KNOWLEDGE EXTRACTION FROM BIBLIOGRAPHIC DATA

Dissertation submitted for the

Degree in Informatics and Telematics

PANAGOPOULOS GEORGE

HAROKOPEIO UNIVERSITY OF ATHENS

JUNE 2014

ABSTRACT

Bibliographic data is of crucial importance to the academic society as it consists of vital material for scholars and researchers. Through the years, various attempts have been conducted in order to exploit this source of information in the best way possible, revealing many scientific novelties during the process, which were later applied for other purposes. On the other hand, it has served as a prosperous field of experimentations for various methodologies in the field of informatics, basically due to its multiformity. As a result bibliographical data are an ideal test bench for combining various data mining methodologies and extracting knowledge.

The objective of this thesis is to introduce a novel approach into the problem of characterizing the scientific impact of authors, not solely based on their publication record, but also on its impact, on the impact of their co-authors and the evolution of the above through the years. As a proof-of-concept of our work, we process the publication record of all authors affiliated with Greek universities as given by the digital library Scopus.

More specifically, for each author, we collect information concerning the number of publications and citations in a yearly basis. In addition to this, we construct two types of collaboration graphs, where each author is connected with his/her co-authors with edges that denote either the strength or the impact of the co-operation. The graphs are build in a yearly basis and the weights on the edges are accumulated in order to aggregate information from previous years. Finally, we apply several graph mining techniques from biology and social network analysis in order to formulate authors' sociability in the aforementioned graphs. We then use these features to create change indexes and depict the evolution of the author's features in time.

In the next step, we cluster together authors with similar publication, citation and collaboration profile using the k-means clustering algorithm and a set of well-known cluster validity measures in order to determine the best number of clusters. The clusters are consequently labeled using the most characteristic features in the description of each cluster.

In order to find the most informative features for each cluster we use statistical methods. Based on these most informative features we illustrate the behavioral pattern of each

cluster, which corresponds to how the publication, impact or collaboration features change over time in each of the authors groups. At last we operate a time series clustering with the dynamic time warping method to cluster the authors based on their timeline and we compare the results in respect to the aforementioned clustering.

CONTENTS

LIST OF IMAGES

LIST OF TABLES

1. INTRODUCTION

Some introductory information about the purpose and the philosophy of our research.

1.1 AIMS

The purpose of this thesis is to create a methodology for the analysis of researchers' performance in publishing articles, being cited and collaborating with other researchers. The methodology is tested on authors affiliated with Greek universities and uses information from the Scopus database. However, any other group of researchers and any other bibliographical database (with publication and citation information per article and year) could be used instead. The features examined for each author, are related to his/her publications, co-authorships and citations, as well as their evolution in time. We combine these information in a series of metrics, some are very popular and other are novel, which in tandem constitute metrics of the significance of a researcher's work. We consequently group researchers based on these metrics.

1.2 TECHNIQUES

The methodologies that we employed into representing and extracting knowledge from our data, are well established ways that are commonly used in knowledge extraction and data mining tasks. The use of graphs is ubiquitous, in many forms such as multigraph, weighted and power graphs, from which we extract certain measures of author importance. In addition bibliographical indices such as papers per year and citations are commonly used, originals and modified to include time penalizations. This thesis is characterized by the essence of time and how it reacts with the characteristics and the indices of an author.

Our analysis consists of these steps:

1. Data collection

   It is bibliographical information retrieval, based on the digital library Scopus. The material concerns works from Greek institutions and the respective authors. The basic methodology in this stage is combining structured, given data with web crawling through the pages of the digital library, and storing information about the targeted papers and authors.
2. Data pre-processing

   The data was distinguished in years, then for each year two types of co-authorship graphs were constructed, to capture an author's social "power" and a co-authorship's impact. Every yearly graph contained information of the respective year and of the previous years'

graphs , modified appropriately to achieve penalization by oldness in the edges' weights. The information at the edges' weights of each graph enclosed the citations, number of co-authors and paper's oldness on the one hand, the number of papers on the other. From these graphs and our dataset, we extracted a number of features for each year to capture the instance of the author's profile at that year. The features discussed various forms of the citations, the papers, the impact of an authors collaborations, his significance based on his position in the graph, his co-authors' and extended community's strength etc. To calculate these we applied certain graph mining methods like power graph and eigenvector analysis. We then use these features to create change indices and depict the evolution of the author's features in time.

3. Knowledge extraction

   The authors' categorization is the desirable outcome of this analysis. It was achieved using the dataset with the change indices and the K-means clustering algorithm. The right number of clusters was determined by running experiments with the algorithm and measuring a set of clustering validity indices, like Dunn index (Dunn, 1973), Davies-Bouldin (Davies & Bouldin, 1979), average within group sum of squares and average distance between cluster centroids. The tagging of the clusters was based on prominent values of their features. In addition a feature selection process was conducted using singular value decomposition in an aggregated form of the clustered dataset, in order to define the most impactful on clustering characteristics. Finally, a dataset with the authors' time series of the most impactful feature was build, to manage a time series clustering using the dynamic time warping measure and partitioning around medoids. The results were evaluated in respect to the first clustering and a rate of success was recorded.

In summary, the main contributions of this dissertation are:

- Construction of Greek affiliation authors dataset using sophisticated methods. For example part of this phase demanded the identification of unidentified authors. The identified co-authors of an unidentified author, allowed us to construct an id resolution mechanism, which tallies a candidate id for an unidentified author, based on its collaboration frequency with the identified co-authors.
- The oldness factor of each collaboration, and its weights analogically reduction. The oldness factor depicted in bibliographical measures (e.g. citations retrieved penalized by year). Change indices on bibliographical measures.
- Evaluation between time series clustering and simple clustering of change indices.

1.3 DISSERTATION STRUCTURE

The thesis is organized as follows: Section 2 presents some rudimentary concepts that the reader should be familiarized with, in order to fully comprehend the stages of the analysis. Section 3 presents the theoretic aspect of the methods that we used, some of them being already renowned techniques in the field of informatics while others are newer concepts and the rest are ideas that we propose. Section 4 describes the methods in a more practical manner, with respect to the implementation details. Section 5 bestows some examples in order to better understand some of the thesis' critical notions. Section 6 winds up the research and gives pointers to future work.

2. BACKGROUND

This section provides some fundamental information on the technologies and procedures which we employed and is necessary, in order to understand the remaining of the thesis and the analysis we followed.

2.1 SCOPUS

Our analysis focuses on certain Greek universities (see Appendix 7.1) research activity of the last fifteen years. In our attempt to ensure that our work's results will be as reliable as possible, we gathered our data from one of the most eminent sources of bibliographical data, the digital library Scopus.

2.1.1 ABOUT

Scopus is a bibliographic database, which contains containing titles, authors and citations for academic journal articles. It covers approximately 21,000 titles (journals and conferences) from over 5,000 publishers, concerning scientific, technical, medical, and social sciences. It is owned by Elsevier, which is an esteemed publishing company, and is available online by subscription. Scopus search incorporates searches of scientific web pages through Scirus, another Elsevier product (Kulkarni, Aziz, Shams, & Busse, 2009). Moreover, Scopus offers author profiles, containing affiliations, number of publications and their bibliographic data such us references, the number of citations each published document has received as well as entrenched analytics to present a general picture of the author’s career. It has registration advantages, like observing an author’s change in time and calculating the respective h-index. In comparison with other similar libraries, Scopus offers 20% more coverage than Web of Science, covers a wider range of journals than PubMed and has more consistent results than Google scholar, which provides more inadequate, less often updated citation information. (Falagas, Pitsouni, Malietzis, & Pappas, 2008) (Erten, Harding, Kobourov, Wampler, & Yee, 2004)

2.1.2 FACILITIES AND LIMITATIONS

Scopus provides publication data using various methods (e.g. API, structured file, browsing). The limit for downloading article information in a single csv file is 20000 publications. One can download multiple csv files, if for example downloads one file per institution. The features of every publication are the year that it was published, the name of the authors (without ids), the ISSN

and the title of the journal, the number of pages, the total citations from publication since 1997, the citations per year from 1998 onwards. Though large in volume, this dataset lacked in a vital substance, an author identifier, rendering the information deficient. That was due to the fact that in case of synonymies we were exposed into mixing information of two or more different authors into one.

2.2 CO-AUTHORSHIP GRAPHS

Collaboration graphs (Odda, 1979) are widely used in the field of mathematics, social sciences and informatics, mainly in the domain of social network analysis. Collaboration graph is a graph structure, where the vertices are the persons who collaborate with each other and an edge connecting them shows a collaborative relationship between them. In co-authorship graphs the relationship depicted by the edge can be co-writing of a paper or of many papers. In general, someone can argue that the use of graphs is quite versatile and may serve into capturing many information of a network between authors.

2.2.1 EDGE WEIGHTS

The weight of the edge can represent various things as well. The citations that the paper has gotten, or the amount of papers that the two authors have co-authored, are a few apparent examples. The weight of the edge is usually proportional to the nature of the graph, meaning that a simple graph with citations of a paper in the edge, will turn into a multigraph **,**when the two authors write a new paper , since a new edge between them will be added. On the other hand, if we keep a single edge for each co-author pair, then we either loose the information of the second publication, or we modify the weight of the edge into e.g. the sum of citations that the two co-authors have gotten. An alternative way used in many cases is a hyper graph (O'Madadhain, Hutchins, & Smyth, 2005). A graph containing the same vertices as in the aforementioned cases but with hyper edges, can show a co-authorship of a paper with simultaneous edges between its authors.

2.2.2 EDGE DIRECTION

Generally edges can be directed or undirected. When referring to co-authorship, they are usually undirected due to their mutuality and because someone can not strictly define the direction of the

edge. Directed edges in the graph, can depict the citations that an author has made to papers of another author. This pattern is more recognized in another popular bibliographic graph, a citation network (An, Janssen, & Milios, 2004). In this case the vertices can depict papers and the edges depict the references that a paper has to another paper. It can help reveal author homogeneity, similarity between papers (Martyn, 1964) etc.

2.2.3 GRAPH-BASED INDICES

Typical indices for bibliographic graphs have been borrowed from social network analysis and graph mining, such as normalized degree (Borgatti & Everett, 1999) , betweenness (Freeman, 1977)etc. Recently more special formulas have been proposed for specific use in bibliographical graphs. Eigenfactor (West & Wiseman, The Eigenfactor Metrics, 2008) the importance of an article or a journal based on the amount and rate of citations it receives from influential articles or journals. Though proposed initially for citation networks, it has been applied to authors too (West, Jensen, Dandrea, Gordon, & Bergstrom, 2013) with the same philosophy.

2.3 POWER GRAPHS

Power graph analysis is mainly used in the field of bioinformatics to visualize complex networks without losing information (Royer, Reimann, Andreopoulos, & Schroeder, 2008). This methodology has been successfully applied to co-authorship networks in the past, (Tsatsaronis, et al., 2011) (Varlamis & Tsatsaronis, 2012).

2.3.1 DEFINITION

Power graph identifies basic motifs such us star, clique and biclique in the graph and uses them to construct a compact depiction of the graph. This is achieved using power nodes, which is a circle enclosing nodes or power nodes, and power edges, the edges between power nodes. The transformation of the motifs into power nodes is depicted at Image 1 and follows this methodology:

- Bicliques are two sets of nodes with an edge connecting every node in the one set with every node in the other. In a power graph, a biclique is depicted as two power nodes consisting of the nodes in the initial two sets, and a power edge between them.
- Cliques are a set of nodes with an edge from every node to every other node. In a power graph, a clique is represented by a power node containing all the authors that connect to each other, with a loop.

- Stars are one set of nodes and one other node, with an edge between every node in the set and the distinct node. In a power graph, a star is represented by a power edge between a regular node and a power node that has all the nodes to which the regular node is connected to in the regular graph.

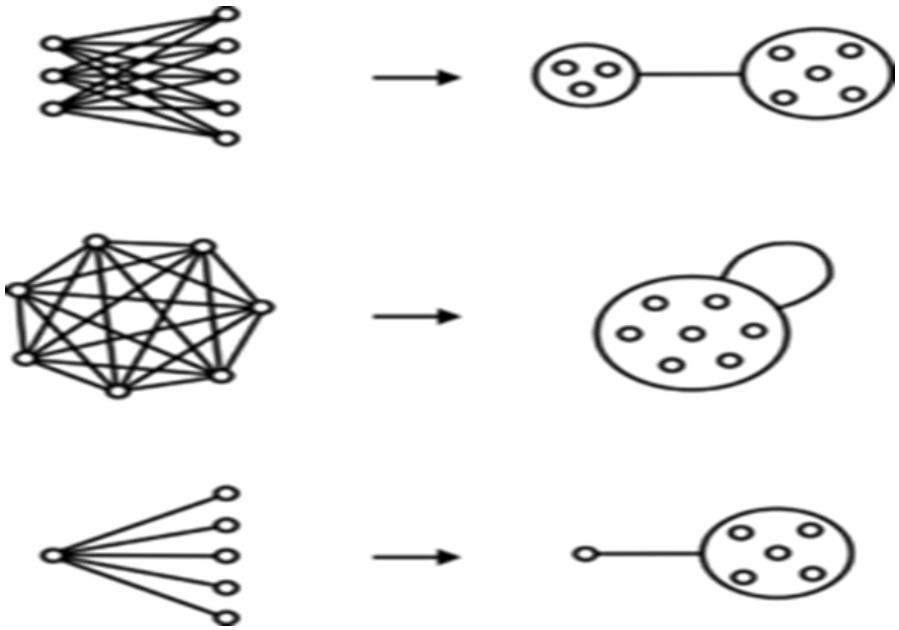


Image 1: GRAPH MOTIFS IN POWER GRAPH REPRESENTATION

2.4 CLUSTERING

Cluster analysis is the procedure in which a group of objects is arranged in smaller groups, based on the similarity of the objects inside each sub group and their dissimilarity with the objects in the rest of the groups. These groups are called clusters. It is a common technique used in data mining, machine learning, statistical data analysis etc.

2.4.1 CLUSTERING ALGORITHMS

Clustering algorithms can be classified according to their cluster models.

- Centroid models: Centroid based models depict the clusters as mean vectors, which may not be an object in the data set clustered e.g. K-means (MacQueen, 1967).

- Connectivity models: Models using connectivity, create groups based on the distance connectivity of the observations e.g. Hierarchical (Johnson, 1967).
- Distribution models: Clusters are defined exploiting statistical distributions of the data. (Xiaofei, Deng, Yuanlong, Hujun, & Jiawei, 2011)
- Density models: Creates the clusters based on the density of the observation values in the data space e.g. DBSCAN (Martin, Kriegel, Sander, & Xu, 1996).

2.4.2 CLUSTER VALIDITY

Cluster validity is a term referring to the degree that the clusters produced by the clustering procedure, are indeed an existing structure in the initial dataset, hence the clustering was successful. The validity metric chosen in many cases depend on the analysis' data and intentions (Halkidi, Batistakis, & Vazirgiannis, 2001). This means that the metrics used for a labelled dataset clustering (external), is different to the ones used for a non-labelled (internal). Similarly the clustering evaluation in cases of algorithm efficiency comparison, is different from the cases where the analysis aims at adequate clusters. Generally internal cluster validity measures traverse the clusters in the data space and measure the values of formulas based on the distances of the observation, the centroids, the medoids, the structure, the size or the density of the clusters. E.g. Silhouette (Rousseeuw, 1987) .External measures, on the other hand, use data labelled with the right clusters (supervised method). The clustering algorithm runs and the results are evaluated based on how close they are to the predetermined labels e.g. Rand Measure (Rand, 1971) .

3. METHODS

Specific methods that we embodied, modified or created during our research.

3.1 DEFINITIONS

The metrics that we formulated in the rest of the Methods stage, contain specific measures that should be clarified:

- $P_i = Set\ of\ author's\ i\ papers.$
- $t_n = The\ year\ we\ examine.$
- $t_0 = Year\ 1998, the\ year\ were\ our\ dataset\ starts.$
- $tz = Year\ 2005, the\ year\ were\ our\ dataset\ ends.$
- $pap_i(j) = Number\ of\ papers\ the\ author\ i\ had\ writen\ during\ year\ j.$
- $cit(i,j) = Citations\ that\ paper\ i\ recieved\ at\ year\ j.$
- $aut(i) = Number\ of\ authors\ that\ co-wrote\ paper\ i.$
- $year(i) = Publication\ year\ of\ paper\ i.$
- $N_x = Set\ of\ nodes\ neighbor\ to\ node\ x.$
- $E(i,x) = The\ between\ x\ and\ i\ edge's\ weight\ in\ the\ quality\ graph$
- $PN_x = Set\ of\ power\ nodes\ neighbor\ to\ power\ node\ x.$
- $PC_x = Set\ of\ power\ nodes\ containing\ power\ node\ x.$
- $PW(x) = Weight\ of\ power\ node\ x.$
- $PE(y,x) = Weight\ of\ power\ edge\ connecting\ power\ node\ x\ and\ power\ node\ y.$
- $Cl = Set\ of\ clusters.$
- $c_i = Cluster\ i'scentroid.$
- $d(i,j) = Distance\ between\ point\ i\ and\ point\ j\ in\ data\ space.$
- $\lambda =$ $Eigenvector\ centrality\ constant, which\ depends\ on\ the\ choice\ of\ normalization.$

3.2 GRAPHS

The use of collaborative graphs enabled us to represent author information in various dimensions, like time resistance and social significance. For this purpose we took advantage of the flexibility the graphs provide into the properties of edge weights, as well as the fact that the graphs expanded as the years passed, producing important clues about an author's activities.

3.2.1 GRAPH EVOLUTION

The co-authorship graphs were evolutionary. This means that each year did not just contain the collaborations established in that year, but accumulates all collaborations **until** that year. Hence, if two authors have co-authored one paper in 1999 and two in 2000, the graph of 2000 will contain the paper of 1999 and the two papers that they have co-authored in 2000. Same thing applies for the citation a collaboration had gotten. The difference in this case lies in the citations the paper is depicted to have taken in each year. The citations are cumulative, meaning that as the time passes, the references made towards a paper from other papers can only increase or stay stable, because references cannot be erased. The graphs created for each year depict the impact of the paper in a given time, thus it is essential to take into consideration the citations gotten until that year, not just the ones made during that year. For example, a paper written in 1999 will be represented in the 1999- graph with the citation that the paper got at 1999 . In the 2000-graph, the citations taken into account will be the ones made until 2000, which means the citations of 1999 **and** the citations of 2000. Same applies for the rest of the years.

3.2.2 EDGE WEIGHTS

While the graph structure captures the social impact of an author, the weight of the edges contained another vital substance in our analysis, the actual impact of a collaboration. A co-authorship is mapped to an edge between the two authors in the yearly graphs. This implies that this edge should enclose the information of all the papers the two authors have co-authored in until the year of the yearly graph. This knowledge aggregation is of crucial matter, since it must take into account every aspect of each of the papers' impact. Best suited to the research's needs, were a pair of graphs with different edge weights, so as to capture as many dimensions of the problem as possible.

3.2.2.1 QUANTITY EDGE WEIGHT

The edge weight of the quantity graph represents the co-authorship volume (CV) of two authors x, y is the amount of times author x and author y have co-authored in a paper until time t_n (the number of papers).

$$CV(x,y) = \sum_{\forall i \in P_x \cap P_y} 1 \qquad , year(i) <= t_n \tag{1}$$

3.2.2.2 QUALITY EDGE WEIGHT

The edge weight of the quality graph represents the co-authorship impact (CI) of two authors x, y is the aggregated impact of a list of papers co-written by the two authors and it is defined as:

$$CI(x,y) = \sum_{\forall i \in P_x \cap P_y} \frac{\alpha * \sum_{j=t_0}^{t_n} cit(i,j) + \beta}{aut(i) * \left(1 + t_n - year(i)\right)} \qquad ,year(i) <= t_n \qquad (2)$$

The essence of this model is that a paper's impact in a given time $t_n$is analogous to the citation it has had until this time, reverse analogous to the number of co-authors that took part in writing it and to the year it was published minus t_n, which depicts how old is the paper at time t_n. The +1 at the denominator covers the case where $t_n = year(i)$ (when the paper is written in the current year t_n). For every two authors , the sum of all of their paper's impact, is the impact of their coauthorship at a given time t_n. The choice of values α and β, denotes the interest on the impact of an author's work (α) or on the quantity of his/her publications (β). In our experiments, we decide to set $\alpha = 0{,}7$ and $\beta = 0{,}3$ but this definitely needs further experimental justification.

3.2.3 GRAPH FEATURES

During the process of composing datasets with author features, to employ them during clustering, we used graph mining techniques to capture the social impact of the author. These techniques, coming from social network analysis and bioinformatics, resulted in particular indices describing knowledge for each author, which were later set as author characteristics. These features capture every aspect that denotes the strength the author holds in the network. Due to the time penalization of the edge weights, an author's co-authorship's impact is measured in a quite accurate way. In addition, the power of the author's co-authorship group as well as his extended community (the co-authors' co-authors) are included, from both the quantity and the quality perspective. Obvious measures such as the number of co-authors and the position of the author in the network are also taken into consideration.

3.2.3.1 FEATURES IN QUALITY GRAPH

The graph containing the quality essence of a collaboration was chosen to extract author characteristics from, mainly because the impact of the collaboration is best depicted in the quality

aspect. Indices that have to do with the structure of the graph, are the same in both the quality and the quantity graph.

3.2.3.1.1 EIGENVECTOR CENTRALITY

The eigenvector centrality (Bonacich, Factoring and weighting approaches to status scores and clique identification, 1972) (Bonacich, Some unique properties of eigenvector centrality, 2007) of an author x ($Eigen(x)$)corresponds to the impact the author had based on his position in the co-authorships graph at a given time. Based on the fact that an edge with an eminent author is more important to the score in question than an edge with a not so successful author, eigenvector centrality calculation assigns initial equal values to every author and then recalculates them based on the connections each node has. This iterative process ends when the values converge.

$$Eigen(x) = \frac{1}{\lambda}\sum_{t\in N_x} Eigen(t) \quad (3)$$

3.2.3.1.2 DEGREE

Degree (Borgatti & Everett, 1999) of an author x ($Deg(x)$)is the amount of edges he has in a given time, hence the amount of co-authors he has.

$$Deg(x) = |N_x| \quad (4)$$

3.2.3.1.3 COLLABORATIVE WEIGHT

An author's x collaborative weight (CLW) is the sum of the author's edges' weights. The interpretation of this is the general quality of the author's collaborations in a given time.

$$CLW(x) = \sum_{\forall i\in N_x} E(i,x) \quad (5)$$

3.2.3.2 POWER GRAPH

Our graphs were particularly dense. Power graph analysis is specialized in extracting knowledge from dense graphs, mainly in bioinformatics. We applied this methodology in order to extract some extra information about our authors. Both graphs, quantity and quality, were transformed and mined. Hence, there were created pairs of the same feature, each corresponding to the value in the quality and the quantity power graph respectively. The information that we can extract from a power graph, regards to an author's collaboration with strong individuals or groups and it is

analogical to the weight of the collaboration itself. We can also get a feel of the extended co-authorship society that the author belongs to. Thus we can expect increased features on authors who belong to an eminent co-authorship group or are part of a general successful scientific society, like an eminent institution.

