacy risk level.

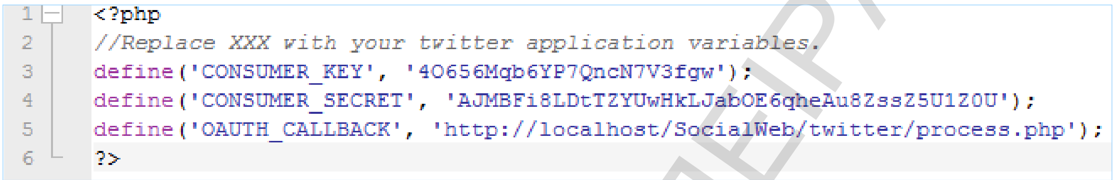
In the next sections of this chapter, the basic implementation components of the application shall be analyzed in detail.

7.2 Twitter Login Process

When the user enters the application, it is required to login with his Twitter account. This module uses session variables to temporarily store user information and consists of two basic PHP files: configuration, login and process files. Abraham William's Twitter PHP Library was also used, which is widely known for the effectiveness and simplicity it provides.

Configuration File

Configuration file stores the Twitter customer key, secret and callback URL. These application variables are provided by Twitter, during the process of creating a new application in Twitter.

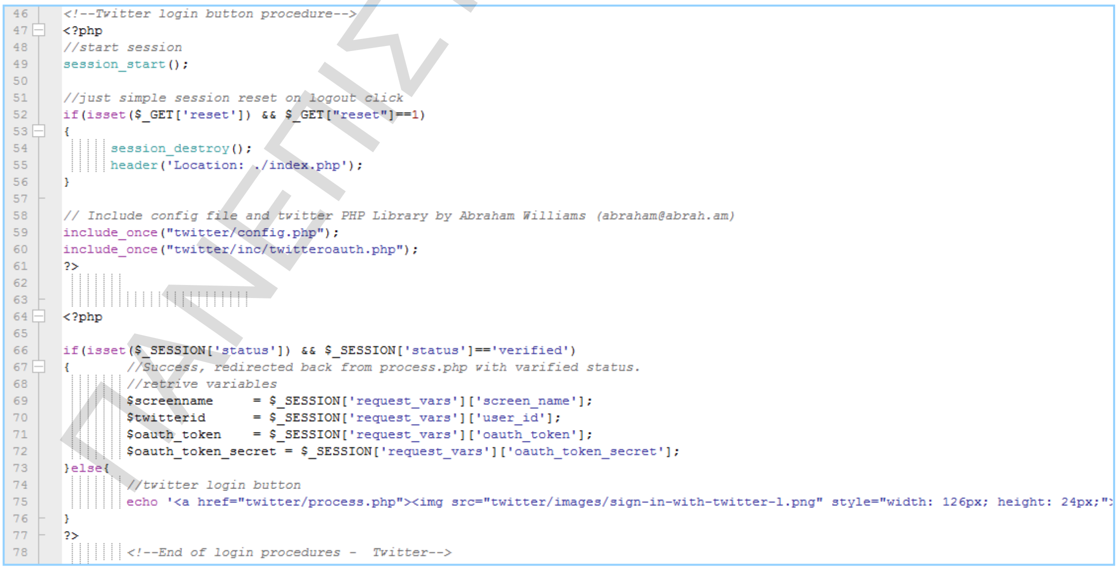


```php
<?php
//Replace XXX with your twitter application variables.
define('CONSUMER_KEY', '40656Mqb6YP7QncN7V3fgw');
define('CONSUMER_SECRET', 'AJMBFi8LDtTZYUwHkLJabOE6qheAu8ZssZ5U1Z0U');
define('OAUTH_CALLBACK', 'http://localhost/SocialWeb/twitter/process.php');
?>
```

Figure 33: Configuration PHP file.

Login File

Login file is actually the home page and contains a login button. When the user clicks on login button, he is redirected to the process file, which sends the user to the Twitter Auth page to obtain a request token. Afterwards, the user is redirected back to process file. In case of a successful authorization, process file sets details in session variables.

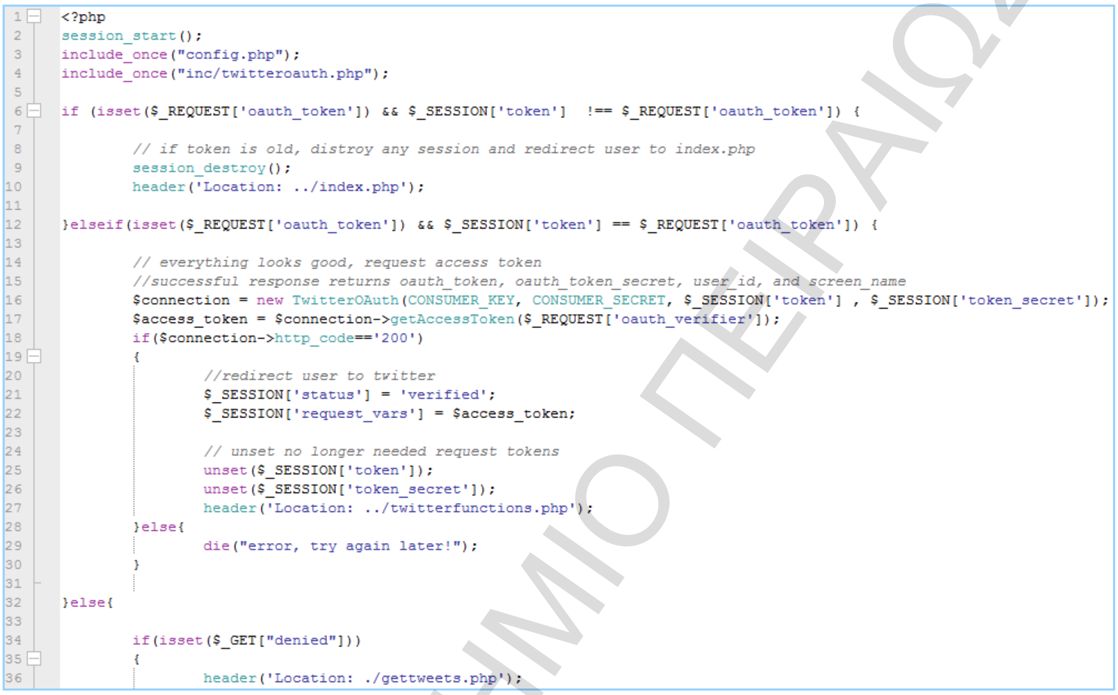


```php
<!--Twitter login button procedure-->
<?php
//start session
session_start();

//just simple session reset on logout click
if(isset($_GET['reset']) && $_GET["reset"]==1)
{
        session_destroy();
        header('Location: ./index.php');
}

// Include config file and twitter PHP Library by Abraham Williams (abraham@abrah.am)
include_once("twitter/config.php");
include_once("twitter/inc/twitteroauth.php");
?>


<?php

if(isset($_SESSION['status']) && $_SESSION['status']=='verified')
{       //Success, redirected back from process.php with varified status.
        //retrive variables
        $screenname     = $_SESSION['request_vars']['screen_name'];
        $twitterid      = $_SESSION['request_vars']['user_id'];
        $oauth_token    = $_SESSION['request_vars']['oauth_token'];
        $oauth_token_secret = $_SESSION['request_vars']['oauth_token_secret'];
}else{
        //twitter login button
        echo '<a href="twitter/process.php"><img src="twitter/images/sign-in-with-twitter-1.png" style="width: 126px; height: 24px;">
}
?>
        <!--End of login procedures -  Twitter-->
```

Figure 34: Login Process.

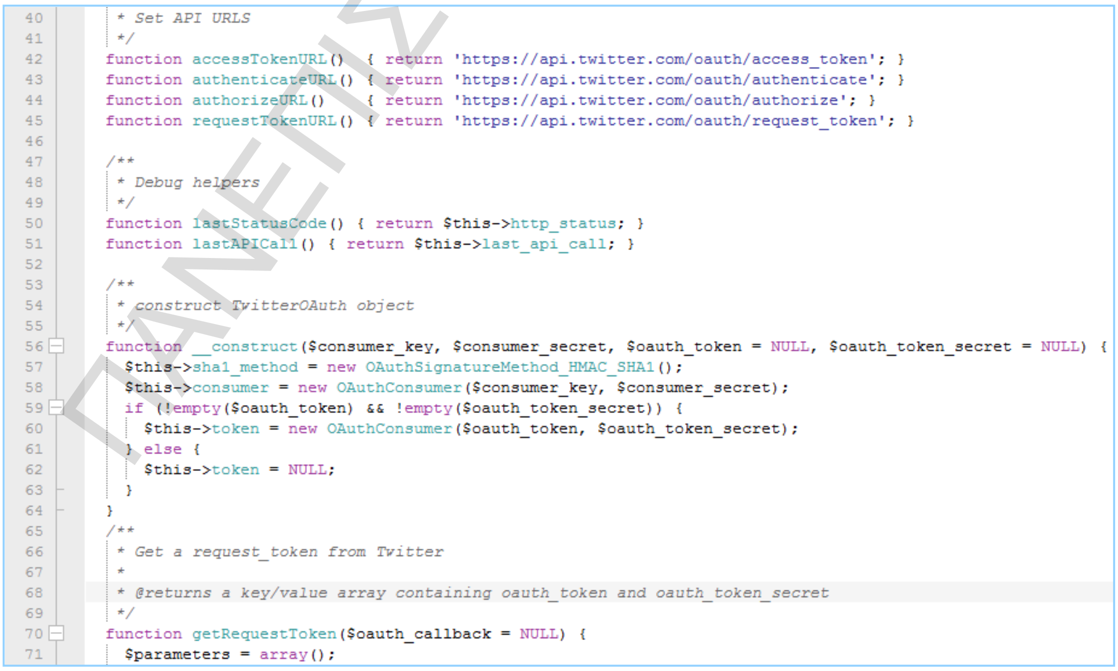
Process File

This file is used to perform a comparison among variables and redirects the user back and forth. Thanks to this file the user is able to obtain a request token that is passed to Twitter Authorize page as oauth_token parameter. Once user signs in, user is authenticated and returned to the callback URL.



```php
<?php
session_start();
include_once("config.php");
include_once("inc/twitteroauth.php");

if (isset($_REQUEST['oauth_token']) && $_SESSION['token']  !== $_REQUEST['oauth_token']) {

        // if token is old, distroy any session and redirect user to index.php
        session_destroy();
        header('Location: ../index.php');

}elseif(isset($_REQUEST['oauth_token']) && $_SESSION['token'] == $_REQUEST['oauth_token']) {

        // everything looks good, request access token
        //successful response returns oauth_token, oauth_token_secret, user_id, and screen_name
        $connection = new TwitterOAuth(CONSUMER_KEY, CONSUMER_SECRET, $_SESSION['token'] , $_SESSION['token_secret']);
        $access_token = $connection->getAccessToken($_REQUEST['oauth_verifier']);
        if($connection->http_code=='200')
        {
                //redirect user to twitter
                $_SESSION['status'] = 'verified';
                $_SESSION['request_vars'] = $access_token;

                // unset no longer needed request tokens
                unset($_SESSION['token']);
                unset($_SESSION['token_secret']);
                header('Location: ../twitterfunctions.php');
        }else{
                die("error, try again later!");
        }

}else{

        if(isset($_GET["denied"]))
        {
                header('Location: ./gettweets.php');
```

Figure 35: Process PHP File.



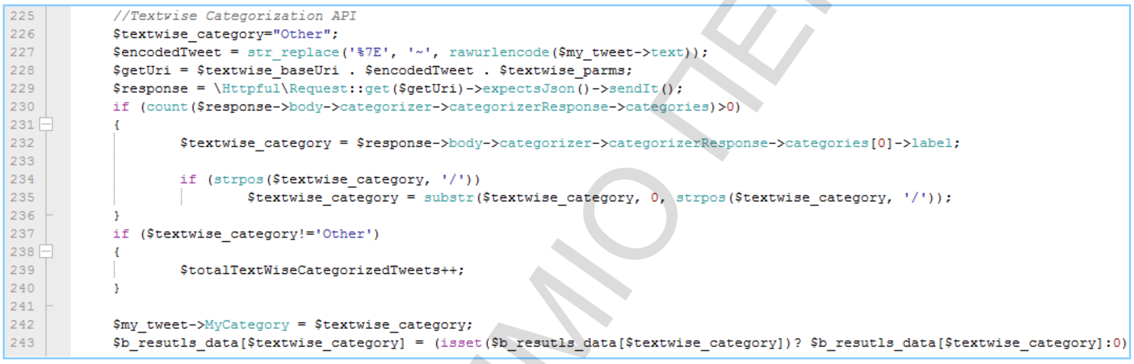
```php
 * Set API URLS
 */
function accessTokenURL()  { return 'https://api.twitter.com/oauth/access_token'; }
function authenticateURL() { return 'https://api.twitter.com/oauth/authenticate'; }
function authorizeURL()    { return 'https://api.twitter.com/oauth/authorize'; }
function requestTokenURL() { return 'https://api.twitter.com/oauth/request_token'; }

/**
 * Debug helpers
 */
function lastStatusCode() { return $this->http_status; }
function lastAPICall() { return $this->last_api_call; }

/**
 * construct TwitterOAuth object
 */
function __construct($consumer_key, $consumer_secret, $oauth_token = NULL, $oauth_token_secret = NULL) {
  $this->sha1_method = new OAuthSignatureMethod_HMAC_SHA1();
  $this->consumer = new OAuthConsumer($consumer_key, $consumer_secret);
  if (!empty($oauth_token) && !empty($oauth_token_secret)) {
    $this->token = new OAuthConsumer($oauth_token, $oauth_token_secret);
  } else {
    $this->token = NULL;
  }
}
/**
 * Get a request_token from Twitter
 *
 * @returns a key/value array containing oauth_token and oauth_token_secret
 */
function getRequestToken($oauth_callback = NULL) {
  $parameters = array();
```

Figure 36: Abraham William's PHP Library.

