UNIVERSITY OF PIRAEUS - DEPARTMENT OF INFORMATICS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

MSc «Advanced Informatics and Computing Systems - Software Development and Artificial Intelligence»
ΠΜΣ «Προηγμένα Συστήματα Πληροφορικής – Ανάπτυξη Λογισμικού και Τεχνητής Νοημοσύνης»

MSc Thesis

Μεταπτυχιακή Διατριβή

Thesis Title: Τίτλος Διατριβής:	**Co-Authorship Networks Analysis** Ανάλυση Δικτύων Συν-Συγγραφέων
Student's name-surname: Ονοματεπώνυμο φοιτητή:	**EMMANOUIL ADAMOPOULOS** ΕΜΜΑΝΟΥΗΛ ΑΔΑΜΟΠΟΥΛΟΣ
Father's name: Πατρώνυμο:	**GRIGORIOS** ΓΡΗΓΟΡΙΟΣ
Student's ID No: Αριθμός Μητρώου:	**ΜΠΣΠ/18001**
Supervisor: Επιβλέπων:	**Dionisios Sotiropoulos, Assistant Professor** Διονύσιος Σωτηρόπουλος, Επίκουρος Καθηγητής

Οκτώβριος 2021/
October 2021

3-Member Examination Committee

Τριμελής Εξεταστική Επιτροπή

Dionisios Sotiropoulos
Assistant Professor

Διονύσιος Σωτηρόπουλος
Επίκουρος Καθηγητής

Georgios Tsihrintzis
Professor

Γεώργιος Τσιχριντζής
Καθηγητής

Evangelos Sakkopoulos
Assistant Professor

Ευάγγελος Σακκόπουλος
Επίκουρος Καθηγητής

Υπεύθυνη Δήλωση:
Έχω διαβάσει και κατανοήσει τους κανόνες του ΠΜΣ που περιέχονται στον Οδηγό Συγγραφής ΔΕ και ιδιαίτερα όσα συνιστούν λογοκλοπή. Δηλώνω ότι η παρούσα διπλωματική εργασία αποτελεί προϊόν αποκλειστικά δικής μου προσπάθειας, υπό την καθοδήγηση του επιβλέποντος καθηγητή, ενώ για όλες τις πηγές που χρησιμοποιήθηκαν περιλαμβάνονται οι αντίστοιχες αναφορές.

Abstract

Sociology is the foundation stone of the contemporary research field of Social Network Analysis. In 1830's this field was introduced for the first time (Linton, 2004) and this was the beginning of a new area of science. Since then, Social Network Analysis has been in the center of scientists' concern and much work is being done until today. The goal of this scientific field is the identification of a community's structure, which is represented in the form of a graph, with vertices connected with edges. There are several mathematical models which express the structure of networks and they are presented in this work thoroughly. Scientists use these models to extract valuable information of Networks, such as identifying the most central node or finding the shortest path through which an information could pass from one node to another. With the late advancements in computing processing and data analysis techniques, research can be applied to large Networks with many participants. In this work a historical overview of Social Network Analysis is presented. The basic characteristics of a Network are noticed and available digital tools of analysis are referenced. An analysis, using these mathematical models, has been conducted on three Networks that were extracted through a forty-year scholarly literature of journal DEA-related articles between 1978 and 2016 (Emrouznejad & Yang, 2018). The results and conclusions of this analysis are presented by the end of this work and future work that could be done is also proposed.

Περίληψη

Η Κοινωνιολογία είναι ο θεμέλιος λίθος του σύγχρονου ερευνητικού πεδίου της Ανάλυσης Κοινωνικών Δικτύων. Τη δεκαετία του 1830 έγινε η πρώτη σχετική εισήγηση (Linton, 2004), γεγονός που αποτέλεσε και την απαρχή του συγκεκριμένου επιστημονικού πεδίου. Έκτοτε, η Ανάλυση Κοινωνικών Δικτύων βρίσκεται στο επίκεντρο του επιστημονικού ενδιαφέροντος και μέχρι σήμερα συνεχίζεται να παράγεται με εντατικούς ρυθμούς γνώση επί του θέματος. Στόχος του πεδίου αποτελεί η αναγνώριση της δομής μιας κοινότητας, η οποία μπορεί να αναπαρασταθεί με τη μορφή ενός γράφου, στον οποίο υπάρχουν κορυφές συνδεδεμένες με δεσμούς. Υπάρχουν διάφορα μαθηματικά μοντέλα που εκφράζουν τη δομή ενός δικτύου, τα οποία περιγράφονται εκτενώς στη παρούσα εργασία. Οι επιστήμονες χρησιμοποιούν αυτά τα μοντέλα με σκοπό να εξάγουν πολύτιμη πληροφορία σχετικά με τα δίκτυα, όπως για παράδειγμα η εύρεση του πιο κεντρικού κόμβου ή η εύρεση του συντομότερου μονοπατιού, μέσα από το οποίο μπορεί μια πληροφορία του δικτύου να περάσει από τον ένα κόμβο σε έναν άλλο. Με την εξέλιξη της υπολογιστικής δύναμης και των μεθόδων ανάλυσης δεδομένων, είναι δυνατή πλέον η ανάλυση εκτενών κοινωνικών δικτύων με μεγάλο αριθμό συμμετεχόντων. Στην παρούσα εργασία παρουσιάζεται μια ιστορική διαδρομή στο επιστημονικό πεδίο της Ανάλυσης Κοινωνικών Δικτύων. Σημειώνονται τα βασικά χαρακτηριστικά ενός δικτύου και αναφέρονται τα διαθέσιμα εργαλεία ανάλυσης δικτύων. Κάνοντας χρήση των μαθηματικών μοντέλων, πραγματοποιήθηκε ανάλυση πάνω σε τρία δίκτυα που εξήχθησαν από τη βιβλιογραφία άρθρων-κειμένων όπως αυτή διαμορφώθηκε στο διάστημα σαράντα χρόνων μεταξύ του 1978 και 2016 γύρω από την Ανάλυση Δεδομένων Περιβάλλοντος (Data Envelopment Analysis) (Emrouznejad & Yang, 2018). Τα αποτελέσματα και συμπεράσματα αυτής της ανάλυσης αναφέρονται στο τέλος της παρούσας εργασίας και στη συνέχεια προτείνεται μελλοντική έρευνα που μπορεί να πραγματοποιηθεί.

1. Introduction to Social Networks

1.1 A historical overview in the scientific literature

In international bibliography, Social Network Analysis can be defined as a tool that describes four features of Social Sciences: structural intuition, systematic relational data, graphic images and mathematical or computational models (Linton, 2004). Sociologists were the first to conduct research on social structures and networks and were those to put the foundations in the scientific field of Social Network Analysis (SNA).

According to Freeman C. Linton, the first to introduce such intuition to the need of social analysis and research was Isidore Auguste Marie François Xavier Comte, back in 1830's who was the first to develop sociology as a scientific field. He was a pioneer in Social Sciences who introduced for the first time the term "*Sociology*" to the literature, trying to identify and specify the uncovered laws of society. In his later work, Comte tried to present the ways that social actors are interconnected into a social system: "*Families become tribes and tribes become nations".* He used structural terms that can be found in recent literature of SNA (Linton, 2004).

After Comte several considerable sociologists followed, who conducted research in micro- or macro-societies, having the goal to identify the links between social actors and investigate patterns of social connections in order to describe social phenomena. Apropos, some of them are Ferdinand Tönnies (1855/1936), who tried to characterize a social system considering the personal and direct links between individuals who shared common values and beliefs and Emile Durkheim (1893/1964), who identified links which produced in terms of solidarity between suppressed social units. Georg Simmel (1858/1918) whose basic work was to explore the "*patterning of interaction",* went further with his terminology, using contemporary terms in his analysis, found in literature until nowadays, such as "*system of relations*" and a "*network of lines between men.*"(Linton, 2004)

Later on, a series of experiments conducted by Stanley Milgram in 1967, called the "Small World", is supposed to be the first experimental approach in SNA. In this experiment one can find the key notions of Graph Theory and several techniques which still are used to identify key characteristics of a Network (Milgram, 1967). In this particular work one can find terms like "*Common pathways", "Average path length", "Target Person"* and others, which clearly refer to the later research field of SNA.

Ithiel de Sola Pool and Manfred Kochen (Pool and Kochen, 1978), put some basic mathematical models in the study of "*Social Contact"* and described the fundamental analytical methods for SNA. Kochen had a background in Physics and was a PhD Mathematician, a fact that probably contributed to the investigation of the topic via mathematical scope. In their work they incorporate Graph-theoretic models to describe the network, and statistical models to identify the shortest path between randomly selected individuals in large networks or identify the number of their common acquaintances. Another important aspect of their work is the definition of "*stratum",* in which they provide models to count the probability of contact between units that coexist in the same stratum of a network (Linton, 2004).

By the end of the 20th century, SNA started to draw the attention of scientists and researchers even more, who came from other scientific fields, which resulted in a significant growth in the number of published scientific articles. From data that were retrieved from 3 databases (Sociological Abstracts, Medline Advanced and PsycINFO) for the period 1963–2000 on cumulative published articles in SNA it was shown that after 1981 began the development of this scientific field, as more scientists, other than Sociologists, began to use this method for research (Figure 1.1.1) (Otte & Rousseau, 2002).

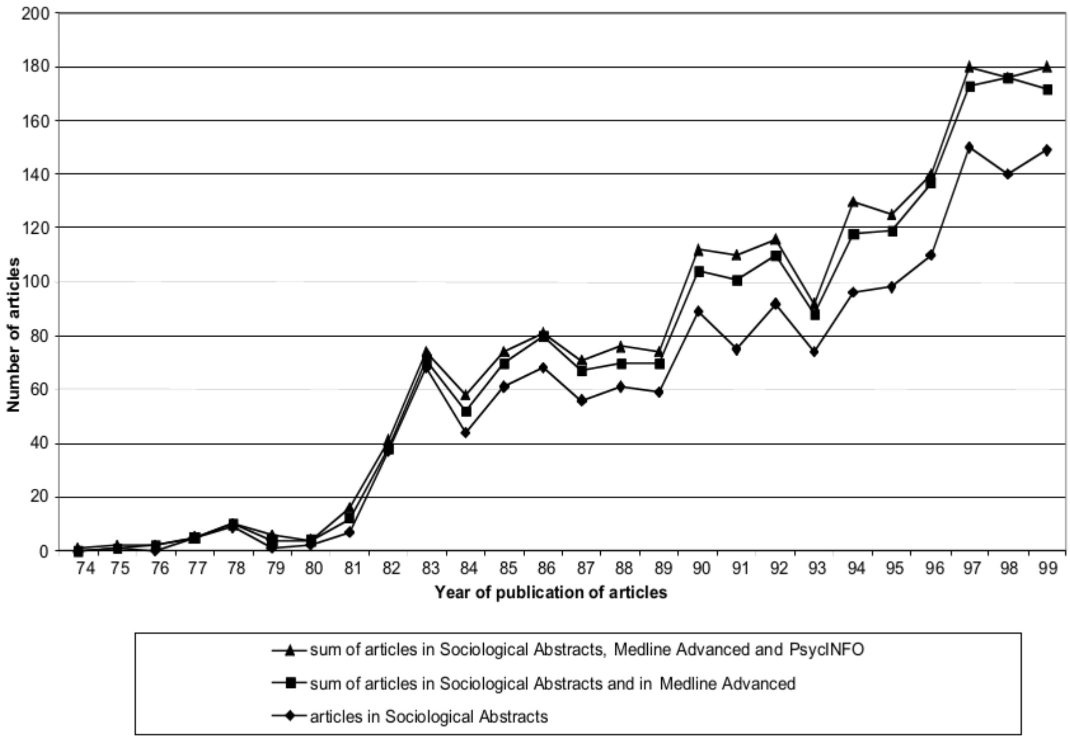


Figure 1.1.1. Sum of SNA articles from the three databases as measured between 1974 - 2000

Furthermore, one reason to contribute to that growth was the foundation of the International Network for Social Network Analysis (INSNA) by Barry Wellman in 1978. This act institutionalized the field and brought together researchers, by offering them a platform in which they could find news of the current research and web-based services that helped them find scholarly articles and other content, related to the field (Otte & Rousseau, 2002).

Within the first years of the 20th century, dedicated software led to the increase of interest in Social Network Analysis within the global scientific community. In 2002 the best-known UCInet was released. Other software releases are Cytoscape, GraphVis, igraph, Pajek. A tool for data visualization and Network visualisation is Gephi as well, which is used in the current dissertation.

In recent years extended research is being conducted in many scientific fields in order to investigate different topics. Studies on large scale networks have provided insights into the topologies of networks. Two important physicists, Barabási (Barabási and Albert, 1999) and Watts (Watts and Strogatz, 1998), have contributed a lot with their work. Several models have been developed with the one of Watts-Strogatz to be considered as a very important work. This model suggests a single parameter model, which interpolates between an ordered finite dimensional lattice and a random graph (Albert and Barabási, 2002)

1.2 Social Networks Characteristics and Techniques of Analysis

1.2.1 Representation of Social Networks and their Basic Characteristics

Every Social Network consists of a finite set of vertices which are connected between each other with links. These vertices can represent a wide variety of entities (e.g. persons, countries, organizations, scientific articles, authors, terms etc.) and the links between them, which are also referred to as edges, represent relationships. Each scientific field has its own terminology for these basic components of Social Networks (Figure 1.2.1.1), but the actual structure remains always the same. Thus, a Social Network is usually easy to be represented as a graph and is often regarded to and analyzed as such (Tabassum, Souza Fernandes Pereira, Fernandes & Gama, 2018).

points	lines	field
vertices	edges, arcs	math
nodes	links	computer science
sites	bonds	physics
actors	ties/relations	sociology

Figure 1.2.1.1 Social Network terminology variation among different scientific fields

So, each Social Network is actually a graph and its two main components are:

- the vertices, each of which represent an individual entity, that may or may not be connected to another participant in the Network
- the edges, which represent the relationships that exist between the participants of the Network.

Consequently, a graph G, through a mathematical perspective, may consist of a non-empty set of vertices (V(G)) and a set of edges (E(G)), being defined as G = (V(G), E(G)). Two types of graphs exist, the directed and the undirected graphs. If all edges of a graph are represented with arrows (also in literature referred to as arcs), pointing from one vertex to one or multiple other vertices, this graph is called directed, and it has the ability to represent the orientation of the different relationships that appear within the Network. In theory, if $E_{12}$is the arc and v_1 and $v_2$are vertices of the graph, such that $E_{12} = (v_1, v_2)$, then E_{12}connects $v_1$and v_2, having $v_1$to be the initial vertex or tail and the second vertex v_2 the terminal vertex or head. In contrast, a graph whose edges are simple lines, does not contain such information and it is called undirected (Tabassum, Souza Fernandes Pereira, Fernandes & Gama, 2018).