3.2.3.2.1 POWER NODE WEIGHT

The power node weight (PN_{weight}) that the author x belongs to, in the quality and quantity power node. Which stands for the authors close community impact in the graph.

$$PN_{weight}(x) = PW(x) \tag{6}$$

3.2.3.2.2 POWERCLIQUE WEIGHT

The power node clique weight (PN_{clique}) that the authors belonged to, in the quality and the quantity power node. Which stands for the weight of the extended community of the author, meaning the co-authors that his co-authors have and their bonds. For each power node , its clique weight is defined as

$$PN_{clique}(x) = \sum_{\forall i \in PN_x} PE(i,x) * \mathrm{P}W(i) + \sum_{\forall j \in PC_x} PW(j) \tag{7}$$

3.3 TIME EFFECT

As mentioned above time plays an important role in our analysis. Each section of our knowledge extraction process has a time related aspect. In every case, we tried to interpret in the best way the dimension of time, either with conservative methods, or with novel approaches. From our perspective, as the age of an occurrence increases, its impact becomes weaker. That derives from the fact that as time passes more innovative ideas surpass the older ones, leaving the latter rest in history. Of course in science every opinion is useful and can be reexamined, that is why we apply our theory in citations too, this way the really successful old works stay on top. It is important to clarify that this theory is exploited in categorizing authors based on their work, not the ideas themselves. The novelty lies in the time penalization method, which is not so common at bibliographical research. The common approach for oldness is an N-year index, in which case the incidents in N last years are taken into account. A more sophisticated technique can reveal

information like an author's dynamic or his endurance in time, that could not be clearly derived with the conservative perceptions.

3.3.1 TIME PENALIZED FEATURES

The yearly datasets consisted of author characteristics in every year. A part of these characteristics consisted of graph related metrics while the rest had to do with the independent measurements of an author. These values depicted the author success as an individual, without taking into consideration the social parameter, similar to the well-known productivity measure h-index (Hirsch, 2005) .

3.3.1.1 SUM OF PAPERS

The papers sum (P_{sum}) that an author x had written until a given time t_n.

$$P_{sum}(x) = \sum_{i=t_0}^{t_n} pap_x(i) \qquad (8)$$

3.3.1.2 PAPERS PENALIZED BY OLDNESS

The amount of papers an author x had written until a given time t_0 penalized by oldness ($P_{penalized}$). The formula used for this is

$$P_{penalized}(x) = \sum_{i=t_0}^{t_n} \frac{pap_x(i)}{(1 + t_n - i)} \qquad (9)$$

The intuition behind this, is that the amount of papers an author has written is more important when he wrote them closer to the current time and not the distant past.

3.3.1.3 CURRENT PAPERS

The amount of papers an author x wrote at time t_n, showing how active is the author at current time (P_now).

$$P_{now(x)} = pap(t_n) \qquad (10)$$

3.3.1.4 SUM OF CITATIONS

The sum of the citations (Cit_sum) that the author x's papers had gotten until time t_n.

$$Cit_{sum}(x) = \sum_{i=t_0}^{t_n} \sum_{\forall j \in P_x} cit(i,j) \quad (11)$$

3.3.1.5 CITATIONS PENALIZED BY OLDNESS

The sum of the citations that the author x's papers had received, penalized by oldness ($Cit_{penalized}$):

$$Cit_{penalized}(x) = \sum_{i=t_0}^{t_n} \sum_{\forall j \in P_x} \frac{cit(i,j)}{(1+i-year(j))} \quad (12)$$

This can be better understood with an example. If we are examining time 2004 and an author has written an article at 2001, gathering these citations: 2 at 2001, 20 at 2002, 30 at 2003 and 15 at 2004, this article has 2/(1+2004-2001) + 20/(1+2004-2002)+30/(1+2004-2003) +15/1=37,16 citations. Now if we examine the article at 2005, and the citations have become: 2 at 2001, 20 at 2002, 30 at 2003, 15 at 2004 and 3 at 2005, this article has 2/(1+2005-2001) + 20/(1+2005-2002)+30/(1+2005-2003) +15/(1+2005-2004)+3/1=25,9 citations

The essence of this is that the citations an article has had, have more impact at time t_n as they had been earned closer to it, because an old citation may not be valid by the current time. It makes sense that the second case has less impact than at the first case because in the first case, the paper has obtained more citations closer to the time t_n, which means that it is up to date in contrast to the second case where it has gotten 3 citations at its time, showing that it is not so current anymore. This is an intuitive way to capture the meaning of temporal impact and observe its progression in time.

3.3.1.6 CURRENT CITATIONS

The sum of the citations that the authors' papers got at time t_n, showing how much impact do the authors paper have at current time (Cit_{now}).

$$Cit_{now}(x) = \sum_{\forall j \in P_x} cit(t_n,j) \quad (13)$$

3.3.2 CHANGE INDICES

To include the change of the aforementioned features, the dataset that we used to cluster the authors combined the yearly datasets, exploiting a change index of 5 measures, for each feature, to capture their evolution and the nature of each author's features changes through the years. The change index contained 4 change metrics and one metric represented the overall level of the feature. In the formulas below, $f(i)$ is the value of an examined feature in year i.

3.3.2.1 MIN & MAX FEATURE CHANGE

The maximum and minimum change that the feature has undergone ($minC$ & $maxC$), showing the peak and the lowest deviation the feature have had.

$$\text{minC} = \min\big(f(i) - f(i-1)\big) \tag{14}$$

$$maxC = \max\big(f(i) - f(i-1)\big) \tag{15}$$

$$i \in [t_0, t_z]$$

3.3.2.2 LAST FEATURE CHANGE

The last change of a feature ($lastC$) to depict the author's dynamic in this dimension at that time.

$$lastC = f(t_z) - f(t_z - 1) \tag{16}$$

3.3.2.3 SUM OF FEATURE CHANGES

The sum of the features changes ($sumC$), representing its stability and the nature of its rate (positive-negative).

$$sumC = \sum_{j=t_0}^{t_z} f(i) - f(i-1) \tag{17}$$

3.3.2.4 FEATURE VALUE

The aforementioned measures depict the characteristics of the change that the feature has undergone. In order to determine the value of the feature through time we create a last index Feature Value ($featVal$), which varies depending on the nature of the feature:

- Penalized Features (papers, citations, normalized, graph features): The last value. Since these features are decreasing based on the years by definition, the last value depicts the overall value of the author's characteristic.
- Cumulative features (papers, citations until time N): The last value divided by the number of years the author has occurred in. It represents the average value.
- Temporal features (papers, citations at current time): The sum of values in every year, divided by the number of year an author has occurred in, summed with its trend. The trend of a feature is measured as the gradient (inverse tangent of slope, in radians) of the linear regression fitted in the feature's values in time.

3.4 FEATURE TABLE

The following table summarizes the features that we implemented and have defined until now.

Table 1: AUTHOR FEATURE TABLE

Name	Origin	Explanation	Method Pointer
P_{sum}	Database	The amount of papers an author had written until the year of the dataset.	3.2.1.1
$P_{penalized}$	Database	The amount of papers an author had written penalized by oldness.	3.2.1.2
P_{now}	Database	The amount of papers an author wrote at the year of the dataset.	3.2.1.3
Cit_{sum}	Database	The sum of the citations that the authors' papers had gotten until the year of the dataset.	3.2.1.4

$Cit_{penalized}$	Database	The sum of the citations that the authors' papers had gotten , penalized by oldness.	3.2.1.5
Cit_{now}	Database	The sum of the citations that the authors' papers got at the year of the dataset.	3.2.1.6
$Eigen$	Quality Graph	The eigenvalue of the author in the quality graph of the dataset's year.	3.1.3.1.1
Deg	Quality Graph	The degree of each author in the quality graph of the dataset's year.	3.1.3.1.2
CLW	Quality Graph	The sum of the author's edges weight in the quality graph of the dataset's year.	3.1.3.1.3
WPN_{weight}	Quality Power graph	The weight of the power node that the author belonged to, in the quality power graph of the dataset's year.	3.1.3.2.1
WPN_{clique}	Quality Power graph	The clique weight of the power node that the authors belonged to, in the quality power graph of the dataset's year.	3.1.3.2.2

SPN_{weight}	Quantity Power graph	The weight of the power node that the author belonged to, in the quantity power graph of the dataset's year.	3.1.3.2.1
SPN_{clique}	Quantity Power graph	The clique weight of the power node that the authors belonged to, in the quantity power graph of the dataset's year.	3.1.3.2.2

3.5 AUTHOR CLUSTERING

The categorization of the authors is indisputably a clustering task. To determine the optimal number of clusters in the evolution dataset, we run a series of experiments using clustering validity indices and then we use the result to run a clustering algorithm. The clustering aims at identifying specific patterns in the change rate or the average values of the author's characteristics. It then proceeds into discriminating the authors based on these patterns and their feature dissimilarity. In addition an aggregated dataset is constructed, with more meaningful features and a feature selection process acknowledges the impact each of them would have in clustering procedure. Then, it uses the top feature to run a time series clustering based on its time lines. We than compare the similarity of the two clustering results and evaluate the success rate of the latter. Finally we use the results from the first clustering to reveal the features each cluster stands out at, in order to characterize it, classifying the authors in it. The workflow in image 2 depicts the line we followed.

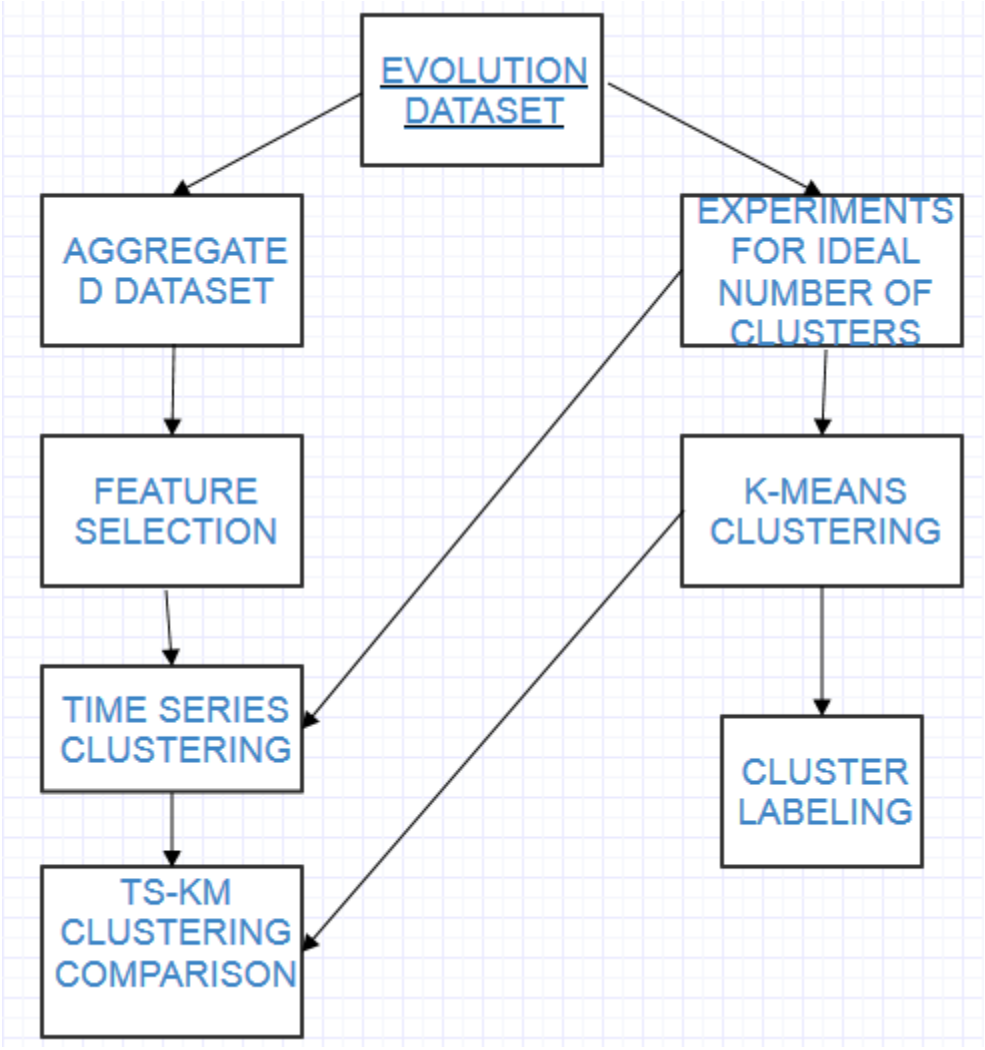


Image 2: CLUSTERING STAGE FLOWCHART

3.5.1 CLUSTER VALIDITY

The evaluation measures are used to determine the efficiency of the clustering. Since the clustering is unsupervised (without labels), indices from internal validation were applied.

3.5.1.1 AVERAGE WITHIN CLUSTER SUM OF SQUARES

It is the sum of the average sum of squares of every cluster divided by the number of clusters. The average sum of squares is defined as the sum of distances between each point of the cluster and the cluster centroid squared , divided by the number of points. Its meaning regards to the average inconsistency of the clusters, hence the less the better.

$$avg(WCSS) = \frac{\sum_{\forall j \in Cl} \frac{\sum_{\forall x \in j} d(x,\ c_j)^2}{|j|}}{|Cl|} \quad (18)$$

3.5.1.2 AVERAGE DISTANCE BETWEEN CLUSTER CENTROIDS

The average intercluster distance is the average of the pairwise distances between every cluster centroid ($avg(DBCC)$). It represents the average distance between the clusters and since we want the cluster to be as distinct as possible, this index should be as high as possible.

$$avg(DBCC) = \frac{\sum_{i \in Cl, j \in Cl, i \neq j} d(c_i, c_j)}{Cl * (Cl-1)/2} \quad (19)$$

3.5.1.3 DAVIES-BOULDIN

This index is formulated as:

$$DB = \frac{1}{n} \sum_{i=1}^{n} max_{i \neq j} \left(\frac{\frac{\sum_{\forall x \in i} d(x,\ c_i)^2}{|i|} + \frac{\sum_{\forall y \in j} d(y,\ c_j)^2}{|j|}}{d(c_i, c_j)} \right) \quad (20)$$

The intuition of this model is that the desired algorithm should produce clusters with low intra cluster (numerator) and high inter cluster similarity (denominator), which means that the smallest the Davies-Bouldin the better. We want small intra cluster distances to capture the fact that the points of a group should be close to the respective centroid. On the other hand we need high between clusters distance, in order for the groups to be as more distinct as possible.

3.5.1.4 DUNN

The model is defined us:

$$DU = min_{1 \leq i \leq n} \left\{ min_{1 \leq j \leq n, i \neq j} \left\{ \frac{d(c_i, c_j)}{max_{1 \leq k \leq n} \frac{\sum_{\forall p \in k} (x_p - c_k)^2}{|k|}} \right\} \right\} \quad (21)$$

Its aim is to identify dense and well separated clusters. It is based in minimum intercluster and maximum intracluster distance. The minimum intercluster distance is measured as the minimum

distance between two given cluster centroids in any of the features. The maximum intracluster distance is defined as the maximum within group sum of squares in any cluster. Having the same essence of Davies Bouldin, but reversed, we want it to be as high as possible.

3.5.1 K-MEANS ALGORITHM

The algorithm used for our project is k-means. K-means algorithm is an iterative partitioning approach based on a given distance and a given number of clusters K. It starts with K random centroids and then uses the given distance and the values of each observation to assign the observation to the closest centroid. It thereafter re calculates the centroids for each group, using the distances between the values of each observation in that group. Than each observation is re assigned to the closest of the new centroids. The iteration goes on until the centroids are stabilized, meaning that they do not change through the iterations. The distance metric used was Euclidian distance, which is recommended for continuous values.

3.5.2 FEATURE SELECTION

The feature selection process is a rudimentary mechanism in data mining and analytical statistics to produce a subset of the initial data set, highlighting some features that are of greater importance or influence to the model you want to use or to the analysis you want to do. Many algorithms and formulas are recommended in, proportional to the purpose and the nature of the analysis. A feature selection algorithm usually includes search for feature subsets and evaluation indices for each subset. When the model is supervised, most suitable are measures such us Pearson's correlations coefficient (Pearson, 1895) and mutual information (Manning, Raghavan, & Schutze, 2008) in relation to the features and the class. In case of unsupervised learning the selection is based on the feature values characteristics, like their deviation or their correlation among them.

3.5.2.1 DATA COMPRESSION

For this part, we tried to create a more compact dataset, with more intuitive columns then the ones referring to the change of the characteristics (3.2.2). We achieved that by aggregating these change features, in favor of revealing the most impactful of the initial 13 author characteristics stated at 3.1.3 and 3.2.1 .We chose these features as they can become easier to interpret, in contrast to the change features which are too complicated and overanalyzed for such a task. The aggregation to one feature was achieved by grouping the change features derived from the same feature and its

changes. For example the 5 features originated from the initial feature P_{now}(3.3.1.3) will be aggregated like this:

$$aggrInd = \left(minC(P_{now}) + maxC(P_{now}) + lastC(P_{now}) + sumC(P_{now})\right) * featVal(P_{now}) \quad (22)$$

This formula can be explained by taking into account that all the change features are equally important to the value that represents the feature. In this way we split equally the weight of the feature representation through time into the change factor and the actual value factor.

3.5.2.2 SINGULAR VALUE DECOMPOSITION

Singular value decomposition is an efficient statistical technique to decompose a dataset with matrix decomposition, find the singular values explaining most of the variance and then retransform it back with only these values, achieving dimension reduction. The decomposition of a dataset X looks like this:

$$X = U * D * V^T \quad (23)$$

Where the columns of the right singular vector V are orthogonal to the columns of the left singular vector U and D is a diagonal matrix with the singular values. Each singular value in the D explains a percentage of variance in the dataset. Each one of the right singular vectors, shows which columns contribute to the variance of the corresponding singular values. In this way, we can extract columns that contribute the most to the variance of the dataset, hence to the clustering too. This can be achieved by keeping the singular values that explain a large amount of variation and then finding the values of each column in the corresponding right singular vectors. Summing the right singular vector values, for each "strong" singular value, results in a vector with one value for each column, depicting the amount of contribution this feature has to the most of the dataset's variance. Keeping the ones over a certain threshold, will reveal a set of characteristics that are important into the clustering process of an author.

3.5.3 TIME SERIES CLUSTERING

Having a time related set of datasets, we can represent each of the features as vectors in time. An author's vector, shows the progression of the respective characteristic through the time examined. For one feature, this results in a new dataset of a vector per author, with number of columns proportional to the amount of time units examined. We can then cluster the authors with a common

algorithm like partitioning around medoids (Kaufman & Rousseeuw, 1987) , based on their vector similarity, using dynamic time warping (DTW) (Berndt & Clifford, 1994) as a distance measure, instead of Euclidean distance or other conventional continuous distance measures.

3.5.3.1 DYNAMIC TIME WARPING

DTW is an algorithm for measuring similarity between two temporal sequences, regardless of time or speed difference. This means that a time series can be compared with another time series even if their lengths differ. This makes it perfect for our case, since we want to cluster authors that appear in different time points throughout the time span that we examine, thus creating vectors of feature values with various lengths. To find the similarity between the two feature vectors from two different authors, DTW warps the two sequences in the time dimension to determine their similarity separated from non-linear variations in the time dimension.

3.5.3.2 PARTITIONING AROUND MEDOIDS

PAM is a clustering approach similar to the k-means algorithm in 3.3.1. Both of them aim at diminishing the distance between points in the same cluster breaking the dataset into group. Both of them are iterative and recomputed their clusters in every iteration until the points in the clusters converge. Their main difference regards to the fact that PAM uses existing data points as cluster centers, in contrast to k-means which uses points representing the mean values of the points in the cluster. In addition, PAM works with an arbitrary matrix of distances between observations, which was the main reason why we chose it to work with the DTW distance.

3.5.4 CLUSTER LABELING

The clusters that are produced from a clustering procedure, are formed due to certain patterns in the data. Namely in our case, the authors are categorized based on their features' behavior. This has as a result certain clusters to show special values in some of the features, in contrast to the rest. This property was exploited in our analysis, to name each cluster based on specific pattern of its features. The value of a cluster's feature, correspond to the centroid of the cluster in that feature.

4. IMPLEMENTATION

The application of the methods, explanation of the code and apposition of the technologies used.

4.1 CRAWLING

The information in the csv files provided by Scopus was deficient, as it lacked a way to identify authors. That was due to the fact that our paper dataset contained only names of authors, not identifiers , thus in case of synonymies we were exposed into mixing information of two or more different authors into one. We attempted to cover this gap using a combination of the given csvs and a sequence of web crawling techniques.

4.1.1 CRAWLING THE PAPERS OF THE INSTITUTIONS

We first attempted to gather as much papers as possible, with identified authors. For that aim, we crawled the pages of the papers that belonged to each of our institutions in descending order based on the paper's citation. The reason behind this order is that we intended to gather the papers with the biggest impact, as Scopus limited us into visiting only 2000 papers for each institution. The number of the authors in papers depicted in that page is up to 10. In case of authors being more then 10, the author list has a symbol (…) to express that there are more authors for that paper. The crawler used the java libraries httpclient[1] to fetch the page and htmlcleaner[2] to clean the html code. We then proceeded into obtaining the title of each paper, its corresponding authors, their ids based on the link in the author's name which led to authors' Scopus page, using xpath (Harth, Umbrich, & Decker, 2006)on the clean html code. In addition to this, we kept an auto increment id in our dataset for every paper, and a 'more' field to determine whether it has more authors then shown or not. At this point we had two datasets. The dataset Scopus gave us with 134.567 papers with unidentified authors and the crawled dataset with 32.352 papers and a percentage of authors identified.