7.3 Categorization and Profiling Process

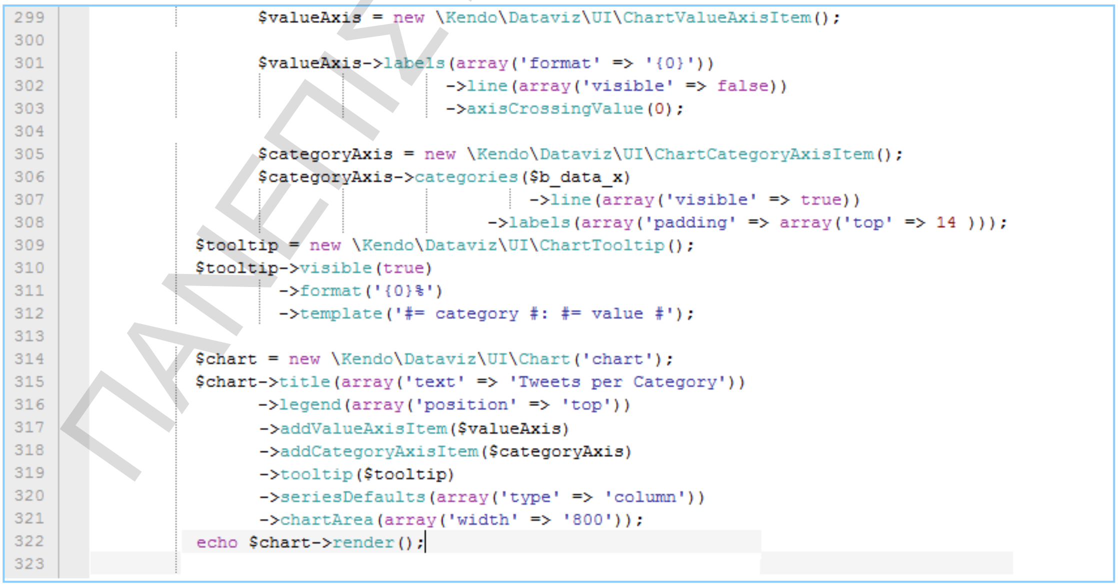
The profiling module bears the responsibility for the development of the user’s profile after the classification process of the retrieved information. In our work, TextWise categorization was used to classify tweets, which in our case constitute the retrieved information, into a certain number of predefined categories.

Categorization process is the key for handling and organizing text data. A commonly accepted text categorization method is TextWise, which analyzes any input text and returns a set of relevant topic categories as a weighted list[80]. The provided API allows developers to extract detailed Semantic key concepts from text data of any size, with a high degree of accuracy.



```
//Textwise Categorization API
$textwise_category="Other";
$encodedTweet = str_replace('%7E', '~', rawurlencode($my_tweet->text));
$getUri = $textwise_baseUri . $encodedTweet . $textwise_parms;
$response = \Httpful\Request::get($getUri)->expectsJson()->sendIt();
if (count($response->body->categorizer->categorizerResponse->categories)>0)
{
        $textwise_category = $response->body->categorizer->categorizerResponse->categories[0]->label;

        if (strpos($textwise_category, '/'))
                $textwise_category = substr($textwise_category, 0, strpos($textwise_category, '/'));
}
if ($textwise_category!='Other')
{
        $totalTextWiseCategorizedTweets++;
}

$my_tweet->MyCategory = $textwise_category;
$b_resutls_data[$textwise_category] = (isset($b_resutls_data[$textwise_category])? $b_resutls_data[$textwise_category]:0)
```

Figure 37: TextWise Categorization API.



```
      $valueAxis = new \Kendo\Dataviz\UI\ChartValueAxisItem();

      $valueAxis->labels(array('format' => '{0}'))
                        ->line(array('visible' => false))
                        ->axisCrossingValue(0);

      $categoryAxis = new \Kendo\Dataviz\UI\ChartCategoryAxisItem();
      $categoryAxis->categories($b_data_x)
                                ->line(array('visible' => true))
                            ->labels(array('padding' => array('top' => 14 )));
$tooltip = new \Kendo\Dataviz\UI\ChartTooltip();
$tooltip->visible(true)
        ->format('{0}%')
        ->template('#= category #: #= value #');

$chart = new \Kendo\Dataviz\UI\Chart('chart');
$chart->title(array('text' => 'Tweets per Category'))
      ->legend(array('position' => 'top'))
      ->addValueAxisItem($valueAxis)
      ->addCategoryAxisItem($categoryAxis)
      ->tooltip($tooltip)
      ->seriesDefaults(array('type' => 'column'))
      ->chartArea(array('width' => '800'));
echo $chart->render();
```

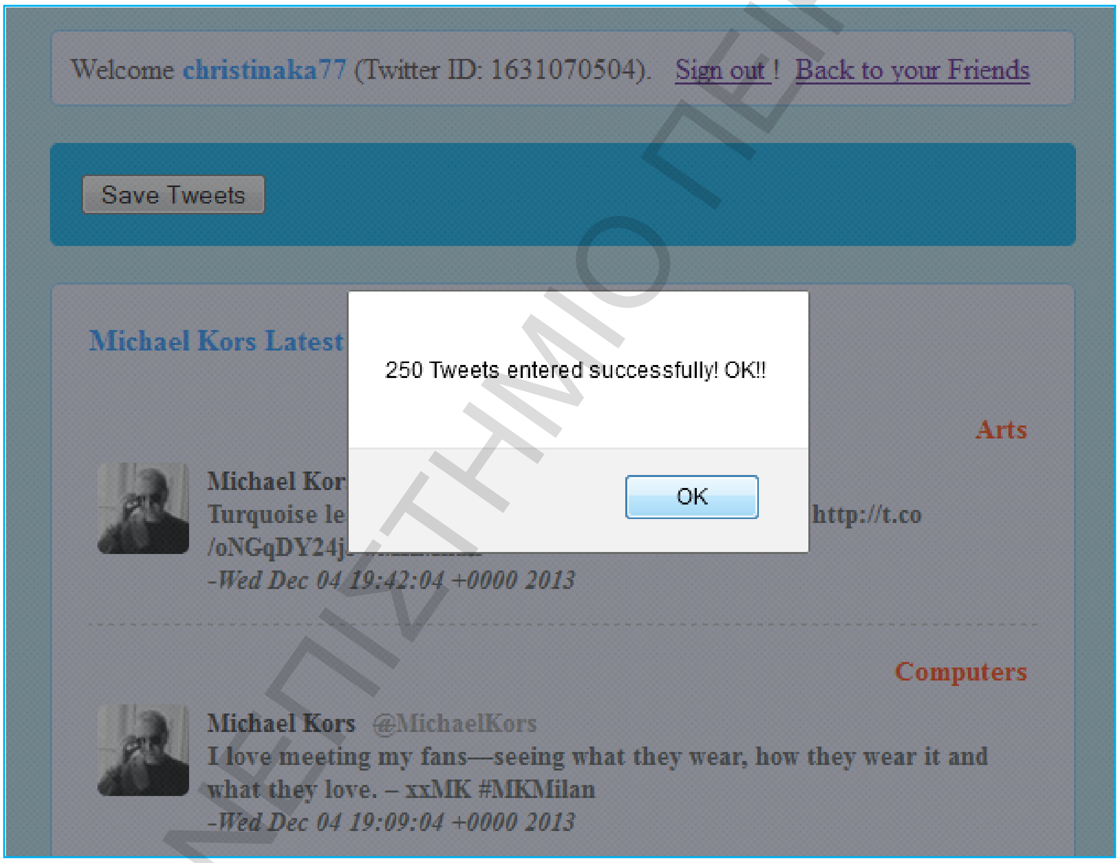
Figure 38: Graph Chart Deployment.

[80] http://www.textwise.com/api.

7.4 Database Design

The database design is essential for developing an application of high performance. The database was developed in a way that overcomes the issue of duplicate records, by enhancing at the same time our ability to maintain the stored information.

Information that was retrieved from Twitter is classified, and afterwards, is stored into the database. In the page we obtain the tweets of friends (people we follow), there is a button “Save Tweets”. By clicking this button, the tweets of the chosen user are automatically stored in the database.



Welcome christinaka77 (Twitter ID: 1631070504). Sign out ! Back to your Friends

Michael Kors Latest

250 Tweets entered successfully! OK!!

Arts

Michael Kor
Turquoise le http://t.co
/oNGqDY24j
-Wed Dec 04 19:42:04 +0000 2013

Computers

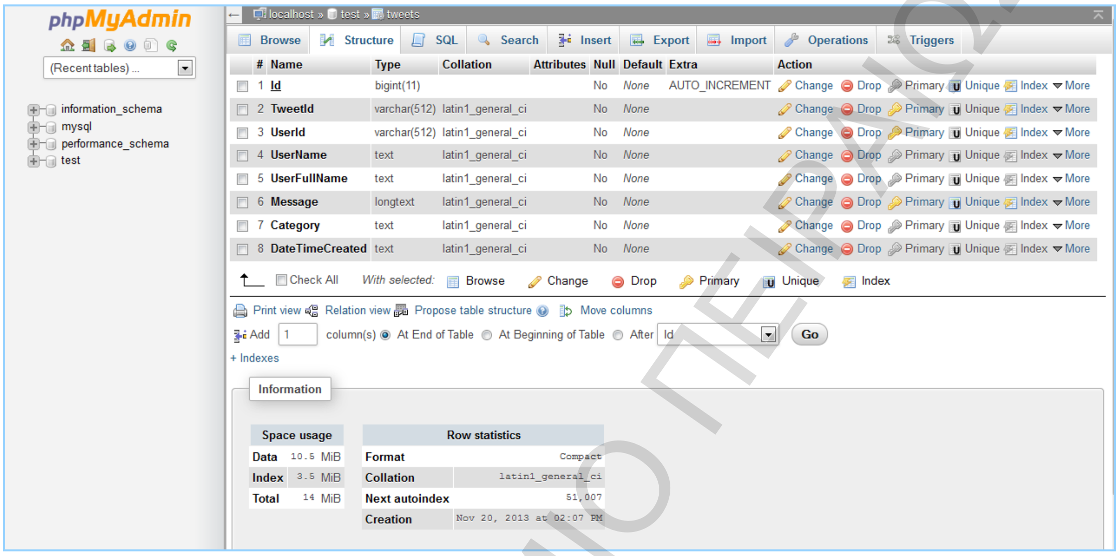
Michael Kors @MichaelKors
I love meeting my fans—seeing what they wear, how they wear it and what they love. – xxMK #MKMilan
-Wed Dec 04 19:09:04 +0000 2013

Figure 39: Process of Storing Tweets in the Database.

For the purposes of our project, we developed a table named “tweets”, which consists of the following fields:

- Id
- TweetId
- UserId

- Username
- UserFullName
- Message
- Category
- DateTimeCreated



	#	Name	Type	Collation	Attributes	Null	Default	Extra	Action
□	1	Id	bigint(11)			No	None	AUTO_INCREMENT	Change Drop Primary Unique Index More
□	2	TweetId	varchar(512)	latin1_general_ci		No	None		Change Drop Primary Unique Index More
□	3	UserId	varchar(512)	latin1_general_ci		No	None		Change Drop Primary Unique Index More
□	4	UserName	text	latin1_general_ci		No	None		Change Drop Primary Unique Index More
□	5	UserFullName	text	latin1_general_ci		No	None		Change Drop Primary Unique Index More
□	6	Message	longtext	latin1_general_ci		No	None		Change Drop Primary Unique Index More
□	7	Category	text	latin1_general_ci		No	None		Change Drop Primary Unique Index More
□	8	DateTimeCreated	text	latin1_general_ci		No	None		Change Drop Primary Unique Index More

□ Check All *With selected:* Browse Change Drop Primary Unique Index

Print view Relation view Propose table structure Move columns

Add 1 column(s) ● At End of Table ○ At Beginning of Table ○ After Id Go

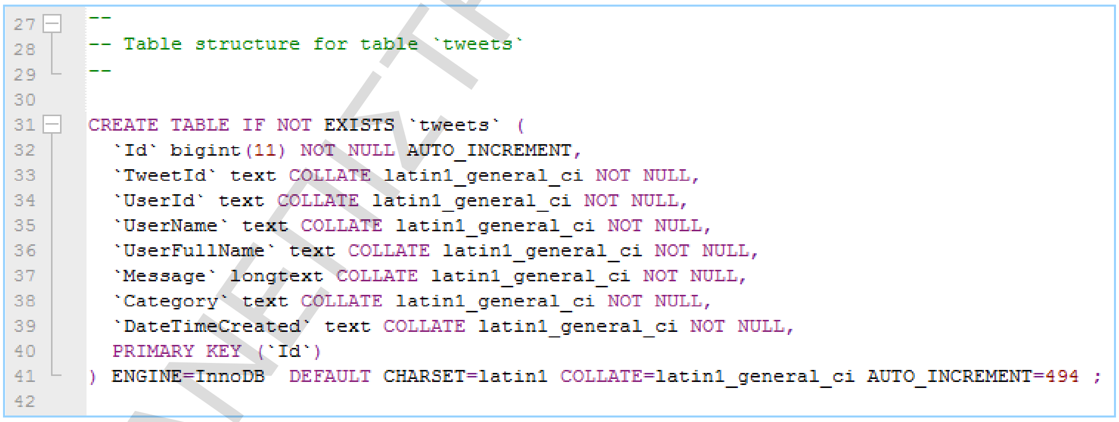
+ Indexes

Information

Space usage	
Data	10.5 MiB
Index	3.5 MiB
Total	14 MiB

Row statistics	
Format	Compact
Collation	latin1_general_ci
Next autoindex	51,007
Creation	Nov 20, 2013 at 02:07 PM

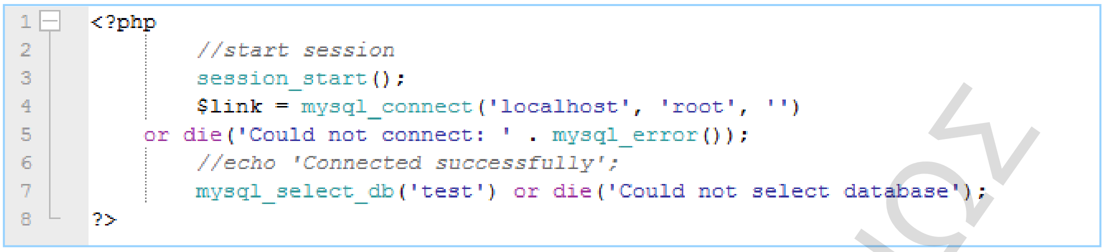
Figure 40: Structure of the Table "tweets" (a).



```
--
-- Table structure for table `tweets`
--

CREATE TABLE IF NOT EXISTS `tweets` (
  `Id` bigint(11) NOT NULL AUTO_INCREMENT,
  `TweetId` text COLLATE latin1_general_ci NOT NULL,
  `UserId` text COLLATE latin1_general_ci NOT NULL,
  `UserName` text COLLATE latin1_general_ci NOT NULL,
  `UserFullName` text COLLATE latin1_general_ci NOT NULL,
  `Message` longtext COLLATE latin1_general_ci NOT NULL,
  `Category` text COLLATE latin1_general_ci NOT NULL,
  `DateTimeCreated` text COLLATE latin1_general_ci NOT NULL,
  PRIMARY KEY (`Id`)
) ENGINE=InnoDB  DEFAULT CHARSET=latin1 COLLATE=latin1_general_ci AUTO_INCREMENT=494 ;
```

Figure 41: Structure of the Table "tweets" (b).