Two basic metrics of graph are defined, one which is related to the number of vertices in the graph and the other to the number of edges, as follows:

- |V(G)| = n, which is called the order of the graph
- |E(G)| = m, which is called the size of the graph

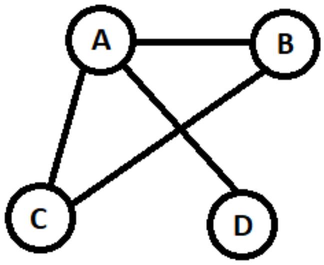
It is proved mathematically that the maximum size of the graph must be $m_{max} = \frac{n(n-1)}{2}$, if it is an undirected graph and $m_{max} = n(n - 1)$if it is a directed graph (Tabassum, Souza Fernandes Pereira, Fernandes & Gama, 2018).

In the family of directed graphs, an additional information may be provided, which has to do with the strength of the relations between its entities, being expressed through an associated weight $w \in \Re_0^+$ to each arc. These graphs are called weighted graphs and are very rich in information, because they express the importance of each connection in the Network. Furthermore, this type of graphs can be cyclic, i.e. graphs containing closed loops of edges or "ring" structures, or acyclic (e.g. trees).

There are several ways of storing a graph as a data structure and in literature two are the main types; lists and matrices. For sparse graphs, lists, such as incidence lists and adjacency lists, are used in order to reduce storage space. Matrix structures can be of different types, depending on the need of metrics that one may want to express for a graph. Some examples of such types of matrices are incidence matrices, adjacency matrices or sociomatrices, Laplacian matrices (contain both adjacency and degree information) and distance matrices (identical to the adjacency matrices with the difference that the entries of the matrix are the lengths of the shortest paths between pairs of vertices) (Tabassum, Souza Fernandes Pereira, Fernandes & Gama, 2018).

In order to have an in depth view of how the representation of Networks work, regarding the most common way to represent a graph with a matrix, that is an adjacency matrix, two cases could be analyzed. In the first case, an undirected and unweighted graph, when represented with an adjacency matrix, this matrix would be a symmetric matrix. In particular, the cells (a_{ij}) between connected vertices (v_1 and v_2) would have the value of 1 (a_{ij}= 1) and those between unconnected ones would have the value of 0 (a_{ij}= 0). All cells on the diagonal will be 0, as a_{ij}= a_{ji} (Figure 1.2.1.2). On the contrary, the cells of the adjacency matrix for a weighted and directed graph would have values, such that a_{ij}= $[0, max(w)]$, where w is the weight associated with the respective arc between v_1 and v_2. In this case, by definition it turns the fact that this matrix would not be necessarily symmetric, due to the existence of closed loops, which may happen to one vertex (as for an example, in a co-authorship network an author, who would be the vertex i, may cite herself in her own article, referring to another work of hers. In this case $a_{ii} = w_{ii}$) (Figure 1.2.1.3).

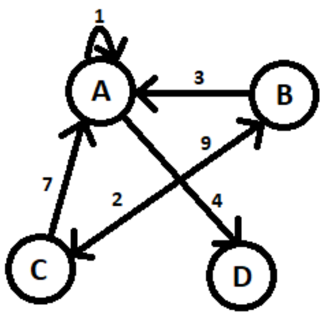
VERTEX	A	B	C	D
A	0	1	1	1
B	1	0	1	0
C	1	1	0	0
D	1	0	0	0



VERTEX	VERTEX
A	B
A	C
A	D
B	C

Figure 1.2.1.2 Representation of an undirected and unweighted graph (center) as adjacency matrix (left) and adjacency list (right)

VERTEX	A	B	C	D
A	1	0	0	4
B	3	0	2	0
C	7	9	0	0
D	0	0	0	0



VERTEX	VERTEX	WEIGHT
A	A	1
A	D	4
B	A	3
B	C	2
C	A	7
C	B	9

Figure 1.2.1.3 Representation of a directed and weighted graph (center) as adjacency matrix (left) and adjacency list (right)

1.2.2 Social Network Analysis Mathematical Models

In Social Network Analysis it is important to have an insight about the conditions under which the links between the participants were formed and in which way they interact with each other within the Network. Therefore, the two main goals are to, first, analyze the role of each participant in the Network and, second, to identify the social system that generated this particular Network (Tabassum, Pereira, Fernandes & Gama, 2018). For these reasons, several useful metrics have been developed and here, some of the most important ones are presented.

The metrics that are analyzed herein can be categorized into two different scales. The first one is the node-level analysis, which can be measured by a node's centrality and it is a way to understand the qualitative value of a node's position in the Network. The second scale is the network-level analysis, which can be measured by geodesic paths between every set of nodes in the Network and provides the researcher with rich information to assess the overall structure of the Network and give insights about the important properties of the underlying social phenomena (Tabassum, Pereira, Fernandes & Gama, 2018).

1.2.2.1 Node-level metrics

1.2.2.1.1 Degree Centrality

Degree centrality is represented as the total number of direct contacts one actor of a Network has, thus, the count of its direct neighbours in the Network. The degree centrality k of a node v can be measured either by the count of the neighbouring nodes in the graph or by the adjacency matrix of Network.

In undirected Social Networks, this can be calculated as follows

$$k_v = \left|N_V\right|, 0 \leqslant k_i \leqslant n \ , \qquad (1.1)$$

where $\left|N_v\right|$ is the neighbourhood of node v and n the total number of graph's nodes

$$k_i = \sum_{j=1}^{n} a_{ij}, \ 0 \leqslant k_i \leqslant n \quad , \qquad (1.2)$$

where a_{ij} is the value of the ij cell in the adjacency matrix.

In directed Networks, degree centrality has two variants: in-degree and out-degree centrality. In-degree centrality is the total count of inbound edges of one node and out-degree is the count of outbound edges. In literature, the degree centrality of directed Social Networks is often called prestige, in order to express the public importance of an actor of the Network. Specifically, in-degree centrality is referred to as support and out-degree as influence and are given by the following equations (Tabassum, Pereira, Fernandes & Gama, 2018):

$$k_i^{+} = \sum_{j=1}^{n} a_{ij} \quad , \quad k_i^{-} = \sum_{j=1}^{n} a_{ij} \ , \qquad (1.3)$$

This metric provides significant information about the Network, because it helps to identify its central, as well as its peripheral actors. If the degree of a node is high, it can be said that the node is important to the Network, in terms that it is connected with many other nodes and thus, holds a major role of channeling the information throughout the Network. In contrast, if the degree of a node is low, that means that the representative actor is at the peripherals of the Network and thus, it is harder for her to be part of the mainstream of information (Freeman, 1978/79).

1.2.2.1.2 Betweenness Centrality

In Social Networks, a common observation is the existence of independent communities (set of two or more connected nodes) that do not share any common link. Thus, no information cannot be transferred between them and so, they stay completely independent. So, finding linking nodes between communities, or more generally, between regions of a Network is of great importance. Betweenness Centrality, often denoted as *b* in literature, can be expressed by the percentage of all shortest, or geodesic paths, that pass through a node that lies between different regions of the Network, with respect to the total count of geodesic paths that connect those regions. Nodes with high Betweenness Centrality are vital connections between those regions and they are often known as “gatekeepers” (Tabassum, Pereira, Fernandes & Gama, 2018). The value of Betweenness Centrality of a node v, standing between two nodes, s and t respectively, can be given by the following equation:

$$b(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad (1.4)$$

where σ_{st} is the number of all geodesic paths that link nodes s and t and $\sigma_{st}(v)$ is the number of which they pass through node v.

This metric can be applied to edges too, in order to identify the way that different regions are connected through actors that communicate but they belong to different regions. In Social Network Analysis it is of high interest to locate such edges, also called bridges, in order to have an insight of the main actors of communication between different regions. The calculation is done by counting the number of all geodesic paths that run along a given edge, similarly to the aforementioned and as indicates the following equation; (Tabassum, Pereira, Fernandes & Gama, 2018).

$$b(e) = \sum_{s,v \in V(G)} \frac{\sigma_{sv}(e)}{\sigma_{sv}} \qquad (1.5)$$

1.2.2.1.3 Closeness Centrality

Betweenness centrality can be said that is related to the control of information. Another perspective of centrality has to do with the ability of a node to avoid such control by others in the Network, so it is referred to its closeness to the rest of the participants in it. Closeness Centrality expresses the “independence” of a node in a Network (Leavitt (1951)) and it can be said that in order to spread an information across the entire Network in optimum time, then it must originate in the most central point of the Network (Bavelas (1948)). Several ways have been proposed for calculating Closeness Centrality but one straightforward model has been given by Sabidussi (1966), proposing its calculation by summing all geodesic paths from the node, whose centrality is examined, to all other actors of the Network. It is actually the inverse calculation that one might expect, since the result increases as nodes lie far apart and so, if $d\ (u,\ v)$is the number of all edges that connect node u with node v then Closeness Centrality of node v is:

$$c_v^{-1} = \sum_{i=1}^{n} d(u,v) \qquad (1.6)$$

where n is the number of all nodes in the Network.

The above equation can be applied only to connected graphs, since if one and only one node exists, which may be unconnected to the rest of the Network, then every node would have an infinite distance from at least one other node, so (Freeman 1978/79):

$$\sum_{i=1}^{n} d(u,v) = \infty \qquad (1.7)$$

The results of the aforementioned equation (1.6) can not be compared between Networks of different sizes, since the respective values depend on the number of the existing nodes in the Network. Thus, removing the factor of size the following equation is given:

$$c_v = \frac{n-1}{\sum_{i=1}^{n} d(u,v)} \qquad (1.8)$$

1.2.2.1.4 Eigenvector Centrality

The Eigenvector Centrality is closely related to the Degree Centrality that was previously mentioned. This metric is calculated in a similar way, by counting the direct connections of an actor inside a Network, but taking into account the importance of these connected actors. In terms of influence, Eigenvector Centrality may introduce another factor that might be taken into account, since in Social Network Analysis it is important to know whether a major influencer is close to the examined actor or, reversively, if this actor is connected to critical transmitters of information in the Network. The respective equation is as follows:

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} a_{ij}x_j \qquad (1.9)$$

where x_i is the Eigenvector Centrality of node i, λ is a constant, also known as eigenvalue, a_{ij} is the value of the respective cell of the adjacency matrix and last x_j is the centrality of node j.

Defining that $x = (x_1, x_2, ..., x_n)$ is the centrality vector, then the equation can be written in vector notation, as follows (Newman, 2006):

$$\lambda \cdot x = A \cdot x \qquad (1.10)$$

1.2.2.2 Network-level Analysis

1.2.2.2.1 Diameter and Radius

In order to define Diameter and Radius, it is important to define eccentricity. Eccentricity of a vertex u (*e(u)*) of a graph is the maximum distance between the vertex u and any other vertex in the graph.

$$e(u) = max\, d(u, v),\ u, v \in V \qquad (1.11)$$

Eccentricity could be thought of as the maximum distance a vertex has from the most distant vertex in the graph.

Radius of a Network is defined as the minimum eccentricity between two vertices.

$$r = min\,\{e : v \in V\} \qquad (1.12)$$

In addition, the diameter of a Network is defined as the maximum eccentricity (Tabassum, Pereira, Fernandes & Gama, 2018).

$$d = max\,\{e : v \in V\} \qquad (1.13)$$

1.2.2.2.2 Average Geodesic Distance

The Average Geodesic Distance helps to get an idea of the average distance of all nodes inside a Network. It can be found in literature with the symbol l and is defined as follows:

$$l = \frac{1}{\frac{1}{2}n(n-1)}\sum_{i\geq j} d(i, j) \qquad (1.14)$$

where d(i,j) is the distance between vertex i and j and $\frac{1}{2}n(n - 1)$ expresses the total number of edges that are possible in a Network of n number of vertices (Tabassum, Pereira, Fernandes & Gama, 2018).

1.2.2.2.3 Average Degree

The average degree of a Network can be found by the average value of all nodes' degrees and is expressed mathematically as follows (Tabassum, Pereira, Fernandes & Gama, 2018):

$$\overline{k} = \frac{1}{n}\sum_{i=1}^{n} k_i \quad (1.15)$$

1.2.2.2.4 Modularity

A necessary and important measure that can be extracted from a Network is the number of communities that exist inside. The term “community” refers to the small subgroup of vertices that can be found into a Network, in which there is a dense interconnection between the participants and a sparser connection with other actors or communities into the same Network (Newman, 2006). The tool to quantify the essence of community arrangement in a Network is modularity. The modularity is given by subtracting the number of total edges between nodes of a community and the number of them in an equivalent Network in which its edges have been placed randomly. This is a widely accepted way to search for such communities in a Network and the higher the modularity is, the higher the possibility of a well detected community. Modularity can take negative values as well.

2. Co-authorship / Bibliographic Networks

2.1 Literature review

2.1.1 Historical overview

Co-authorship networks have always drawn the interest of researchers in the field of Social Networks Analysis, particularly of those whose research regards social or natural sciences. It is understood the value of this kind of networks, as the information they carry, regarding the cooperations (co-authorships on a research paper) and the frequency they take place in time, expresses relationships of true communication, since the authors choose their colleagues with whom they cooperate with, in order to produce a research project (Kumar, 2015).

As mentioned in a previous chapter (see chapter 1.1) most of the literature produced in the 20th century, regarding the Social Network Analysis research field, had a main concern about social network structures and communications between its actors. The majority of them had an empirical and inherent comprehension of these structures. Although these studies tend to identify patterns and characteristics in Social Networks and have much to offer in international literature, firstly, they have some limitations. First, most of them they are “*labor intensive”,* probably because it was harder in the 20th century or earlier for anyone to find data and therefore, the networks were limited to the number of actors they had and second, they comprise of uncertainty in terminology, as one actor might comprehend terms like “friendship”, “acquaintance” etc. differently than another in the same network (Newman, 2001).

These issues can be of significant importance when research is applied on social networks, particularly when the researcher tries to identify patterns in human interaction and study the links in the scope of society. This was something that several researchers took into consideration and so, they tried to analyze networks which concluded more data and were more precise in terms of connectedness, such as power grids and computer networks. But again, although these networks can be seen as a product of social life and human activity and so, in a way, they could offer some conclusions, they do not refer to actual human interaction, naked of bias in terminology. (Newman, 2001).

Newman (2001) focused on these biases and tried to overpass them by applying research to co-authorship networks, where "*precise definition of acquaintance is possible"*. In his paper he proposes that such networks can be of great value, because he assumes that there exist strong social bonds between authors that collaborate to produce research. Thus, the outcomes of the study can be valuable in order to examine social structures.

Barabási *et al.* (2002) recognized an extensive pool of data in co-authorship networks, focusing on databases which contained literature of an 8-year period (1991-98) in the field of mathematics and neuro-science. He and his colleagues came to further the research of Newman and tried to explore the evolution of these social networks by finding the underlying "*structural mechanisms"* that define their topology.