4.1.2 AUTHOR PAGE CRAWLING AND ID MATCHING

[1] http://hc.apache.org/httpcomponents-client-ga/index.html

[2] http://htmlcleaner.sourceforge.net/index.php

At this point we needed to tally the authors in papers in the crawled dataset, to those of the corresponding paper in the given dataset, and secondly store information for each author distinctively. For start we tried to identify the authors in papers in the given dataset, that we encountered during the 4.1.1 process and had no 'more' authors then shown in the web page, so every author in this paper was identified. This was done, marking each paper in the given dataset with the identifier of the same paper (based on year and title) in the crawled dataset. After that, for each paper we tallied the authors in the former dataset with the authors in the latter, together with their ids. During this process we had the name and the author's id, and we used them to store information about him/her, using two crawlers. The first one used the author's id to fetch the authors page in the same manner as in the 4.1.1 and used xpath to draw his/hers full name, affiliation, city-country and field of expertise. The second crawler would emerge when the first one could not find the page. It exploited the Scopus search for the author's name and searched the result list. The target is the author that has the same id in his link, as the id of the author that we initially tried to crawl. It keeps the same information for the author as the former crawler. The authors that a crawled paper with 'more' lacked, could be found in the given dataset, but not identified. Thus, for each paper in the crawled dataset , we registered the ids of the authors that existed in it with the corresponding authors in the respective paper of the given dataset in the same manner as above. For the rest of the authors in the papers of the given dataset (which are the 'more' authors of the web page) we first made a query to our authors table based on the name and if we found one and only one author with that name, we registered that id for that author in the paper. If that wouldn't work, we proceeded by crawling the Scopus search result page for this author's name and if the result was only one and its name corresponded to the name we were looking for, we would draw the id, affiliation, city-country and field of expertise. In case neither of these would work, the rest of the authors were registered with a mark, to distinguish them as unidentified.

4.1.3 ID RESOLUTION USING COLLABORATIVE FILTERING

The process described in 4.1.2 resulted into 29.463 papers in the given dataset with all authors or some of them identified. In order to identify the rest of the names in the incomplete papers, we constructed a more sophisticated crawler which used the co-authorships of an author's name to define who he most probably is in the following way:

For each author in every paper with unidentified authors in the given dataset, we search the result set of the Scopus author search for his name, and kept ids and information for the first five that

have similar names to the one we are looking for. Lexicographical similarity between two names was defined as their Levenshtein Distance being less than 2. (Levenshtein, 1966)

The Levenshtein Distance formula is:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0, else \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} \end{cases} \quad (24)$$

After keeping those five or less candidate authors, we made a query in our given dataset for each one. The query counted the number of occurrences that our candidate id had with the known authors of this paper (or conjunction) . The intuition behind this is that an unidentified author X with 5 possible ids who has written a paper with identified Y (Yid) and Z (Zid), has more chances that his id is the one that has collaborated more with Yid or Zid. It is the case where the author is already in our database but has synonymies so we cannot identify him. If there is a tie, the id that was first at the result list is chosen, as Scopus itself presents the results in a descending accuracy manner, so the first one is most likely the corresponding id.

4.1.4 IDENTIFICATION OF AUTHORS THAT WERE NOT IN THE CRAWLING PAPERS

In the last part, it is important to be stated that we split the dataset into three sub datasets based on the publication year of a paper, in order for it to be more easily manipulated and have fault tolerance, as this process was the most time consuming and came through some changes and debugging. That resulted in three sub tables with 39.100 (p.year<2003), 38.030 (p.year>=2003&p.year<=2008) and 41.684 (p.year>2008) rows respectively. Having completed the papers in the given dataset that corresponded to the crawled dataset, the rest (~100.000 paper) consisted entirely of unknown author names. In order to identify them we followed the following procedure:

For every paper we queried our author database for a one-one correspondence based on the name of the author, and register the respective id to the searched author name. If there were more than one or none results, we continued by crawling the Scopus author search result set, and drawing the author information mentioned above, if and only if the result was one and the name was similar. If either of them worked, we ended up with some of the authors in the paper identified, and we proceeded using collaborative filtering for the rest, as mentioned in the previous part, with the

difference that in this case we searched the three sub tables and sum the count of occurrences in each one for each candidate id. If neither of them worked, we tallied the name of the author with the first author that had the same name, from the Scopus author search result set.

4.1.5 DATABASE

The final schema of our database was:

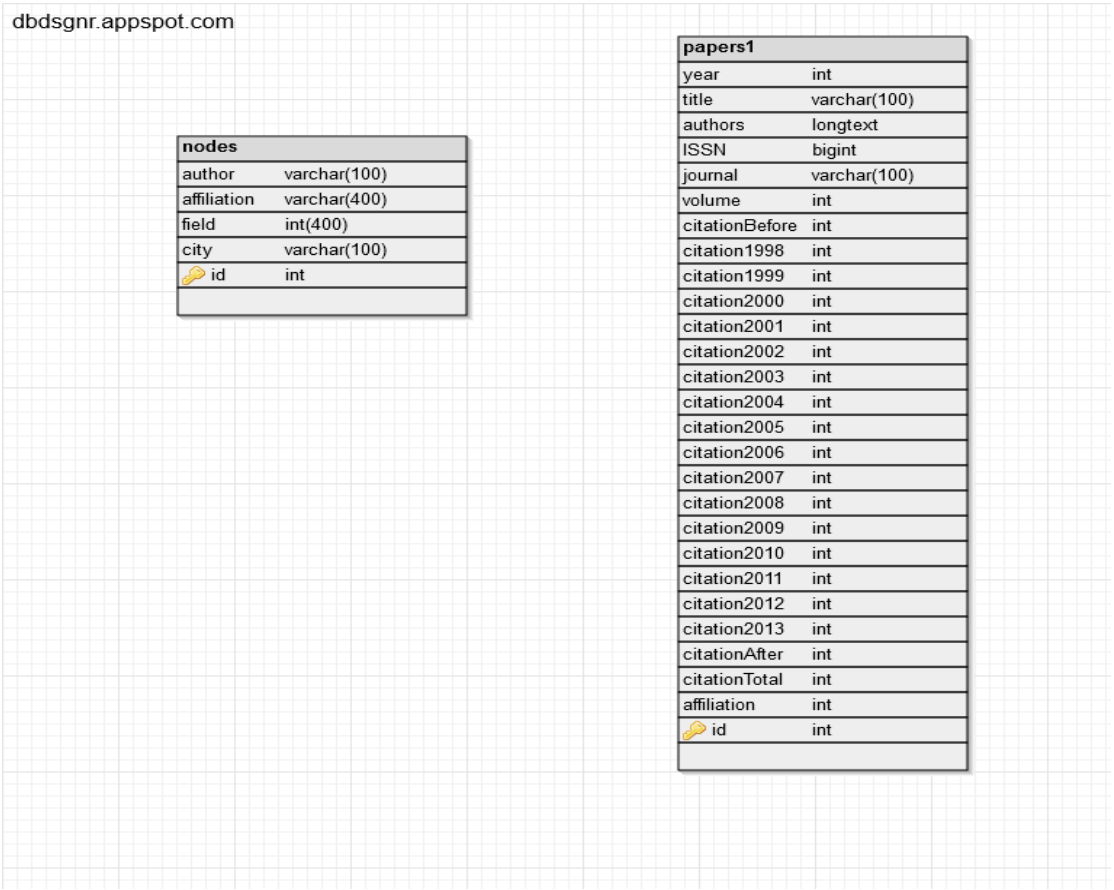


Image 3: DATBASE SCHEMA (tables papers 2 & 3 are not depicted as they have the same schema as papers1)

Table 2:TABLES ROW SIZE

Table Name	Number of Rows	Description

Nodes	132030	Information for the authors
papers1	35232	The papers written in 1998 to 2002
papers2	41898	The papers written in2003 to 2008
papers3	41475	The papers written in 2009 to 2013

4.2 COLLABORATION GRAPHS

1. With our database papers complete in terms of author identification, we were able to construct **collaboration** graphs to extract social features for every author. During this stage we found obstacles like the volume of our dataset, that restricted us from deploying all of our data or apply ordinary methods used in similar cases.

4.2.1 GRAPH PRE-PROCESSING

For every year, for every paper, we draw edges from every author to every other author in the paper, avoiding self-edges, keeping for every edge as weights the number of co-authors in the paper, its publication year and the citation the paper has taken from its publication year until the graph's year. The graph was stored in an edge list format in .txt files. An edge between two authors may occur more than once (when the two authors have written more than one paper the same year). If the two authors id100 and id101, were in id100,id101 order in a paper and in id101,id100 order in another paper, the undirected property of the edges is lost, as the latter will be identified as a different edge from the former, when is the same edge, but different collaboration (there are different weights), hence these will be aggregated in the next step. To avoid this, we use the fact that the ids of the authors are numbers. We always posit the smaller number of the two ids in the left part of the edge, and the larger on the second. In this way an edge between id100 and id101, will always be in this order, even if the author list in a paper has this order id101, id100.

Example:

EDGE	7004227072	7402561616	4	1997	3
EDGE	6603835699	7003379290	3	1996	2
EDGE	6603590870	7004227072	0	1998	2
EDGE	6603590870	7004227072	6	1996	2

The .edg files that resulted from this preprocessing were of larger volume than a normal pc can manipulate (ranging from 900Mb to over 3Gb). Thus we had to adapt a different approach and endeavor a graph compression.

4.2.2 GRAPH COMPRESSION

The graphs that resulted from 4.2.1 were multigraphs (Balakrishnan and Ranganathan 2012) in the sense that a vertex can have more than one edge with another vertex, as each edge depicts one paper they have co-authored in. Hence we can aggregate each "paper" edge to form one unique co-authorship edge for a pair of vertices. Due to the large volume of our graphs and time restraints, it was impossible to use every graph created. We utilized the graphs corresponding to the authors who have written before and during the years 1998 to 2005. Authors who have only been publicized before 1998 was not involved. Again our computational limits rendered a conventional approach manipulating each whole graph with a programming language framework not an option. Thus the problem was reached by a different point of view. Firstly we took advantage again of the fact that the ids are numbers ,and sorted each graph file using external sort java package https://code.google.com/p/externalsortinginjava/.In this manner the edges to be aggregated were gathered one after another. For example

Before		**After**
id1⇔id2		id1⇔id2
id1⇔id3		id1⇔id2

id1⇔ id3		id1⇔id3
id1⇔id4		id1⇔id3
id1⇔id2		id1⇔id4

From this point on, it was easy to aggregate collaborations into forming an edge which depicts a whole coauthorship relation between two authors in a given year. We read the files line by line, keeping the two ids and aggregated the weights of each co-authorship in one of the two ways delineated at 3.1.2. The edge would change when either the first or the second part of the line read (which corresponded to the ids of the collaboration) changed. In that point the two former ids and the weight aggregated till then formed a new edge and were stored in a txt file. Than, a new aggregated edge would start, with the new ids, until the ids read would change again.

4.2.3 BRIEF GRAPH ANALYSIS

The operations stated at 4.2.1 had as outcome a set of 14 graphs (one quantity and one quality for each year). We proceed into giving some graph analytics in order to better interpret the structure of the graphs. We chose to analyze the weighted by quality graphs, because though having the same structure and form to the quantity one, its edges capture more information. It is vital to reenact the evolution of basic graph measures such as vertices, edges and the mean degree, which is how many edges a vertex has in average. The plots are shown in image 4.

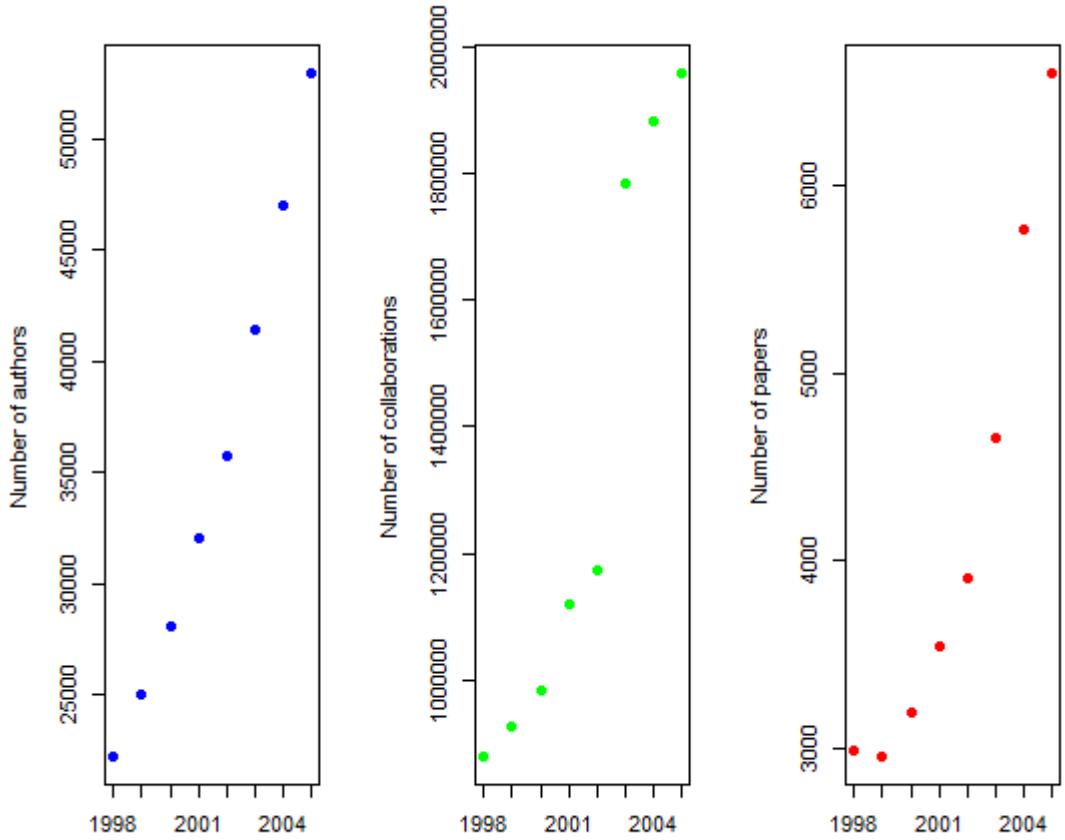


Image 4: GRAPH INDICES PLOTS

To further explain these plots, the plot of the authors (blue) shows that the increase in vertices is almost linear, implying a stable increase in authors. That, combined with the fact that the number of papers (red) in 1998-1999, have as an outcome a low and decreasing productivity level per author in the first two years. Moreover, initially the number of new collaborations increase too, but in a lower rate than authors, expressing a decreasing new collaborations per author index. This means that papers produced in season 1999 had a larger number of co-authors than those written in 2000, because in the former the papers decrease while in the second they increase. The number of papers per year is increasing after 1999, with proportionally the same rate as the growth of authors, hence the scientific community becomes proportionally productive after 1999. At year 2001 the growth of edges has a little boost, while the increase of vertices stays stable, thus the mean collaborations per author does not decline as much as the previous years, were the new collaborations were few but new authors emerged vastly . The year 2002 has the lowest edge escalation while the upsurge in number of papers and vertices stay steady, hence we can conclude that this year was socially ‘ill’, with many authors who wrote by themselves or in close small

groups. In contrast to 2002, the social factor explodes 2003, because of the huge boost (~50%) in edges and a substantial in papers (~20%). Here we witness a great rise in the new collaborations an author conducts and a descent in the paper he writes. This may be explained either as a stronger bonding in the scientific community, which resulted in new ideas and inspiration, or an emerging of new sciences where each paper demands work of many authors. The escalation in papers keeps rising from now on, surpassing the one of authors, marking a great an prosperous era in terms of productivity. The new collaborations continue to increase, but in a rhythm close to their starting one and since the authors continue to grow more quickly, becoming more than double than their initial amount, the reduction of mean collaboration per author falls again. Thus we can conclude that the speed of new authors emerging in science as well as the production of papers written, are generally significant larger than the expansion of the social factor, denoting that authors tend to write with people who have collaborated already, in contrast to newer ones.

4.3 CLUSTERING

The clustering procedure is the peak of our research, since it provided us with the desirable outcome, a categorization of the authors. It is made up of many stages, through which we traversed, alternating between Java programming language and R statistical language.

4.3.1 FEATURES

The features that were extracted are the essence of the information that we wanted to capture for every author. They also represent the novel metrics and methods we adapted in order to best depict the required information. We distinguish them based on their relationship with time, hence the dataset they are included in.

4.3.1.1 YEARLY DATASETS

The yearly datasets consisted of the main attributes that in our opinion characterize an author. They are distinguished based on their origin and purpose.

4.3.1.1.1 AUTHOR FEATURES

The independent features described in 3.2.1, are calculated directly from the database. Mysql queries fetch and penalize the respective values asked, and java manipulates them and stores them in csv files, resulting in 7 datasets with 7 columns each (id, features in 3.2.1).

4.3.1.1.2 SOCIAL FEATURES

The graph features delineated in 3.1.3.1 are measured using the R library igraph[3]. We load each years quality graph, calculate the 3 features and store them in 7 corresponding csv files, with 4 columns each (id, features at 3.1.3.1).

4.3.1.1.3 POWER GRAPH FEATURES

To construct a power graph from a weighted graph we had to use the powergraph.jar[4]. That procedure was quite slow and pretentious for our dense graphs, being the main reason why we stopped at year 2005 rather than exploiting the full potential of our data until the year 2013. Nevertheless the power graph.jar had as a main input one of our graphs in .edg format (quantity & quality, every year) and as an output a .bbl format. To create the specific indices from 3.1.3.2 we build a java parser which exploited the JUNG [5] framework. The power nodes were handled as regular nodes, while there were a specific class defined for power edges. The weight of the power node was easy to calculate as it just demanded to tally a power node's weight to each of the nodes inside him, which corresponded to authors. The power cliques weight was calculated by finding the neighbors of a power node and summing their weights, each multiplied with the respected edge's weight (edge connecting the power node and the neighbor). This number was then added to the sum of weights of the power nodes containing the examined power node, which were discovered recursively. This procedure produced 7 datasets with 5 columns each (id, features stated at 3.1.3.2)

4.3.1.2 EVOLUTION DATASET

For the construction of the evolution dataset, we first build the partial change datasets between the years. Namely, we iterate through the consecutive yearly datasets and detract the value of a feature in the examining dataset, to the corresponding one in the previous year's dataset. This results in 6 change datasets, to respective years 1999-1998, 2000-1999 etc. After that we go over these 6 datasets in accordance to their respective years in chronologically ascending series, and calculate for each author and each of his features, the four indices mentioned in 3.2.2.1, 3.2.2.2, 3.2.2.3. We

[3] http://igraph.org/r/

[4] http://www.biotec.tu-dresden.de/research/schroeder/powergraphs/download-command-line-tool.html

[5] http://jung.sourceforge.net/

also keep track of the average value of the feature from the initial yearly datasets as mentioned in 3.2.2.4. The average is calculated as stated in 3.2.2.4 based on the feature's nature. For the temporal features, the trend is defined as the inverse tangent of the slope of the linear regression for the feature's values in each respective year, in radians. To calculate this we made use of the Math [6]java library. If an author is encountered for the first time during the iteration, his values are stored as his first change, like detracting with zero.

[6] http://commons.apache.org/proper/commons-math/

4.3.2 ALGORITHM

In order to determine the right number of clusters, we used specific indices as stated in 3.3.2. The R algorithm (see Appendix 0) runs K-means 98 times and calculates these indices. We made use of the clusterSim[7] R library, from which we used the index.DB function to calculate the Davies Bouldin index for each clustering. The indices were stored in a matrix 98*5 , one column for the possible number of clusters, the second for the Average Within Cluster Sum of Square, the third for the Average Distance Between Cluster Centroids, the Fourth for the Davies Bouldin and the fifth for the Dunn index. We made plots of the pairwise multiplication of these indices ,(WCSS*DaviesBouldin) at image 5 and (DBCC*Dunn) at image 6, to show how they progress in regard to the number of clusters so as to determine the best number of clusters. 7 is the optimum number , since it is the case with the most efficient combination of high WCSS*Davies-Bouldin and low DBCC*Dunn. This stands because 7 is one of the highest cases in the red (high) plot, with only 4 and 5 surpassing it, but at the same time having the lowest value of the three in the blue (low) plot. We also pursued to have as more clusters as possible, under a logical context, to fully exploit the variation of our features.

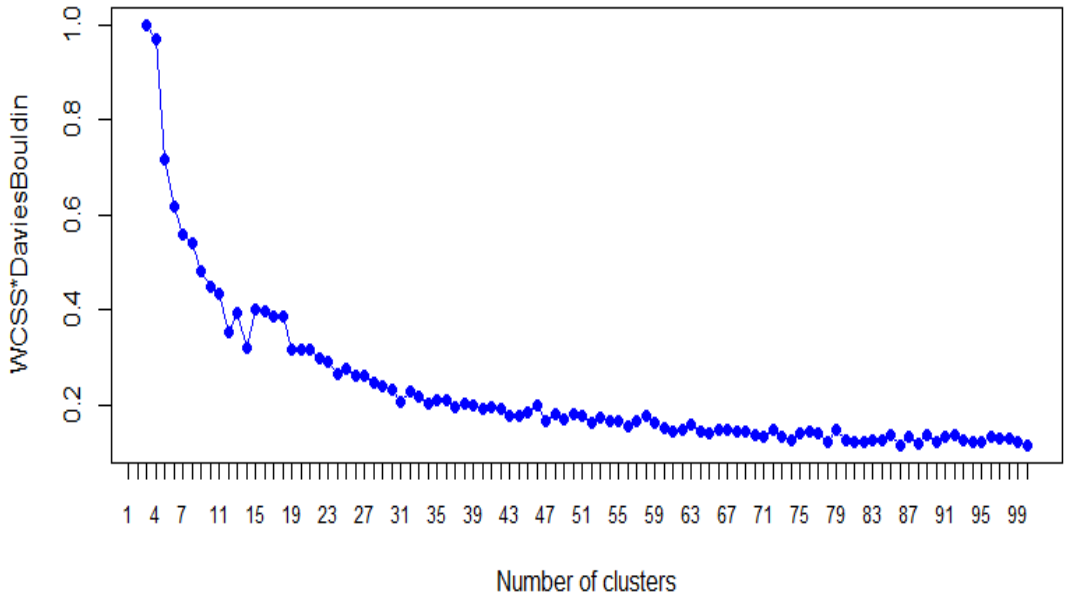


Image 5:WCSS*DAVIESBOULDIN PLOT

[7] http://cran.r-project.org/web/packages/clusterSim/index.html

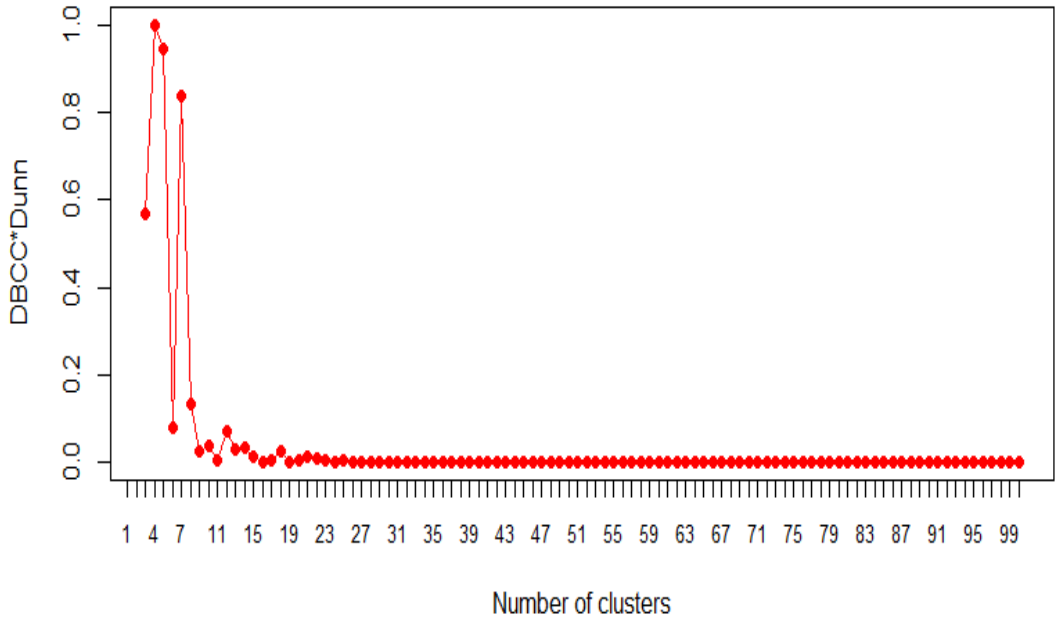


Image 6: DBCC* DUNN PLOT

4.3.3 FEATURE SELECTION

One crucial part of our analysis is revealing the most important features that an author has to maximize in order to increase his success. To achieve that, it was necessary to build a more compact dataset than the evolution dataset, which consisted of 65 features, in favor of revealing the most impactful of the initial 13 features in the yearly datasets. The method for the compression is defined at 3.3.4.1 and the main reason was to distinguish more meaningful characteristics, as mentioned also in 3.3.4.1. After the aggregation of the dataset, we apply singular value decomposition in the 13 feature dataset to distinguish the most influential columns (Wall, Rechtsteiner, & Rocha, 2003) in the manner described at 3.3.4.2. Image 7 depicts the singular values of the dataset, and it is easy to see that the 4 first enclose the most variance, specifically approximately the 85% of it.