Therefore, the quality of the database design impacts software maintenance, as well as application performance and scalability. Meeting the requirements set during the initial stage of the design process contributes in avoiding coding errors or time loss. Also, a badly designed database will require more complex SQL code which will perform slowly[81].



```
<?php
        //start session
        session_start();
        $link = mysql_connect('localhost', 'root', '')
    or die('Could not connect: ' . mysql_error());
        //echo 'Connected successfully';
        mysql_select_db('test') or die('Could not select database');
?>
```

Figure 42: Establishing a Database Connection.

The architectural design of our database corresponds to the following scheme:

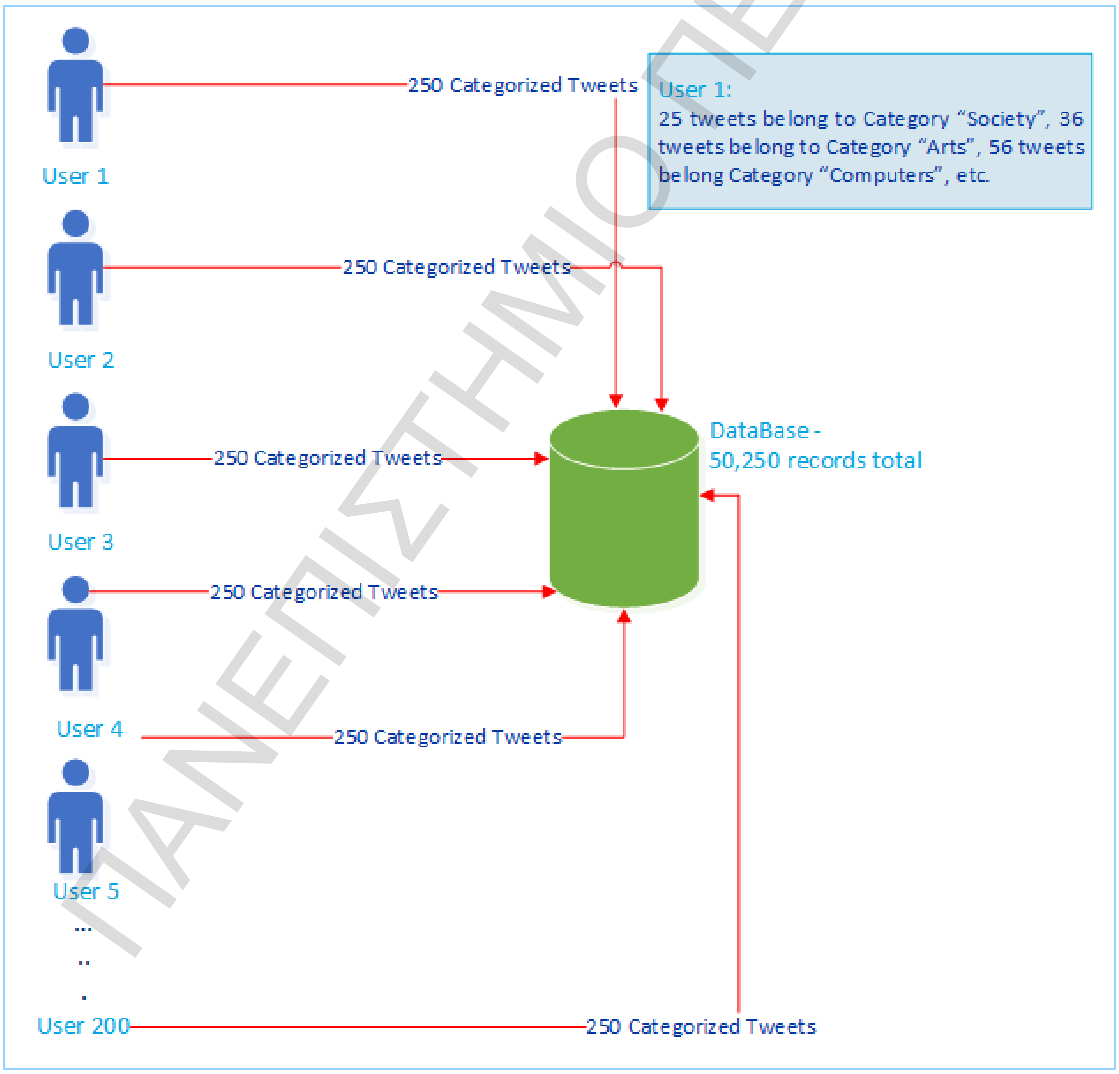
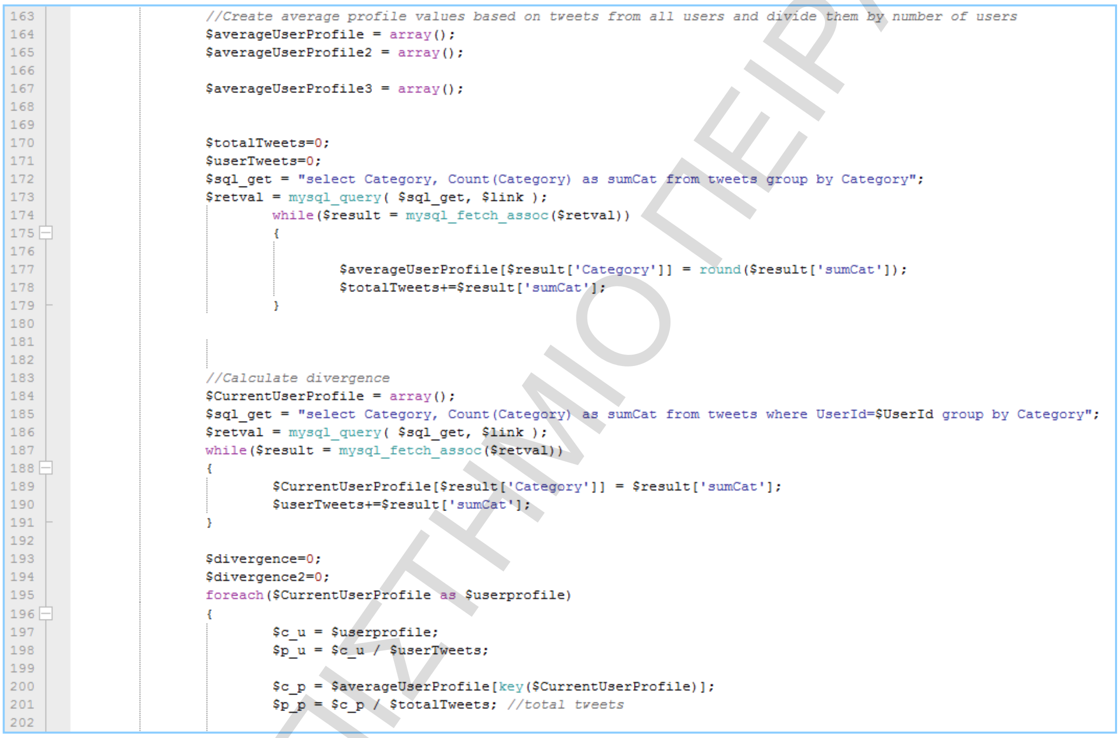


Figure 43: Process of Storing Tweets.

[81] Paul Nielsen, SQL Server MVP, *Database Design ROI*.

7.5 Process of Applying Information-Theoretic Parameters

A number of information-theoretic criteria were used to measure the privacy of user profiles. In the context of our work, profiling was considered as a dynamic process where data are retrieved, prepared, correlated and finally applied. The main goal of profiling procedure is the detection of the individual characteristics of a particular user. The risk of profiling raises severe concerns, therefore privacy metrics were employed in order to provide a certain level of protection to users, through the process of identifying and quantifying the potential privacy risks.



```
//Create average profile values based on tweets from all users and divide them by number of users
$averageUserProfile = array();
$averageUserProfile2 = array();

$averageUserProfile3 = array();


$totalTweets=0;
$userTweets=0;
$sql_get = "select Category, Count(Category) as sumCat from tweets group by Category";
$retval = mysql_query( $sql_get, $link );
        while($result = mysql_fetch_assoc($retval))
        {

                $averageUserProfile[$result['Category']] = round($result['sumCat']);
                $totalTweets+=$result['sumCat'];
        }



//Calculate divergence
$CurrentUserProfile = array();
$sql_get = "select Category, Count(Category) as sumCat from tweets where UserId=$UserId group by Category";
$retval = mysql_query( $sql_get, $link );
while($result = mysql_fetch_assoc($retval))
{
        $CurrentUserProfile[$result['Category']] = $result['sumCat'];
        $userTweets+=$result['sumCat'];
}

$divergence=0;
$divergence2=0;
foreach($CurrentUserProfile as $userprofile)
{
        $c_u = $userprofile;
        $p_u = $c_u / $userTweets;

        $c_p = $averageUserProfile[key($CurrentUserProfile)];
        $p_p = $c_p / $totalTweets; //total tweets

```

Figure 44: Calculating Average User Profile and Divergence Value.

8 Conclusions and Future Work

8.1 Concluding Remarks

In the era of networking the need for privacy protection is more intense than ever. A variety of approaches aimed at protecting users' privacy have gained considerable momentum in the literature review. Privacy enhancing technologies serve exactly this purpose. The methods they implement engage cryptography schemes, probabilistic mechanisms, multicast routing, multiple proxies or pseudonyms, all attempting to respond to more accurate and sophisticated attacks. However, despite the efforts, there still exists an ambiguity regarding their effectiveness and capacity to fulfill privacy requirements.

Therefore, the fact that user' online activities are susceptible to be traced, poses new privacy challenges. Social networking sites, such as Twitter, require special attention with regard to privacy, as they may aim to ascertain users' unique interests and preferences. Most users also have fundamental gaps in understanding the way applications handle their personal information, thereby enabling personalized systems collecting and correlating information users disclose, while interacting with these applications. Consequently, the detection of individual characteristics allows the adversary to extract a profile of interests for a particular user.

In our work, a web-based application was presented as an attempt of designing a system that provides the user a certain level of privacy protection by quantifying the profiling risk he deals with, depending on the content of the tweets he publishes. In accordance with the objectives set in the project, our tool is consisted of a group of different components, all of them forming an integrated framework for profiling users and measuring their privacy risk in Twitter. Also, both information-theoretic quantities, employed as our fundamental privacy criteria, were of great importance in the process of measuring the level of privacy of the information that a user wishes to protect.

8.2 Future Research Directions

In this section, we provide possible improvements and research directions based on the outcomes presented in this thesis. At first, the model of our user profile was developed according to a predefined set of categories of interest, derived from the classification technique TextWise. However, the process of analyzing text to extract meaningful information is quite demanding. Even though TextWise was an effective tool for performing categorization tasks, the mechanisms it uses are not known. Thus, we are unaware of the text mining algorithms and mathematical concepts it employs. Perhaps, Naïve Bayes classification would exhibit more precise results, since it relies on Bayesian decision theory and analysis; nevertheless, more time would be needed to develop an accurate training data set, while the objectives of our work did not encompass the development of a language processing component.

In addition, our experimental results were based on a data sample of fifty thousand tweets. A larger data set would enrich our study and provide more integrated outcomes regarding the user profiling process, and therefore the privacy risk level of Twitter users. A line for further research could also be to extend our understandings of privacy measurement through other social networking sites, such as Facebook and LinkedIn, which incorporate stricter security and privacy settings.

Also, the adversary model adopted by our case scenario refers to the attempts of the privacy attacker to develop profiles that identify a user, in the sense of distinguishing him from the rest of the population. However, there is breeding ground for further enhancements in this scenario. The adversary could also attempt to classify a particular user into a predefined group, with respect to his profile of interests. The categorization of a user as a member of a specific group of users could be a challenging issue to deal with, since Twitter does not allow for the moment to classify a user by employing individual characteristics, such as age, sex, ethnic background, location, etc.

Another pending work could also be the implementation or the modification of the existing privacy enhancing mechanisms according to the privacy risk level of a user. A system that quantifies user privacy in combination with PETs could provide a secure and effective solution regarding the protection of users' privacy.