Both Newman and Barabasi *et.al* did not use centrality measures to characterize their networks. Simultaneously to Barabasi's *et. al* study, Otte and Rousseau (2002) introduced Social Network Analysis as a field in sociology that helps in forming a strategy for investigating social structures. They involved centrality measures in their study (closeness, degree and betweenness) in order to find the most central authors in the networks.

Co-authorship networks have been studied ever since extensively in both natural and social sciences. After Newman and Otte and Rousseau, many followed, such as Moody (2004) Quatman and Chelladurai (2008), Yan et al. (2010), Racherla and Hu (2010) Lewison and Markusova (2010) and Uddin et al (2012) (Kumar, 2015).

2.1.2 Relevant work to Co-authorship Networks

Apart from co-authorship networks, research is being conducted using different indicators, focusing on connecting entities in order to construct networks that could help extract social structures. These indicators are citations, co-citations, author co-citations, bibliographic coupling and co-word networks analysis (Kumar, 2015).

Citation analysis focuses on the total number of citations one research entity (journal, author, article etc.) has received or on the connection between articles that have been cited by each other. Co-citation analysis may produce a network by relating all articles which exist in a single reference list of a research paper. When two documents cite one or more common research papers then it is said that they are "bibliographically coupled". In co-word networks two words are connected when they co-exist in the same document. All of the above ways of network formations try to investigate research patterns and trends, rather than associate authors, as co-authorship networks do (Kumar, 2015). Though, recent studies try to conclude all the aforementioned techniques, along with co-authorship networks, in order to reveal generic patterns and create an overview of the landscape in research productivity. (Biscaro and Giupponi, 2014)

Another attempt to express the value of cooperation between authors has been raised within the mathematics community, expressed by Erdös' number. Erdös number finds its roots in the Hungarian mathematician Paul Erdös, who spent his later years of his life in producing a big number (1401 in total) of scientific papers with his colleagues. Erdös number shows the proximity of a mathematician to Erdös and specifically, a mathematician has the number 1 when she has published a paper with Erdös, the number 2 when she has published one with an author who has published a paper with Erdös and so on (Newman, 2001).

3. Dataset Generation

In this work, Social Network Analysis has been conducted on three Networks which have been generated after an intensive data analysis applied to a forty-year scholarly literature of journal DEA-related articles between 1978 and 2016 (Emrouznejad & Yang, 2018). The dataset was provided in PDF format which was manipulated with data analysis techniques written in the Python programming language.

3.1 Stages of Analysis

3.1.1 First stage of Analysis

Python is a powerful programming language for data analysis purposes which provides tools to manipulate TXT files in a fast and efficient way. In order to extract the data provided with the maximum accuracy, it was required to convert the PDF file to a TXT. The objective of the analysis was to extract all unique authors' names found in file as well as the titles of the articles and finally the adjacency list of co-authors.

The list of citations in the dataset was written in a standard pattern, in which the authors' names are mentioned first, separated with commas and the article title follows. The challenge was to produce a code script which will read correctly the aforementioned, taking into account the particularities found in several instances (for example, some authors have names which consist of more than two words).

The product of the above analysis was three datasets; The “unique authors list”, the “authors adjacency list” and the “articles’ titles list”.

3.1.2 Second stage of Analysis

Previous work, which conducted analysis on the same dataset but with the objective to pair them with their respective abstracts and citations, had produced two additional datasets. These datasets were processed in order to extract only the entries that were valid, i.e the entries which carried data and were not empty. The products of the process were two datasets which included the articles of the initial dataset which had valid abstracts and valid citations list respectively. Their respective authors' adjacency lists were produced as well.

The dataset with valid citation lists was used to produce the articles’ Network. Through each article’s list of citations, only those titles which existed in the initial dataset were kept. Thus, a directed and non weighted Network has been created.

In addition, using these two datasets, two more were created, which contain the respective co-authors’ directed Networks.

3.1.3 Third stage of Analysis

The third and final stage of analysis was conducted on the three final datasets, two directed and weighted and one directed and non-weighted Networks. These Networks are:

- valid_abstract_author (directed - weighted)
- valid_citation_authors (directed - weighted)
- valid_citation_titles (directed - non weighted)

Social Network Analysis was conducted making use of the free software Gephi, which helped to visualize the Networks and produce the respective results of the current research.

4. Experimental Results

A Social Network Analysis has been conducted on three datasets, which were produced by the main dataset, as described in Chapter 3. There were two levels of analysis, one in the Nde level of the Network and one in the Network level. The results that are presented here were extracted through the reports given from the Data Laboratory of Gephi software. The visualization of the results and the Networks are provided within Gephi and with the installation of the plugin "SigmaExporter", which provides additional visualization tools to present and export a Network.

4.1 valid_abstract_authors Dataset

This directed and weighted graph contains 11730 nodes, which are the authors of all titles of the initial dataset, which are accompanied with valid abstracts' file, as described in chapter 3.

4.1.1 Node Level Analysis

4.1.1.1 Degree centrality

Below are presented the results that were extracted after the degree centrality analysis. The results show that the maximum value of degree centrality found is 120 nodes, where 119 are out-degree and 1 in-degree. The majority of the participants of the Network have a relatively low degree centrality value and specifically 11161 out of 11730 (95%) nodes in total have equal or lower degree centrality of 10 nodes. The average degree centrality is 1.92. Three graphs representing the degree distribution are given below (Figures 4.1.1.1.1, 4.1.1.1.2, 4.1.1.1.3).

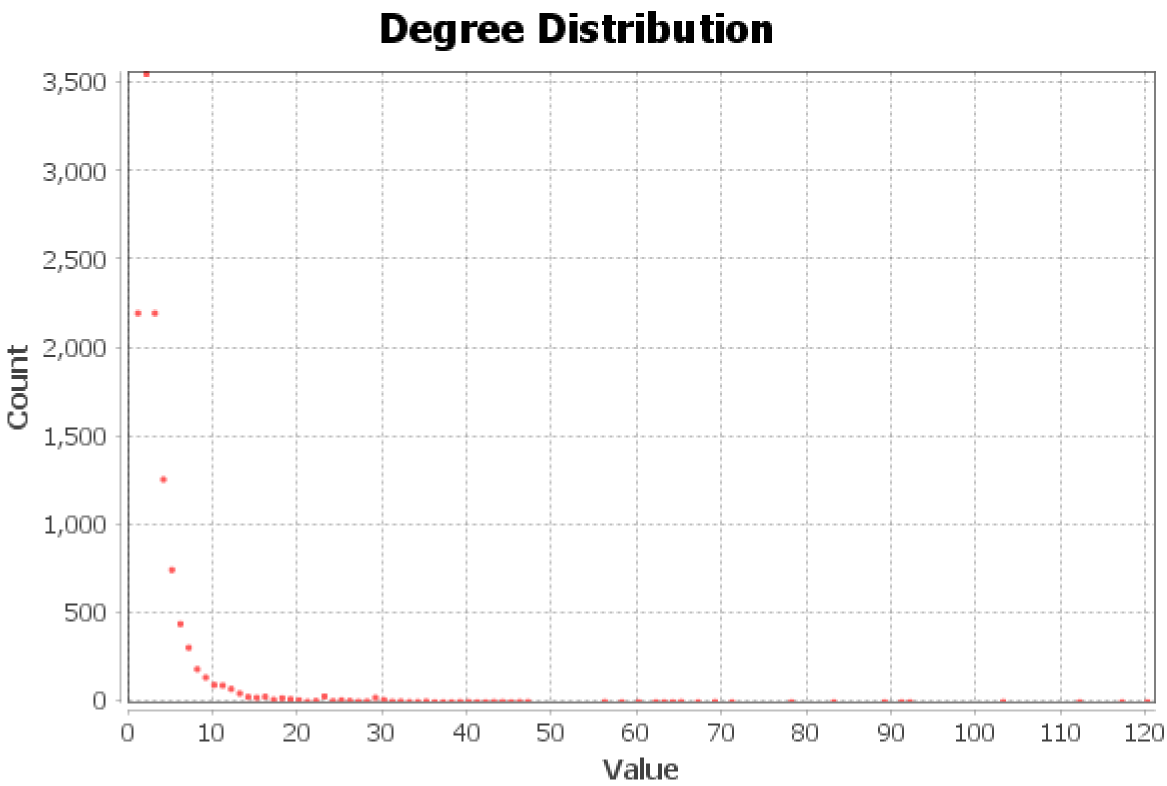


Figure 4.1.1.1.1 Degree centrality distribution - valid_abstract_authors dataset

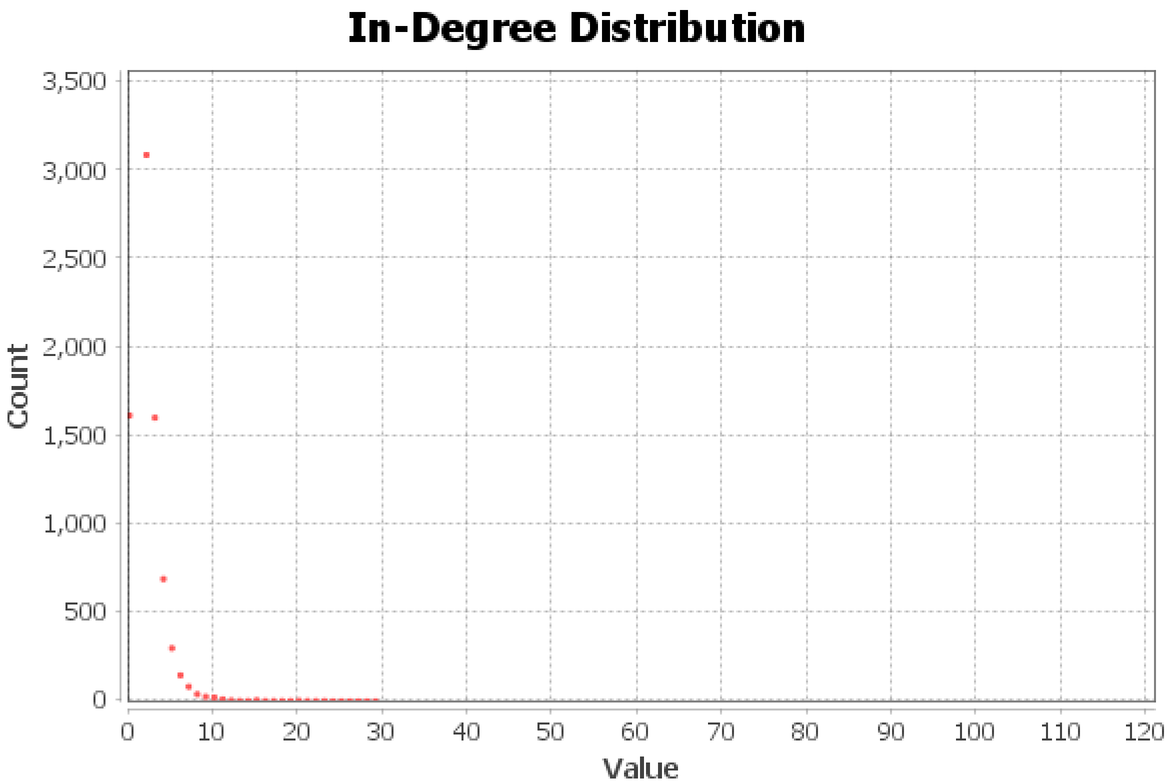


Figure 4.1.1.1.2 In - Degree centrality distribution - valid_abstract_authors dataset

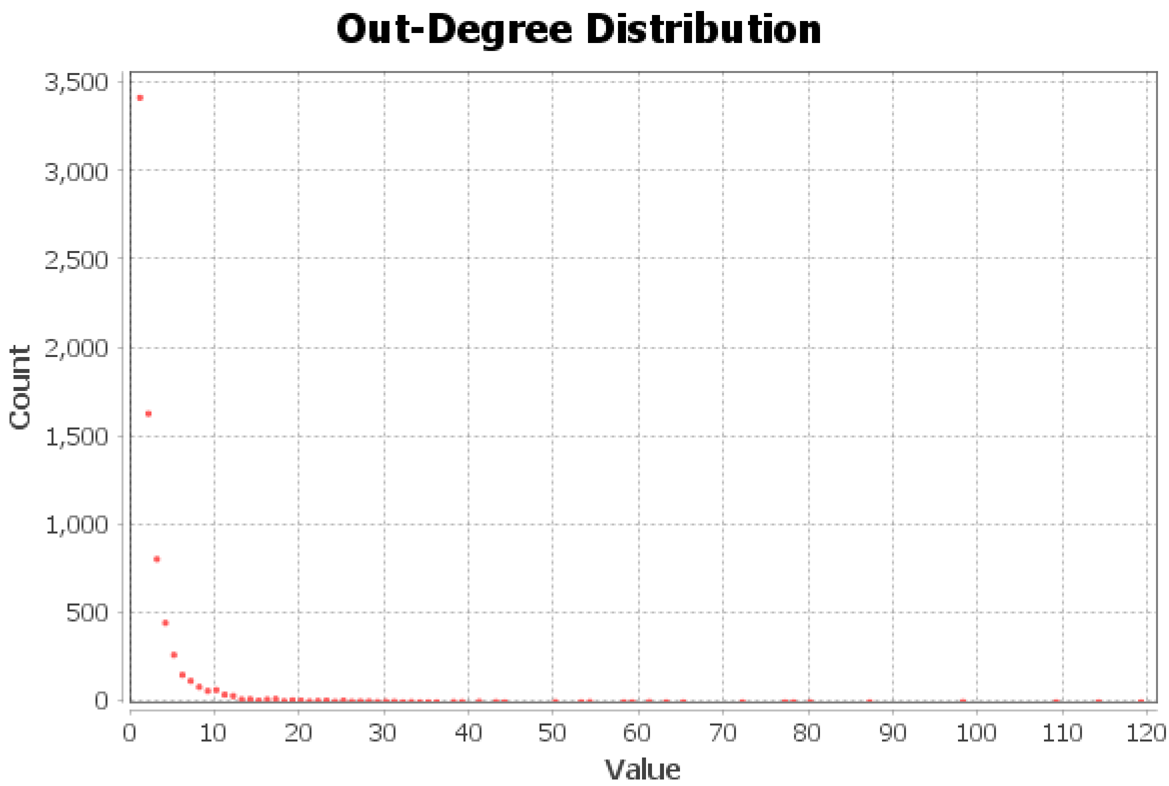


Figure 4.1.1.1.3 Out - Degree centrality distribution - valid_abstract_authors dataset

4.1.1.2 Betweenness centrality

As can be seen in the betweenness centrality graph that follows (Figure 4.1.1.2.1), a large number of vertices have a relatively low betweenness centrality and there are only a few with higher values. In particular, 462 vertices have betweenness centrality over 1000, 195 take values within 500 and 1000 and the rest (11075 or 94.4%) are below 500.