Thus we need to consider the first four singular values and keep the features having the biggest impact on them, based on the corresponding right singular vectors. Image 8 shows the plot with the 13 features and their values. According to this plot, the citations that an author takes every year summed with the corresponding trend in radiance, is the most important characteristic of an author, followed by the number of papers penalized by time and his number of papers by year.

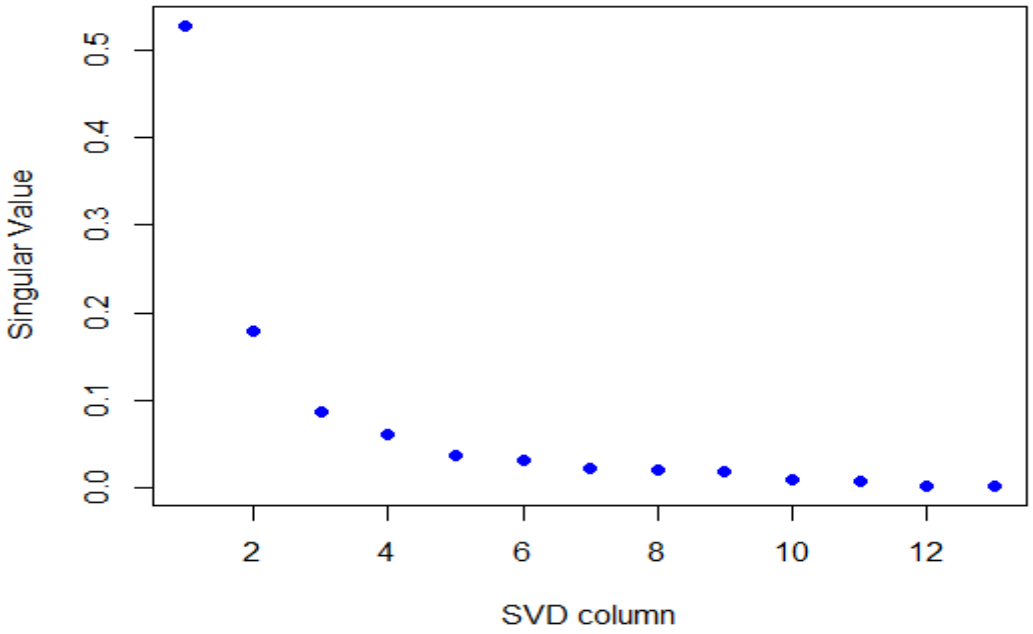


Image 7:AGGREGATED DATASET SINGULAR VALUES (PERCENTAAGE OF DEVIATION DEPICTED BY EACH VECTOR)

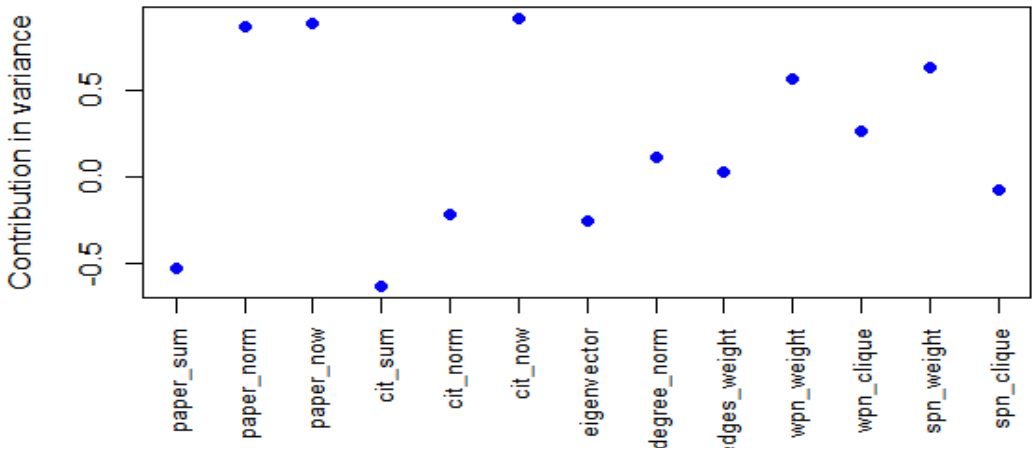


Image 8:FEATURE IMPACT ON CLUSTERING

4.3.4 TIME SERIES CLUSTERING

As explained in 3.3.5 time series clustering is applied in a set of time series, in order to categorize them. This means that we cannot use every feature of our dataset, because that would give us 13 different time series for every author, which will be compared to the corresponding 13 time series of the rest of the authors, hence we will have 13 different clustering results and 13 different clusters to which the author will belong to. The time series chosen to best represent the activity of each author, was the one corresponding to the timeline of the most influential feature that resulted from 4.3.3, namely the every year citations (3.2.1.6).

4.3.4.1 TIME SERIES DATASET

The dataset build contained the 43.567 authors, and 9 columns, the first representing the author id and the rest the values of the selected feature in each of the years 1998-2005. If an author would emerge in a year after 1998, than the values in the previous years would be set to NA. This is feasible, because the DTW distance measure, analysed in 3.3.5.1 enables the comparison of time series of different lengths.

4.3.4.2 CLUSTERING

The algorithm employed for this task was PAM (partitioning around medoids) for the reasons delineated at 3.3.5.2, as implemented in the R library fpc[8]. Another R library which we used was dtw[9], to define the DTW distance between the time series for the PAM clustering. The number of clusters was set to 7, as this is the most possible number of author groups existing in our dataset, according to our analysis at 4.3.2.

4.3.4.3 CLUSTERING COMPARISON

One interesting aspect of our research was to determine how will the time series clustering results agree with the ones from the k-means clustering. To achieve that we found the similarity of each of the time series clusters to the k-means clusters, in terms of common authors. The purpose of this is to tally each cluster of the time series clustering to one and only one of the k-means clustering, and count how many authors were clustered right. We build a matrix representing the

[8] http://cran.r-project.org/web/packages/fpc/index.html

[9] http://dtw.r-forge.r-project.org/

common authors between each cluster, rows represent the time series clusters and columns the k-means.

Table 3: TIME SERIES CLUSTERS VS K-MEANS CLUSTERS

Name/ Size	1/ 3958	2/ 1132	3/ 2846	4/ 394	5/ 1161	6/ 544	7/ 33532
1 / 5283	658	48	388	56	230	99	3804
2 / 25921	1974	1001	1510	0	101	0	21335
3 / 9233	1075	67	784	0	90	0	7217
4 / 1721	232	12	140	0	262	0	1075
5 / 743	19	4	24	3	478	115	100
6 / 287	0	0	0	2	0	285	0
7 / 378	0	0	0	333	0	45	0

We continue with this algorithm:

1. Initialize sum=0
2. While matrix has more than two columns
3. Find the biggest value in the matrix, its row and its column
4. Add it to the sum
5. Delete the column and the row

6. return to 2

We implemented this algorithm in R (see Appendix 7.3). The conclusion is that we can achieve ~55% of right author grouping, just by taking into account the most influential feature. That is the 2/14 of the initial information used (id and cit_now).

4.3.5 LABELING

In this stage, we tried to name the clusters based on their characteristics, as mentioned in 3.3.3. To accomplish this, we first created two thresholds for every of the 65 feature in the dataset. For each feature, its vector of the 7 cluster centroids was used, to extract the .85 percentile and the 0.15 percentile of it. The n-th percentile of an observation variable is the value that cuts off the first n percent of the data values when it is sorted in ascending order, like a high or low oriented average. These represented a guide to whether a cluster's centroid is high or low, in that feature. Having two vector of high and low thresholds, consisted of 65 values each, we run through a cluster's centroids to find out the features that differentiate. We than characterize each cluster based on the marked features, as follows:

- Cluster 1 (394 authors): Highest number of average papers and citations with increasing rate but low minimum changes. High power clique weight but moderate power node. ➔ Powerful authors belonging to high community.
- Cluster 2 (33532 authors): Low maximum and high minimum changes in social indices. In other words steady, moderate increase in the social aspect. Important dynamics in in paper fertility, however average citations. ➔ Average author.
- Cluster 3 (1161 authors): Steady change indices in papers and high values in last changes of citations. ➔ Cluster of increasingly successful authors, but still socially and productively moderate.
- Cluster 4 (2846 authors) *[10]: Low last change in all social metrics, otherwise similar behavior to cluster 1. ➔ Group of sudden socially severed, with average papers and citations.
- Cluster 5 (544 authors): Substantial average and max weight rate in quality and quantity power nodes, with high dynamic in citations. ➔ Rising citation receivers, associated to impactful groups.

[10] * When conclusions could not be derived, the threshold span increased in 0.8 and 0.2 percentile, and recalculated the clusters' predominant features.

- Cluster 6 (3958 authors): The smallest sum of changes in all social metrics, depicting a steady decreasing behavior in the social aspect. ➔ Group of stable socially withering, neutral authors.
- Cluster 7 (1132 authors): Considerable sum and last change in power node weight showing increasing dynamic in close community. Limited number of papers and citations with low increase rate. ➔ Inferior authors belonging to an upcoming co-authorship group.

5. EXAMPLES

5.1 CRAWLING

The actual procedure of crawling contained some important technical details that it is should be mentioned. We give examples of every kind of Scopus pages that we parsed and how we manipulated the urls.

5.1.1 INSTITUTIONS PAPER LIST

USED IN: 4.1.1

URL :Initial page:

http://www.Scopus.com/results/results.url?cc=10&sort=cp-f&src=... sdt=a&sl=15&s=AF-ID%2860012296%29&ss=cp-f&ws=r-f&ps=r-f&cs=r-f&origin=resultslist&zone=resultslist

the rest of the pages:

http://www.Scopus.com/results/results.url?sort=cp-f&src=...sdt=a&sl=15&s=AF-ID%2860012296%29&cl=t&**offset=201**&origin=resultslist &ss=cp-f…

http://www.Scopus.com/results/results.url?sort=cp-f&src=... sdt=a&sl=15&s=AF-D%2860012296%29&cl=t&**offset=401**&origin=resultslist&ss=cp-f…

It is easy to see that the main difference is the offset parameter, which is not included in the first page, but it is in the second and it increases by 200 for each next page. Thus all we had to do is modify the url into containing the particular offset and increase it by 200 for every next page, for 10 pages.

Image:

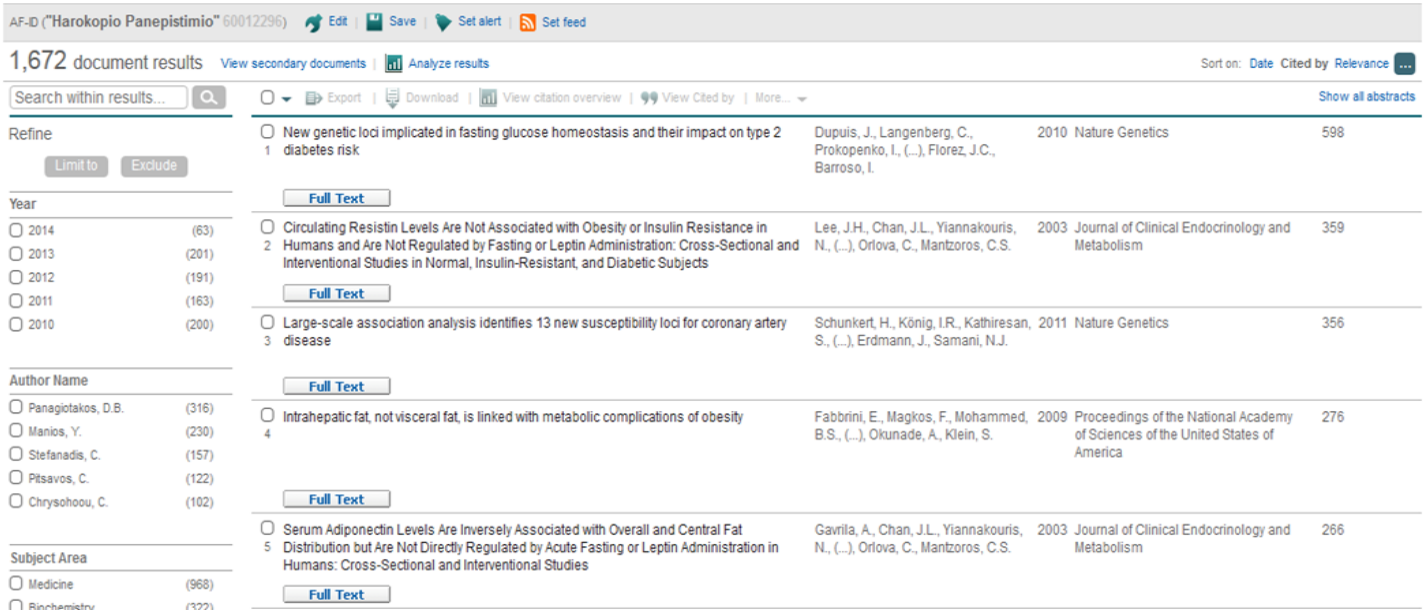


Image 9 :SCOPUS INITIAL PAGE OF AFFILIATION'S PUBLICATIONS

5.1.2 AUTHOR PAGE

USED IN: 4.1.2, 4.1.3, 4.1.4

URL: http://www.Scopus.com/authid/detail.url?authorId=**6603228762**

6603228762 is the id of the author for whom we want to get information, to parse a page of another author, we changed the id to the respective author's id.

Image:

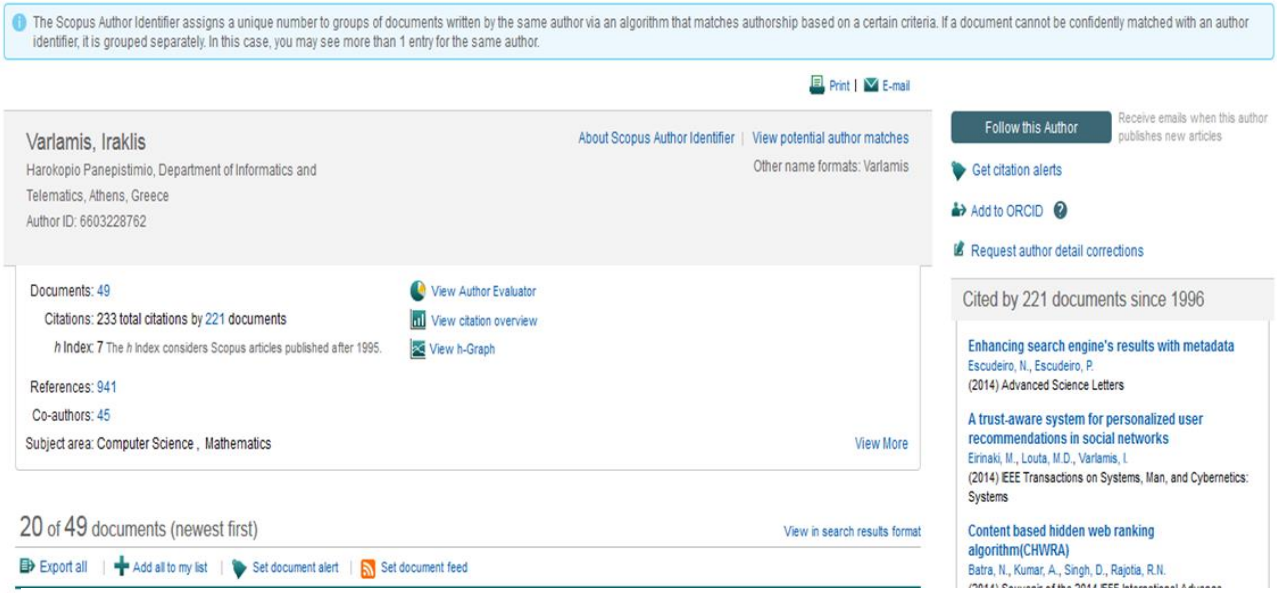


Image 10: SCOPUS AUTHOR PAGE

5.1.3 SCOPUS AUTHOR SEARCH RESULT PAGE

USED IN: 4.1.2, 4.1.3, 4.1.4

URL:http://www.Scopus.com/results/authorNamesList.url?sort=count-...=AUTH--LAST--NAME%28**Varlamis**%29+AND+AUTH--FIRST%28I.%29&st1=**Varlamis**&st2=**I.**&selectionPageSearch=anl&reselectAuthor...

The ‘Varlamis’ and ‘I.’ correspond to the name and the initial of the author searched. For another author, you replace them with the author’s name and initial.

Image:

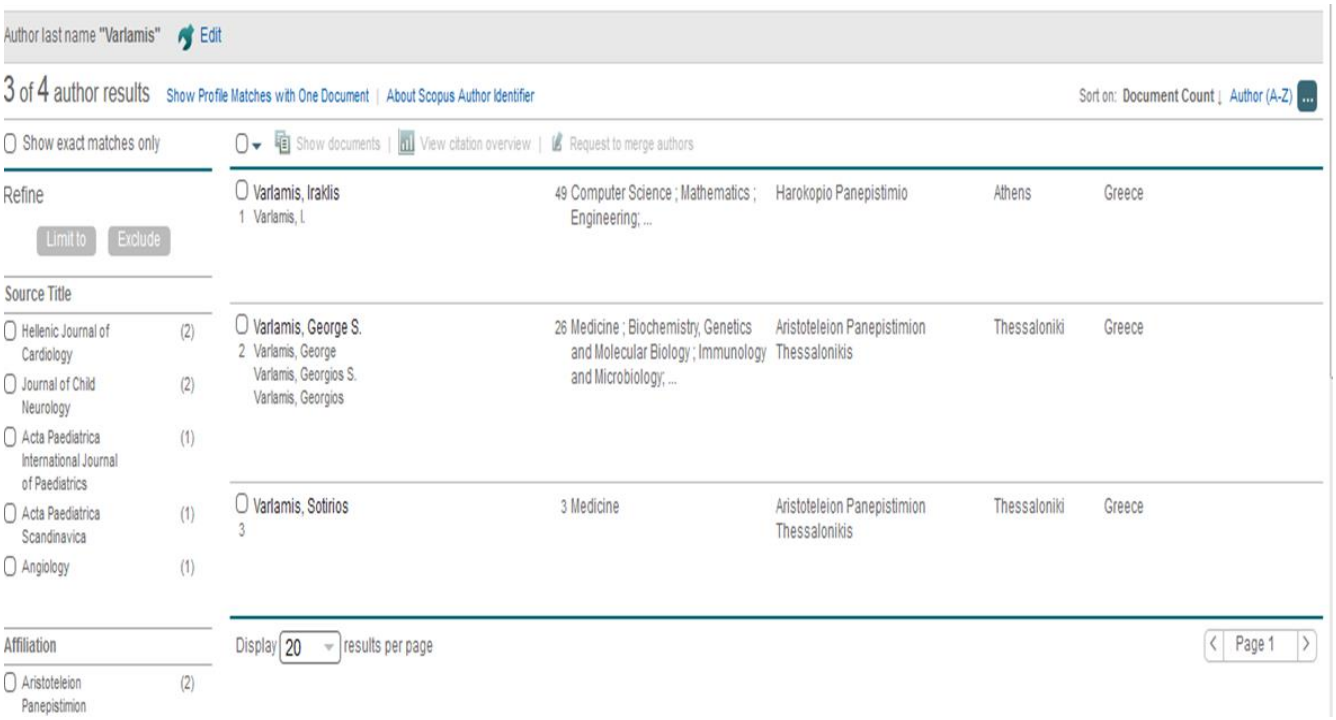


Image 11: SCOPUS AUTHOR SEARCH RESULT PAGE

5.2 GRAPH EXAMPLE

In order to comprehend the full information that the graphs contain, we view thoroughly an author's network as an example, from the quality graph. The author with the most impactful co-authorships (biggest sum of edge weights) during year 1998 is Resvanis L.K. with Scopus id=7004604742. His collaboration "neighborhood" consists of 1547 co-authors and the mean edge weight is 0.1608 and is depicted at image 12. The strength of the co-authorship is displayed as color and as edge length meaning the brighter and the closer a neighbor is to the central author, the more impact the co-authorship has.