9 References

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, Ke Tao, *Twitter-Based User Modeling for News Recommendations*, Web Information Systems, Delft University of Technology.

[2] Alessandro Acquisti, Ralph Gross, R., *Imagined communities: Awareness, information sharing, and privacy on the Facebook*, PET 2006.

[3] Charu C. Aggarwal, Tarek Abdelzaher, *Integrating sensors and social networks*, Chapter 14.

[4] Ashish T. Bhole, Savita H. Lambole, *Design and Implementation of Distributed Security using Onion Routing*, International Conference & Workshop on Recent Trends in Technology, (TCET) 2012.

[5] Stelios Dritsas, Dimitris Gritzalis, Costas Lambrinoudakis, *Protecting privacy and anonymity in pervasive computing: trends and perspectives*, Volume 23, Issue 3, August 2006, Pages 196–210.

[6] Alessandro Acquisti, Stefanos Gritzalis, Costas Lambrinoudakis, Sabrina de Capitani di Vimercati, *Digital Privacy: Theory, Technologies and Practices*, Auerbach Publications, 2008.

[7] G.W. van Blarkom, J.J. Borking, J.G.E. Olk, *Handbook of Privacy and Privacy-Enhancing Technologies - The case of Intelligent Software Agents*, Privacy Incorporated Software Agent (PISA) Consortium, the Hague, the Netherlands, 2003.

[8] Stefan Brands, *Non-Intrusive Identity Management*, McGill School of Computer Science & Credentica, March 23, 2004.

[9] Carter T. Butts, *Social network analysis: A methodological introduction*, Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California, Irvine, California, USA, Asian Journal of Social Psychology (2008), 11, 13–4.

[10] Claude Castelluccia, Arvind Narayanan, *Privacy considerations of online behavioural tracking*, European Network and Information Security Agency (ENISA), October 2012.

[11] Ann Cavoukian, Privacy *by Design, The 7 Foundational Principles, Implementation and Mapping of Fair Information Practices*, Information & Privacy Commissioner, Ontario, Canada.

[12] A. Erola, J. Castella-Roca, A. Viejo, and J. M. Mateo-Sanz, *Exploiting social networks to provide privacy in personalized Web search*, J. Syst., Softw., vol. 84, no. 10, pp. 1734-745, 2011.

[13] Leticia Fernández Franco, *A survey and comparison of anonymous communication systems: anonymity and security*, Universitat Oberta de Catalunya (UOC), June 2012.

[14] David M. Goldschlag, Michael G. Reed, Paul F. Syverson, *Hiding Routing Information*, Workshop on Information Hiding, Cambridge, UK, May, 1996.

[15] R. Gross and A. Acquisti, *Information Revelation and Privacy in Online Social Networks*, Workshop on Privacy in the Electronic Society (WPES), 2005.

[16] Samuel D. Warren and Louis D. Brandeis, *The Right to Privacy*, Harvard Law Review, Vol. 4, No. 5 (Dec. 15, 1890), pp. 193-220.

[17] Edwin T. Jaynes, *On the rationale of maximum-entropy methods*, Proc. IEEE, vol. 70, no. 9, pp. 939{952, Sep. 1982.

[18] Christos Kalloniatis, Evangelia Kavakli, and Stefanos Gritzalis, *Addressing privacy requirements in system design: the PriS method*, September 2008, Volume 13, Issue 3, pp 241-255.

[19] Lucy L. Thomson, Human Rights Electronic Evidence Study, *Admissibility of electronic documentation as evidence in U. S. courts*, December 1, 2011.

[20] Andreas M. Kaplan, Michael Haenlein, *Users of the world, unite! The challenges and opportunities of Social Media*, Kelley School of Business, Indiana University, Business Horizons (2010) 53, 59-68, available online at www.sciencedirect.com.

[21] Stefanos Gritzalis, *Enhancing Web privacy and anonymity in the digital era*, Information and Communication Systems Security Laboratory, Department of Information and Communications Systems Engineering, University of the Aegean, Samos, Greece.

[22] Steven Furnell, Costas Lambrinoudakis, Javier Lopez (Eds.), *Trust, Privacy, and Security in Digital Business*, 10th International Conference, TrustBus 2013, LNCS 8058, pp. 62–73, Springer-Verlag Berlin Heidelberg 2013.

[23] C. Lambrinoudakis, S. Gritzalis, S. Katsikas, *Privacy Protection in Information and Communication Technologies: Technical and Legal Issues*, Papasotiriou, Greece, 2010.

[24] Brian Reed, *The "Mysteries of Human Life": Dealing with an Ambiguous Right to Privacy*, Planned Parenthood v. Casey 505 U.S. 833, 851 (1992).

[25] Michael K. Reiter, Aviel D. Rubin, *Crowds: Anonymity for Web Transactions*, AT&T Labs-Research.

[26] Stanley Wasserman, Katherine Faust, *Social Network Analysis: Methods and Applications, Structural analysis in the social sciences*, ENG and New York: Cambridge University Press, 1994.

[27] Douglas A. Luke and Jenine K. Harris, *Network Analysis in Public Health: History, Methods and Applications*, Department of Community Health, School of Public Health, Saint Louis University, Annual Review of Public Health, (2007), 28:69–93.

[28] Behnam Hajian, Tony White, *Modelling Influence in a Social Network: Metrics and Evaluation*, School of Computer Science, Carleton University, Ottawa, Canada.

[29] José Antonio Estrada Jiménez, *Implementation of a Firefox Extension that Measures User Privacy Risk in Web Search*, Master Thesis, Department of Telematics Engineering, Universitat Politècnica de Catalunya, 2013.

[30] Ana Fernanda Rodríguez Hoyos, *Evaluation of the privacy risk for online search and social tagging systems*, Master Thesis, Department of Telematics Engineering, Universitat Politècnica de Catalunya, 2013.

[31] Lee Humphreys, Phillipa Gill, Balachander Krishnamurthy, *Privacy on twitter-How much is too much? Privacy issues on Twitter*.

[32] Benjamin Greschbach and Sonja Buchegger, *Friendly Surveillance – A New Adversary Model for Privacy in Decentralized Online Social Networks*.

[33] John Lowry, Rico Valdez, Brad Wood, *Adversary Modeling to Develop Forensic Observables*.

[34] Erika McCallister, Tim Grance, Karen Scarfone, Recommendations of the National Institute of Standards and Technology, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, NIST Special Publication 800-122, April 2010.

[35] Lilian Mitrou, Maria Karyda, *EU´s Data Protection Reform and the right to be forgotten - A legal response to a technological challenge?*

[36] D. Rebollo-Monedero, J. Forné, *Optimal query forgery for private information retrieval*, IEEE Trans. Information Theory, vol. 56, no. 9, pp. 4631-4642, 2010.

[37] D. Rebollo-Monedero, J. Parra-Arnau, J. Forné, *An Information-Theoretic Privacy Criterion for Query Forgery in Information Retrieval*, Department of Telematics Engineering, Technical University of Catalonia (UPC), arXiv:1111.4045v1 [cs.IT] 17 Nov 2011.

[38] D. Rebollo-Monedero, J. Forné, J. Domingo-Ferrer (2012), *Query Profile Obfuscation by Means of Optimal Query Exchange between Users*, IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 5, pp. 641-654, Sept.-Oct. 2012

[39] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, *Measuring the Privacy of User Profiles in Personalized Information Systems*, Future Generation Computer Systems (FGCS), 2013.

[40] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forné, D. Rebollo-Monedero (2013), *Privacy-Preserving Enhanced Collaborative Tagging*, IEEE Trans. Knowl. Data Eng., Published http://dx.doi.org/10.1109/TKDE.2012.248.

[41] D. Rebollo-Monedero, J. Forné, A. Solanas, A. Martinez-Ballesté, *Private Location-Based Information Retrieval through User Collaboration*. Computer Communications, 33 (6): 762-774, 2010.

[42] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, J. L. Muñoz, O. Esparza (2012), *Optimal Tag Suppression for Privacy Protection in the Semantic Web*, Elsevier Data & Knowl. Eng., 81: 46-66.

[43] C. Tripp Barba, L. Urquiza Aguiar, M. Aguilar Igartua, J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, E. Pallarès (2013), *A Collaborative Protocol for Anonymous Reporting in Vehicular Ad Hoc Networks*, Computer Standards & Interfaces. Volume 36, Issue 1, November 2013, Pages 188–197

[44] Javier Parra Arnau, *Privacy protection of user profiles in personalized information systems*, Dissertation for the Degree of Doctor of Philosophy, Dept. of Telematics Engineering, Universitat Politècnica de Catalunya, September 2013.

[45] Michael Hay, Gerome Miklau, David Jensen, Philipp Weis, Siddharth Srivastava, *Anonymizing Social Networks*, Computer Science Department Faculty Publication, University of Massachusetts – Amherst, 2007.

[46] Omar Hasan, Benjamin Habegger, Lionel Brunie, Nadia Bennani, Ernesto Damiani, *A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case*. In IEEE 2nd International Congress on Big Data, Santa Clara, CA, USA, 2013.

[47] Aniket Kate, Greg M. Zaverucha, Ian Goldberg, *Pairing-Based Onion Routing with Improved Forward Secrecy*, David R. Cheriton School of Computer Science, University of Waterloo.

[48] Ian Goldberg, *Privacy Enhancing Technologies for the Internet III: Ten Years Later*, David R. Cheriton School of Computer Science, University of Waterloo.

[49] Muhamed Ilyas, Dr. R. Vijayakumar, *LPM: A distributed architecture and algorithms for location privacy in LBS*, International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.

[50] Konstantina Vemou and Maria Karyda, *A Classification of Factors Influencing Low Adoption of PETs Among SNS Users*, Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece.

[51] Andreas Krisch, *RFID Privacy Issues*, Contribution to the RFID Expert Group Meeting, 10 July 2007.

[52] Paul Nielsen, *SQL Server MVP*, Database Design ROI.

[53] Femi Olumofin, Ian Goldberg, *Revisiting the Computational Practicality of Private Information Retrieval*, Cheriton School of Computer Science, University of Waterloo.

[54] Meredith L. Patterson, Len Sassaman, *Subliminal Channels in the Private Information Retrieval Protocols*. In: Proceedings of the 28th Symposium on Information Theory in the Benelux, 2007.

[55] Andreas Pfitzmann, Marit Hansen, *Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management – A consolidated proposal for terminology*, 2008.

[56] Fabrizio Sebastiani, *Text Categorization*, Universita di Padova, Padova, Italy.

[57] Yun Shen, Siani Pearson, *Privacy Enhancing Technologies: A Review*, HP Laboratories, HPL-2011-113.

[58] Agusti Solanas, Josep Domingo-Ferrer, Antoni Martınez-Ballest, *Location Privacy in Location-Based Services: Beyond TTP-based Schemes*, 1st International Workshop on Privacy in Location-Based Applications (PILBA 2008) within 13th European Symposium on Research in Computer Security (ESORICS 2008), Malaga, Spain, Oct 2008. ISBN: 1613-0073.

[59] Eran Toch, Yang Wang, Lorrie Faith Cranor, *Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems*, Published online: 10 March 2012, DOI 10.1007/s11257-011-9110-z.

[60] Sergey Yekhanin, *Locally Decodable Codes and Private Information Retrieval Schemes*, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, July 2007.

[61] *Social networks analysis, theory and applications*, PDF generated at Mon, 03 Jan 2011.

[62] Smart Card Alliance, *Privacy and Secure Identification Systems: The role of smart cards as a privacy-enabling technology*, ID-03001, February 2003.

[63] Privacy Office, U.S. Department of Homeland Security Washington, *Privacy Technology Implementation Guide*, August 16, 2007.

[64] The Pew Research Center's Internet and American Life Project, April 26 – May 22, 2011 Spring Tracking Survey.

Electronic References

[65] Data protection and privacy laws, https://www.privacyinternational.org/issues/data-protection-and-privacy-laws.

[66] MozillaZine, User Tracking, http://kb.mozillazine.org/User_tracking.

[67] The Office of the Data Protection Commissioner, Law on data protection-Data Protection Acts, http://dataprotection.ie/ViewDoc.asp?fn=%2Fdocuments%2Flegal%2FLawOnDP.htm&CatID=7&m=l

[68] Pinsent Masons, International law firm, http://www.out-law.com/page-413.

[69] Free Haven, http://www.freehaven.net/.

[70] Crazy engineers, http://www.crazyengineers.com.

[71] Naive Bayesian Text Classification, http://www.drdobbs.com/architecture-and-design/naive-bayesian-text-classification/184406064.

[72] Python Course, Text classification in Python, http://www.python-course.eu/text_classification_python.php.

[73] Sign in with Twitter, http://www.saaraan.com/2012/08/sign-in-with-twitter-using-php.

[74] Twitter, https://twitter.com/i/connect.

[75] Twitter developers, https://dev.twitter.com/docs.

[76] Textwise, http://www.textwise.com/.