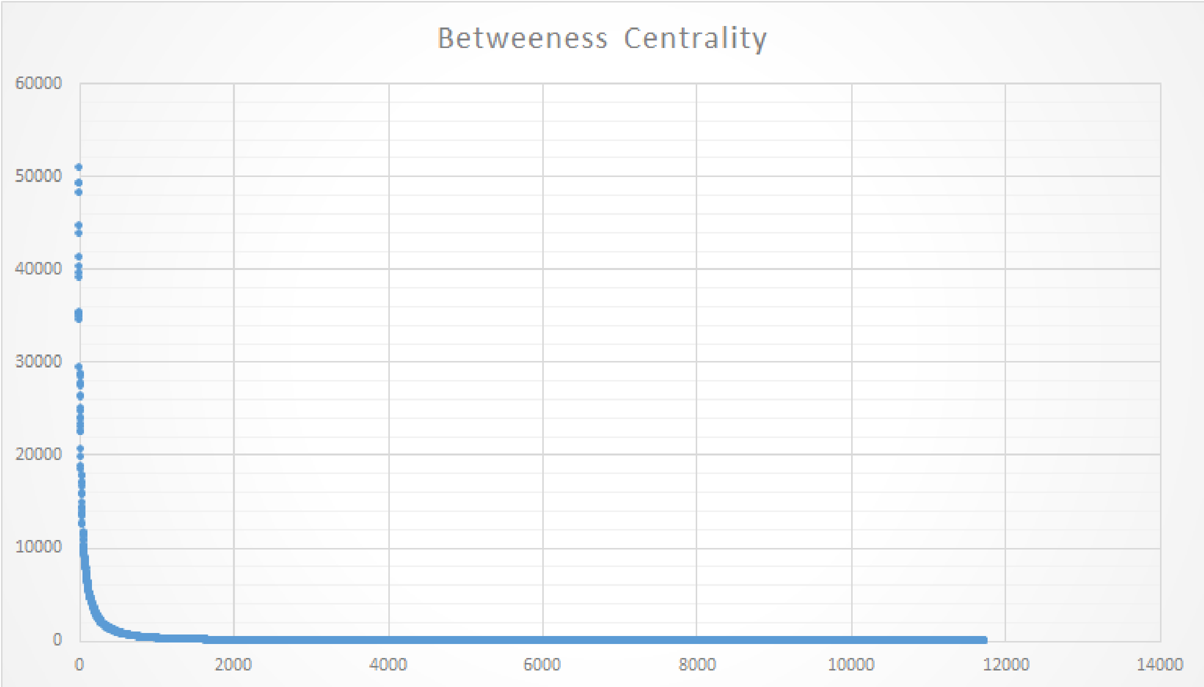


Figure 4.1.1.2.1 Betweenness centrality distribution - valid_abstract_authors dataset.

4.1.1.3 Closeness centrality

6976 out of 11730 nodes in total take values in closeness centrality over 0 to 1 and the other 4754 (40%) have no value. Below is given the respective graph of closeness centrality distribution among the Network vertices (Figure 4.1.1.3.1).

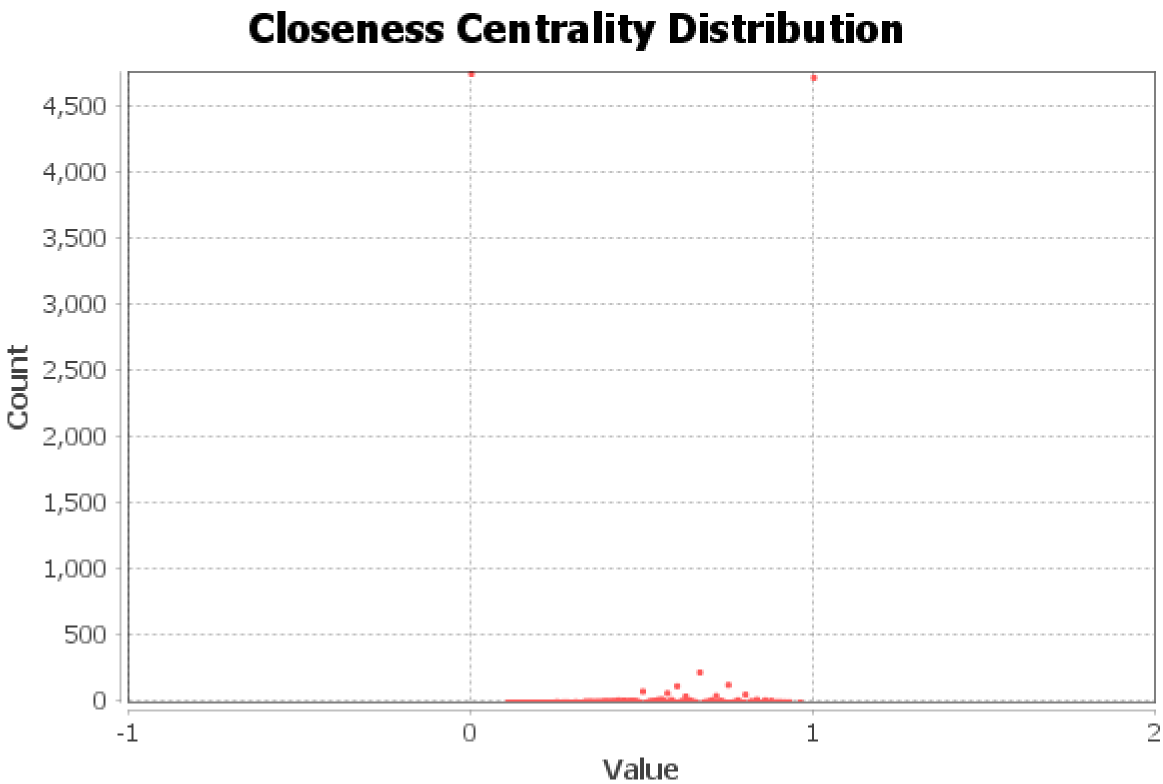


Figure 4.1.1.3.1 Closeness centrality distribution - valid_abstract_authors dataset.

4.1.1.4 Eigenvector centrality

The vertices of the Network being examined show low values of Eigenvector centrality. Specifically, only 55 nodes have Eigenvector centrality's value that range between 0.1 and 1 and the vast majority falls into a range of 0.1 to 0. The respective graph that represents this distribution follows (Figure 4.1.1.3.1).

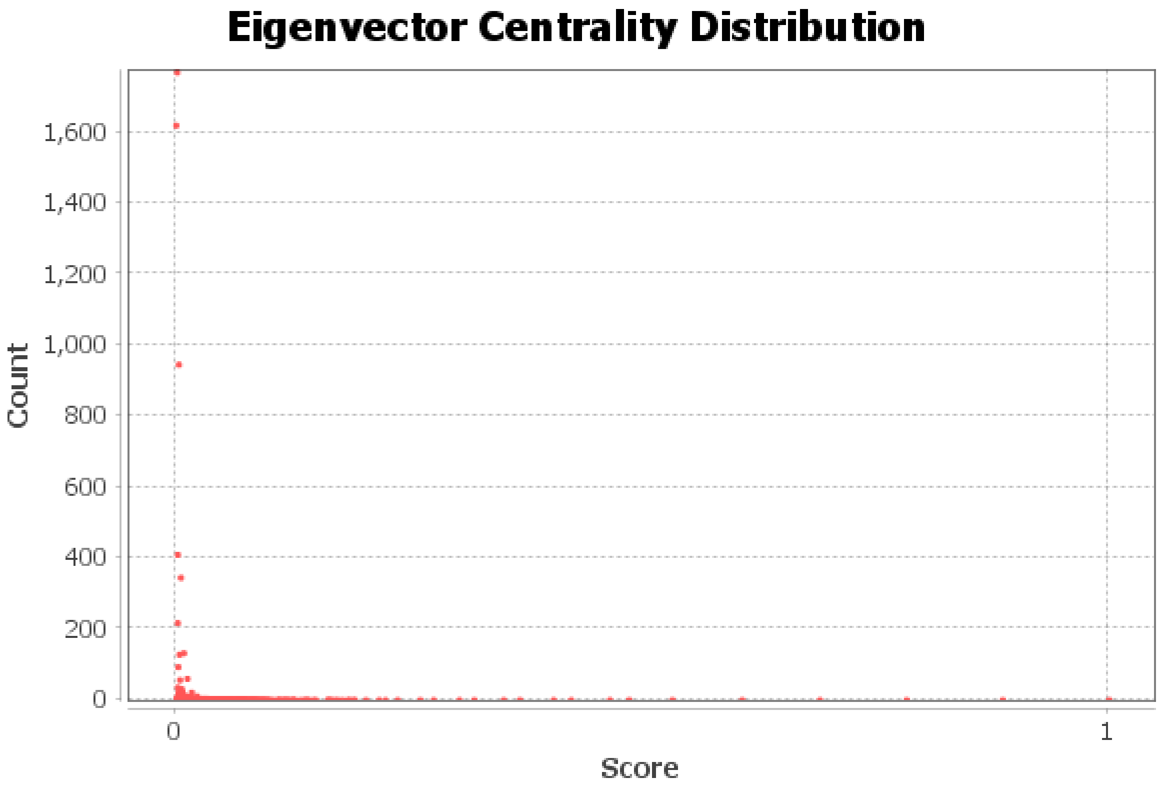


Figure 4.1.1.3.1 Eigenvector centrality distribution - valid_abstract_authors dataset.

4.1.2 Network Level Analysis

4.1.2.1 Network Diameter

The maximum node eccentricity found in the Network is 17, which gives its Diameter as well. As it can be seen, a large portion of vertices has eccentricity greater than 5 (618 vertices out of 11730 or 5% of the total Network), while 60% of the total Network's nodes have eccentricity with values over 0. The average value of eccentricity is 1.26. The Network's eccentricity distribution is depicted in the following graph (Figure 4.1.2.1.1)

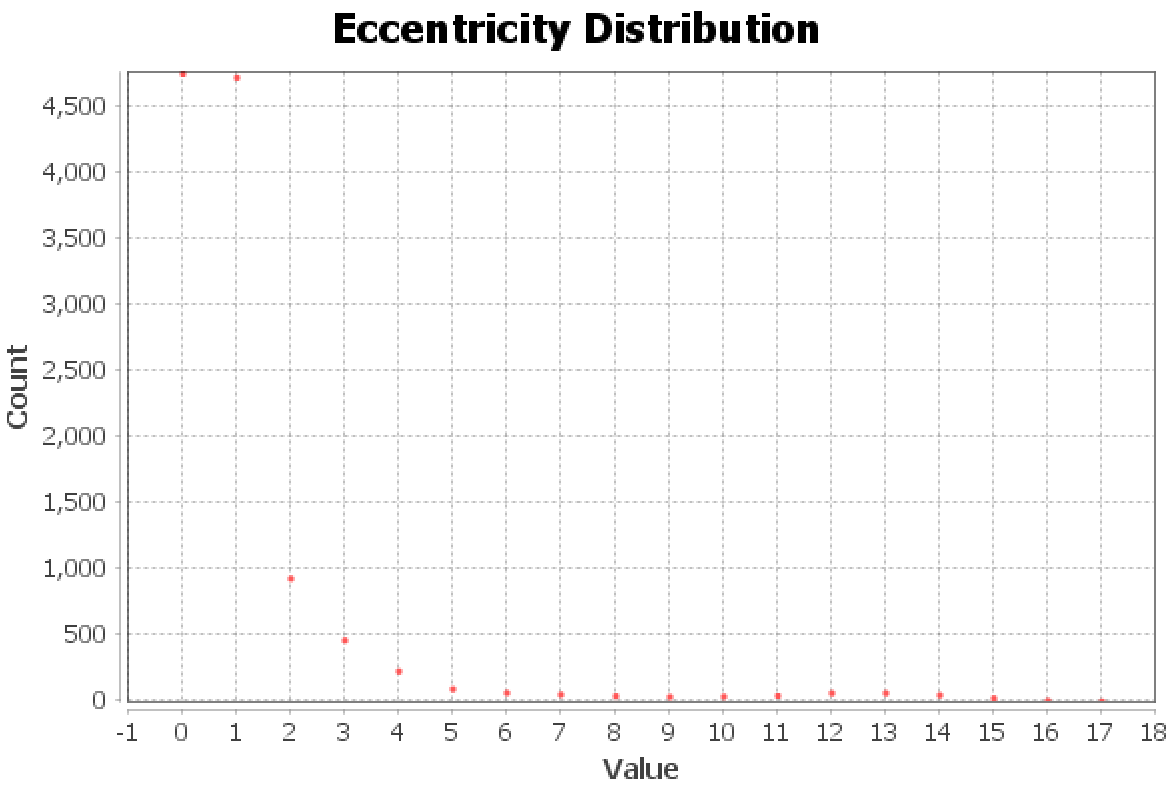


Figure 4.1.2.1.1 Eccentricity distribution - valid_abstract_authors dataset

4.1.2.2 Network Average Geodesic Distance

The average geodesic distance of the Network is 5.55.

4.1.2.3 Network Average Degree

The average degree of the Network is 1.92.

4.1.2.4 Network Modularity

The calculation of the Network Modularity has been done using the option “randomize” which produces a better Network decomposition but increases the computational time, the edge weights have been used for the purpose and finally the resolution has been set to 1.0. The modularity of the Network is 0.954 and the number of all communities found in it is 1979. Below follows the respective graph of the communities’ size distribution (Figure 4.1.2.4.1).

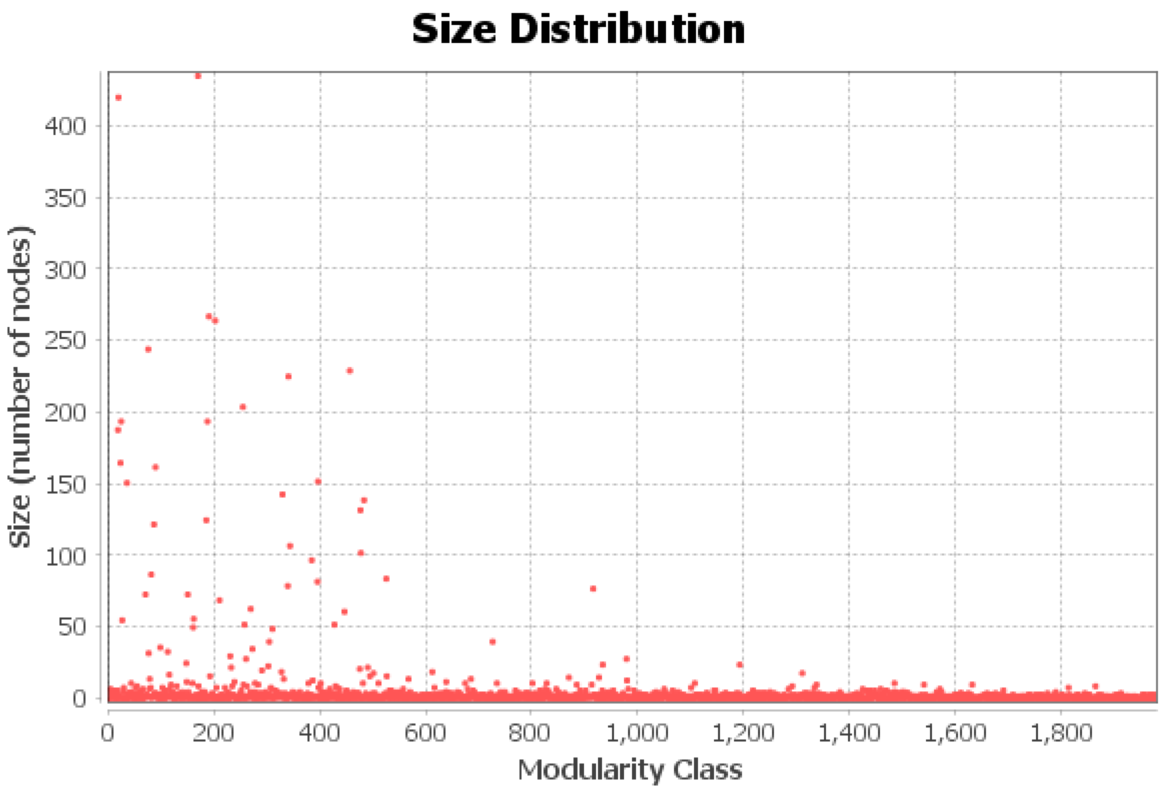


Figure 4.1.2.4.1 Size distribution - valid_abstract_authors dataset

4.2 valid_citation_authors Dataset

This directed and weighted graph contains 11536 nodes, which are the authors of all titles of the initial dataset, which are accompanied with valid citations' file, as described in chapter 3.