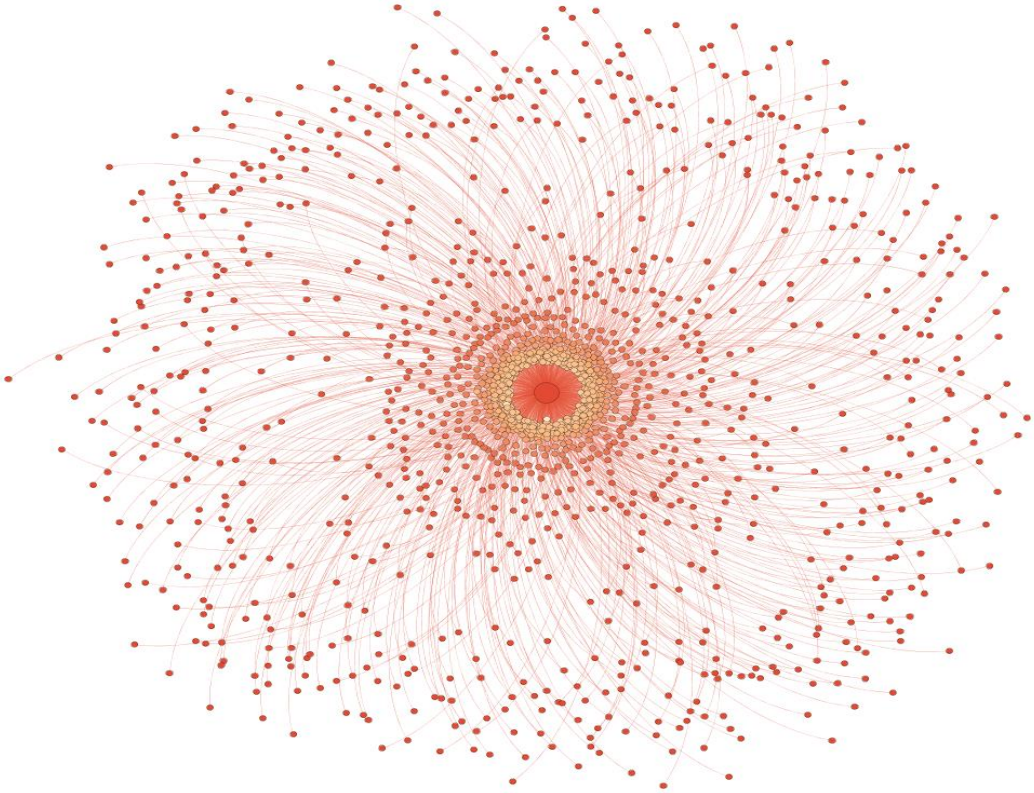


Image 12: NETWORK OF TOP 1998 AUTHOR

Keeping track of the same author, we examined his "neighborhood" at year 2001. It now consists of 1660 coauthors with a mean weight edge of 0.222. To depict the difference between the two networks and the collaborations, we combined them in one graph, separating the main author as two different persons but keeping the same edges between his coauthors, as depicted in image 13**.** In addition the two nodes representing the main author were drawn in the opposite ends of the image to distinguish them better. The left node is the author during 2001 and the right one is during 1998.

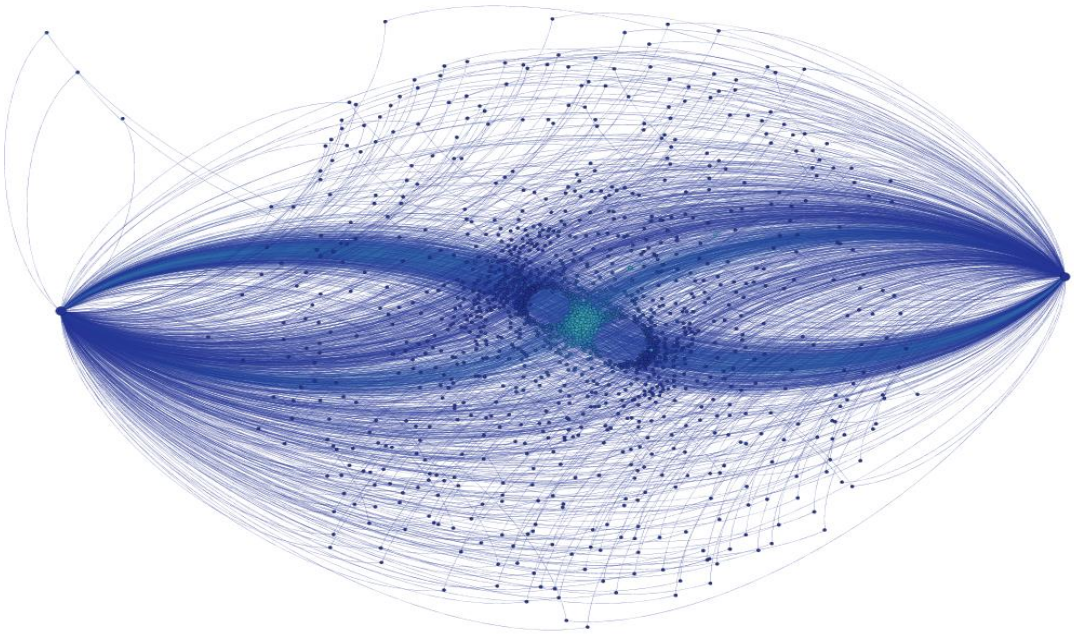


Image 13: TOP 1998 AUTHOR IN 1998 (RIGHT) AND 2001 (LEFT)

Again the color and the position of the co-authors is proportional to the impact of the co-authorship. So we see that stronger coauthorships, which are more central and have nodes of lighter color, were kept strong through time. Someone can also derive that the rest of the co-authorships that are wide spread, were more occasional or less successful. We distinguished three edges from the rest of the network (upper left). Assigning edge width proportional to the edge weight and taking a closer and less “elegant” look at the graph, in image 14, we can presume why.

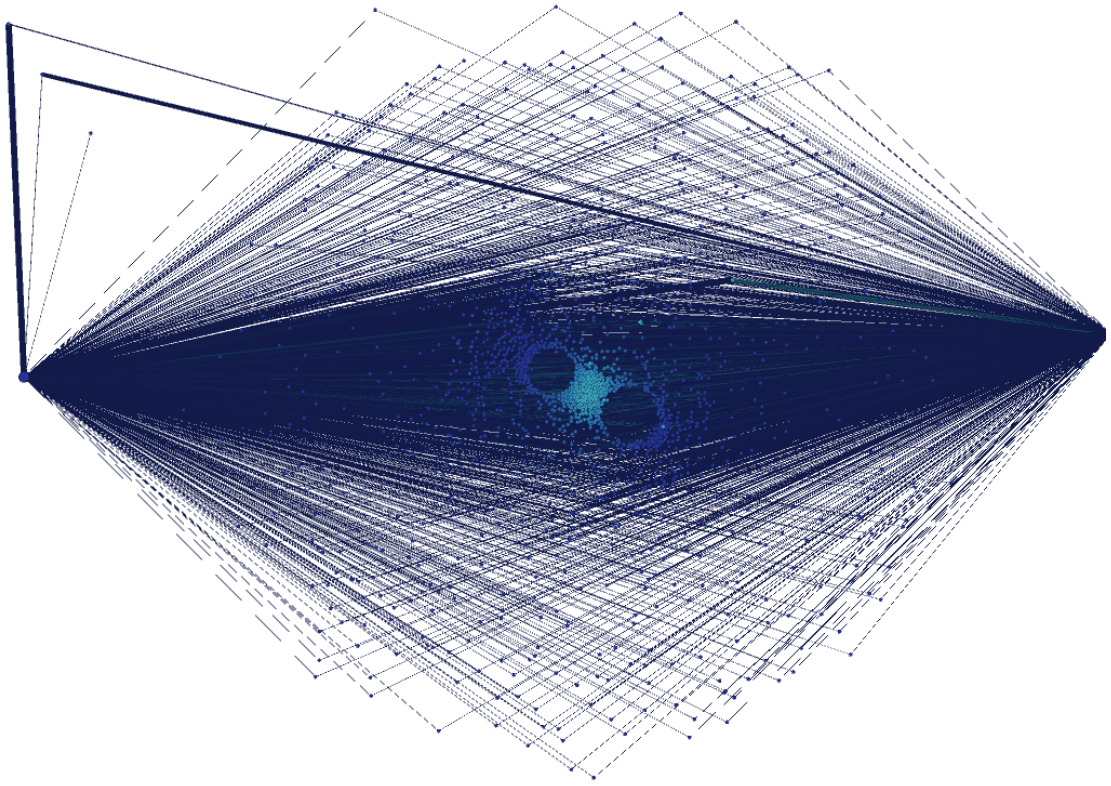


Image 14: IMAGE 13, WITH COAUTHORSHIP WEIGHT DEPICTED ON EDGE THICKNESS

The three edges depict a different aspect of our implementation. Firstly (the higher one) shows how a coauthorship can be strengthened through time, as the two authors write together, that is why the edge is "heavier" during 2001 than during 1998, the actual difference between the edge weights is 0.21 (0.44 at 1998 and 0.65 at 2001).That edge depicts a collaboration of the main author with the author Stanescu, C. Scopus id= 24175473500.

Secondly (the second one) depicts how a coauthorship can tabefy through time. It is the opposite situation to the above, as the edge had a weight of 0.25 at 1998 and a weight of 0.20 at 2001. The edge depicts a collaboration of the main author with the author Sebastia, A. Scopus id= 55994682800.

The third node depicts a new coauthorship with the author Starinsky, N. Scopus id=6602773097. A node with just one edge can only be observed coming from the left part of the graph, meaning that there is no coauthorship that occurred until 1998 and is absent at 2001.

6. CONCLUSIONS & FUTURE WORK

In this thesis we endeavor an author categorization as well as general knowledge extraction, exploiting various aspects of the information we can derive from an author's work and collaborations. We crawled bibliographic data with time span, from the Scopus digital library and covered any inconsistencies by applying a technique similar to collaborative filtering. Graph representation and mining techniques allowed us to capture the social, individual and time related facets of an author's impact, simultaneously extracting various information with regard to the rate of publications and new authors, the top authors and their co-authorships' endurance in time, the power of successful co-authoring communities etc.. During this process new metrics characterizing an author evolved, introducing time penalization in Bibliographical, Power Graph and Social Network Analysis, to capture the temporal character of a success or a collaboration. These features were deployed into building yearly datasets and then we captured the evolution of each author's feature, with certain indices. These indices made up the data to which we applied K-means algorithm and clustered the authors. The number of the clusters was defined by experimenting and using well established clustering validity measures. Cluster labeling was conducted depending on the feature characteristics of each cluster, and the 7 groups of authors which derived, differentiated mostly in belonging community's power, citation rate and social impact. Furthermore, an attempt to expose the most influential to the clustering features, was done, by applying singular value decomposition. The results displayed that citations an author receives each year and the trend they follow is the most important of an authors features, with the number of papers penalized by time and the average number of papers by year as second and third. Each author had a timeline with values of the peak feature, which were used to perform time series clustering with the dynamic time warping measure as distance. The resulted clusters were tallied with the clusters of the aforementioned clustering, showing that almost 55 percent of the clustering was the same, meaning that the authors grouped together in the first place, were grouped together again. This means that time series clustering encloses a big percentage of information and can be used in the features for more efficient clustering with less data required.

Our plans for the future focus on constructing a classification mechanism, which will classify an author to a respective group, given the required features. Moreover it is important to work with the whole crawled dataset, as it encloses a substantial larger amount of information, in terms of social and bibliographical metrics. Finally, under exploring more thoroughly the time series clustering,

a combination of more than one feature's time series may prove to be significant in the clustering's success.

7. APPENDICES

7.1 GREEK AFFILIATIONS

Greek higher educational institutions that are included in Scopus.

University of Athens
Aristoteleion Panepistimion Thessalonikis
Ethniko Metsovio Polytechnio
Polytechnion Kritis
Panepistimion Patron
Panepistimio Kritis
Panepistimion Ioanninon
Dimokrition Panepistimion Thrakis
Panepistimio Thesalias
Panepistimion Aegaeou
Geoponiko Panepistimion Athinon
Panepistimion Pireos
Ikonomikon Panepistimion Athinon
Panepistimion Makedonias
University of Peloponnese
Harokopio Panepistimio
Hellenic Open University
Ionian Panepistimion
Panteion Panepestimion Ikonomikon kai Politicon Epistimon
University of Central Greece

7.2 R CLUSTERING EXPERIMENTS CODE

```
setwd("C:/Users/Administrator/Desktop/thesis")
library(clusterSim)##for index.DB
data=read.csv("final.csv")
ids=data[,1]##keep the ids of the authors
data=data[,-1]
cls=data.frame(matrix(nrow=98,ncol=5))
names(cls)=c("number","avg(within groups sum of squares)","avg(distance between
cluster centroids)","avg(db)","avg(dunn)")
                        for(p in 3:100){

kClust=kmeans(data,centers=p) ##kmeans clustering with different number of clusters
wgss=c()
for(i in 1:length(kClust$withinss)){
wgss=c(wgss,kClust$withinss[i]/kClust$size[i]) }##normalized within group sum of
squares        dbcc=0
dun=c()
for(i in 1:ncol(kClust$centers)){
inter=c()
h=0
for(j in 1:nrow(kClust$centers)){
for(k in j:nrow(kClust$centers)){
if(k!=j){
inter=c(inter,abs(kClust$centers[j,i]-kClust$centers[k,i]))##pairwise distance between
cluster centroids
h=h+1##number of distances
}
}
}
dbcc=dbcc+sum(inter)/h##distance between cluster centroids for each feature
dun=c(dun,inter)##intercluster distances
                }
cls[p-2,1]=p
cls[p-2,2]=sum(wgss)/p##average sum of squares through clusters
cls[p-2,3]=dbcc/ncol(kClust$centers)##average distance between centroids through
columns        cls[p-2,4]=index.DB(data,kClust$cluster,centrotypes="centroids",p=2)$DB
##Davies-Bouldin        cls[p-
2,5]=min(dun)/max(wgss)##minimum(intercluster)/maximum(intracluster)
}
```

Εικόνα 1: setwd("C:/Users/Administrator/Desktop/thesis")
library(clusterSim)##for index.DB
data=read.csv("final.csv")
ids=data[,1]##keep the ids of the authors
data=data[,-1]
cls=data.frame(matrix(nrow=98,ncol=5))

7.3 R CLUSTERING COMPARISON CODE&OUTPUT

```
n=read.csv("tsPam.csv")##time series clusters
o=read.csv("kmeans.csv")##k-means clusters

ComAuthors=matrix(nrow=7,ncol=7)
for(j in 1:7){
  TsIds=as.factor(n[n[,2]==j,1])##ids of authors in j cluster of timeSeries
  for(i in 1:7){
    KIds=as.factor(o[o[,2]==i,1])##ids of authors in i cluster of kmeans
    count=0
    for(s in KIds){
      if(s %in% TsIds){
        count=count+1##count the common authors
      }
    }
    ComAuthors[j,i]=count
  }
}
print(ComAuthors)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]  [,7]
## [1,]  658   48  388   56  230   99  3804
## [2,] 1974 1001 1510    0  101    0 21335
## [3,] 1075   67  784    0   90    0  7217
## [4,]  232   12  140    0  262    0  1075
## [5,]   19    4   24    3  478  115   100
## [6,]    0    0    0    2    0  285     0
## [7,]    0    0    0  333    0   45     0
```

```
ComAuthors=rbind(c(1,2,3,4,5,6,7),ComAuthors)##corresponding TS clusters
ComAuthors=cbind(c(0,1,2,3,4,5,6,7),ComAuthors)## >> kmeans clusters
sum=0
while(ncol(ComAuthors)>2){
  k=which(ComAuthors==max(ComAuthors),arr.ind=TRUE)
  #sprintf("%s TsCluster corresponds to %s KCluster with %s Common Authors", ComAuthors[k[1],1] ,ComAuthors[1,k[2]],max(ComAuthors))
  print(paste(paste(paste(ComAuthors[k[1],1]," TsCluster corresponds to "),paste(ComAuthors[1,k[2]]," KCluster with ")),paste(max(ComAuthors)," common authors")))
  sum=sum+max(ComAuthors)
  ComAuthors=ComAuthors[-k[1],]##delete the row
  ComAuthors=ComAuthors[,-k[2]]##delete the column
}
```

```
## [1] "2  TsCluster corresponds to  7  KCluster with  21335  common authors"
## [1] "3  TsCluster corresponds to  1  KCluster with  1075  common authors"
## [1] "5  TsCluster corresponds to  5  KCluster with  478  common authors"
## [1] "1  TsCluster corresponds to  3  KCluster with  388  common authors"
## [1] "7  TsCluster corresponds to  4  KCluster with  333  common authors"
## [1] "6  TsCluster corresponds to  6  KCluster with  285  common authors"
```

```
sum=sum+ComAuthors[2,2] ##add the last value
print(paste("the percentage of right clustering is",sum*100/nrow(n)))
```

```
## [1] "the percentage of right clustering is 54.8730661525042"
```

8. BIBLIOGRAPHY

An, Y., Janssen, J., & Milios, E. (2004). Characterizing and mining the citation graph of the computer science literature. *nowledge and Information Systems, 6(6)*, 664-678.

Balakrishnan, R., & Ranganathan, K. (2012). *A textbook of graph theory.*

Berndt, D., & Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. *KDD workshop Vol. 10. No. 16.*

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology 2.1*, 113-120.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*.

Borgatti, S., & Everett, M. (1999). The centrality of groups and classes. *The Journal of Mathematical Sociology 23(3)*, 181-201.

Davies, D., & Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence (2)*, 224-227.

Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. 32-57.

Erten, C., Harding, P., Kobourov, S., Wampler, K., & Yee, G. (2004). GraphAEL: Graph animations with evolving layouts. *Graph Drawing*.

Falagas, M., Pitsouni, E., Malietzis, G., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB Journal 22 (2)*, 338–42.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry* , 35-41.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems 17(2-3)*, 107-145.

Harth, A., Umbrich, J., & Decker, S. (2006). Multicrawler: A pipelined architecture for crawling and indexing semantic web data. *The Semantic Web-ISWC 2006*, 258-271.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America 102.46*, (σσ. 16569-16572).

Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika 32.3* , 241-254.

Kaufman, L., & Rousseeuw, P. (1987). Clustering by means of Medoids. *Statistical Data Analysis Based on the L1 Norm*, σσ. 405-416.

Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of Citations in Web of Science, Scopus, and Google Scholar for Articles Published in General Medical Journals. *JAMA 302 (10)*, 1092–6.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady Vol. 10*.

MacQueen, J. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Berkeley Symposium on Mathematical Statistics and Probability Vol. 1. No. 281-297*. University of California Press.

Manning, C., Raghavan, P., & Schutze, H. (2008). *An Introduction to Information Retrieval.* Cambridge University Press.

Martin, E., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining Vol. 96*, (σσ. 226-331).

Martyn, J. (1964). Bibliographic coupling. *Journal of Documentation 20.4*, 236.

Odda, T. (1979). On properties of a well-known graph or what is your Ramsey number? Topics in graph theory. *Annals of the New York Academy of Sciences 328.1*, 166–172.

O'Madadhain, J., Hutchins, J., & Smyth, P. (2005). Prediction and ranking algorithms for event-based network data. *CM SIGKDD Explorations Newsletter 7.2*, 23-30.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58(347-352)*, 240–242.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66(336)*, 846–850.

Rousseeuw, P. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics 20*, 53-65.

Royer, L., Reimann, M., Andreopoulos, B., & Schroeder, M. (2008). Unraveling Protein Networks with Power Graph Analysis. *PLoS Computational Biology 4(7)*.

Tsatsaronis, G., Varlamis, I., Torge, S., Reimann, M., Norvag, K., Schroeder, M., & Zschunke, M. (2011). How to Become a Group Leader? or Modeling Author Types Based on Graph Mining. *Research and Advanced Technology for Digital Libraries*, 15-26.

Varlamis, I., & Tsatsaronis, G. (2012). Mining Potential Research Synergies from Co-Authorship Graphs using Power Graph Analysis. *International Journal of Web Engineering and Technology 7(3)*, 250 - 272.

Wall, M., Rechtsteiner, A., & Rocha, L. (2003). Singular value decomposition and principal component analysis. *A practical approach to microarray data analysis*, 91-109.

West, J., & Wiseman, M. (2008). The Eigenfactor Metrics. *Journal of Neuroscience 28 (45)*, 11433–11434.

West, J., Jensen, M., Dandrea, R., Gordon, G., & Bergstrom, C. (2013). Author-Level Eigenfactor Metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology 64(4)*, 787-801.

Xiaofei, H., Deng, C., Yuanlong, S., Hujun, B., & Jiawei, H. (2011). Laplacian regularized gaussian mixture model for data clustering. *Knowledge and Data Engineering, IEEE Transactions 23(9)*, 1406-1418.

ΑΚΟΛΟΥΘΕΙ ΜΙΑ ΕΚΤΕΤΑΜΕΝΗ ΠΕΡΙΛΗΨΗ ΤΟΥ ΠΑΡΑΠΑΝΩ ΚΕΙΜΕΝΟΥ ΣΤΑ ΕΛΛΗΝΙΚΑ

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΒΙΒΛΙΟΓΡΑΦΙΚΑ ΔΕΔΟΜΕΝΑ

Εργασία που υποβάλλονται για το

Πτυχίο Πληροφορικής και Τηλεματικής

ΠΑΝΑΓΌΠΟΥΛΟΣ ΓΕΩΡΓΙΟΣ

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΟΥΝΙΟΣ 2014

ΠΕΡΙΛΗΨΗ

Τα βιβλιογραφικά δεδομένα είναι ζωτικής σημασίας για την ακαδημαϊκή κοινωνία που αποτελείται από ζωτική ύλη για μελετητές και ερευνητές. Με τα χρόνια, διάφορες προσπάθειες έχουν γίνει, προκειμένου να εκμεταλλευτούν αυτή την πηγή των πληροφοριών με τον καλύτερο δυνατό τρόπο, αποκαλύπτοντας πολλές επιστημονικές καινοτομίες κατά τη διάρκεια αυτής της διαδικασίας, οι οποίες αργότερα εφαρμόστηκαν για άλλους σκοπούς. Από την άλλη πλευρά, έχει υπηρετήσει ως ένα ευημερών πεδίο πειραματισμών σχετικά με διάφορες μεθοδολογίες στον τομέα της πληροφορικής, κυρίως λόγω της πολυμορφίας του. Ως αποτέλεσμα τα βιβλιογραφικά δεδομένα είναι μια ιδανική πάγκο δοκιμών για τον συνδυασμό διαφόρων δεδομένων εξόρυξη μεθοδολογίες και εξόρυξη γνώσης.

Ο στόχος αυτής της εργασίας είναι να εισαγάγει μια νέα προσέγγιση στο πρόβλημα του χαρακτηρισμού της επιτυχίας των συγγραφέων, όχι μόνο με βάση τις δημοσίευσες τους , αλλά και με τις επιπτώσεις της συνεργασίας τους με άλλους συγγραφείς και την εξέλιξη των παραπάνω στο χρόνο. Ως απόδειξη της έννοιας της εργασίας μας, έχουμε επεξεργαστεί τις δημοσιεύσεις όλων των συγγραφέων που συνδέονται με ελληνικά πανεπιστήμια όπως δόθηκε από την ψηφιακή βιβλιοθήκη Scopus.