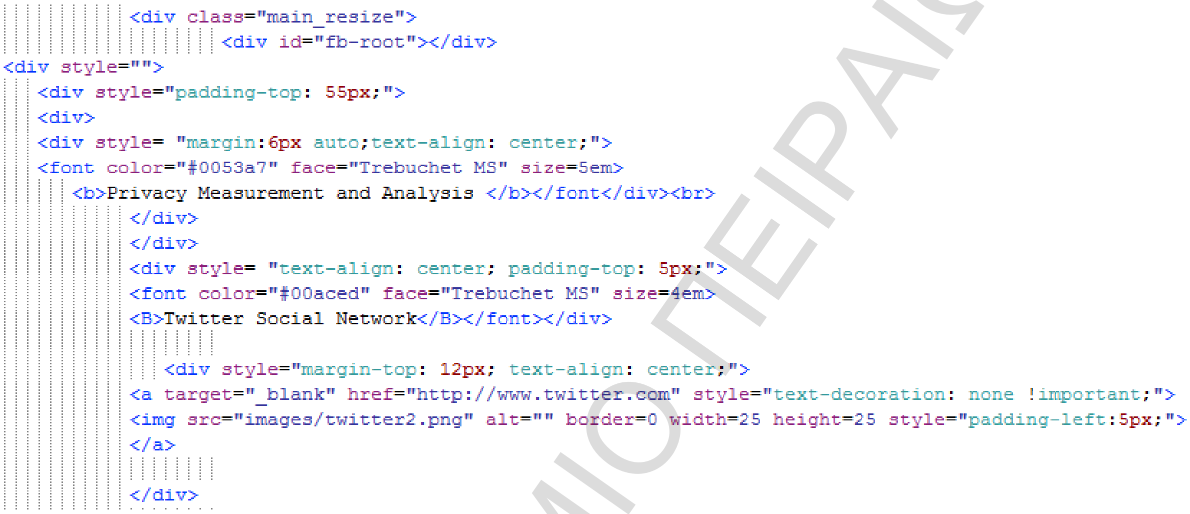
[77] TextWise API, http://www.textwise.com/api.

[78] Wikipedia, the free encyclopedia, www.en.wikipedia.org/.

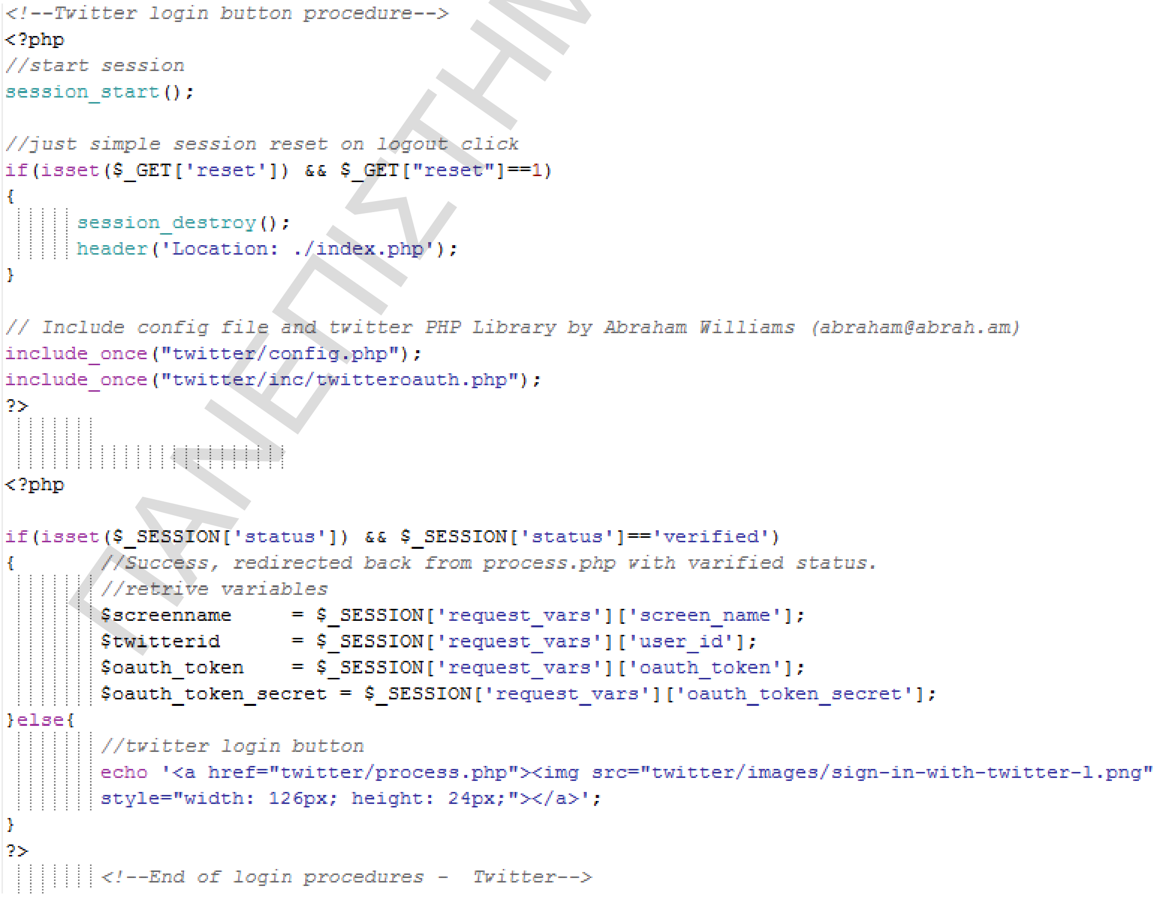
10 Appendices

This chapter contains a sample of PHP code for basic structural components that were developed to fulfill the objectives of this master thesis.

10.1 Index Page



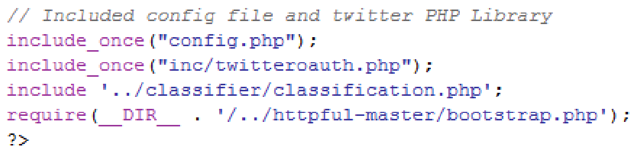
```
           <div class="main_resize">
                   <div id="fb-root"></div>
<div style="">
   <div style="padding-top: 55px;">
   <div>
   <div style= "margin:6px auto;text-align: center;">
   <font color="#0053a7" face="Trebuchet MS" size=5em>
      <b>Privacy Measurement and Analysis </b></font</div><br>
           </div>
           </div>
           <div style= "text-align: center; padding-top: 5px;">
           <font color="#00aced" face="Trebuchet MS" size=4em>
           <B>Twitter Social Network</B></font></div>

              <div style="margin-top: 12px; text-align: center;">
           <a target="_blank" href="http://www.twitter.com" style="text-decoration: none !important;">
           <img src="images/twitter2.png" alt="" border=0 width=25 height=25 style="padding-left:5px;">
           </a>

           </div>
```



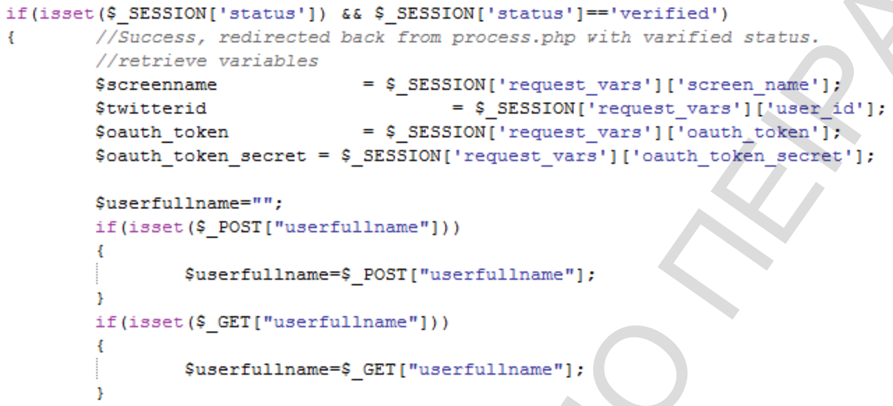
```
<!--Twitter login button procedure-->
<?php
//start session
session_start();

//just simple session reset on logout click
if(isset($_GET['reset']) && $_GET["reset"]==1)
{
      session_destroy();
      header('Location: ./index.php');
}

// Include config file and twitter PHP Library by Abraham Williams (abraham@abrah.am)
include_once("twitter/config.php");
include_once("twitter/inc/twitteroauth.php");
?>


<?php

if(isset($_SESSION['status']) && $_SESSION['status']=='verified')
{       //Success, redirected back from process.php with varified status.
        //retrive variables
        $screenname     = $_SESSION['request_vars']['screen_name'];
        $twitterid      = $_SESSION['request_vars']['user_id'];
        $oauth_token    = $_SESSION['request_vars']['oauth_token'];
        $oauth_token_secret = $_SESSION['request_vars']['oauth_token_secret'];
}else{
        //twitter login button
        echo '<a href="twitter/process.php"><img src="twitter/images/sign-in-with-twitter-l.png"
        style="width: 126px; height: 24px;"></a>';
}
?>
        <!--End of login procedures -  Twitter-->
```

10.2 Twitter Functions

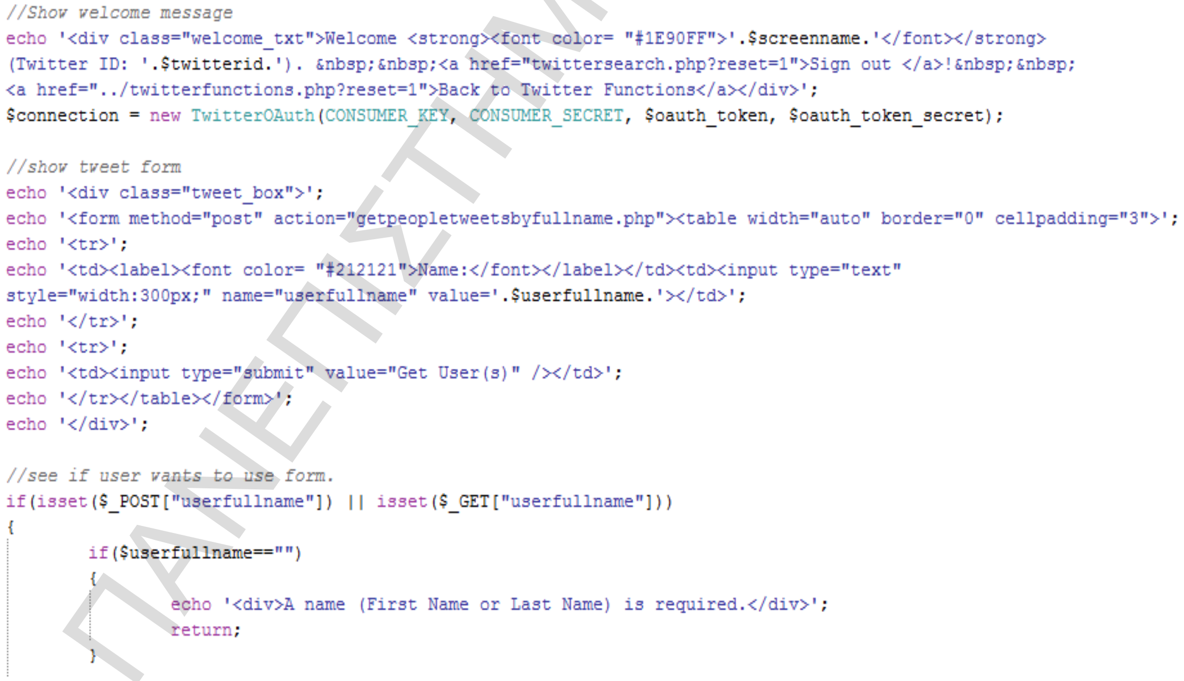
10.2.1 Getting Users based on different Criteria (followers, friends, full name, screen name)



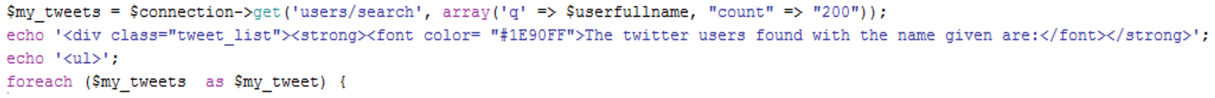
```
// Included config file and twitter PHP Library
include_once("config.php");
include_once("inc/twitteroauth.php");
include '../classifier/classification.php';
require(__DIR__ . '/../httpful-master/bootstrap.php');
?>
```



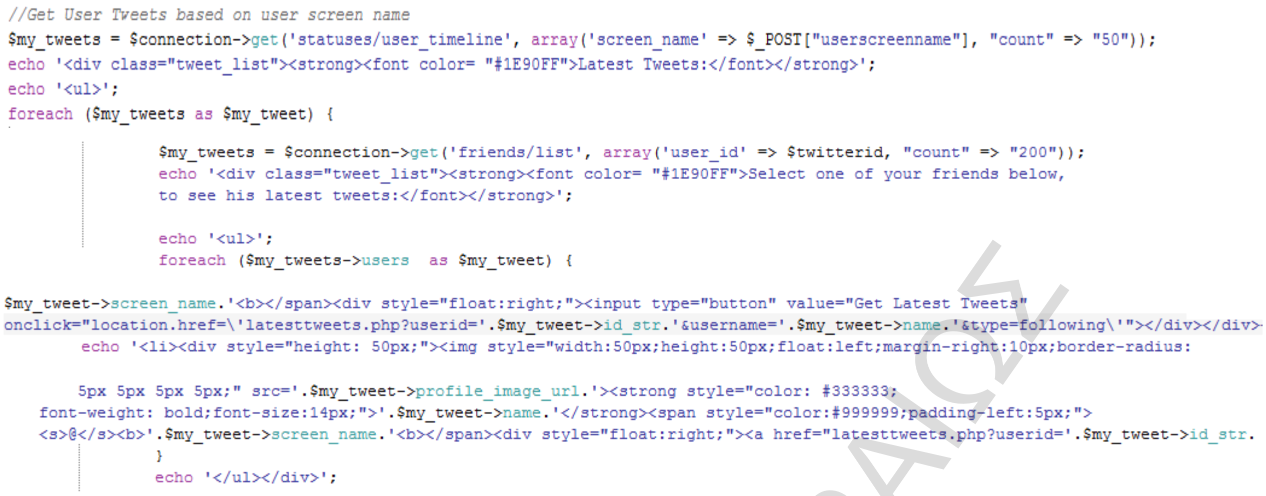
```
if(isset($_SESSION['status']) && $_SESSION['status']=='verified')
{       //Success, redirected back from process.php with varified status.
        //retrieve variables
        $screenname             = $_SESSION['request_vars']['screen_name'];
        $twitterid                      = $_SESSION['request_vars']['user_id'];
        $oauth_token            = $_SESSION['request_vars']['oauth_token'];
        $oauth_token_secret = $_SESSION['request_vars']['oauth_token_secret'];

        $userfullname="";
        if(isset($_POST["userfullname"]))
        {
                $userfullname=$_POST["userfullname"];
        }
        if(isset($_GET["userfullname"]))
        {
                $userfullname=$_GET["userfullname"];
        }
```



```
//Show welcome message
echo '<div class="welcome_txt">Welcome <strong><font color= "#1E90FF">'.$screenname.'</font></strong>
(Twitter ID: '.$twitterid.'). &nbsp;&nbsp;<a href="twittersearch.php?reset=1">Sign out </a>!&nbsp;&nbsp;
<a href="../twitterfunctions.php?reset=1">Back to Twitter Functions</a></div>';
$connection = new TwitterOAuth(CONSUMER_KEY, CONSUMER_SECRET, $oauth_token, $oauth_token_secret);

//show tweet form
echo '<div class="tweet_box">';
echo '<form method="post" action="getpeopletweetsbyfullname.php"><table width="auto" border="0" cellpadding="3">';
echo '<tr>';
echo '<td><label><font color= "#212121">Name:</font></label></td><td><input type="text"
style="width:300px;" name="userfullname" value='.$userfullname.'></td>';
echo '</tr>';
echo '<tr>';
echo '<td><input type="submit" value="Get User(s)" /></td>';
echo '</tr></table></form>';
echo '</div>';

//see if user wants to use form.
if(isset($_POST["userfullname"]) || isset($_GET["userfullname"]))
{
        if($userfullname=="")
        {
                echo '<div>A name (First Name or Last Name) is required.</div>';
                return;
        }
```