4.2.1 Node Level Analysis

4.2.1.1 Degree centrality

Below are presented the results that were extracted after the degree centrality analysis. The results show that the maximum value of degree centrality found is 118 nodes, where 117 are out-degree and 1 in-degree. The majority of the participants of the Network have a relatively low degree centrality value and specifically 10912 out of 11536 (95%) nodes in total have equal or lower degree centrality of 10 nodes. The average degree centrality is 1.9. Three graphs representing the degree distribution are given below (Figures 4.2.1.1.1, 4.2.1.1.2, 4.2.1.1.3).

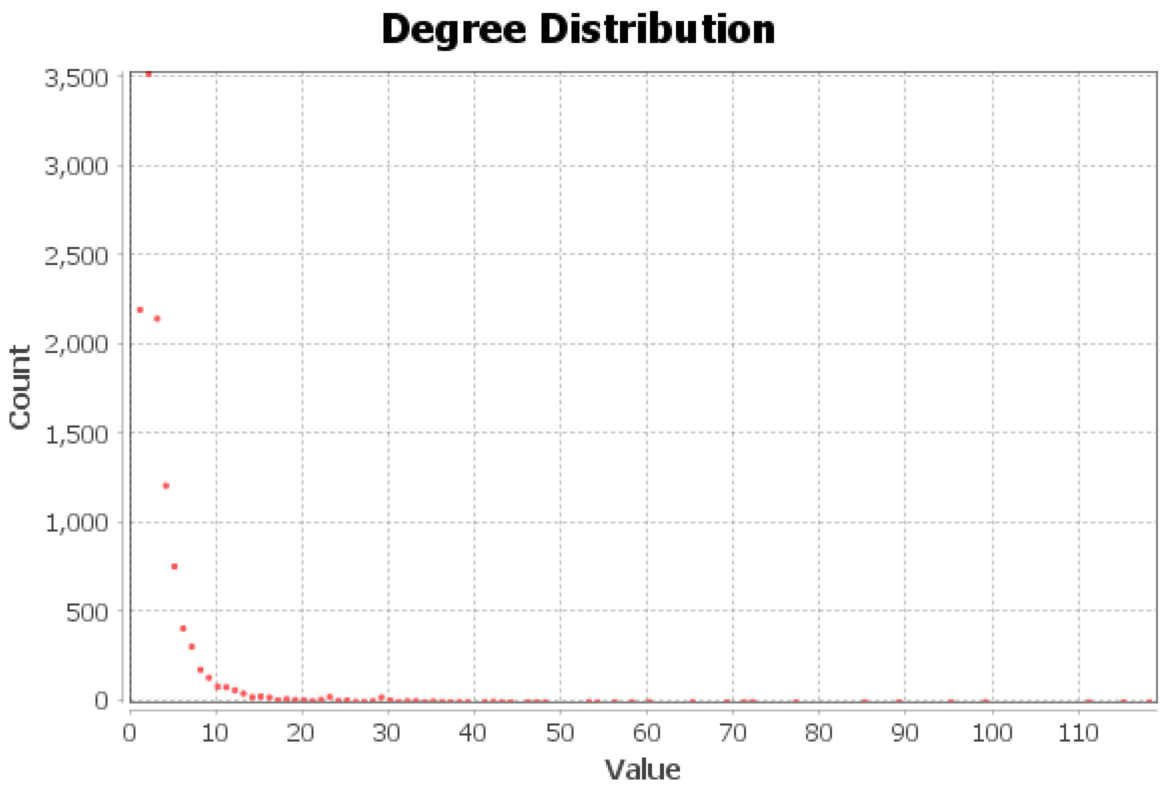


Figure 4.2.1.1.1 Degree centrality distribution - valid_citation_authors dataset

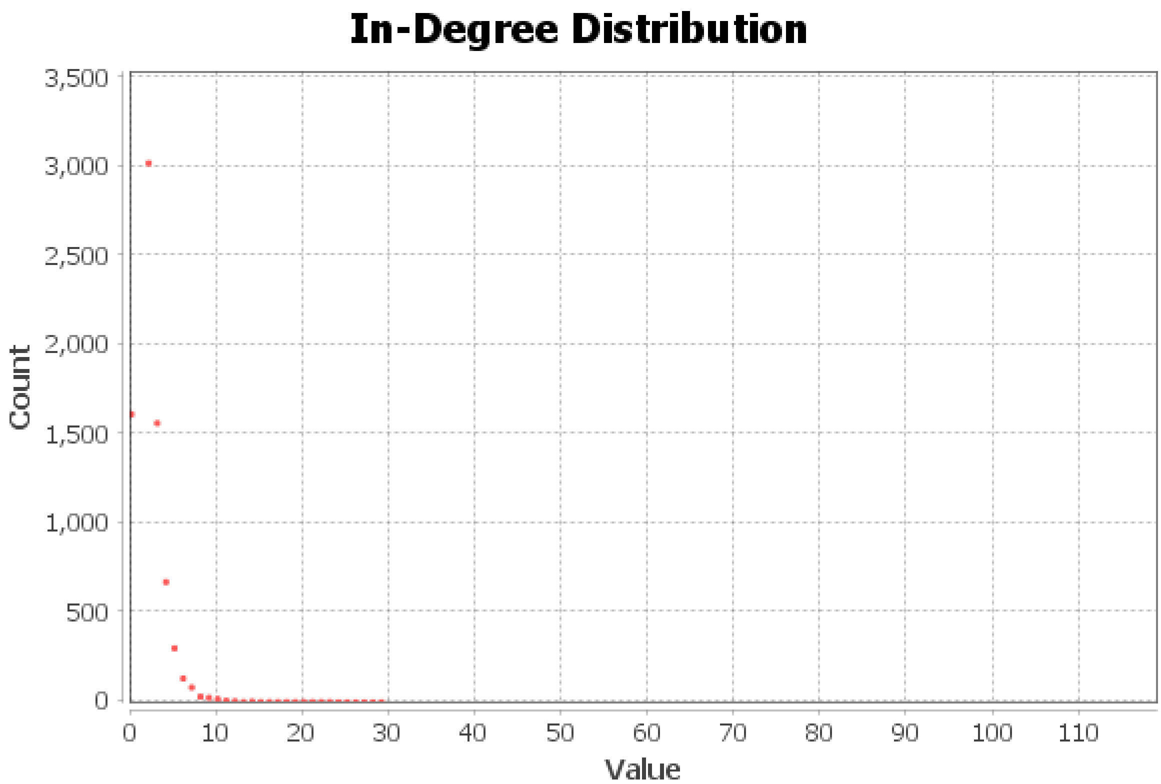


Figure 4.2.1.1.2 In - Degree centrality distribution - valid_citation_authors dataset

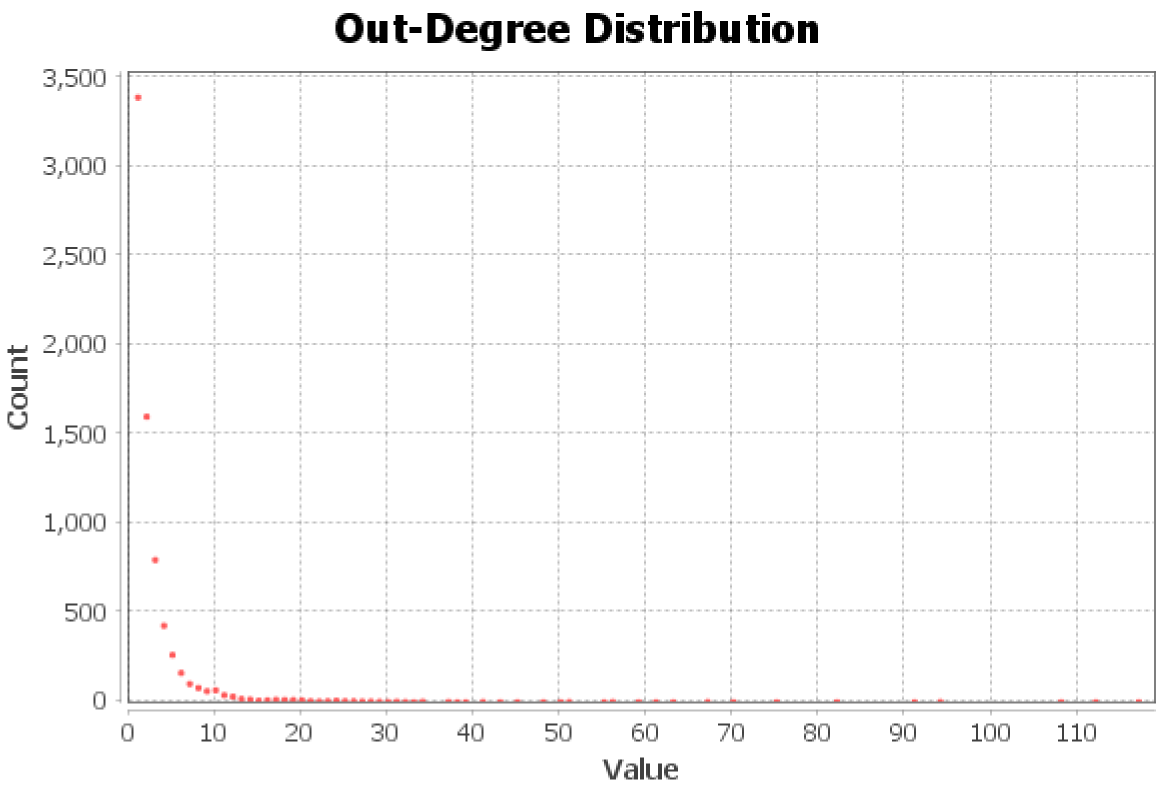


Figure 4.2.1.1.3 Out - Degree centrality distribution - valid_citation_authors dataset

4.2.1.2 Betweenness centrality

As can be seen in the betweenness centrality graph that follows (Figure 4.2.1.2.1), a large number of vertices have a relatively low betweenness centrality and there are only a few with higher values. In particular, 427 vertices have betweenness centrality over 1000, 180 take values within 500 and 1000 and the rest (10930 or 94.7%) are below 500.

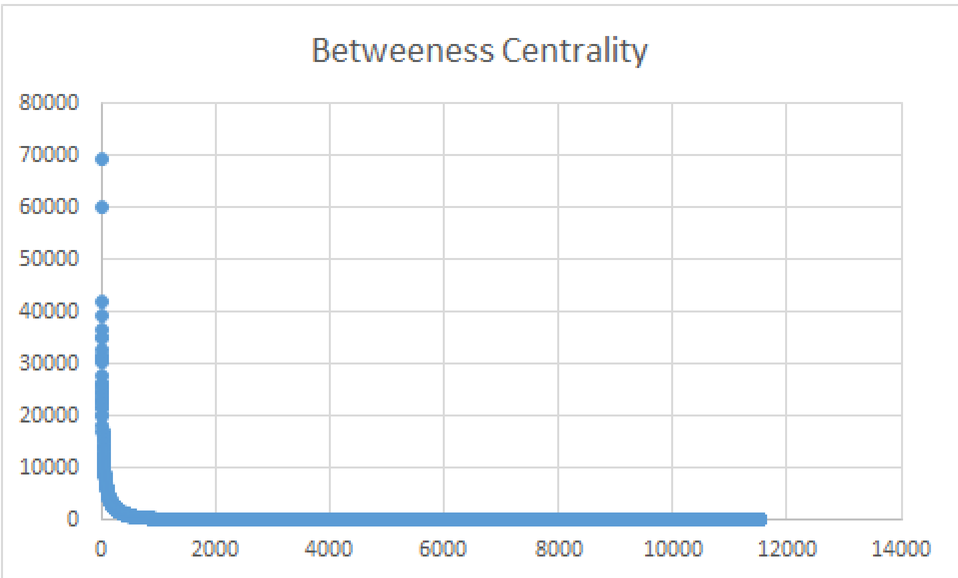


Figure 4.2.1.2.1 Betweenness centrality distribution - valid_citation_authors dataset

4.2.1.3 Closeness centrality

6837 out of 11536 nodes in total take values in closeness centrality over 0 to 1 and the other 4700 (41%) have no value. Below is given the respective graph of closeness centrality distribution among the Network vertices (Figure 4.2.1.3.1).