Πιο συγκεκριμένα, για κάθε συγγραφέα, συλλέγουμε πληροφορίες σχετικά με τον αριθμό των δημοσιεύσεων και τις παραπομπές σε ετήσια βάση. Εκτός από αυτό, έχουμε κατασκευάσει δύο τύπους γράφων συνεργασίας, όπου κάθε συγγραφέας είναι συνδεδεμένο με τον/την συν-συγγραφείς με ακμές που υποδεικνύουν είτε τη δύναμη ή την επίδραση της συνεργασίας. Οι γράφοι έχουν κατασκευαστεί σε ετήσια βάση, καθώς και τα βάρη των ακμών να είναι αθροιστικά με σκοπό τη συγκέντρωση των πληροφοριών από τα προηγούμενα χρόνια. Τέλος, εφαρμόζουμε πολλές τεχνικές εξόρυξης γράφων από την βιολογία και την ανάλυση κοινωνικών δικτύων ώστε να ορίσουμε την κοινωνικότητα των συγγραφέων στους προαναφερθείς γράφους. Έπειτα χρησιμοποιούμε αυτά τα χαρακτηριστικά για να δημιουργήσουμε δείκτες αλλαγών για να συλλάβουμε την εξέλιξη των χαρακτηριστικών συγγραφέων στο χρόνο.

Σε επόμενη φάση, ομαδοποιούμε τους συγγραφείς με παρόμοιο αριθμό δημοσιεύσεων, αναφορών και προφίλ συνεργασιών χρησιμοποιώντας τον k-means αλγόριθμο ομαδοποίησης και μια λίστα από γνωστές μετρικές για αξιολόγηση

ομαδοποίησης ώστε να καθορίσουμε τον καλύτερο αριθμό των ομάδων. Έπειτα δημιουργούνται επιγραφές για κάθε ομάδα με βάση τα πιο εξέχων χαρακτηριστικά τους.

Για να ανακαλύψουμε τα πιο σημαντικά σε πληροφορία χαρακτηριστικά χρησιμοποιούμε στατιστικές μεθόδους. Τέλος κάνουμε μια ομαδοποίηση με βάση χρονολογικές σειρές χρησιμοποιώντας dynamic time warping για να ομαδοποιήσουμε τους συγγραφείς με βάση το χρονολόγιο και συγκρίνουμε τα αποτελέσματα με αυτά της κανονικής ομαδοποίησης.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

Μερικές εισαγωγικές πληροφορίες για τους σκοπούς και τη φιλοσοφία των ερευνών μας.

1.1 ΣΤΟΧΟΙ

Ο σκοπός αυτής της εργασίας είναι η δημιουργία μιας μεθοδολογίας για την ανάλυση των ερευνητικών επιδόσεων σε δημοσίευση άρθρων, αναφορών και συνεργασίες με άλλους ερευνητές. Η μεθοδολογία ελέγχεται σε συγγραφείς συνδεδεμένους με ελληνικά πανεπιστήμια και χρησιμοποιεί πληροφορίες από τη βάση δεδομένων Scopus. Ωστόσο, οποιαδήποτε άλλη ομάδα ερευνητών και οποιαδήποτε άλλα βιβλιογραφική βάση δεδομένων (με δημοσίευση και υποβληθείσες πληροφορίες ανά άρθρο και έτος) θα μπορούσε να χρησιμοποιηθεί. Τα χαρακτηριστικά που εξετάστηκαν για κάθε συγγραφέα, σχετίζονται με τις δημοσιεύσεις του, τις συνεργασίες, τις αναφορές, καθώς και την εξέλιξή αυτών στο χρόνο. Έχουμε συνδυάσει αυτές τις πληροφορίες σε μια σειρά από μετρικές, άλλες πολύ δημοφιλείς και άλλες πρωτότυπες, οι οποίες, αποτελούν στοιχεία μετρήσεων για τις επιδόσεις ενός ερευνητή.

1.2 ΤΕΧΝΙΚΕΣ

Οι μεθοδολογίες που χρησιμοποιούνται στην εξόρυξη γνώσης από τα δεδομένα μας, είναι δημοφιλείς τρόπους που χρησιμοποιούνται συνήθως στις εργασίες εξόρυξης δεδομένων. Η χρήση των γράφων είναι παντού, σε διάφορες μορφές, όπως multigraph, με βάρη η power graphs, από τις οποίες αποσπούμε ορισμένε μέτρα σημαντικότητας ενός συγγραφέας. Επιπλέον βιβλιογραφική δείκτες όπως άρθρα ανά έτος και αναφορές χρησιμοποιούνται, είτε ως είθισται είτε τροποποιημένα ώστε να περιλαμβάνουν μείωση σε σχέση με την παλαιότητα. Η διατριβή αυτή χαρακτηρίζεται από την ουσία του χρόνου και πώς επηρεάζει τα χαρακτηριστικά και τους δείκτες ενός συγγραφέα.

Η ανάλυση μας αποτελείται από τα παρακάτω βήματα:

4. Συλλογή δεδομένων

   Είναι ανάκτηση βιβλιογραφικών πληροφοριών, με βάση την ψηφιακή βιβλιοθήκη Scopus. Το υλικό αφορά έργα από Ελληνικά ιδρύματα και τους αντίστοιχους συγγραφείς. Η βασική μεθοδολογία σε αυτό το στάδιο είναι ο συνδυασμός δομημένων δεδομένων με web crawling μέσα από τις σελίδες της

ψηφιακής βιβλιοθήκης, και η αποθήκευση πληροφοριών σχετικά με τα στιχειοθετημένα άρθρα και συγγραφείς.

5. Προ-επεξεργασία δεδομένων

   Τα δεδομένα χωρίζονται σε χρόνια, έτσι για κάθε έτος δύο τύποι co-δημιουργός γράφων κατασκευάστηκαν, έναν για την καταγραφή της κοινωνικής "δύναμης" ενός συγγραφέα και ένα για την ποιότητα των συνεργασιών. Κάθε ετήσιος γράφος περιέχει πληροφορίες του αντίστοιχου έτους και των προηγούμενων, τροποποιημένο κατάλληλα ώστε να επιτυγχάνετε η μείωση λόγο παλαιότητα στις ακμές. Οι πληροφορίες στις ακμές περικλείουν αναφορές, αριθμό συν-συγγραφέων και παλαιότητα αφενός, αφετέρου τον αριθμό των άρθρων που γράψαν οι δύο συγγραφείς. Από αυτούς τους γράφους και την βάση δεδομένων, θα βγουν ορισμένα χαρακτηριστικά για κάθε έτος που αποτυπώνουν το προφίλ του συγγραφέα το συγκεκριμένο έτος. Τα χαρακτηριστικά αφορούν τις διάφορες μορφές των αναφορών, των άρθρων, των επιπτώσεων μιας συνεργασίας, της σημασίας του συγγραφέα λόγω της θέσης του στο συγγραφικό δίκτυο κλπ. Για να υπολογιστούν αυτά, χρειάστηκε να εφαρμοστούν ορισμένες μέθοδοι εξόρυξης γράφων όπως powergraph και eigenvector ανάλυση. Ύστερα χρησιμοποιούμε μετρικές αλλαγών σε αυτά για να αιχμαλωτίσουμε την εξέλιξη του ερευνητή στο χρόνο.

6. Εξόρυξη γνώσης

   Η συσταδοποίηση (clutering) των συγγραφέων είναι το επιθυμητό αποτέλεσμα της ανάλυσης μας. Αυτό επιτεύχθηκε μέσω του dataset με τους δείκτες αλλαγής και με αλγόριθμο K-means. Ο ιδανικός αριθμός κατηγοριών καθορίστηκε εκτελώντας πειράματα με τον αλγόριθμο και μετρώντας ένα σύνολο από δείκτες αξιολόγησης, όπως ο δείκτης Dunn (Dunn 1973) , Davies-Bouldin (Davies and Bouldin 1979), η μέση απόσταση από το κέντρο ενός cluster και η απόσταης μεταξή των cluster. Ο χαρακτηρισμός των clusters βασίστηκε σε επιφανείς τιμές των χαρακτηριστικών τους. Επίσης μια διαδικασία επιλογής χαρακτηριστικών διεξήχθη, χρησιμοποιώντας της στατιστική τεχνική singular value decomposition σε μια απλουστευμένη μορφή του dataset, προκειμένου να καθορίσουν τις πιο εντυπωσιακά στο clustering χαρακτηριστικά. Τέλος, κατασκευάστηκε ένα dataset με τις χρονοσειρές των συγγραφέων για το πιο εξέχων χαρακτηριστικό, για να γίνει μια συσταδοποίηση με χρονοσειρές χρησιμοποιώντας το dynamic time warping ως μέτρο απόστασης των

χρονοσειρών και τον αλγόριθμο partitioning around medoids. Τα αποτελέσματα αξιολογούνται σε σχέση με την πρώτη συσταδοποίηση και καταγράφετε το ποσοστό επιτυχίας.

Συνοπτικά, οι κυριότερες συνεισφορές της παρούσας εργασίας:

- Κατασκευή dataset με τους ερευνητές των Ελληνικών ιδρυμάτων χρησιμοποιώντας εξελιγμένες μεθόδους. Για παράδειγμα ένα μέρος αυτής της φάσης απαιτούσε την αναγνώριση αγνώστων συγγραφέων. Οι αναγνωρισμένοι συν-συγγραφείς ενός αγνώστου συγγραφέα, μας επέτρεψαν να κατασκευάσουμε ένα μηχανισμό ταυτοποίησης για την επίλυση, που αντιστοιχεί έναν υποψήφιο id για έναν άγνωστο συγγραφέα, με βάση τη συχνότητα συνεργασίας με τους αναγνωρισμένους συν-συγγραφείς.
- Ο παράγοντας της παλαιότητας σε κάθε συνεργασίας, και η αναλογική μείωση στα βάρη των γράφων. Η απεικόνιση της παλαιότητας σε βιβλιογραφικά μέτρα (π. χ. αναφορές τιμωρούμενες ανά έτος). Δείκτες αλλαγής σε βιβλιογραφικές μέτρα.
- Αξιολόγηση μεταξύ συσταδοποίησης με χρονοσειρές και απλής συσταδοποίησης με δείκτες αλλαγών.

1.3 ΔΟΜΗ ΕΡΓΑΣΙΑΣ

Η πτυχιακή είναι οργανωμένη ως εξής: Ενότητα 2 παρουσιάζει ορισμένες στοιχειώδεις έννοιες με τις οποίες που ο αναγνώστης θα πρέπει να είναι εξοικειωμένος, προκειμένου να κατανοήσει πλήρως τις φάσεις της ανάλυσης. Κεφάλαιο 3 παρουσιάζεται το θεωρητικό υπόβαθρο των μεθόδων που αξιοποιήθηκαν, ορισμένες από τις οποίες είναι ήδη γνωστές και επιφανείς τεχνικές στον τομέα της πληροφορικής, άλλες είναι νεότερες και τα υπόλοιπα είναι ιδέες που προτείνουμε. Κεφάλαιο 4 συνοψίζονται τα συμπεράσματα της έρευνας και δίνονται κατευθύνσεις για μελλοντική εργασία.

2. ΤΑ ΒΑΣΙΚΑ

Αυτή η ενότητα παρέχει ορισμένες βασικές πληροφορίες σχετικά με τις τεχνολογίες και διαδικασίες, τις οποίες χρησιμοποιήσαμε και είναι αναγκαίες, για την κατανόηση του υπόλοιπου της διατριβής.

2.1 SCOPUS

Η ανάλυση επικεντρώνεται στην ερευνητική δραστηριότητα ορισμένων ελληνικών πανεπιστήμιων τα τελευταία δεκαπέντε χρόνια. Στην προσπάθειά μας να εξασφαλίσουμε ότι τα αποτελέσματα της εργασίας μας θα είναι όσο το δυνατόν πιο αξιόπιστα, συγκεντρώθηκαν στοιχεία από μια από τις πιο επιφανής και αξιόλογες πηγές των βιβλιογραφικών δεδομένων, η ψηφιακή βιβλιοθήκη Scopus.

2.1.1 ΓΕΝΙΚΑ

Το Scopus είναι μία βιβλιογραφική βάση δεδομένων, η οποία περιέχει τίτλους, συγγραφείς, και αναφορές για τα ακαδημαϊκά άρθρα. Καλύπτει περίπου 21.000 τίτλους (περιοδικά και συνέδρια) πάνω από 5.000 εκδότες, όσον αφορά τις επιστημονικές, τεχνικές και ιατρικές και κοινωνικές επιστήμες. Είναι στην ιδιοκτησία της Elsevier, που είναι μια αξιοσέβαστη εκδοτική εταιρεία, και είναι διαθέσιμη στο διαδίκτυο με συνδρομή. Η αναζήτηση στο Scopus ενσωματώνει αναζητήσεις επιστημονικών ιστοσελίδες μέσω του Scirus, ένα άλλο προϊόν της Elsevier (Kulkarni, et al. 2009). Επιπλέον, το Scopus προσφέρει προφίλ συγγραφέωνλ, που περιέχουν τα ιδρύματα σε οποία ανήκουν, τον αριθμό των δημοσιεύσεων και βιβλιογραφικές μετρικές όπως οι αναφορές μας, ο αριθμός των αναφορών σε κάθε έγγραφο καθώς και analytics για να παρουσιάσει μια γενική εικόνα της καριέρας του συγγραφέα. Σε σύγκριση με άλλες ανάλογες βιβλιοθήκες, το Scopus προσφέρει 20% μεγαλύτερη κάλυψη από το Web της επιστήμης, καλύπτει ένα ευρύτερο φάσμα των εφημερίδων από PubMed και έχει πιο συνεπή αποτελέσματα από το Google Scholar, που παρέχει πιο ανεπαρκείς, λιγότερο συχνά ενημερώμένες πληροφορίες (Falagas, et al. 2008) (Erten, et al. 2004).

2.1.2 ΛΕΙΤΟΥΡΓΙΕΣ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ

Το Scopus προσφέρει δεδομένα άρθρων, χρησιμοποιώντας διάφορες μεθόδους (π. χ. API, δομημένα αρχεία, περιήγηση). Το όριο για τη λήψη αρχείου με πληροφορίες σε

μορφή csv είναι 20.000 δημοσιεύσεις.. Τα χαρακτηριστικά της κάθε δημοσίευσης είναι το έτος που δημοσιεύθηκε, το όνομα των συγγραφέων (χωρίς αναγνωριστικά id), το ISSN και ο τίτλος της εφημερίδας, ο αριθμός των σελίδων, οι συνολικές αναφορές από τη δημοσίευση ως το έτος 1997, οι αναφορές ανά έτος από το 1998 και μετά. Αν και μεγάλο σε όγκο, από αυτό το σετ δεδομένων έλειπε μια ζωτική ουσία, τα id των συγγραφέων, καθιστώντας τις πληροφορίες ελλιπείς. Αυτό οφείλεται στο γεγονός ότι στην περίπτωση συνωνυμιών, θα μπορούσαμε να μπερδέψουμε πληροφορίες για δύο διαφορετικούς ερευνητές σε έναν.

2.2 ΣΥΝ-ΣΥΓΓΡΑΦΙΚΟΙ ΓΡΑΦΟΙ

Οι γράφοι συνεργασίας (Odda 1979) χρησιμοποιούνται ευρέως στον τομέα των μαθηματικών, των κοινωνικών επιστημών και της πληροφορικής, κυρίως στον τομέα της ανάλυσης κοινωνικών δικτύων. Γράφος συνεργασίας είναι μια δομή γράφου, όπου οι κόμβοι αναπαριστούν άτομα που συνεργάζονται μεταξύ τους και η ακμή που τους συνδέει δείχνει μια συνεργατική σχέση μεταξύ τους. Σε γενικές γραμμές, μπορεί κάποιος να ισχυριστεί ότι η χρήση των γράφων είναι αρκετά ευέλικτη και μπορεί να χρησιμεύσει στην αποτύπωση πολλών πληροφοριών για ένα δίκτυο μεταξύ συγγραφέων.

2.2.1 ΒΑΡΗ ΑΚΜΩΝ

Το βάρος των ακμών μπορεί να αντιπροσωπεύει διάφορα πράγματα. Οι αναφορές που έχουν λάβει μέχρι στιγμής, η τον αριθμό των άρθρων που έχουν συγγράψει δύο ερευνητές είναι δύο προφανή παραδείγματα. Το βάρος της ακμής είναι συνήθως ανάλογο με τη φύση του γράφου, με την έννοια ότι ένας απλός γράφος με βάρος τις αναφορές ενός άρθρου στις ακμές, θα μετατραπεί σε ένα multigraph, όταν οι δύο συγγραφείς γράψουν ένα καινούργιο άρθρο , δεδομένου ότι μια νέα ακμή μεταξύ τους θα προστεθεί. Από την άλλη πλευρά, αν διατηρούμε μία ακμή για κάθε συγγραφικό ζεύγος, τότε είτε χάνουμε τα στοιχεία της δεύτερης δημοσίευση, ή τροποποιούμε το βάρος της ακμής π. χ. το άθροισμα των αναφορών που οι δύο συγγραφείς έχουν πάρει. Ένας εναλλακτικός τρόπος που χρησιμοποιείται σε πολλές περιπτώσεις είναι ένα hyper γράφημα (O'Madadhain, Hutchins και Smyth 2005) .

2.2.2 ΚΑΤΕΥΘΗΝΣΗ ΑΚΜΗΣ

Γενικά οι ακμές μπορεί να είναι με κατευθυνόμενες η κατευθυνόμενες. Όταν μιλάμε για συνεργατικούς γράφους, συνήθως είναι μη κατευθυνόμενοι γιατί η ακμή είναι αμοιβαία. Οι κατευθυνόμενες ακμές σε έναν γράφο μπορούνε να υποδεικνύουν τις αναφορές που έχει κάνει ένα άρθρο σ ένα άλλο. Αυτό το μοτίβο αναγνωρίζεται περισσότερο σε ένα άλλο δημοφιλές βιβλιογραφικό γράφο, ένα δίκτυο αναφορών (An, Janssen και Milios 2004) Στην περίπτωση αυτή, οι κόμβοι είναι άθρα και οι ακμές είναι αναφορές του ενός άρθρου στο άλλο. Μπορεί να βοηθήσει στην αποκάλυψη συγγραφικής ομοιογένειας, ομοιότητα μεταξύ εγγράφων (Martyn 1964) κλπ.

2.2.3 ΔΕΙΚΤΕΣ ΓΡΑΦΩΝ

Είναι οι κυρίως τυπικοί δείκτες για της ανάλυσης κοινωνικών δικτύων όπως normalized degree (Borgatti και Everett 1999) , betweenness (Freeman 1977) κλπ. Πρόσφατα pio ειδικές φόρμουλες έχουν προταθεί για συγκεκριμένη χρήση σε βιβλιογραφικούς γράφους. Το Eigenfactor (West και Wiseman, The Eigenfactor Metrics 2008), η σημασία της ενός άρθρου ή ενός περιοδικού με βάση το ποσό και το ποσοστό των αναφορών που λαμβάνει από έγκριτα άρθρα ή εφημερίδες. Μολονότι αρχικά προτάθηκε για δίκτυα αναφορών, έχει εφαρμοστεί και σε συγγραφικά δίκτυα (West, Jensen, και συν. 2013) (με την ίδια φιλοσοφία).

2.3 POWER GRAPHS

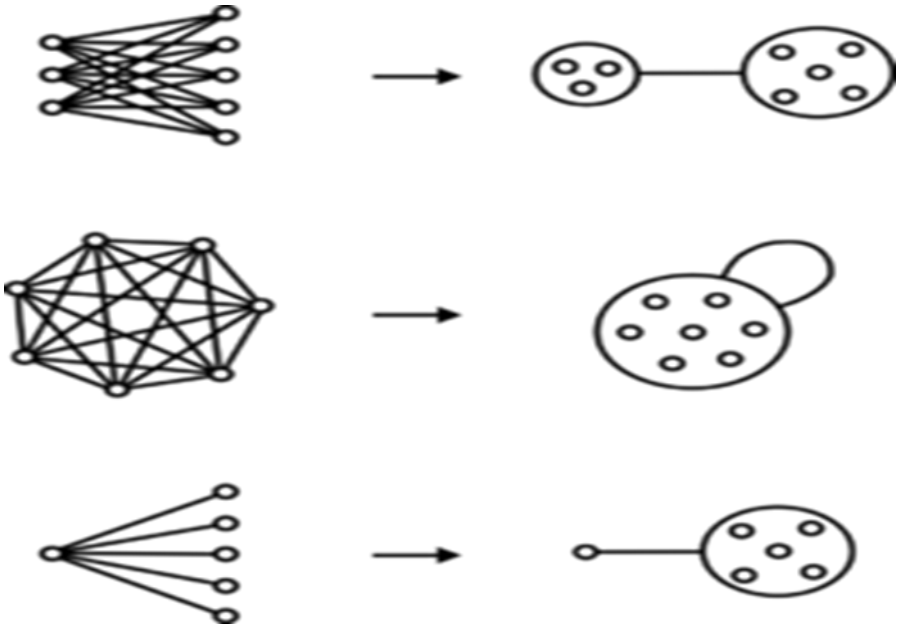
Η power graph ανάλυση χρησιμοποιείται κυρίως στον τομέα της βιοπληροφορικής για οπτικοποίηση πολύπλοκων δικτύων χωρίς την απώλεια πληροφοριών (Royer, και συν. 2008). Αυτή η μεθοδολογία έχει εφαρμοστεί με επιτυχία και σε συγγραφικά δίκτυα στο παρελθόν, (Tsatsaronis, και συν. 2011) (Varlamis και Tsatsaronis 2012).