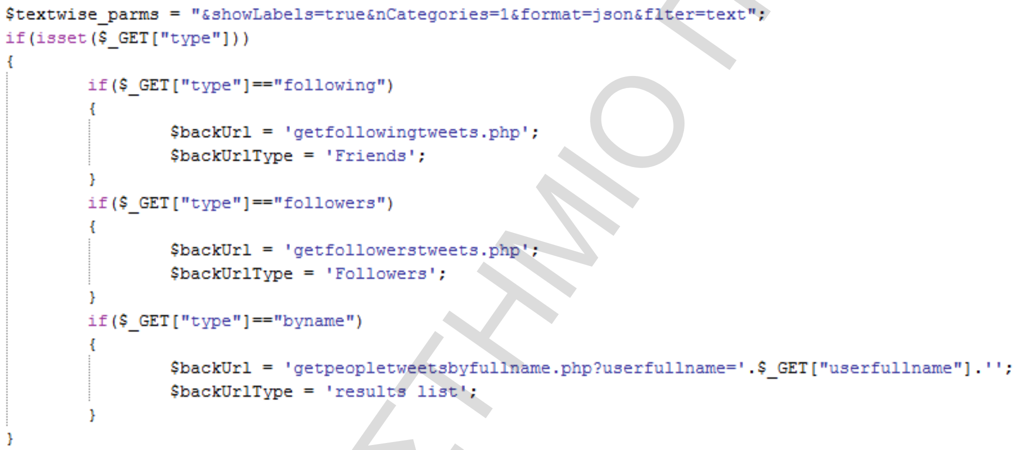


```
$my_tweets = $connection->get('users/search', array('q' => $userfullname, "count" => "200"));
echo '<div class="tweet_list"><strong><font color= "#1E90FF">The twitter users found with the name given are:</font></strong>';
echo '<ul>';
foreach ($my_tweets  as $my_tweet) {
```

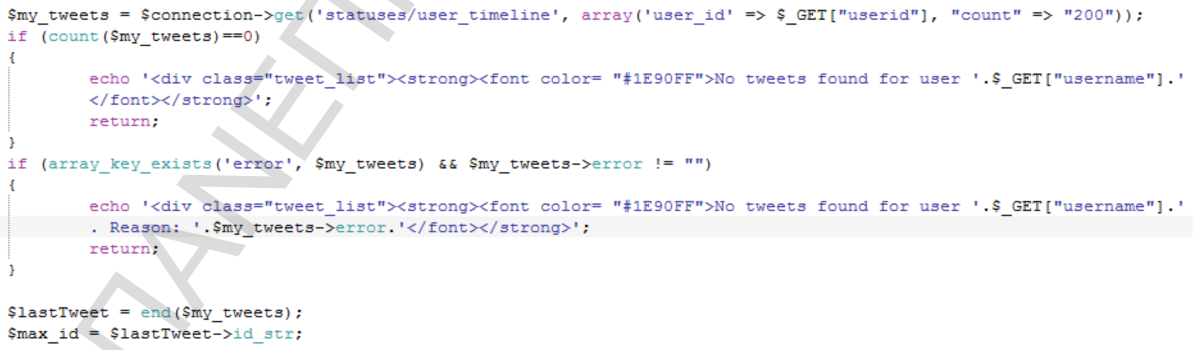


```
//Get User Tweets based on user screen name
$my_tweets = $connection->get('statuses/user_timeline', array('screen_name' => $_POST["userscreenname"], "count" => "50"));
echo '<div class="tweet_list"><strong><font color= "#1E90FF">Latest Tweets:</font></strong>';
echo '<ul>';
foreach ($my_tweets as $my_tweet) {

                $my_tweets = $connection->get('friends/list', array('user_id' => $twitterid, "count" => "200"));
                echo '<div class="tweet_list"><strong><font color= "#1E90FF">Select one of your friends below,
                to see his latest tweets:</font></strong>';

                echo '<ul>';
                foreach ($my_tweets->users  as $my_tweet) {

$my_tweet->screen_name.'<b></span><div style="float:right;"><input type="button" value="Get Latest Tweets"
onclick="location.href=\'latesttweets.php?userid='.$my_tweet->id_str.'&username='.$my_tweet->name.'&type=following\'"></div></div>
        echo '<li><div style="height: 50px;"><img style="width:50px;height:50px;float:left;margin-right:10px;border-radius:

       5px 5px 5px 5px;" src='.$my_tweet->profile_image_url.'><strong style="color: #333333;
    font-weight: bold;font-size:14px;">'.$my_tweet->name.'</strong><span style="color:#999999;padding-left:5px;">
    <s>@</s><b>'.$my_tweet->screen_name.'<b></span><div style="float:right;"><a href="latesttweets.php?userid='.$my_tweet->id_str.
                }
                echo '</ul></div>';
```

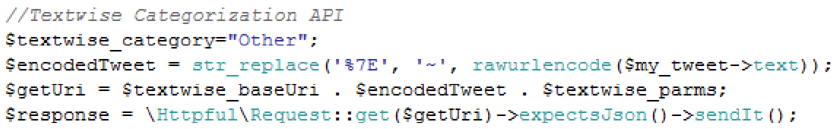
10.3 Latest Tweets Retrieval and Categorization



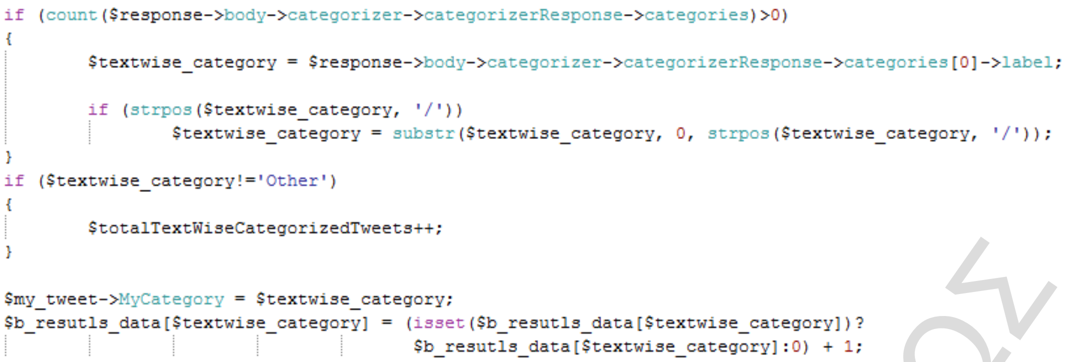
```
$textwise_parms = "&showLabels=true&nCategories=1&format=json&flter=text";
if(isset($_GET["type"]))
{
        if($_GET["type"]=="following")
        {
                $backUrl = 'getfollowingtweets.php';
                $backUrlType = 'Friends';
        }
        if($_GET["type"]=="followers")
        {
                $backUrl = 'getfollowerstweets.php';
                $backUrlType = 'Followers';
        }
        if($_GET["type"]=="byname")
        {
                $backUrl = 'getpeopletweetsbyfullname.php?userfullname='.$_GET["userfullname"].'';
                $backUrlType = 'results list';
        }
}
```



```
$my_tweets = $connection->get('statuses/user_timeline', array('user_id' => $_GET["userid"], "count" => "200"));
if (count($my_tweets)==0)
{
        echo '<div class="tweet_list"><strong><font color= "#1E90FF">No tweets found for user '.$_GET["username"].'
        </font></strong>';
        return;
}
if (array_key_exists('error', $my_tweets) && $my_tweets->error != "")
{
        echo '<div class="tweet_list"><strong><font color= "#1E90FF">No tweets found for user '.$_GET["username"].'
        . Reason: '.$my_tweets->error.'</font></strong>';
        return;
}

$lastTweet = end($my_tweets);
$max_id = $lastTweet->id_str;
```



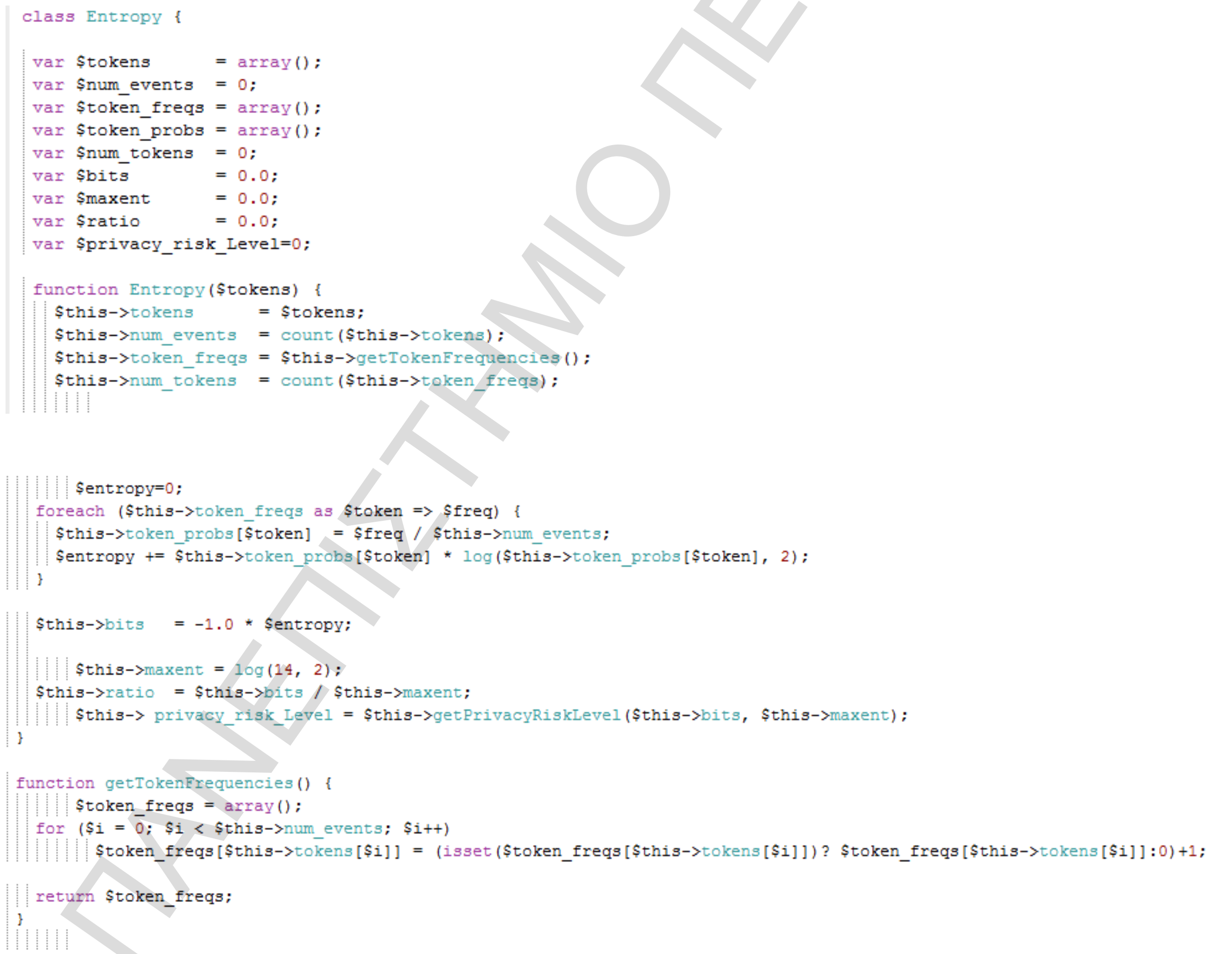
```
//Textwise Categorization API
$textwise_category="Other";
$encodedTweet = str_replace('%7E', '~', rawurlencode($my_tweet->text));
$getUri = $textwise_baseUri . $encodedTweet . $textwise_parms;
$response = \Httpful\Request::get($getUri)->expectsJson()->sendIt();
```



```
if (count($response->body->categorizer->categorizerResponse->categories)>0)
{
        $textwise_category = $response->body->categorizer->categorizerResponse->categories[0]->label;

        if (strpos($textwise_category, '/'))
                $textwise_category = substr($textwise_category, 0, strpos($textwise_category, '/'));
}
if ($textwise_category!='Other')
{
        $totalTextWiseCategorizedTweets++;
}

$my_tweet->MyCategory = $textwise_category;
$b_resutls_data[$textwise_category] = (isset($b_resutls_data[$textwise_category])?
                                       $b_resutls_data[$textwise_category]:0) + 1;
```