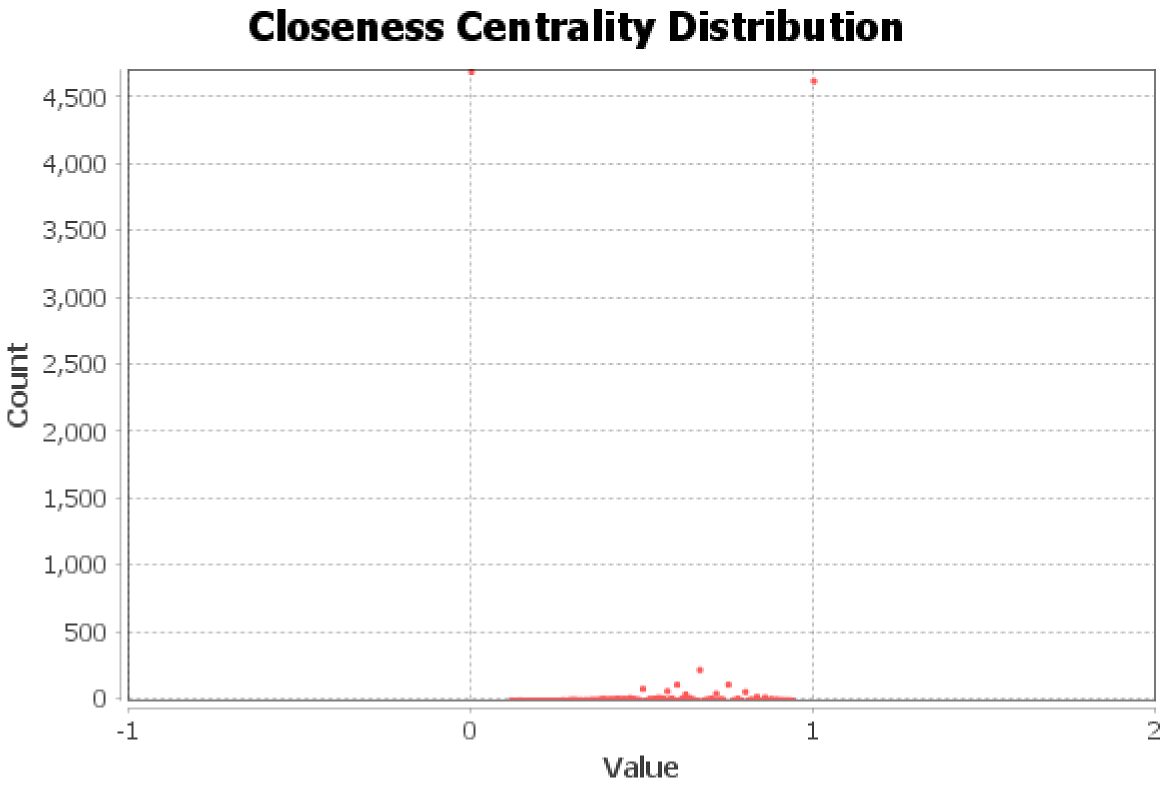


Figure 4.2.1.3.1 Closeness centrality distribution - valid_citation_authors dataset

4.2.1.4 Eigenvector centrality

The vertices of the Network being examined show low values of Eigenvector centrality. Specifically, only 48 nodes have Eigenvector centrality's value that range between 0.1 and 1 and the vast majority falls into a range of 0.1 to 0. The respective graph that represents this distribution follows (Figure 4.2.1.3.1).

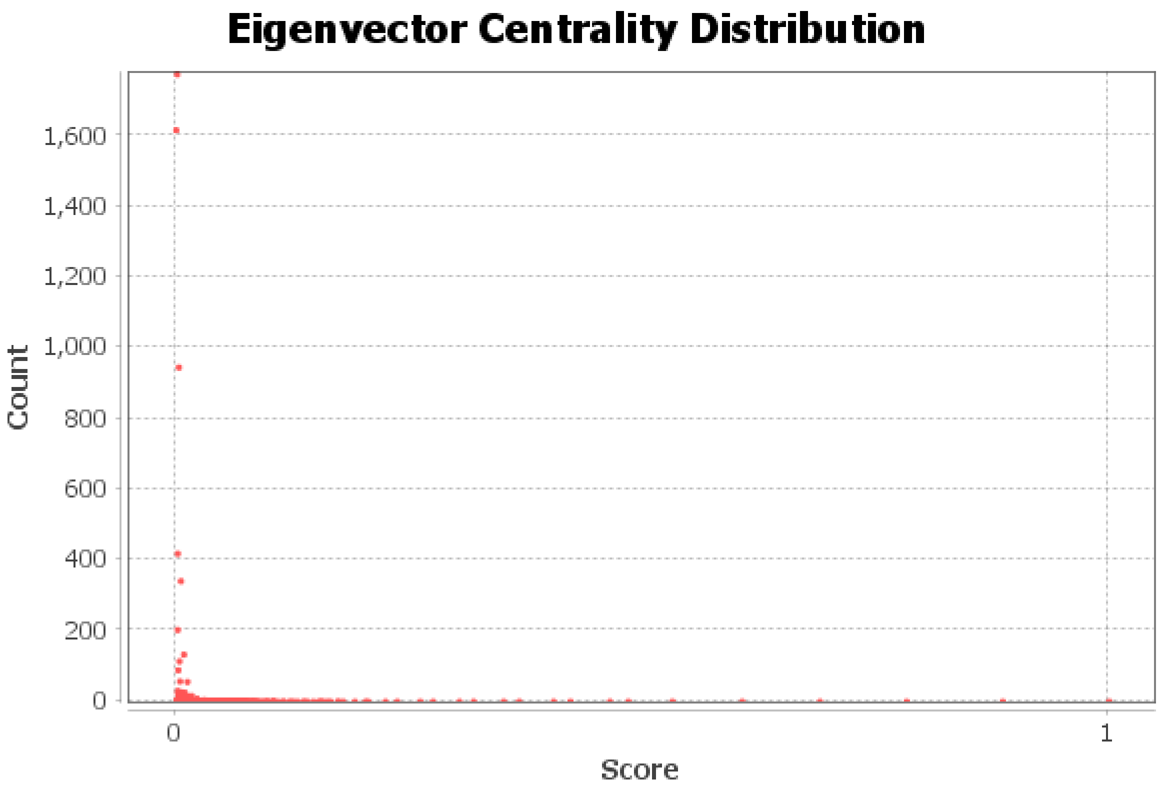


Figure 4.2.1.3.1 Eigenvector centrality distribution - valid_citation_authors dataset

4.2.2 Network Level Analysis

4.2.2.1 Network Diameter

The maximum node eccentricity found in the Network is 17, which gives its Diameter as well. As it can be seen, a tiny portion of vertices has eccentricity greater than 5 (590 vertices out of 11536 or 5% of the total Network), while 59.2% of the total Network's nodes have eccentricity with values over 0. The average value of eccentricity is 1.22. The Network's eccentricity distribution is depicted in the following graph (Figure 4.2.2.1.1).

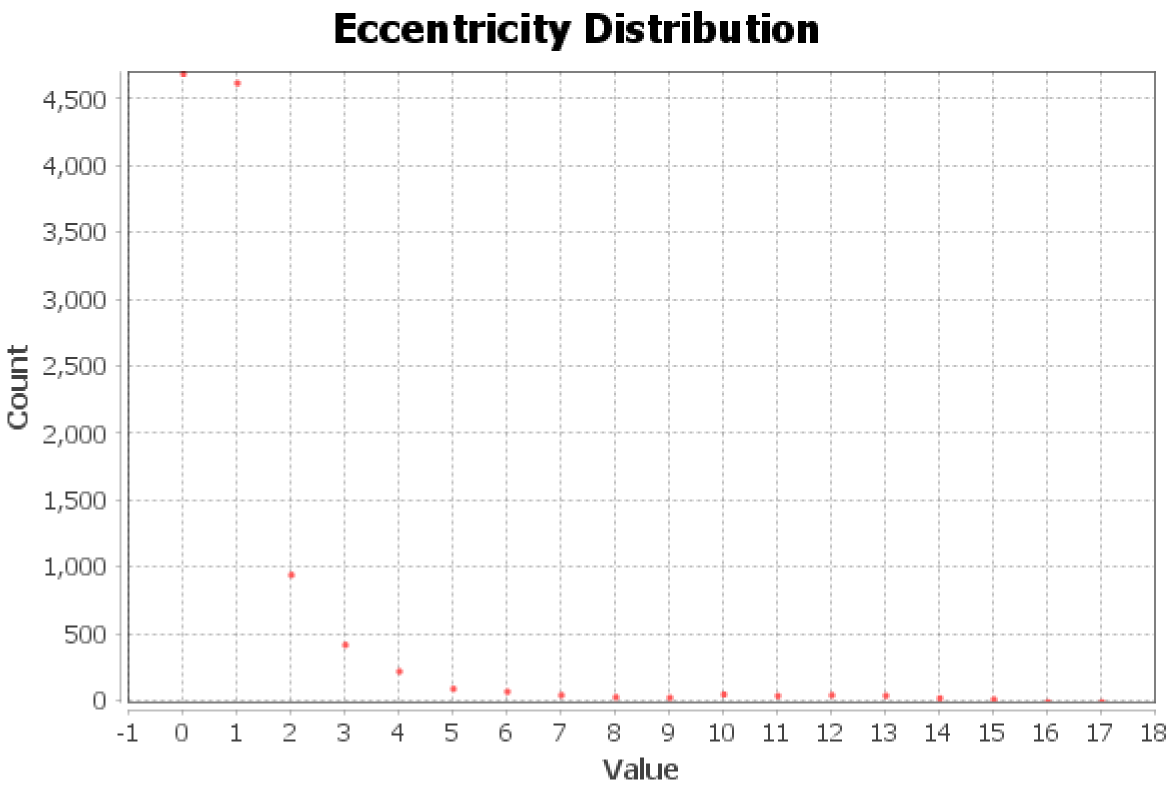


Figure 4.2.2.1.1 Eccentricity distribution - valid_citation_authors dataset

4.2.2.2 Network Average Geodesic Distance

The average geodesic distance of the Network is 5.31.

4.2.2.3 Network Average Degree

The average degree of the Network is 1.9.

4.2.2.4 Network Modularity

The calculation of the Network Modularity has been done using the option "randomize" which produces a better Network decomposition but increases the computational time, the edge weights have been used for the purpose and finally the resolution has been set to 1.0. The modularity of the Network is 0.953 and the number of all communities found in it is 1990. Below follows the respective graph of the communities' size distribution (Figure 4.2.2.4.1).

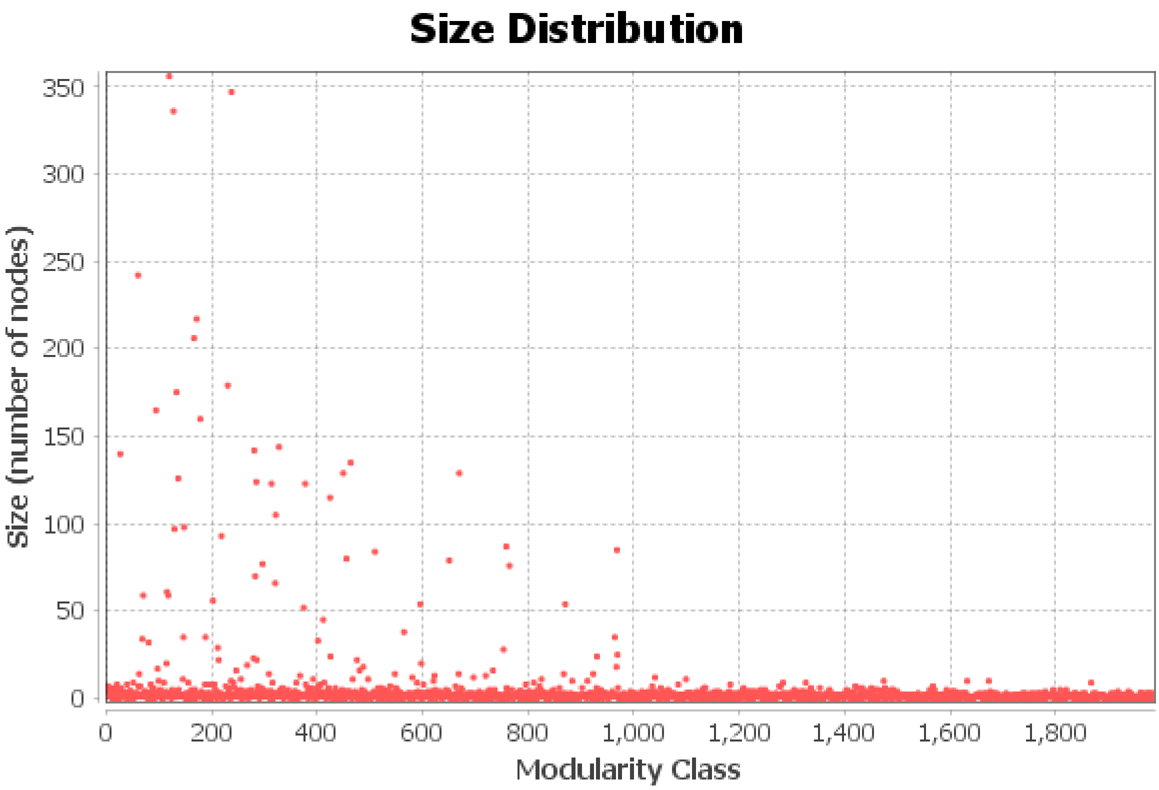


Figure 4.2.2.4.1 Size distribution - valid_citation_authors dataset

4.3 valid_citation_titles Dataset

This directed and weighted graph contains 8079 nodes, which are the titles with valid citations' file, as described in chapter 3.

4.3.1 Node Level Analysis

4.3.1.1 Degree centrality

Below are presented the results that were extracted after the degree centrality analysis. The results show that the maximum value of degree centrality found is 555 nodes, where 546 are out-degree and 9 in-degree. The majority of the participants of the Network have a relatively low degree centrality value and specifically 6668 out of 8079 (83%) nodes in total have equal or lower degree centrality of 20 nodes. Three graphs representing the degree distribution are given below (Figures 4.3.1.1.1, 4.3.1.1.2, 4.3.1.1.3).