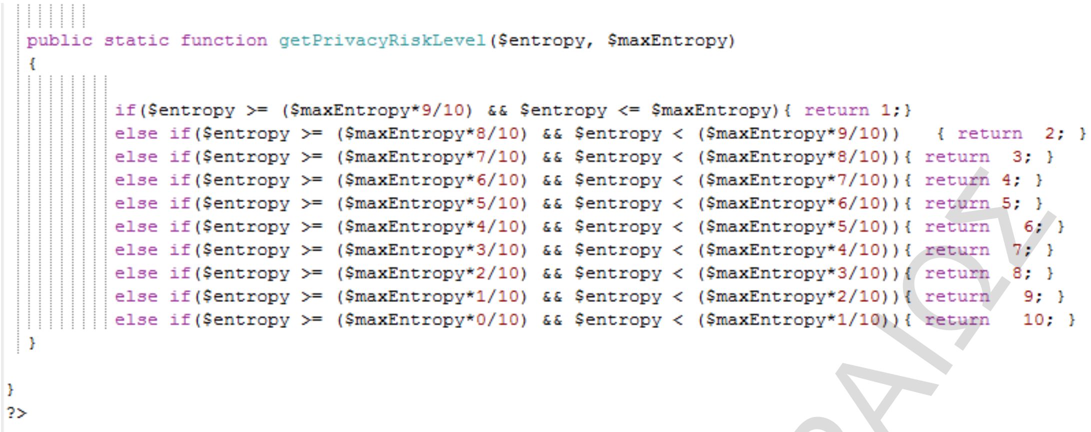
10.5 Measuring User Privacy

10.5.1 Function for calculating Entropy of a User Profile



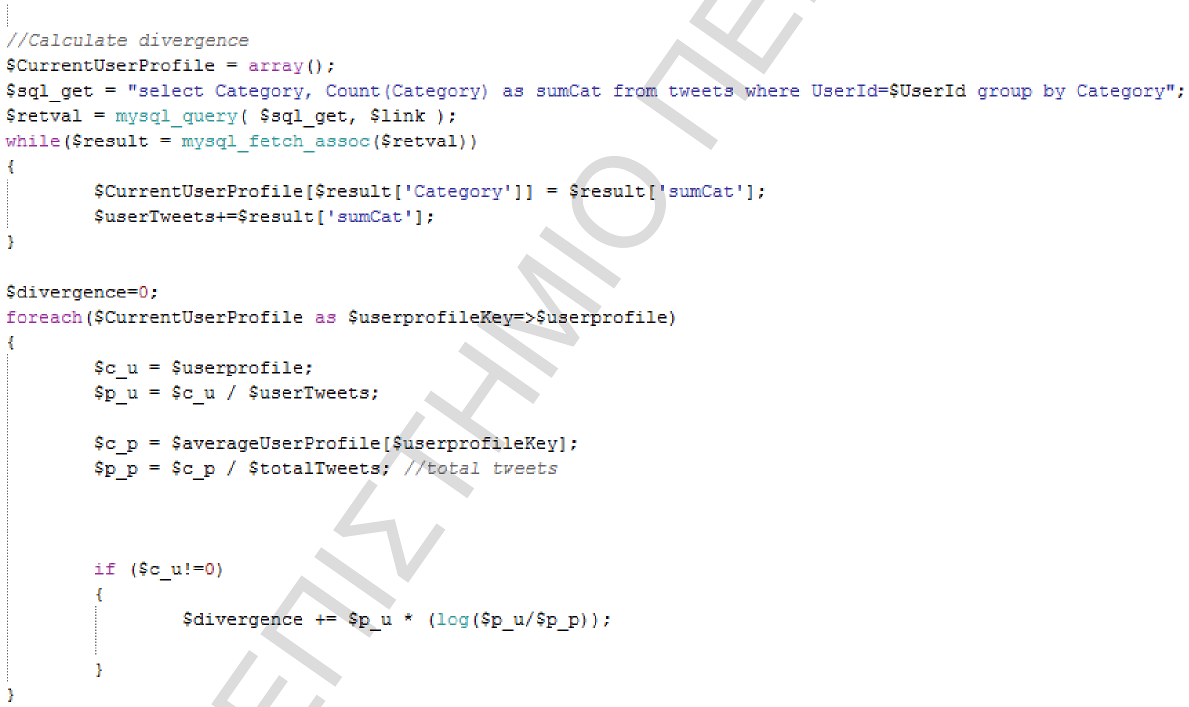
```
class Entropy {

 var $tokens      = array();
 var $num_events  = 0;
 var $token_freqs = array();
 var $token_probs = array();
 var $num_tokens  = 0;
 var $bits        = 0.0;
 var $maxent      = 0.0;
 var $ratio       = 0.0;
 var $privacy_risk_Level=0;

 function Entropy($tokens) {
   $this->tokens      = $tokens;
   $this->num_events  = count($this->tokens);
   $this->token_freqs = $this->getTokenFrequencies();
   $this->num_tokens  = count($this->token_freqs);



       $entropy=0;
  foreach ($this->token_freqs as $token => $freq) {
   $this->token_probs[$token]  = $freq / $this->num_events;
   $entropy += $this->token_probs[$token] * log($this->token_probs[$token], 2);
  }

  $this->bits   = -1.0 * $entropy;

      $this->maxent = log(14, 2);
  $this->ratio  = $this->bits / $this->maxent;
      $this-> privacy_risk_Level = $this->getPrivacyRiskLevel($this->bits, $this->maxent);
}

function getTokenFrequencies() {
      $token_freqs = array();
  for ($i = 0; $i < $this->num_events; $i++)
        $token_freqs[$this->tokens[$i]] = (isset($token_freqs[$this->tokens[$i]])? $token_freqs[$this->tokens[$i]]:0)+1;

  return $token_freqs;
}
```

10.5.2 Function for calculating Privacy Risk Level



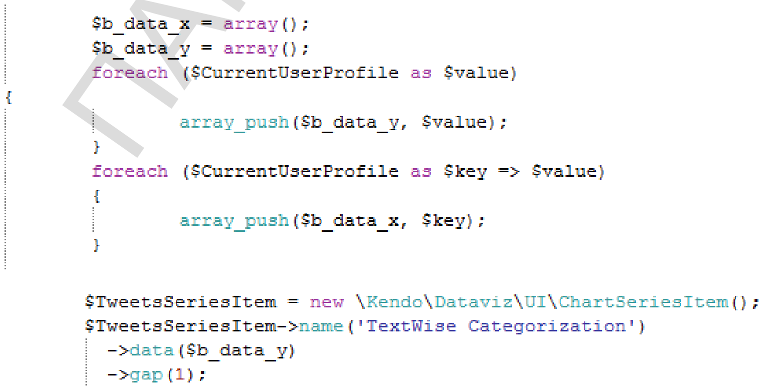
```
  public static function getPrivacyRiskLevel($entropy, $maxEntropy)
  {
          if($entropy >= ($maxEntropy*9/10)  && $entropy <= $maxEntropy){ return 1;}
          else if($entropy >= ($maxEntropy*8/10)  && $entropy < ($maxEntropy*9/10))   { return  2; }
          else if($entropy >= ($maxEntropy*7/10)  && $entropy < ($maxEntropy*8/10)){ return  3; }
          else if($entropy >= ($maxEntropy*6/10)  && $entropy < ($maxEntropy*7/10)){ return 4; }
          else if($entropy >= ($maxEntropy*5/10)  && $entropy < ($maxEntropy*6/10)){ return 5; }
          else if($entropy >= ($maxEntropy*4/10)  && $entropy < ($maxEntropy*5/10)){ return   6; }
          else if($entropy >= ($maxEntropy*3/10)  && $entropy < ($maxEntropy*4/10)){ return  7; }
          else if($entropy >= ($maxEntropy*2/10)  && $entropy < ($maxEntropy*3/10)){ return  8; }
          else if($entropy >= ($maxEntropy*1/10)  && $entropy < ($maxEntropy*2/10)){ return   9; }
          else if($entropy >= ($maxEntropy*0/10)  && $entropy < ($maxEntropy*1/10)){ return   10; }
  }

}
?>
```

10.5.3 Function for Calculating Divergence

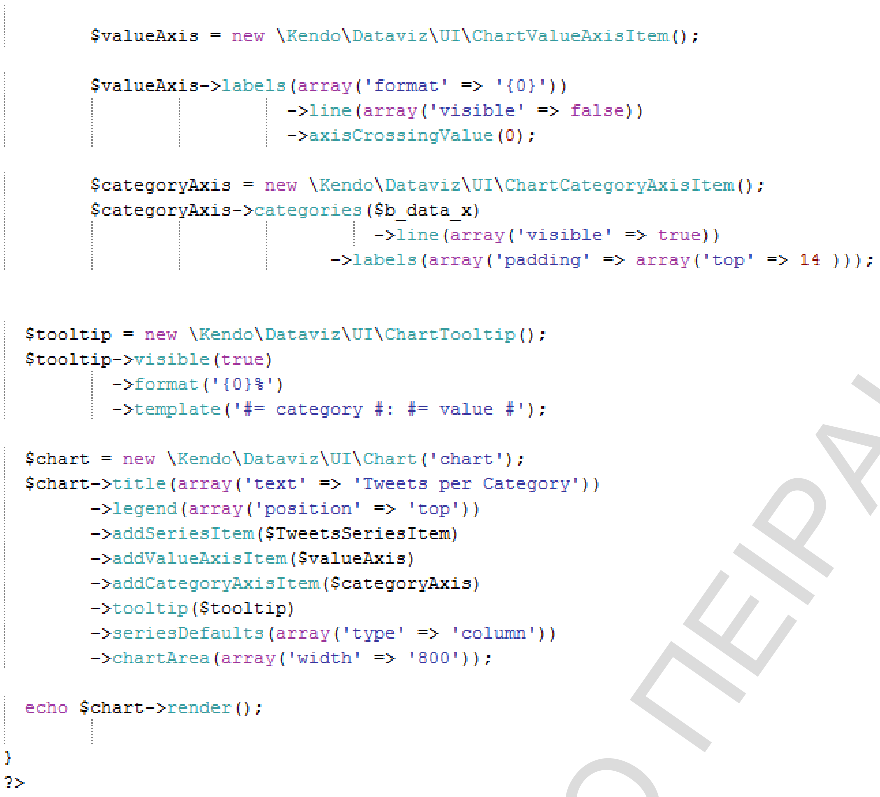


```
//Calculate divergence
$CurrentUserProfile = array();
$sql_get = "select Category, Count(Category) as sumCat from tweets where UserId=$UserId group by Category";
$retval = mysql_query( $sql_get, $link );
while($result = mysql_fetch_assoc($retval))
{
        $CurrentUserProfile[$result['Category']] = $result['sumCat'];
        $userTweets+=$result['sumCat'];
}

$divergence=0;
foreach($CurrentUserProfile as $userprofileKey=>$userprofile)
{
        $c_u = $userprofile;
        $p_u = $c_u / $userTweets;

        $c_p = $averageUserProfile[$userprofileKey];
        $p_p = $c_p / $totalTweets; //total tweets


        if ($c_u!=0)
        {
                $divergence += $p_u * (log($p_u/$p_p));

        }
}
```

10.6 Loading Chart of Interests

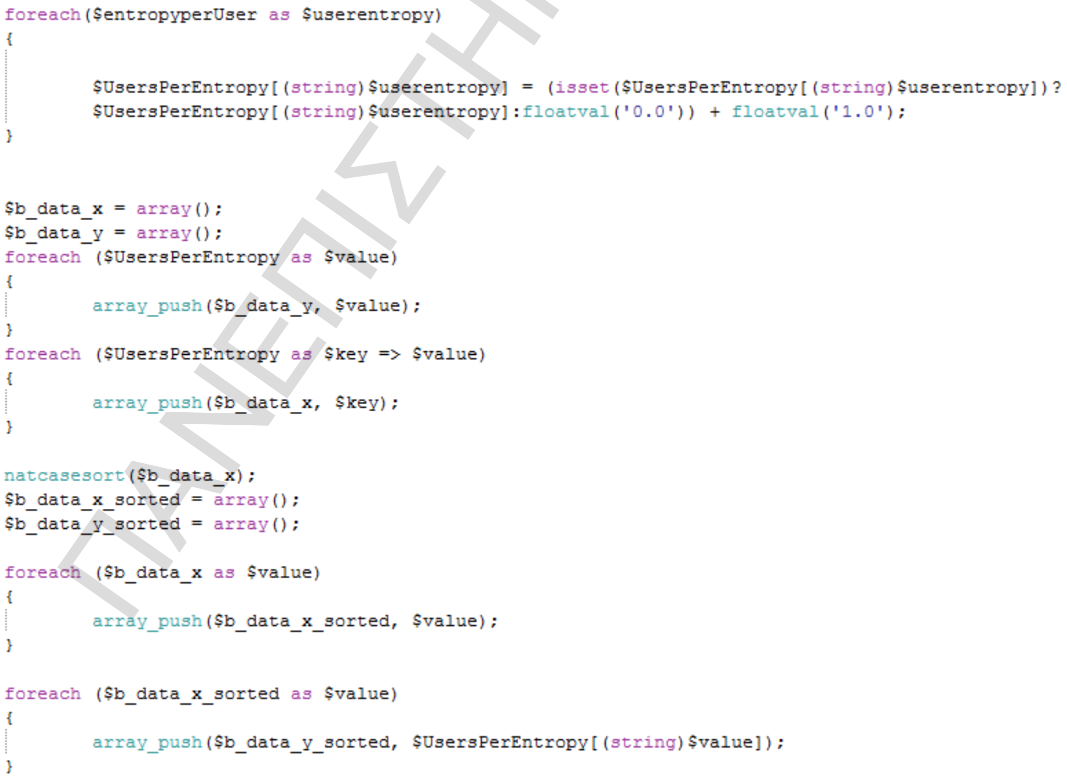


```
        $b_data_x = array();
        $b_data_y = array();
        foreach ($CurrentUserProfile as $value)
{
                array_push($b_data_y, $value);
        }
        foreach ($CurrentUserProfile as $key => $value)
        {
                array_push($b_data_x, $key);
        }

       $TweetsSeriesItem = new \Kendo\Dataviz\UI\ChartSeriesItem();
       $TweetsSeriesItem->name('TextWise Categorization')
         ->data($b_data_y)
         ->gap(1);
```