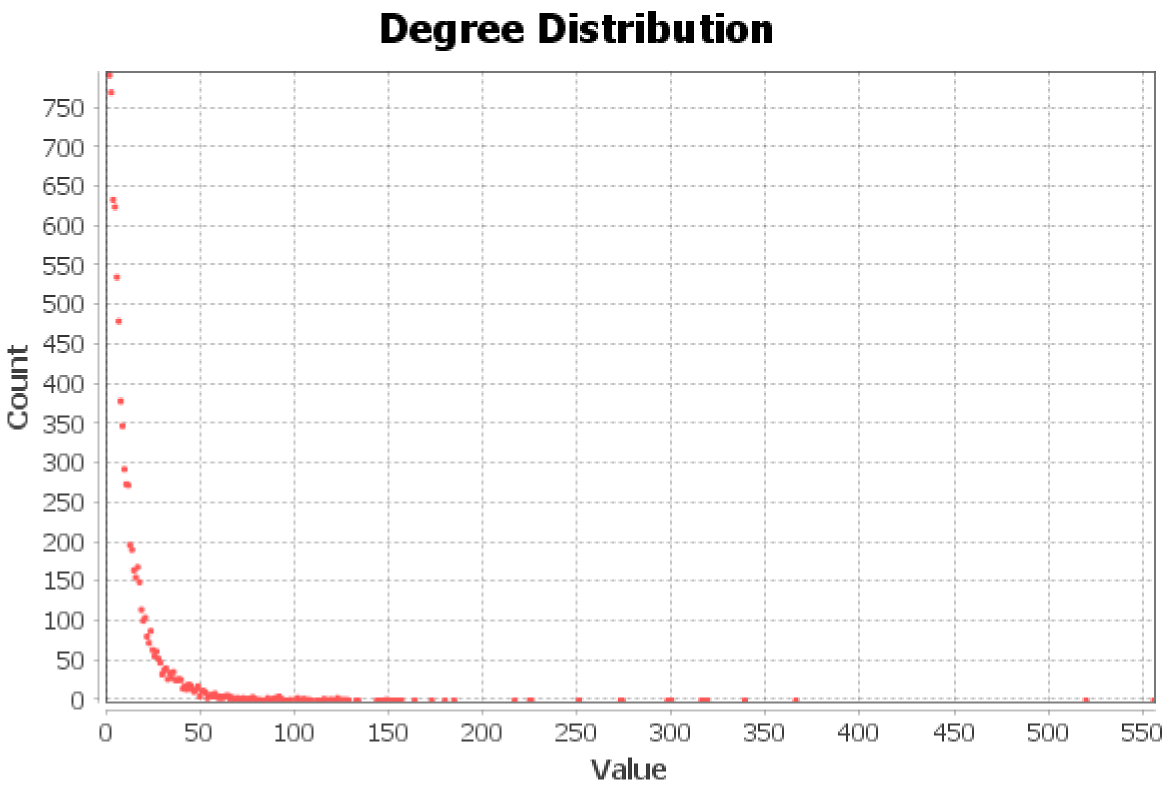


Figure 4.3.1.1.1 Degree centrality distribution -valid_citation_titles Dataset

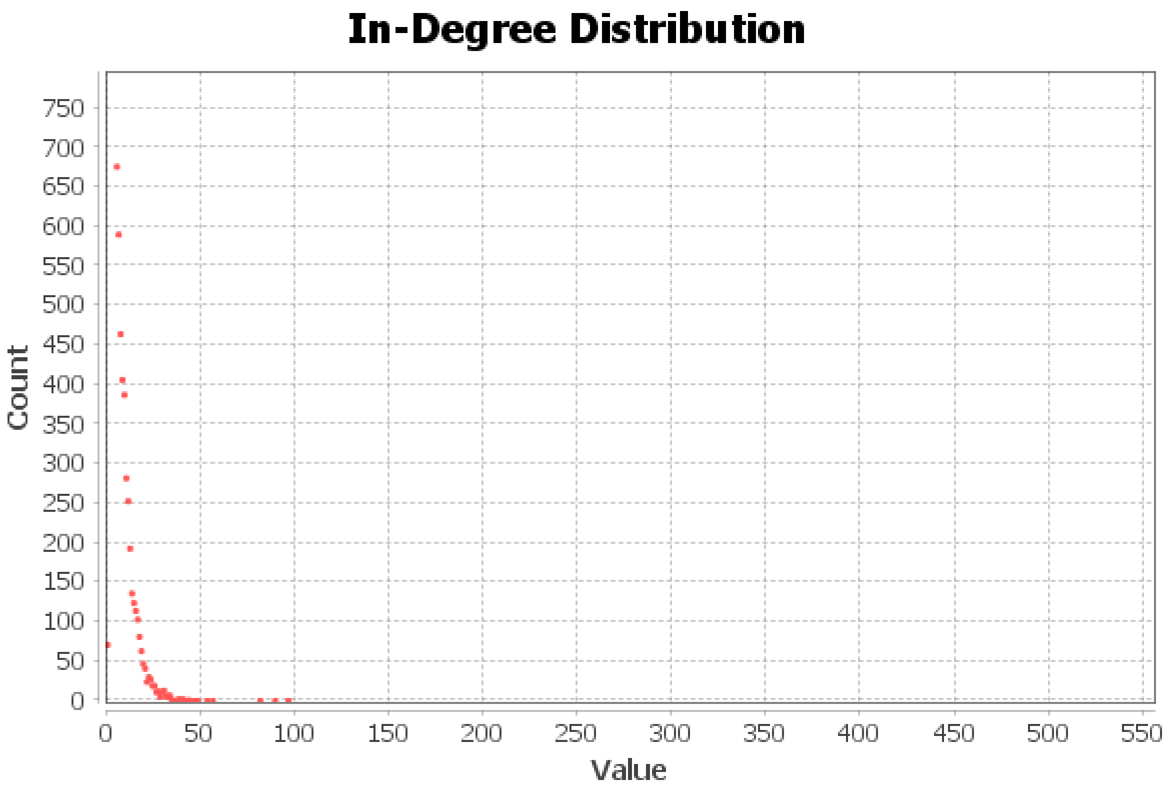


Figure 4.3.1.1.2 In - Degree centrality distribution -valid_citation_titles Dataset

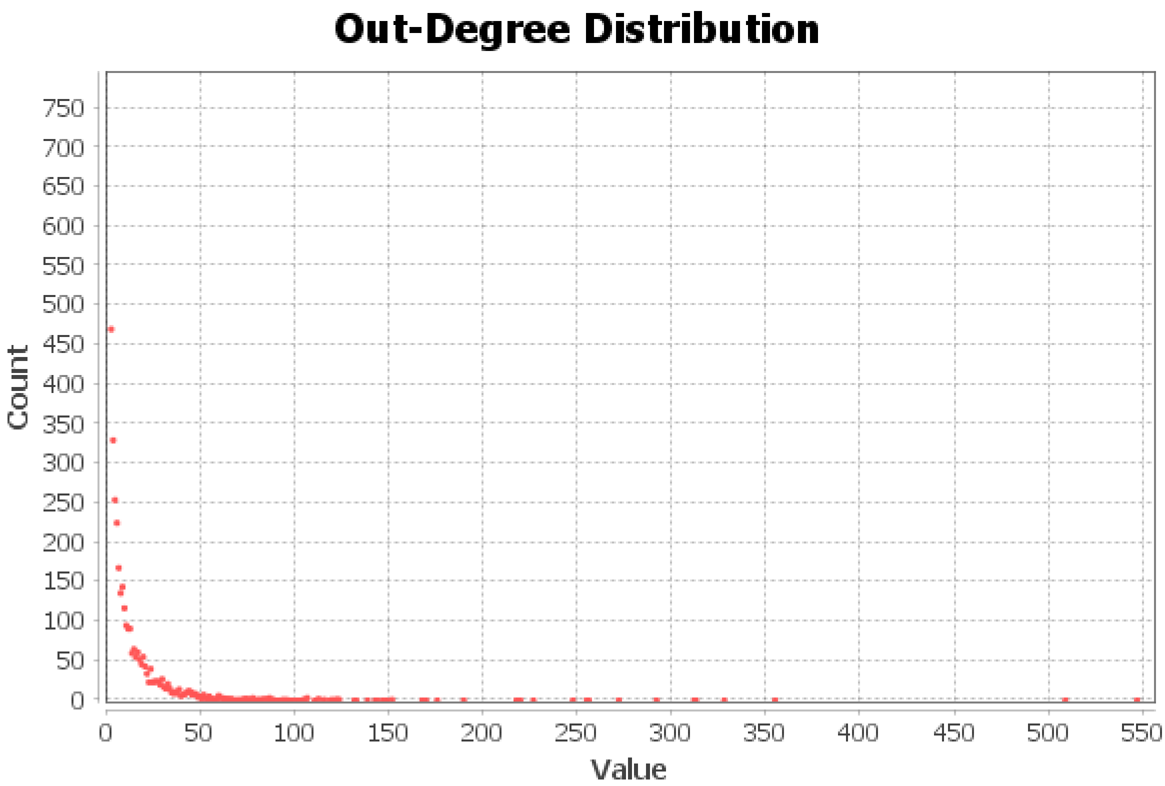


Figure 4.3.1.1.3 Out - Degree centrality distribution - valid_citation_titles Dataset

4.3.1.2 Betweenness centrality

As can be seen in the betweenness centrality graph that follows (Figure 4.3.1.2.1), a large number of vertices have a relatively low betweenness centrality and there are only a few with higher values. A significant point to notice here is the existence of one single node which has betweenness centrality 727288.58 and is the maximum value found among vertices and then follows one vertex that drops this value down to 283779.24 (61% drop). In general, 2976 vertices have betweenness centrality over 1000, 370 take values within 500 and 1000 and the rest (4735 or 58.6%) are below 500.

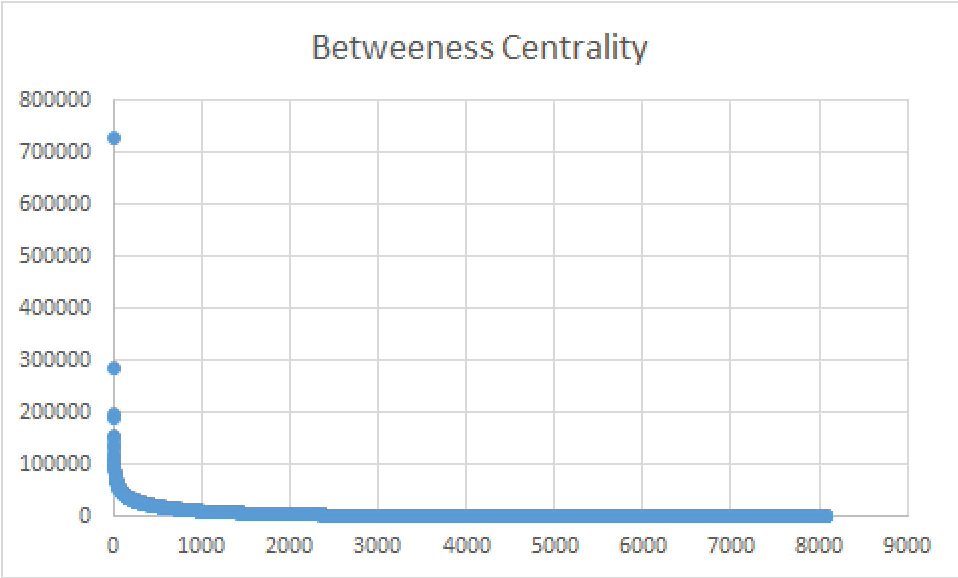


Figure 4.3.1.2.1 Betweenness centrality distribution - valid_citation_titles Dataset

4.3.1.3 Closeness centrality

4074 out of 8079 nodes in total take values in closeness centrality over 0 to 1 and the other 4005 (49.5%) have no value. Below is given the respective graph of closeness centrality distribution among the Network vertices (Figure 4.3.1.3.1).

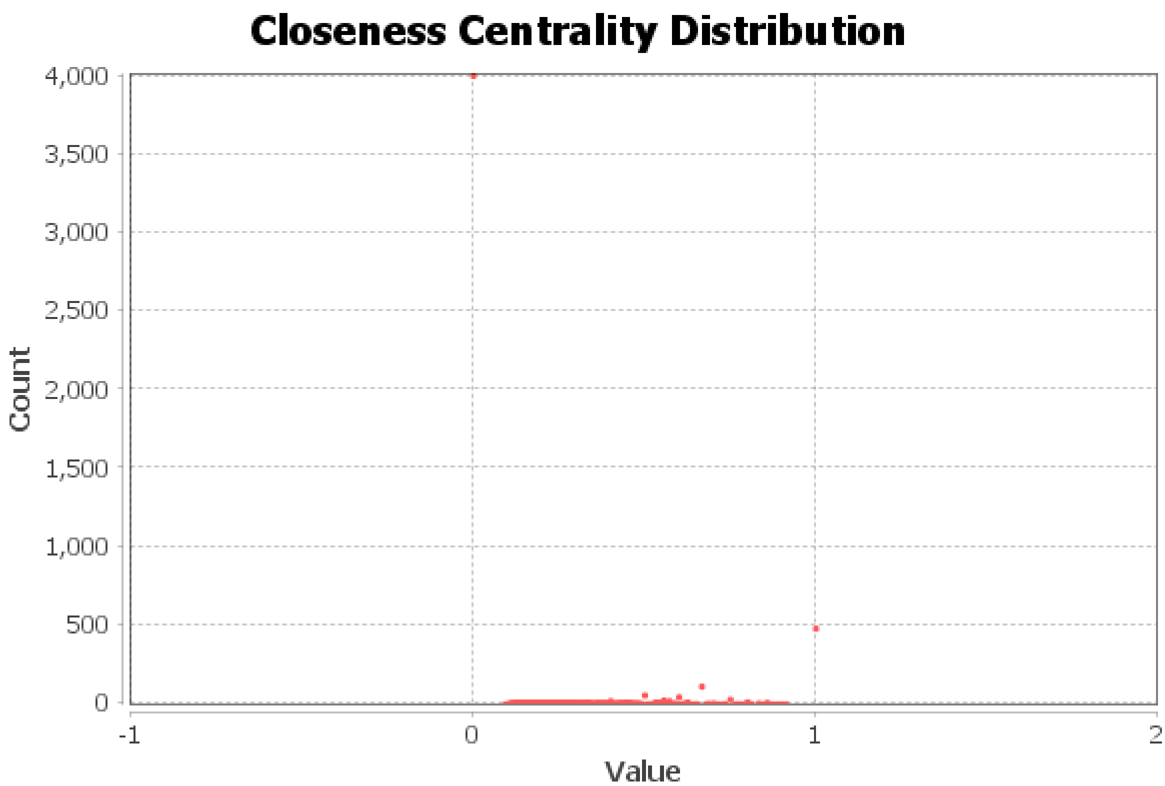


Figure 4.3.1.3.1 Closeness centrality distribution - valid_citation_titles Dataset

4.3.1.4 Eigenvector centrality

The vertices of the Network being examined show low values of Eigenvector centrality. Specifically, only 1029 nodes have Eigenvector centrality's value that range between 0.1 and 1 and the rest 7060 fall into a range of 0.1 to 0. The respective graph that represents this distribution follows (Figure 4.3.1.4.1).