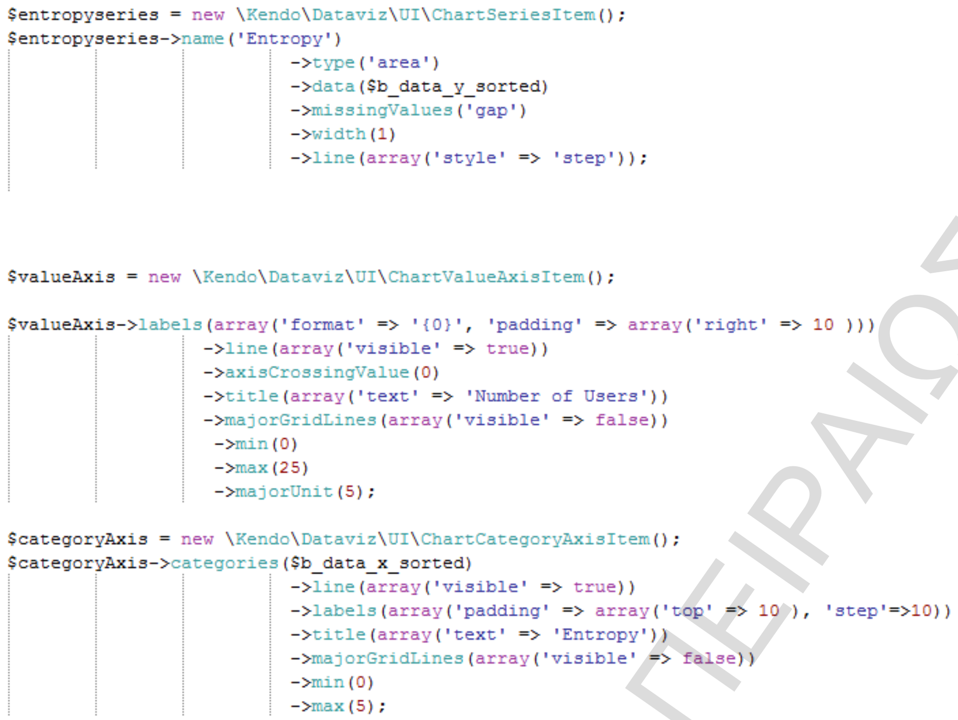


```
        $valueAxis = new \Kendo\Dataviz\UI\ChartValueAxisItem();

        $valueAxis->labels(array('format' => '{0}'))
                          ->line(array('visible' => false))
                          ->axisCrossingValue(0);

        $categoryAxis = new \Kendo\Dataviz\UI\ChartCategoryAxisItem();
        $categoryAxis->categories($b_data_x)
                                  ->line(array('visible' => true))
                              ->labels(array('padding' => array('top' => 14 )));


  $tooltip = new \Kendo\Dataviz\UI\ChartTooltip();
  $tooltip->visible(true)
          ->format('{0}%')
          ->template('#= category #: #= value #');

  $chart = new \Kendo\Dataviz\UI\Chart('chart');
  $chart->title(array('text' => 'Tweets per Category'))
        ->legend(array('position' => 'top'))
        ->addSeriesItem($TweetsSeriesItem)
        ->addValueAxisItem($valueAxis)
        ->addCategoryAxisItem($categoryAxis)
        ->tooltip($tooltip)
        ->seriesDefaults(array('type' => 'column'))
        ->chartArea(array('width' => '800'));

  echo $chart->render();

}
?>
```

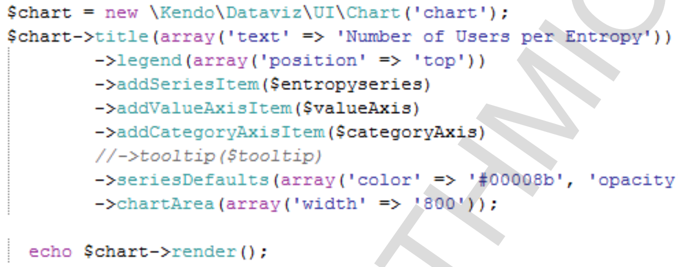
10.7 Loading Histogram of Entropy



```
foreach($entropyperUser as $userentropy)
{

        $UsersPerEntropy[(string)$userentropy] = (isset($UsersPerEntropy[(string)$userentropy])?
        $UsersPerEntropy[(string)$userentropy]:floatval('0.0')) + floatval('1.0');
}


$b_data_x = array();
$b_data_y = array();
foreach ($UsersPerEntropy as $value)
{
        array_push($b_data_y, $value);
}
foreach ($UsersPerEntropy as $key => $value)
{
        array_push($b_data_x, $key);
}

natcasesort($b_data_x);
$b_data_x_sorted = array();
$b_data_y_sorted = array();

foreach ($b_data_x as $value)
{
        array_push($b_data_x_sorted, $value);
}

foreach ($b_data_x_sorted as $value)
{
        array_push($b_data_y_sorted, $UsersPerEntropy[(string)$value]);
}
```

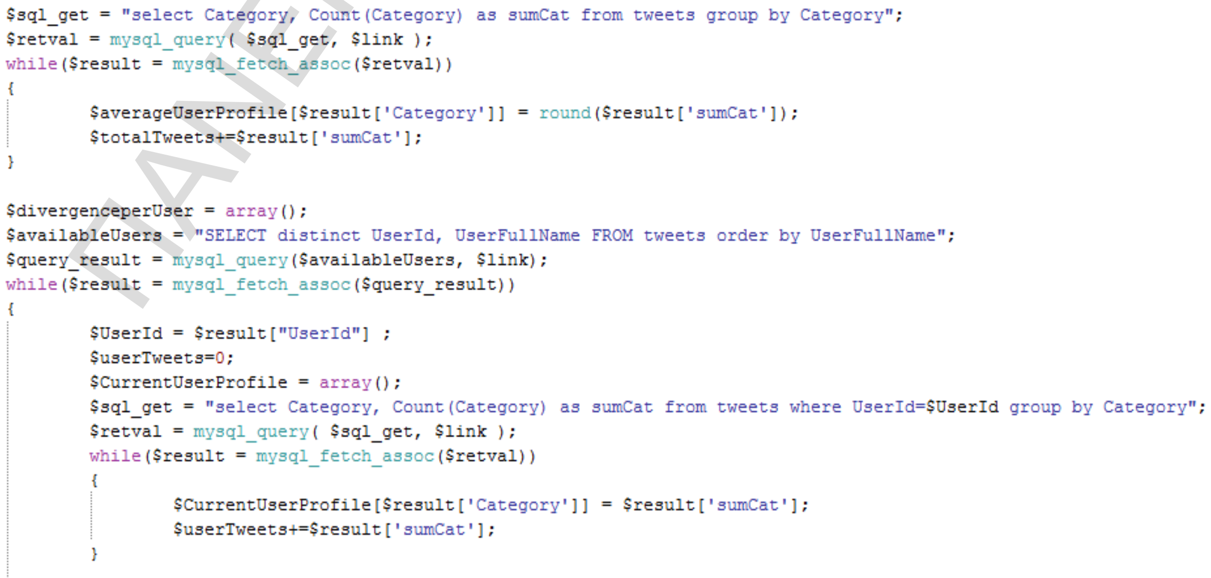


```
$entropyseries = new \Kendo\Dataviz\UI\ChartSeriesItem();
$entropyseries->name('Entropy')
                         ->type('area')
                         ->data($b_data_y_sorted)
                         ->missingValues('gap')
                         ->width(1)
                         ->line(array('style' => 'step'));



$valueAxis = new \Kendo\Dataviz\UI\ChartValueAxisItem();

$valueAxis->labels(array('format' => '{0}', 'padding' => array('right' => 10 )))
                 ->line(array('visible' => true))
                 ->axisCrossingValue(0)
                 ->title(array('text' => 'Number of Users'))
                 ->majorGridLines(array('visible' => false))
                  ->min(0)
                  ->max(25)
                  ->majorUnit(5);

$categoryAxis = new \Kendo\Dataviz\UI\ChartCategoryAxisItem();
$categoryAxis->categories($b_data_x_sorted)
                         ->line(array('visible' => true))
                         ->labels(array('padding' => array('top' => 10 ), 'step'=>10))
                         ->title(array('text' => 'Entropy'))
                         ->majorGridLines(array('visible' => false))
                         ->min(0)
                         ->max(5);
```

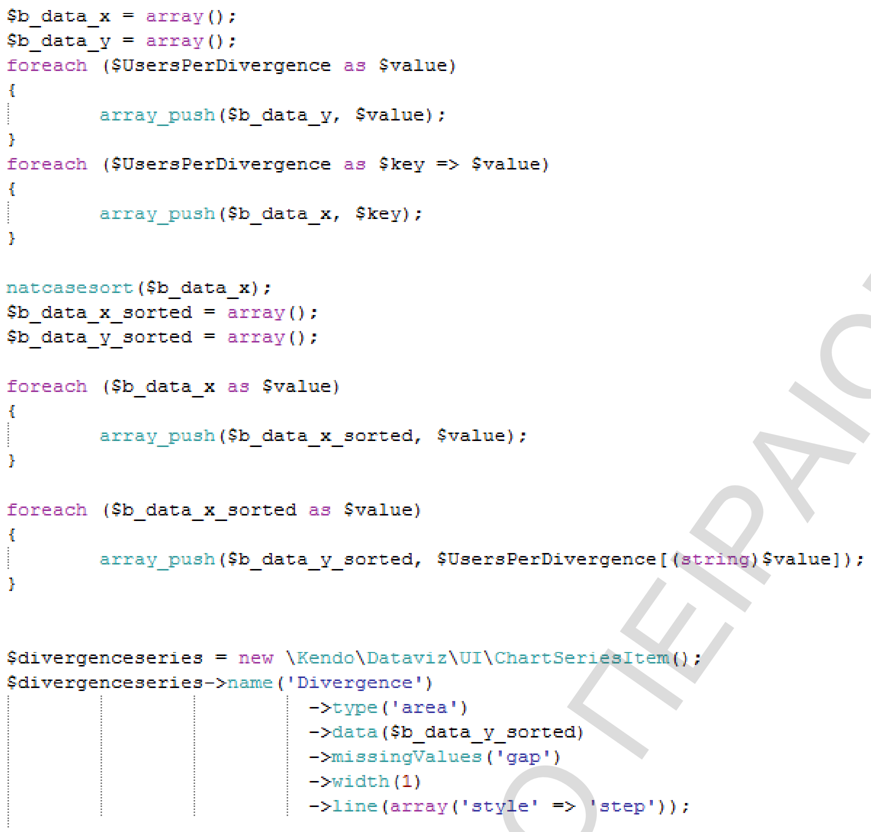


```
$chart = new \Kendo\Dataviz\UI\Chart('chart');
$chart->title(array('text' => 'Number of Users per Entropy'))
        ->legend(array('position' => 'top'))
        ->addSeriesItem($entropyseries)
        ->addValueAxisItem($valueAxis)
        ->addCategoryAxisItem($categoryAxis)
        //->tooltip($tooltip)
        ->seriesDefaults(array('color' => '#00008b', 'opacity'=>'1'))
        ->chartArea(array('width' => '800'));

  echo $chart->render();
```

10.8 Loading Histogram of Divergence



```
$sql_get = "select Category, Count(Category) as sumCat from tweets group by Category";
$retval = mysql_query( $sql_get, $link );
while($result = mysql_fetch_assoc($retval))
{
        $averageUserProfile[$result['Category']] = round($result['sumCat']);
        $totalTweets+=$result['sumCat'];
}

$divergenceperUser = array();
$availableUsers = "SELECT distinct UserId, UserFullName FROM tweets order by UserFullName";
$query_result = mysql_query($availableUsers, $link);
while($result = mysql_fetch_assoc($query_result))
{
        $UserId = $result["UserId"] ;
        $userTweets=0;
        $CurrentUserProfile = array();
        $sql_get = "select Category, Count(Category) as sumCat from tweets where UserId=$UserId group by Category";
        $retval = mysql_query( $sql_get, $link );
        while($result = mysql_fetch_assoc($retval))
        {
                $CurrentUserProfile[$result['Category']] = $result['sumCat'];
                $userTweets+=$result['sumCat'];
        }
```



```
$b_data_x = array();
$b_data_y = array();
foreach ($UsersPerDivergence as $value)
{
        array_push($b_data_y, $value);
}
foreach ($UsersPerDivergence as $key => $value)
{
        array_push($b_data_x, $key);
}

natcasesort($b_data_x);
$b_data_x_sorted = array();
$b_data_y_sorted = array();

foreach ($b_data_x as $value)
{
        array_push($b_data_x_sorted, $value);
}

foreach ($b_data_x_sorted as $value)
{
        array_push($b_data_y_sorted, $UsersPerDivergence[(string)$value]);
}


$divergenceseries = new \Kendo\Dataviz\UI\ChartSeriesItem();
$divergenceseries->name('Divergence')
                          ->type('area')
                          ->data($b_data_y_sorted)
                          ->missingValues('gap')
                          ->width(1)
                          ->line(array('style' => 'step'));
```