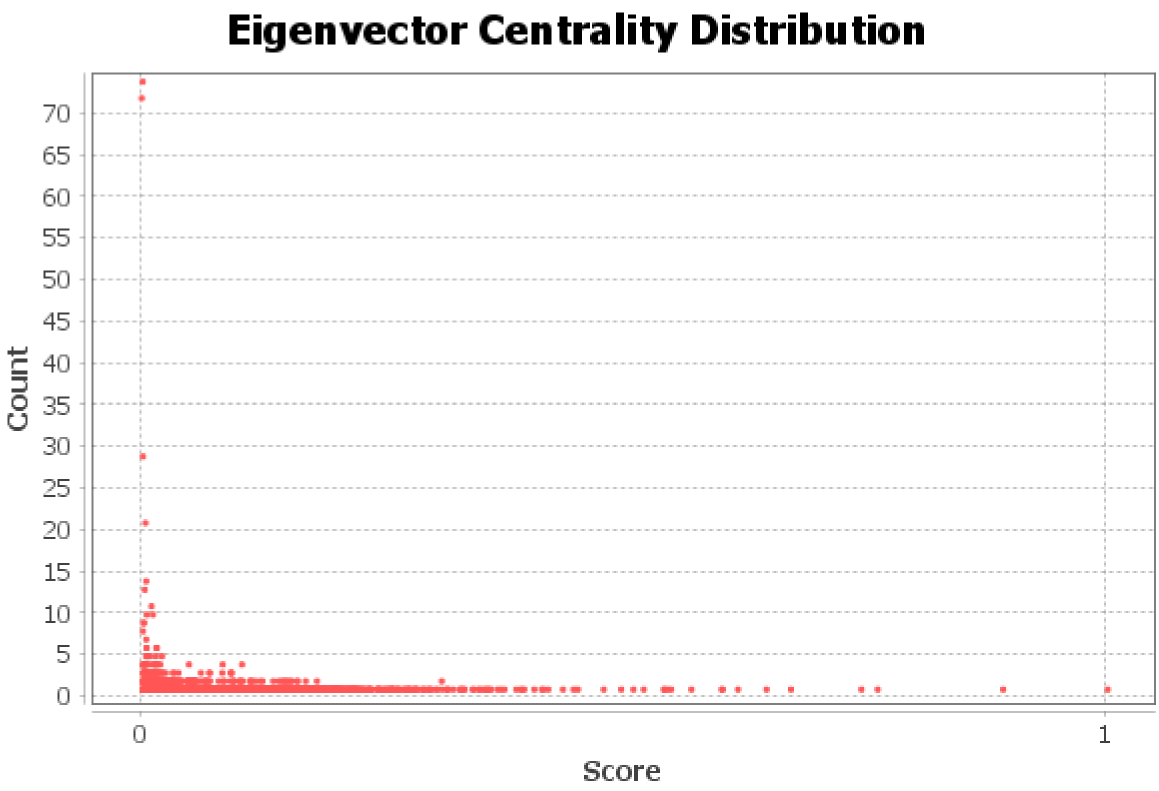


Figure 4.3.1.4.1 Eigenvector centrality distribution - valid_citation_titles Dataset

4.3.2 Network Level Analysis

4.3.2.1 Network Diameter

The maximum node eccentricity found in the Network is 24, which gives its Diameter as well. The eccentricity's values follow a harmonic motiv with one pick on value e = 1 and another one on value e = 12 (Figure 4.3.2.1.1). 434 nodes take eccentricity values over 15 and the rest (7645 or 94.6%) lie below that value. The average eccentricity is 4.53.

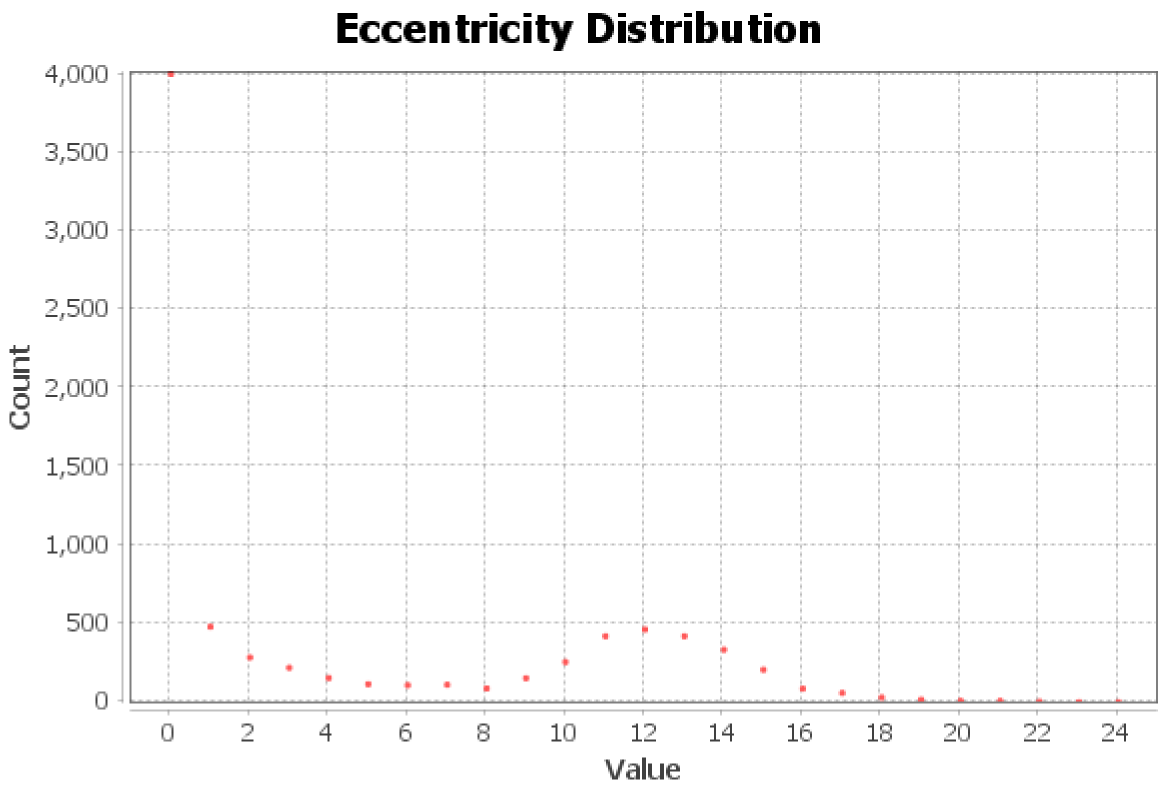


Figure 4.3.2.1.1 Eccentricity distribution - valid_citation_titles Dataset

4.3.2.2 Network Average Geodesic Distance

The average geodesic distance of the Network is 5.02.

4.3.2.3 Network Average Degree

The average degree of the Network is 6.38.

4.3.2.4 Network Modularity

The calculation of the Network Modularity has been done using the option “randomize” which produces a better Network decomposition but increases the computational time, the edge weights have been used for the purpose and finally the resolution has been set to 1.0. The modularity of the Network is 0.52 and 28 communities are found. Below follows the respective graph of the communities’ size distribution (Figure 4.3.2.4.1).

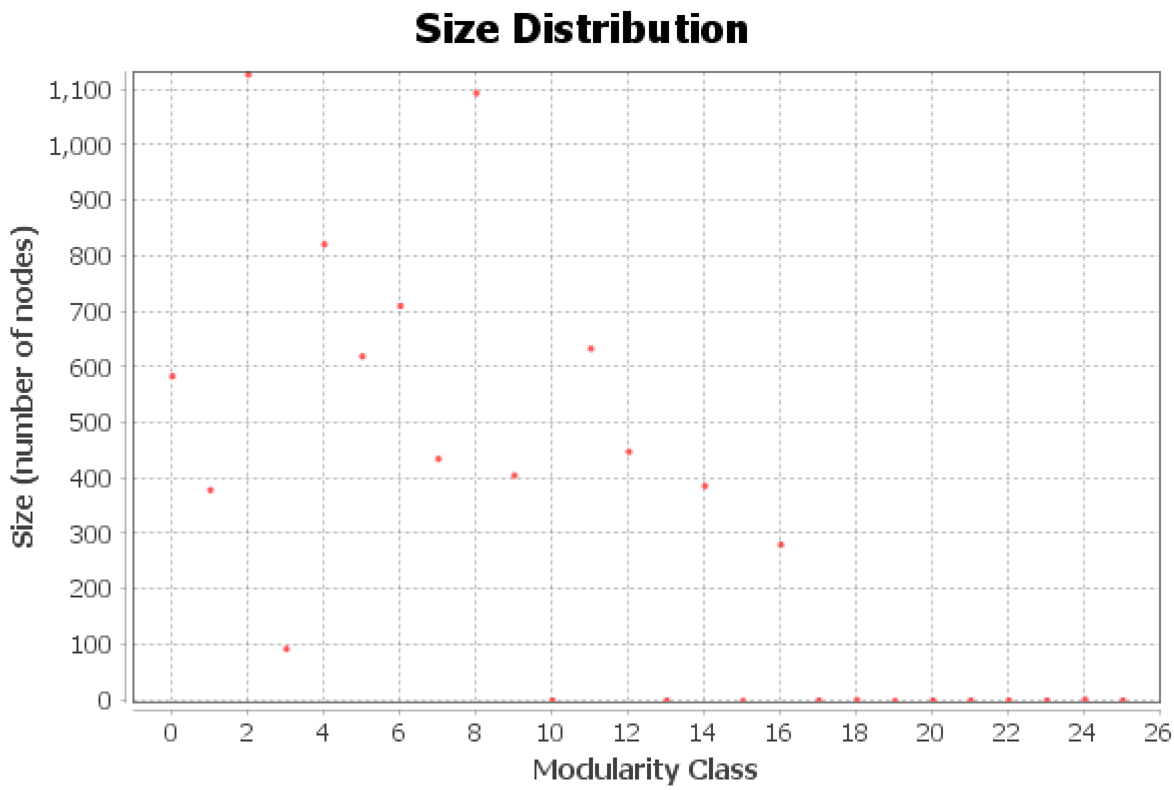


Figure 4.3.2.4.1 Eccentricity distribution - valid_citation_titles Dataset

5. Conclusions & Future Work

In this work there have been examined three co-authorship Networks which were extracted after a data analysis of a forty-year scholarly literature of journal DEA-related articles between 1978 and 2016 (Emrouznejad & Yang, 2018). Each network was imported into Gephi software and after analysis, results were collected both in node and network level analysis.

Both authors' networks, valid_abstract_authors and valid_citation_authors, performed in a similar way and the results of each model are similar. Their average degree centrality is about 1.9, a number that shows that the majority of the networks' actors have low interconnection. Specifically, there are about 8 to 9 authors in each network that have a high number of connections (over 80) and the vast majority have significantly less. In particular, only 5% of authors, in both networks, can be considered as central nodes and the other 95% as peripheral nodes of the network. It can be said that these networks include mainly supporters rather than influencers, because the number of in-bound arcs exceeds the number of out-bound arcs.

A similar pattern can be observed in betweenness centrality distribution of both networks. Again, about 95% of the participants present a relatively low value of betweenness centrality, compared to the highest values of the measurement that have only the remaining 5% of them in the Network. This shows that the informational rootes of the Network pass through a significant low percentage of authors participating in it.

Examining the closeness centrality in both author's networks, it is observed that almost 60% of the participants present zero (0) closeness centrality, demonstrating two sparse

Networks in which most of its nodes are peripherals, self-organized in mini communities which do not connect to the main core. The main core consists of 40% of the total number of the participants. The same conclusion can be extracted by examining the Eigenvector centrality of both Networks.

The Network Level Analysis which has been conducted on both authors' networks presents the same view that we have seen in the aforementioned results. Both Networks can be characterized as sparse connected Networks, in which many small communities are organized in the peripherals, while the core has a diameter of 17. This means that there is not a meaningful interconnection between authors and both Networks have a high modularity.

Regarding the titles' Network, the degree centrality distribution shows a low interconnection between the articles. The articles with the highest degree centrality are the 17% of the total number of the participants. Again, betweenness, closeness and Eigenvector centralities follow a similar pattern, which indicates this low interconnectivity of the participants.

In addition, the Network Level Analysis shows a sparse connected Network, in which many communities are self-organized in the peripherals and less lie in the core of the Network.

The aforementioned results present a low interconnectivity between authors and articles. They show sparse connected Networks with tiny communities in it, which indicates a low level of communication within the Network. An interesting research could be done on these Networks, relative to the level of transferability of information. It would be valuable research if one could apply experiments on these Networks, in order to find the optimal path through which the information could pass to the entire Network. In this work, all Networks were extracted without concluding the time in which each article was written. This information is available though, in the initial dataset. Another possible work that could be done, would be to extract the year of each article and, through this, predict the possibility of each community to connect with the rest in the future.

6. Bibliography

Albert, R. and Barabási, A.L. (2002), “Statistical mechanics of complex networks”, Reviews of Modern Physics, Vol. 74 No. 1, pp. 47-97.

Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics And Its Applications*, *311*(3-4), 590-614. doi: 10.1016/s0378-4371(02)00736-7

Bavelas, A. and D. Barrett (1948). “An experimental approach to organizational communication”. Personnel 27:366-371.

Barabási, A.L. and Albert, R. (1999), “Emergence of scaling in random networks”, Science, Vol. 286 No. 5439, pp. 509-512.

Biscaro, C. and Giupponi, C. (2014), “Co-authorship and bibliographic coupling network effects on
citations”, Plos One, Vol. 9 No. 6.

de Sola Pool, I., & Kochen, M. (1978). Contacts and influence. *Social Networks*, *1*(1), 5-51. doi: 10.1016/0378-8733(78)90011-4

Emrouznejad, A., & Yang, G. (2018). A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socio-Economic Planning Sciences*, *61*, 4-8. doi: 10.1016/j.seps.2017.01.008

Freeman, L. C. (1978/79), Centrality in Social Networks Conceptual Clarification. Social Networks, 1 (1978/79) 215-239

Kumar, S. (2015). Co-authorship networks: a review of the literature. *Aslib Journal Of Information Management*, *67*(1), 55-73. doi: 10.1108/ajim-09-2014-0116

Leavitt, H. J. (1951).“Some effects of communication patterns on group performance”. Journal of Abnormal and Social Psychology 46:38-50.

Lewison, G. and Markusova, V. (2010), “The evaluation of Russian cancer research”, Research Evaluation, Vol. 19 No. 2, pp. 129-144.

Linton C., F. (2002). The Development of Social Network Analysis. *Empirical Press Vancouver, BC Canada*, 10-65.

Milgram, Stanley (1967). "The Small World Problem". Psychology Today. Ziff-Davis Publishing Company.

Moody, J. (2004), “The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999”, American Sociological Review, Vol. 69 No. 2, pp. 213-238.

Newman, M. (2001). The structure of scientific collaboration networks. *Proceedings Of The National Academy Of Sciences*, *98*(2), 404-409. doi: 10.1073/pnas.98.2.404

Newman, M. (2006). Modularity and community structure in networks. *Proceedings Of The National Academy Of Sciences*, *103*(23), 8577-8582. doi: 10.1073/pnas.0601602103

Newman, M. E. J. (2006). Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109–1040

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal Of Information Science*, *28*(6), 441-453. doi: 10.1177/016555150202800601

Quatman, C. and Chelladurai, P. (2008), "The social construction of knowledge in the field of sport
management: a social network perspective", Journal of Sport Management, Vol. 22 No. 6, pp. 651-676.

Racherla, P. and Hu, C. (2010), "A social network perspective of tourism research collaborations",
Annals of Tourism Research, Vol. 37 No. 4, pp. 1012-1034.

Tabassum, S., Souza Fernandes Pereira, F., Fernandes, S., & Gama, J. (2018). Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery*. doi: 10.1002/widm.1256

Uddin, S., Hossain, L., Abbasi, A. and Rasmussen, K. (2012), "Trend and efficiency analysis of co-authorship network", Scientometrics, Vol. 90 No. 2, pp. 687-699.

Watts, D.J. and Strogatz, S.H. (1998), "Collective dynamics of 'small-world' networks", Nature, Vol. 393 No. 6684, pp. 440-442.

Yan, E.J., Ding, Y. and Zhu, Q.H. (2010), "Mapping library and information science in China: a co authorship network analysis", Scientometrics, Vol. 83 No. 1, pp. 115-131.