2.3.1 ΟΡΙΣΜΟΣ

Η power graph ανάλυση αναγνωρίζει βασικά μοτίβα σε έναν γράφο ,όπως αστέρι, κλίκα και biclique και τα χρησιμοποιεί για να κατασκευάσει μια πιο συμπαγής αναπαράσταση του γραφήματος. Αυτό επιτυγχάνεται με τη χρήση power κόμβων, τα οποία είναι ένας κύκλος που περικλείει κόμβους ή κόμβους ισχύος, και power ακμές,

τις ακμές μεταξύ των power κόμβων. Η μεταμόρφωση των μοτίβων σε power κόμβους απεικονίζεται στην Εικόνα 1 και ακολουθεί την εξής μεθοδολογία:

- Το biclique είναι δύο σύνολα των κόμβων με μια ακμή κάθε κόμβου του ενός σετ με κάθε κόμβο στο άλλο. Σε ένα power graph, ένα biclique απεικονίζεται ως δύο power κόμβοι που αποτελούνται από τους κόμβους στα δύο αρχικά σύνολα, και ένα power edge μεταξύ τους.
- Οι κλίκες είναι ένα σύνολο από κόμβους με μια ακμή από κάθε κόμβο σε κάθε άλλο κόμβο. Στον power graph, μια κλίκα αντιπροσωπεύεται από έναν power κόμβο που περιέχει όλους τους συγγραφείς που συνδέονται μεταξύ τους, με μια αυτοακμή.
- Τα αστέρια είναι ένα σύνολο από κόμβους και ένας άλλος κόμβος, με μια ακμή μεταξύ κάθε κόμβου στο πρώτο σύνολο και στον ξεχωριστό κόμβο. Σε ένα power graph, ένα αστέρι αντιπροσωπεύεται από ένα power edge ανάμεσα σε ένα κανονικό κόμβο και έναν power κόμβο που έχει όλους τους κόμβους με τους οποίους είναι συνδεδεμένος ο ξεχωριστός κόμβος στον κανονικό γράφο.



Εικόνα 23: ΜΟΤΙΒΑ ΓΡΑΦΩΝ ΣΤΑ POWER GRAPHS

Εικόνα 24: ΜΟΤΙΒΑ ΓΡΑΦΩΝ ΣΤΑ POWER GRAPHS

2.4 ΟΜΑΔΟΠΟΙΗΣΗ

Ομαδοποίηση, (cluster analysis) είναι η διαδικασία κατά την οποία ένα σύνολο αντικειμένων διασπάται μικρότερες ομάδες, με βάση την ομοιότητα των αντικειμένων στο εσωτερικό κάθε επιμέρους ομάδας και της ανομοιομορφίας τους με τα αντικείμενα στις υπόλοιπες ομάδες. Οι ομάδες αυτές ονομάζονται cluster. Είναι μια κοινή τεχνική που χρησιμοποιείται για εξόρυξη δεδομένων, μηχανική μάθηση, στατιστική ανάλυση δεδομένων κλπ.

2.4.1 ΑΛΓΟΡΙΘΜΟΙ CLUSTERING

Οι αλγόριθμοι ομαδοποίησης ταξινομούνται σύμφωνα με τα μοντέλα των cluster τους.

- Centroid μοντέλα: Στα centroid μοντέλα τα κέντρα των clusters αναπαριστώνται σαν σημεία μέσων τιμών, τα οποία μπορεί να μην είναι ένα αντικείμενο στο σύνολο δεδομένων π. χ. K-means (MacQueen 1967) .
- Connectivity μοντέλα: Μοντέλα με βάση το connectivity, δημιουργούν ομάδες με βάση την απόσταση των παρατηρήσεων π. χ. Hierarchical (Johnson 1967).
- Distribution μοντέλα: Στα distribution μοντέλα Οι συστάδες ορίζονται με την αξιοποίηση στατιστικών κατανομών των δεδομένων (Xiaofei, et al. 2011).
- Density μοντέλα: Τα density μοντέλα δημιουργούν συμπλέγματα με βάση την πυκνότητα των τιμών των παρατηρήσεων, στο χώρο των δεδομένων π. χ. DBSCAN (Martin, et al. 1996).

2.4.2 ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΟΜΑΔΟΠΟΙΗΣΗΣ

Η αξιολόγηση μια συσταδοποίησης αναφέρεται στο βαθμό που οι συστάδες που παράγονται από τη διαδικασία δημιουργίας συμπλεγμάτων, είναι πράγματι μια υπάρχουσα δομή στα αρχικά δεδομένα μας, οπότε η συσταδοποίηση ήταν επιτυχής. Οι μετρικές εγκυρότητας που επιλέγονται σε πολλές περιπτώσεις εξαρτώνται από την φύση της ανάλυσης των δεδομένων και τις προθέσεις (Halkidi, Batistakis and Vazirgiannis 2001). Γενικά εσωτερικές μετρικές αξιολόγησης διασχίζουν τα clusters στο χώρο των δεδομένων και μετράνε τις τιμές των εξισώσεων που βασίζονται στις αποστάσεις μεταξύ των παρατηρήσεων, των κέντρων του cluster, της δομής, του μεγέθους και της πυκνότητας του. Π. χ. Silhouette (Rousseeuw 1987). Τα εξωτερικά μέτρα, από την άλλη πλευρά χρησιμοποιούν δεδομένα που έχουν κατηγοριοποιηθεί

από πριν (επιτηρούμενη μέθοδος μηχανικής μάθησης). Οι αλγόριθμοι συσταδοποίησης τρέχουν και τα αποτελέσματα αξιολογούνται με βάση το πόσο κοντά είναι στις προκαθορισμένες ετικέτες π. χ. Rand Measure (Rand 1971).

3. ΜΕΘΟΔΟΙ

Ειδικές μέθοδοι που θα αξιοποιήθηκαν, τροποποιήθηκαν ή δημιουργήθηκαν κατά τη διάρκεια ερευνών μας.

3.1 ΟΡΙΣΜΟΙ

Οι μετρικές που αναφέρουμε στο υπόλοιπο στάδιο των μεθόδων, περιέχουν ειδικά μέτρα που θα πρέπει να διευκρινιστούν:

- $P_i = Τα\ άρθρα\ του\ ερευνητή\ i.$
- $t_n = Ο\ χρόνος\ που\ εξετάζουμε.$
- $t_0 = Το\ 1998, ο\ χρόνος\ που\ ξεκίναει\ η\ καταγραφή\ των\ δεδομένων\ μας.$
- $tz = Το\ 2005, ο\ τελευταίος\ χρόνος\ που\ εξετάζουμε.$
- $pap_i(j) =$
  $Ο\ αριθμός\ των\ άρθρων\ που\ έχει\ γράψει\ έναν\ ερευνητής\ i\ το\ έτος\ j.$
- $cit(i,j) = Ο\ αριθμός\ των\ αναφορών\ που\ έλαβε\ το\ άρθρο\ i\ το\ έτος\ j.$
- $aut(i) = Ο\ αριθμός\ των\ συγγραφέον\ ενός\ άρθρου\ i.$
- $year(i) = Το\ έτος\ δημοσίευσης\ του\ άρθρου\ i.$
- $N_x = Οι\ γειτονικοί\ κόμβοι\ του\ κόμβου\ x.$
- $E(i,x) =$
  $Το\ βάρος\ της\ ακμής\ μεταξυ\ του\ x\ και\ του\ i\ στον\ ποιοτικό\ γράφο.$
- $PN_x = Οι\ power\ κόμβος\ που\ είναι\ γείτονες\ στον\ power\ κόμβο\ x.$
- $PC_x = Οι\ power\ κόμβο\ που\ περιέχουν\ τον\ power\ κόμβο\ x.$
- $PW(x) = Το\ βάρους\ του\ power\ κόμβο\ x.$
- $PE(y,x) =$
  $Το\ βάρος\ της\ power\ ακμής\ που\ συνδέει\ τον\ power\ κόμβο\ x\ και\ τον\ power\ κόμβο\ y.$
- $Cl = Ενα\ σύνολο\ από\ συστάδες.$
- $c_i = Το\ κέντρο\ της\ i\ συστάδας.$
- $d(i,j) =$
  $Η\ απόσταση\ από\ το\ σημείο\ i\ στο\ σημείο\ j\ στον\ χώρο\ των\ δεδομένων.$
- $\lambda =$
  $Μια\ σταθερά\ του\ eigenvector\ centrality, που\ εξαρτάται\ από\ τον\ τρόπο\ ομαλοποίησης.$

3.2 ΓΡΑΦΟΙ

Η χρήση συνεργατικής γράφων μας επέτρεψε να αποτυπώσουμε κάποιες πληροφορίες των ερευνητών σε διάφορες διαστάσεις, όπως αντοχή στο χρόνο και η κοινωνική σημασία. Για το σκοπό αυτό εκμεταλλεύτηκαμε την ευελιξία που παρέχουν οι γράφοι στις ιδιότητες των βαρών στις ακμές, καθώς και το γεγονός ότι οι γράφοι επεκτείνονται όσο αυξάνουν τα χρόνια, παράγουν σημαντικά στοιχεία για τις δραστηριότητες ενός συγγραφέα.

3.2.1 ΕΞΕΛΙΚΤΙΚΟΣ ΓΡΑΦΟΣ

Οι συν-συγγραφικοί γράφοι εξελίσσονται στον χρόνο. Αυτό σημαίνει ότι ο γράφος κάθε έτους, δεν περιλαμβάνει μόνο τις συνεργασίες αυτού του έτους, αλλά συσσωρεύονται όλες από τις προηγούμενες χρονιές **μέχρι** το τρέχων. Συνεπώς, εάν δύο συγγραφείς έχουν συνεργαστεί σε ένα άρθρο το 1999 και σε δύο το έτος 2000, ο γράφος του έτους 2000 θα περιέχει το άρθρο του έτους 1999 και τα δύο άρθρα που του 2000. Ίδιο πράγμα ισχύει και για τις αναφορές που έχει δεχθεί μια συνεργασία. Οι αναφορές είναι αθροιστικές, πράγμα που σημαίνει ότι, καθώς ο χρόνος περνά, ο αριθμός των αναφορών που γίνονται προς ένα άρθρο από άλλα άρθρα μπορεί να αυξηθεί ή να παραμείνει σταθερός, επειδή οι αναφορές δεν μπορούν να διαγραφούν. Οι γράφοι που δημιουργήθηκαν αποτυπώνουν την σημασία του άρθρου την δεδομένη χρονική στιγμή, συνεπώς είναι σημαντικό να λάβουμε υπόψιν μας και τις αναφορές που έχει πάρει μέχρι εκείνο τον χρόνο. Για παράδειγμα, ένα άρθρο που γράφτηκε κατά το έτος 1999 θα εμφανίζετε στο 1999-γράφο με τον αριθμό των αναφορών που έλαβε το 1999 . Κατά τον 2000-γράφο οι αναφορές που λαμβάνονται υπόψη είναι εκείνες που έγιναν μέχρι το έτος 2000, δηλαδή και οι αναφορές του = 1999 **και** οι αναφορές του 2000. Αυτό ισχύει και για τα υπόλοιπα έτη.

3.2.2 ΒΑΡΗ ΑΚΜΩΝ

Ενώ η δομή του γράφου αποτυπώνει την κοινωνική δύναμη του συγγραφέα, το βάρος των ακμών περιέχει μια άλλη ουσία, ζωτικής ουσία στην ανάλυσή μας, το πραγματικός αντίκτυπο μιας συνεργασίας. Ένα συνεργασία έχει αντιστοιχιστεί σε μια ακμή μεταξύ των δύο συγγραφέων στον ετήσιο γράφο. Αυτό σημαίνει ότι αυτή η ακμή θα πρέπει να περιέχει τα στοιχεία όλων των άρθρων που συνέγραψαν μαζί οι δύο συγγραφείς, μέχρι

το χρόνο που αντιστοιχεί στον γράφο. Αυτή η συνάθροιση γνώσης κρίσιμο ζήτημα, δεδομένου ότι πρέπει να ληφθούν υπόψη κάθε πτυχή του κάθε έγγραφου. Για τις ανάγκες την ανάλυσης, δημιουργήσαμε ένα ζευγάρι γράφων με διαφορετικό βάρος στις ακμές, ώστε να αποτυπώσουμε τις πολλές διαστάσεις του προβλήματος.

3.2.2.1 ΠΟΣΟΤΙΚΟ ΒΑΡΟΣ ΑΚΜΗΣ

Η ακμή βάρος του ποσοτικού γράφου αντιπροσωπεύει τον όγκο της συνεργασίας (CV) των δύο δημιουργών x, y. Αντιπροσωπεύει το πόσες φορές έχουν συνεργαστεί ο ερευνητής x και ο ερευνητής y μέχρι στιγμή t_n .

$$CV(x,y) = \sum_{\forall i \in P_x \cap P_y} 1 \qquad ,year(i) <= t_n \tag{1}$$

3.2.2.2 ΠΟΙΟΤΙΚΟ ΒΑΡΟΣ ΑΚΜΗΣ

Η ακμή βάρος του ποιοτικού γράφου αντιπροσωπεύει την επιτυχία της συνεργασίας (CI) των δύο δημιουργών x, y . Είναι το συνολικό αντίκτυπο μιας λίστα από έγγραφα που συνέγραψαν οι δύο συγγραφείς και ορίζεται ως:

$$CI(x,y) = \sum_{\forall i \in P_x \cap P_y} \frac{\alpha * \sum_{j=t_0}^{t_n} cit(i,j) + \beta}{aut(i) * \left(1 + t_n - year(i)\right)} \qquad ,year(i) <= t_n \tag{2}$$

Η ουσία αυτού του μοντέλου είναι ότι η επιτυχία ενός άρθρου την δεδομένη χρονική στιγμή t_nείναι ανάλογη με τον αριθμό των αναφορών που έχει λάβει είχε μέχρι αυτή την ώρα, αντίστροφος ανάλογη με τον αριθμό των συγγραφέων που συμμετείχαν στην συγγραφή του και με τη διαφορά του έτος που είχε δημοσιευθεί με το t_n, το οποίο απεικονίζει πόσο παλιό είναι το άρθρο στο χρόνο t_n. Το +1 στο παρονομαστή καλύπτει την περίπτωση όπου $t_n = year(i)$ (Όταν το βιβλίο είναι γραμμένο για το τρέχον έτος t_n). Για κάθε δύο συγγραφείς , το άθροισμα της επιτυχίας όλων των κοινών τους άρθρων είναι η επιτυχία της συνεργασίας τους σε μια δεδομένη χρονική στιγμή t_n. Η επιλογή των τιμών α και β, υποδηλώνει το ενδιαφέρον για τις επιπτώσεις του συγγραφέα της εργασίας (α) ή στην ποσότητα του/της δημοσιεύσεις (βήτα). Σε πειράματα μας, αποφασίζουμε να ρυθμιστεί $\alpha = 0{,}7$ Και $\beta = 0{,}3$ Αλλά αυτό σίγουρα χρειάζεται περαιτέρω περάματα.

3.2.3 ΧΑΡΑΧΤΗΡΙΣΤΙΚΑ ΑΠΟ ΤΟ ΓΡΑΦΟ

Κατά τη διάρκεια της διαδικασίας για τη δημιουργία ενός dataset με χαρακτηριστικά των συγγραφέων , προκειμένου να τους κατηγοριοποιήσουμε με βάση αυτά τα χαρακτηριστικά, χρησιμοποιήσαμε τεχνικές εξόρυξης γράφων για να καταγράψουμε τις κοινωνικές επιπτώσεις του ερευνητή. Αυτές οι τεχνικές, που προέρχονται από την ανάλυση κοινωνικών δικτύων και την βιοπληροφορική, οδήγησαν στην δημιουργία δεικτών που περιγράφουν στοιχεία του ερευνητή για τη δύναμη που κατέχει στο δίκτυο. Λόγω της μείωσης με βάση την παλαιότητα της συνεργασίας, οι μετρικές παίρνουν έναν πιο επίκαιρο χαρακτήρα. Επιπλέον, η ισχύς του συγγραφέα στην συγγραφική του ομάδα καθώς και στην εκτεταμένη κοινότητα (οι συν-συγγραφείς των συν-συγγραφέων του) περιλαμβάνονται, τόσο από την ποσοτική όσο και από την ποιητική προοπτική. Προφανή μέτρα, όπως ο αριθμός των συν-συγγραφέων και η θέση του ερευνητή στο δίκτυο λήφθηκαν επίσης υπόψη.

3.2.3.1 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΠΟΙΟΤΙΚΟΥ ΓΡΑΦΟΥ

Επιλέξαμε να εξάγουμε χαρακτηριστικά από τον ποιοτικό γράφο, κυρίως επειδή περιλαμβάνει την επιτυχία μια συνεργασίας. Δείκτες που έχουν να κάνουν με την δομή του γράφου, είναι η ίδιοι και στον ποσοτικό και στον ποιοτικό.

3.2.3.1.1 EIGENVECTOR CENTRALITY

Το eigenvector centrality (Bonacich, Factoring and weighting approaches to status scores and clique identification 1972) (Bonacich, Some unique properties of eigenvector centrality 2007) για έναν συγγραφέα x ($Eigen(x)$) αντιστοιχεί στην δύναμη που έχει ο συγγραφέα λόγο της θέσης του στο συγγραφικό γράφο σε μια δεδομένη στιγμή. Με βάση το γεγονός ότι η ακμή από έναν επιφανή συγγραφέα είναι πιο σημαντική από μια από κάποιον άσημο, το eigenvector εκχωρεί αρχικά ίσες τιμές σε κάθε συγγραφέα και στη συνέχεια της ξαναυπολογίζει εκ νέου βασιζόμενο στις συνεργασίες. Αυτή η επαναληπτική διαδικασία ολοκληρώνεται όταν οι τιμές συγκλίνουν.

$$Eigen(x) = \frac{1}{\lambda}\sum_{t\in N_x} Eigen(t) \qquad (3)$$

3.2.3.1.2 ΒΑΘΜΟΣ

Βαθμός (Borgatti and Everett 1999) για έναν συγγραφέα x $(Deg(x))$ είναι ο αριθμός των ακμών που έχει μια δεδομένη στιγμή, η αλλιώς πόσους συν-συγγραφείς έχει.

$$Deg(x) = |N_x| \quad (4)$$

3.2.3.1.3 ΒΑΡΟΣ ΣΥΝΕΡΓΑΣΙΑΣ

Το βάρος συνεργασίας (CLW) του ερευνητή x είναι το άθροισμα των βαρών των ακμών του συγγραφέα. Η ερμηνεία αυτή είναι η γενική ποιότητα των συνεργασιών του συγγραφέα σε μια δεδομένη στιγμή.

$$CLW(x) = \sum_{\forall i \in N_x} E(i,x) \quad (5)$$

3.2.3.2 POWER GRAPH

Επειδή οι γράφοι μας ήταν ιδιαιτέρως πυκνοί χρησιμοποιήσαμε power graph analysis που είναι ειδικευμένη στην εξόρυξη γνώσης από πυκνούς γράφους, κυρίως στη βιοπληροφορική. Εφαρμόσαμε αυτήν την μεθοδολογία για να βγάλουμε κάποιες επιπλέον πληροφορίες για την γενικότερη εικόνα του δικτύου κάθε συγγραφέα. Και οι δύο γράφοι μετατράπηκαν σε power graphs. Έτσι, δημιουργήθηκαν ζευγάρια του ίδιο χαρακτηριστικού, καθένα από τα οποία αντιστοιχεί στην ποιοτική και την ποσοτική πλευρά αντίστοιχα. Οι πληροφορίες που μπορούμε να εξαγάγουμε από ένα power graph, αφορούν στην συνεργασία ενός συγγραφέα με ισχυρά άτομα ή ομάδες. Μπορούμε, επίσης, να πάρουμε μια αίσθηση της εκτεταμένης συγγραφικής κοινότητας στην οποία ανήκει ο συγγραφέας. Έτσι, μπορούμε να αναμένουμε αυξημένα χαρακτηριστικά για τους δημιουργούς που ανήκουν σε επιφανείς ομάδες ή αποτελούν μέρος μιας γενικά επιτυχημένης επιστημονικής κοινότητας, σαν ένα επιτυχημένο ερευνητικό ίδρυμα.

3.2.3.2.1 ΒΑΡΟΣ POWER ΚΟΜΒΟΥ

Το βάρος (PN_{weight}) του power κόμβου στο οποίο ανήκει ο συγγραφέας x, που συμβολίζει την επιτυχία και την ισχύ του δεσμού της στενής συγγραφικής του ομάδας.

$$PN_{weight}(x) = PW(x) \quad (6)$$

3.2.3.2.2 ΒΑΡΟΣ POWER ΚΛΙΚΑΣ

Το βάρος της power κλίκας (PN_{clique}) αντιπροσωπεύει την δύναμη της εκτεταμένης κοινότητας στην οποία ανήκει ο συγγραφέας, δηλαδή τους συν-συγγραφείς των συν-συγγραφέων του.

$$PN_{clique}(x) = \sum_{\forall i \in PN_x} PE(i,x) * \mathrm{PW}(i) + \sum_{\forall j \in PC_x} PW(j) \qquad (7)$$

3.3 ΕΠΙΔΡΑΣΗ ΤΟΥ ΧΡΟΝΟΥ

Όπως αναφέρθηκε και ανωτέρω, ο χρόνος διαδραματίζει σημαντικό ρόλο στην ανάλυσή μας. Κάθε τμήμα της εξόρυξης γνώσης έχει μια πτυχή σχετική με το χρόνο. Σε κάθε περίπτωση, προσπαθήσαμε να ερμηνεύσουμε με τον καλύτερο τρόπο τη διάσταση του χρόνου, είτε με συντηρητικές μεθόδους, ή με καινοτόμες προσεγγίσεις. Από τη δική μας σκοπιά, όσο ένα συμβάν παλιώνει, τόσο υποβαθμίζετε η επίπτωση του. Αυτό απορρέει από το γεγονός ότι όσο περνά ο καιρός, πιο καινοτόμες ιδέες ξεπερνούν τις παλιές, αφήνοντας τις στην ιστορία. Βεβαίως στην επιστήμη κάθε γνώμη είναι χρήσιμη και μπορεί να επανεξετάζεται. Αυτός είναι ο λόγος για τον οποίο εφαρμόζουμε την θεωρία μας και στις αναφορές που δέχεται ένα άρθρο, και όχι μόνο στο έτος δημοσίευσης του, κατ' αυτόν τον τρόπο τα πραγματικά επιτυχημένα παλιά έργα παραμένουν στην κορυφή, αφού συνεχίζουν να λαμβάνουν αναφορές. Είναι σημαντικό να διευκρινιστεί ότι αυτή η θεωρία επιστρατεύετε για την καλύτερη κατηγοριοποίηση επιστημόνων, και όχι για την αξιολόγηση των ίδιων των έργων αυτών καθαυτών. Η καινοτομία έγκειται στην τιμωρία ανάλογα με την παλαιότητα, που δεν είναι τόσο συχνή σε βιβλιογραφικές έρευνες. Η κοινή προσέγγιση για την παλαιότητα είναι το N-year index, κατά την οποία λαμβάνονται υπόψιν μόνο ότι συνέβη τα N τελευταία χρόνια λαμβάνονται υπόψη. Μια πιο εξελιγμένη τεχνική μπορεί να αποκαλύψει πληροφορίες, όπως την δυναμική του συγγραφέα ή την αντοχή του στο χρόνο, κάτι που δεν θα μπορούσε να προκύψει σαφώς από τις συμβατικές τεχνικές.

3.3.1 ΜΕΙΩΜΕΝΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕ ΒΑΣΗ ΤΗΝ ΠΑΛΑΙΟΤΗΤΑ

Η ετήσια dataset αποτελούνταν από χαρακτηριστικά συγγραφέων για κάθε έτος. Ένα μέρος των χαρακτηριστικών αυτών αποτελείτο από μετρικές γράφων ενώ τα υπόλοιπα

είχαν να κάνουν με τις ανεξάρτητες μετρήσεις ενός συγγραφέα. Αυτές οι μετρικές απεικονίζουν την επιτυχία του συγγραφέα σαν άτομο, χωρίς να ληφθεί υπόψη η κοινωνική παράμετρος, παρόμοια με το γνωστό μέτρο παραγωγικότητας h-index (Hirsch 2005).

3.3.1.1 ΆΘΡΟΙΣΜΑ ΤΩΝ ΕΓΓΡΑΦΩΝ

Οι άθροισμα άρθρων (P_{sum}) που ένας συγγραφέας x έχει γράψει μέχρι μια δεδομένη χρονική στιγμή tn.

$$P_{sum}(x) = \sum_{i=t_0}^{t_n} pap_x(i) \qquad (8)$$

3.3.1.2 ΑΡΘΡΑ ΜΕΙΩΜΕΝΑ ΣΤΟΝ ΧΡΟΝΟ

Ο αριθμός των άρθρων ενός συγγραφέα x που έχει γράψει μέχρι μια δεδομένη χρονική στιγμή t0, μειωμένο με βάση την παλαιότητα ($P\_penalized$).

$$P_{penalized}(x) = \sum_{i=t_0}^{t_n} \frac{pap_x(i)}{(1 + t_n - i)} \qquad (9)$$

Η διαίσθηση πίσω από αυτό, είναι το γεγονός ότι ο αριθμός των άρθρων ενός συγγραφέα είναι πιο σημαντικός όταν τα έχει γράψει πιο κοντά στον τρέχοντα χρόνο και όχι στο μακρινό παρελθόν.

3.3.1.3 ΤΡΕΧΟΝΤΑ ΑΡΘΡΑ

Ο αριθμός των άρθρων που έγραψε ο συγγραφέας x τη στιγμή t_n, που δείχνει πόσο ενεργός είναι ο συγγραφέας στον τρέχοντα χρόνο (P_now).

$$P_{now(x)} = pap_x(t_n) \qquad (10)$$

3.3.1.4 ΆΘΡΟΙΣΜΑ ΤΩΝ ΑΝΑΦΟΡΩΝ

Ο αριθμός των αναφορών (Cit_sum) που ο συγγραφέας x έχει λάβει μέχρι την στιγμή t_n.

$$Cit_{sum}(x) = \sum_{i=t_0}^{t_n} \sum_{\forall j \in P_x} cit(i,j) \quad (11)$$

3.3.1.5 ΑΝΑΦΟΡΕΣ ΜΕΙΩΜΕΝΕΣ ΣΤΟΝ ΧΡΟΝΟ

Το άθροισμα των αναφορών που ο συγγραφέας x έχει λάβει, μειωμένα με βάση την παλαιότητα ($Cit_{penalized}$):

$$Cit_{penalized}(x) = \sum_{i=t_0}^{t_n} \sum_{\forall j \in P_x} \frac{cit(i,j)}{(1 + i - year(j))} \quad (12)$$

Αυτό μπορεί να γίνει καλύτερα κατανοητό με ένα παράδειγμα. Εάν εξετάζουμε το έτος 2004 και ένας συγγραφέας έχει γράψει ένα άρθρο το 2001, και η συλλογή των αναφορών είναι: 2 το 2001, 20 το 2002, 30 το 2003 και 15 το 2004, το άρθρο έχει 2/(1 +2004-2001) + 20/(1 +2004-2002)+30/(1+2004-2003)+15/1 =37,16 αναφορές. Τώρα αν εξετάσουμε το άρθρο στο έτος 2005, και οι μνημονεύσεις έχουν γίνει: 2 το 2001, 20 το 2002, 30 το 2003, 15 το 2004 και 3 το 2005, το άρθρο θα έχει 2/ (1 +2005-2001)+ 20/(1 +2005-2002)+30/(1 +2005-2003) +15/(1 +2005-2004)+3/1 =25,9 αναφορές.

Η ουσία είναι ότι οι αναφορές που έχει λάβει ένα άρθρο, έχουν μεγαλύτερη επίδραση στο χρόνο t_n όταν έχουν κερδηθεί πιο κοντά του, γιατί μια παλιά αναφορά μπορεί να μην είναι πια έγκυρη. Αυτός είναι ένας διαισθητικός τρόπο να συλλάβουμε την έννοια της χρονικής επίπτωσης και να παρατηρήσουμε την εξέλιξη στο χρόνο.

3.3.1.6 ΤΡΕΧΟΥΣΕΣ ΑΝΑΦΟΡΕΣ

Το άθροισμα των οι αναφορών ότι ο συγγραφέας έλαβε το χρόνο t_n, που δείχνουν την επιτυχία που έχουν τα άρθρα του στη σημερινή εποχή (Cit_{now}).

$$Cit_{now}(x) = \sum_{\forall j \in P_x} cit(t_n, j) \quad (13)$$

3.3.2 ΔΕΙΚΤΕΣ ΑΛΛΑΓΗΣ

Για να συλλάβουμε την εξέλιξη των παραπάνω μετρικών, δημιουργήσαμε ένα δείκτη αλλαγής με 5 μετρήσεις, για κάθε χαρακτηριστικό, ώστε να συλλάβουμε τον τρόπο που αλλάζουν στον χρόνο. Ο δείκτης αλλαγής συμπεριλαμβάνει 4 μετρικές αλλαγής και

μια που αναπαριστά το επίπεδο στο οποίο κυμαίνεται η τιμή του χαρακτηριστικού. Στους τύπους παρακάτω, $f(i)$ είναι η τιμή του χαρακτηριστικού που εξετάζουμε, κατά το έτος i.

3.3.2.1 ΕΛΑΧΙΣΤΗ ΚΑΙ ΜΕΓΙΣΤΗ ΑΛΛΑΓΗ

Η μέγιστη και η ελάχιστη αλλαγή που έχει υποστεί το χαρακτηριστικό ($minC$ & $maxC$), που δείχνει την μεγαλύτερη και την μικρότερη απόκλιση του χαρακτηριστικού.

$$\text{minC} = \min\big(f(i) - f(i-1)\big) \tag{14}$$

$$maxC = \max\big(f(i) - f(i-1)\big) \tag{15}$$

$$i \in [t_0, t_z]$$

3.3.2.2 ΤΕΛΕΥΤΑΙΑ ΑΛΛΑΓΗ

Η τελευταία αλλαγή του χαρακτηριστικού ($lastC$) απεικονίζει τη δυναμική του την δεδομένη χρονική στιγμή.

$$lastC = f(t_z) - f(t_z - 1) \tag{16}$$

3.3.2.3 ΆΘΡΟΙΣΜΑ ΤΩΝ ΑΛΛΑΓΩΝ

Το άθροισμα των μεταβολών των χαρακτηριστικών ($sumC$), που αντιπροσωπεύει τη σταθερότητα και τη φύση του ρυθμού μεταβολής του χαρακτηριστικού (θετικά-αρνητικά).

$$sumC = \sum_{j=t_0}^{t_z} f(i) - f(i-1) \tag{17}$$

3.3.2.4 ΤΙΜΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ

Τα μέτρα που αναφέρονται ανωτέρω απεικονίζουν τις αλλαγές στην τιμή του χαρακτηριστικού. Προκειμένου να προσδιορίσουμε την τιμή του χαρακτηριστικού

στον χρόνο δημιουργήσαμε ένα τελευταίο δείκτη ($featVal$), ο οποίος ποικίλλει ανάλογα με τη φύση της του χαρακτηριστικού:

- Τιμωρούμενα στον χρόνο: (μειωμένα άρθρα & αναφορές, χαρακτηριστικά από γράφους): Η τελευταία τιμή του χαρακτηριστικού. Δεδομένου ότι αυτά τα χαρακτηριστικά μειώνονται ανάλογα με τα χρόνια, εξ ορισμού, η τελευταία τιμή απεικονίζει τη συνολική αξία του.
- Αθροιστικά (άρθρα, αναφορές μέχρι τη στιγμή N): Η τελευταία τιμή διαιρούμενη με τον αριθμό των ετών που εμφανίζετε ο συγγραφέας. Αντιπροσωπεύει τη μέση τιμή του χαρακτηριστικού.
- Τρέχοντα χαρακτηριστικά (άρθρα, αναφορές τον τρέχων χρόνο): Το άθροισμα των τιμών του χαρακτηριστικού σε κάθε έτος, διαιρούμενο με τον αριθμό των ετών που εμφανίζετε ο συγγραφέας, αθροιζόμενο με την τάση του. Η τάση ενός χαρακτηριστικού μετριέται ως η κλίση της ευθείας (σε radians) που σχηματίζει η γραμμική παλινδρόμηση των τιμών του χαρακτηριστικού στο χρόνο.

3.4 ΠΙΝΑΚΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Ο παρακάτω πίνακας συνοψίζει τα χαρακτηριστικά που δημιουργήσαμε και έχουν οριστεί μέχρι τώρα.

Πίνακας 1: ΠΙΝΑΚΑΣ ΣΥΓΓΡΑΦΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Όνομα	**Προέλευση**	**Εξήγηση**	**Μέθοδος**
P_{sum}	Βάση Δεδομένων	Ο αριθμός των άρθρων που έχει γράψει ο ερευνητής μέχρι το δεδομένο χρόνο.	3.2.1.1
$P_{penalized}$	Βάση Δεδομένων	Ο αριθμός των άρθρων που έχει γράψει ο ερευνητής μέχρι το δεδομένο χρόνο, μειωμένο σε σχέση με την παλαιότητα.	3.2.1.2

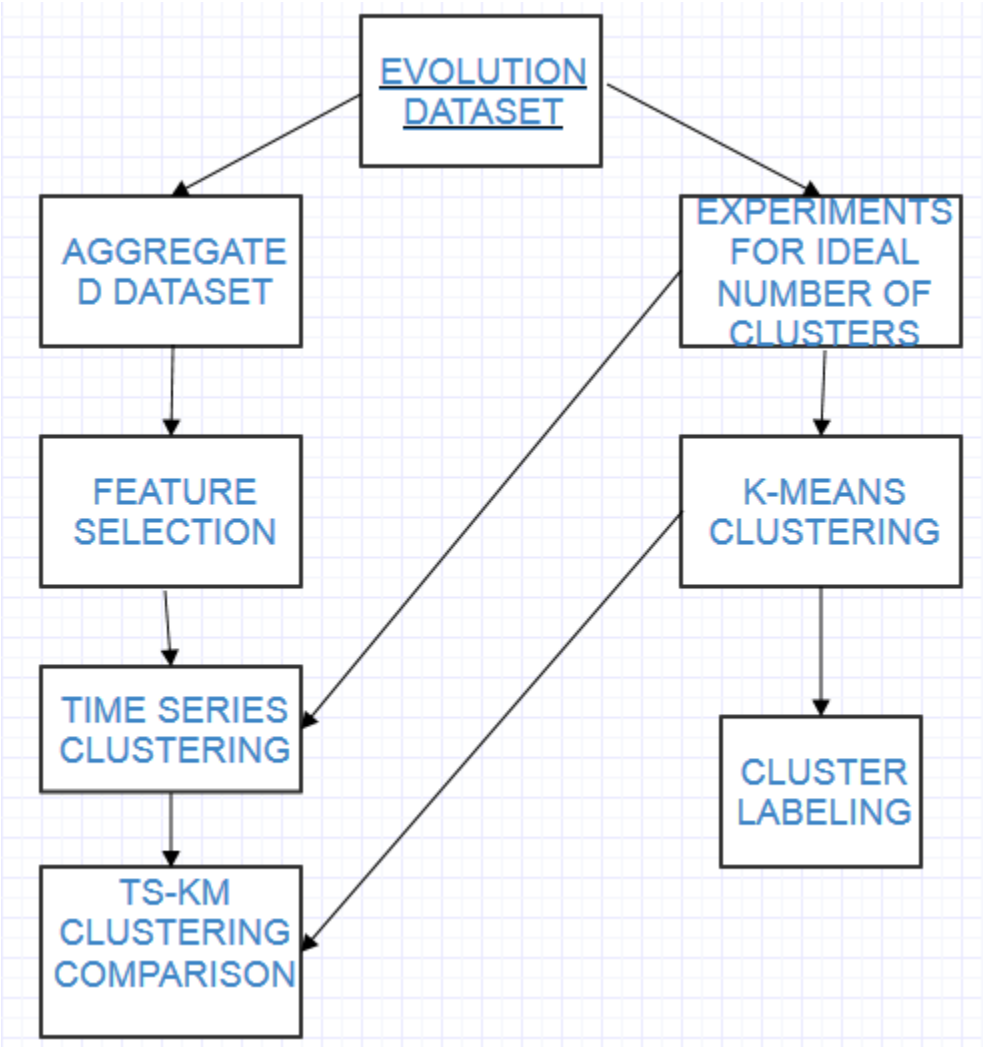
P_{now}	Βάση Δεδομένων	Ο αριθμός των άρθρων που έχει γράψει ο ερευνητής το δεδομένο χρόνο.	3.2.1.3
Cit_{sum}	Βάση Δεδομένων	Το άθροισμα των αναφορών που έχουν λάβει τα άρθρα του ερευνητή μέχρι το δεδομένο χρόνο.	3.2.1.4
Cit_{norm}	Βάση Δεδομένων	Το άθροισμα των αναφορών που έχουν λάβει τα άρθρα του ερευνητή μέχρι το δεδομένο χρόνο, μειωμένο σε σχέση με την παλαιότητα.	3.2.1.5
Cit_{now}	Βάση Δεδομένων	Το άθροισμα των αναφορών που έχουν λάβει τα άρθρα του ερευνητή το δεδομένο χρόνο.	3.2.1.6
$Eigen$	Ποιτικός γράφος	Η eigenvalue του συγγραφέα στον ποιοτικό γράφο του δεδομένου έτους.	3.1.3.1.1
Deg	Ποιοτικός γράφος	Ο βαθμός του συγγραφέα στον ποιοτικό γράφο του δεδομένου έτους.	3.1.3.1.2
CLW	Ποιοτικός γράφος	Το άθροισμα των ακμών του συγγραφέα στον	3.1.3.1.3

		ποιοτικό γράφο του δεδομένου έτους.	
WPN_{weight}	Ποιοτικό power graph	Το βάρος του power κόμβου στον οποίο ανήκει ο συγγραφέας στον ποιοτικό power graph του δεδομένου έτους.	3.1.3.2.1
WPN_{clique}	Ποιοτικό power graph	Το βάρος της κλίκας στην οποία ανήκει ο power κόμβος στον οποίο ανήκει ο συγγραφέας, στον ποιοτικό power graph του δεδομένου έτους.	3.1.3.2.2
SPN_{weight}	Ποσοτικό power graph	Το βάρος του power κόμβου στον οποίο ανήκει ο συγγραφέας στον ποσοτικό power graph του δεδομένου έτους.	3.1.3.2.1
SPN_{clique}	Ποσοτικό power graph	Το βάρος της κλίκας στην οποία ανήκει ο power κόμβος στον οποίο ανήκει ο συγγραφέας, στον ποσοτικό power graph του δεδομένου έτους.	3.1.3.2.2

3.5 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΕΡΕΥΝΗΤΩΝ

Η κατάταξη των συγγραφέων σε ομάδες είναι αναμφισβήτητα μια διαδικασία clustering. Για να προσδιορίσουμε το βέλτιστο αριθμό των clusters στο dataset με τους

δείκτες αλλαγών, εκτελέσαμε μια σειρά από πειράματα με την χρήση δεικτών εγκυρότητας ομαδοποιήσεων και έπειτα χρησιμοποιούμε το αποτέλεσμα για την εκτέλεση ενός αλγορίθμου ομαδοποίησης. Το clustering αποσκοπεί στον προσδιορισμό ειδικών προτύπων στον ρυθμό αλλαγής η στις τιμές των χαρακτηριστικών των ερευνητών. Στη συνέχεια χρησιμοποιεί αυτά τα πρότυπα και την ομοιότητα στα χαρακτηριστικά για να ξεχωρίσει και να ομαδοποιήσει τους ερευνητές. Επιπλέον, κατασκευάσαμε ένα dataset με αθροιστικές στήλες που εκφράζουν πιο ουσιαστικά χαρακτηριστικά και τρέξαμε μια διαδικασία επιλογής χαρακτηριστικών που ξεχωρίζει τα πιο σημαντικά για την ομαδοποίηση χαρακτηριστικά. Στη συνέχεια, χρησιμοποιούμε τις χρονικές σειρές του πιο κορυφαίου χαρακτηριστικού που βγήκε από την προηγούμενη διαδικασία, για να εκτελέσουμε μια συσταδοποίηση με χρονοσειρές. Ύστερα συγκρίνουμε τις ομοιότητες των δύο ομαδοποιήσεων και υπολογίζουμε την επιτυχία του δεύτερου ως προς το πρώτο. Τέλος, χρησιμοποιούμε τα αποτελέσματα από το πρώτο (clustering) για να αποκαλύψουμε τα χαρακτηριστικά στα οποία ξεχωρίζει η κάθε ομάδα κάθε, προκειμένου να την ονοματίσουμε, αντιστοιχίζοντας έτσι τους ερευνητές μέσα σε αυτήν την ομάδα σε συγκεκριμένες συμπεριφορές. Η ροή στην εικόνα 2 απεικονίζει τη γραμμή που ακολουθήσαμε.



3.5.1 ΔΕΙΚΤΕΣ ΕΓΚΥΡΟΤΗΤΑΣ ΟΜΑΔΟΠΟΙΗΣΗΣ

Τα μέτρα αξιολόγησης που χρησιμοποιούνται για να προσδιοριστεί η αποτελεσματικότητα της ομαδοποίησης. Δεδομένου ότι η συσταδοποίηση είναι μη εποπτευόμενη (χωρίς ετικέτες), οι δείκτες που εφαρμόστηκαν είναι εσωτερικής επικύρωσης (internal validity).

3.5.1.1 ΜΕΣΟΣ ΟΡΟΣ ΑΘΡΟΙΣΜΑΤΟΣ ΤΩΝ ΤΕΤΡΑΓΩΝΩΝ ΕΝΤΟΣ ΟΜΑΔΑΣ

Είναι το άθροισμα των μέσων όρων του αθροίσματος των τετραγώνων του κάθε cluster διαιρούμενο με τον αριθμό των cluster. Το μέσο άθροισμα τετραγώνων ορίζεται ως το άθροισμα των αποστάσεων ανάμεσα σε κάθε σημείο του συμπλέγματος και το κέντρο του cluster, διαιρούμενο με τον αριθμό των σημείων. Η έννοια αφορά στην μέση ασυμφωνία των clusters, εξ ου και όσο μικρότερη τόσο καλύτερα.

$$avg(WCSS) = \frac{\sum_{\forall j \in Cl} \frac{\sum_{\forall x \in j} d(x,\ c_j)^2}{|j|}}{|Cl|} \quad (18)$$

3.5.1.2 ΜΕΣΗ ΑΠΟΣΤΑΣΗ ΜΕΤΑΞΥ ΤΩΝ ΚΕΝΤΡΩΝ ΤΩΝ ΟΜΑΔΩΝ

Είναι ο μέσος όρος των ανά δύο αποστάσεων των κέντρων των cluster ($avg(DBCC)$). Αυτό αντιπροσωπεύει τη μέση απόσταση μεταξύ των ομάδων και εφόσον θέλουμε τα cluster σε όσο το δυνατόν μεγαλύτερη αντιδιαστολή, ο δείκτης αυτός θα πρέπει να είναι όσο ψηλά.

$$avg(DBCC) = \frac{\sum_{i \in Cl, j \in Cl, i \neq j} d(c_i, c_j)}{Cl * (Cl - 1)/2} \quad (19)$$

3.5.1.3 DAVIES-BOULDIN

Ο δείκτης αυτός διατυπώνεται ως εξής:

$$DB = \frac{1}{n} \sum_{i=1}^{n} max_{i \neq j} \left(\frac{\frac{\sum_{\forall x \in i} d(x,\ c_i)^2}{|i|} + \frac{\sum_{\forall y \in j} d(y,\ c_j)^2}{|j|}}{d(c_i, c_j)} \right) \quad (20)$$

Το νόημα αυτού του μοντέλου είναι ότι ο επιθυμητός αλγόριθμος θα πρέπει να παράγει ομάδες με χαμηλή εσωτερική απόσταση σε κάθε ομάδα (αριθμητής) και υψηλή εξωτερική απόσταση μεταξύ των ομάδων (στον παρονομαστή), πράγμα που σημαίνει ότι θέλουμε ο δείκτης να είναι μικρός. Θέλουμε μικρή εσωτερική απόσταση για να αποτυπώσετε το γεγονός ότι τα σημεία της ομάδας είναι όμοια. Από την άλλη πλευρά χρειαζόμαστε υψηλή εξωτερική απόσταση, προκειμένου οι ομάδες να είναι όσο τα το δυνατόν πιο διακριτές.

3.5.1.4 DUNN

Το μέτρο αυτό ορίζονται έτσι:

$$DU = min_{1\le i\le n}\left\{min_{1\le j\le n, i\neq j}\left\{\frac{d(c_i,c_j)}{max_{1\le k\le n}\frac{\sum_{\forall p\in k}(x_p-c_k)^2}{|k|}}\right\}\right\} \quad (21)$$

Ο στόχος της είναι να εντοπίσουν πυκνές και καλά διαχωρισμένες συστάδες. Είναι βασισμένη στο ελάχιστο εσωτερική και μέγιστη εξωτερική απόσταση των συστάδων. Η ελάχιστη εσωτερική απόσταση μετριέται ως η ελάχιστη απόσταση μεταξύ δύο δεδομένων κέντρων cluster. Η μέγιστη εξωτερική απόσταση ορίζεται ως το μέγιστο άθροισμα τετραγώνων εντός ομάδας σε κάθε cluster. Έχοντας την ίδια ουσία με Davies Bouldin, αλλά αντίθετη, θέλουμε το Dunn index όσο υψηλότερο γίνετε.

3.5.2 K-MEANS

Ο αλγόριθμος που χρησιμοποιήσαμε στην ανάλυση μας είναι ο k-means. Είναι μια επαναληπτική διαμερισματοποίηση, που βασίζεται σε μια συγκεκριμένη απόσταση και ένα συγκεκριμένο αριθμό από συστάδες Κ. Ξεκινά με Κ τυχαία κέντρα και στη συνέχεια χρησιμοποιεί τη συγκεκριμένη απόσταση και τις τιμές των χαρακτηριστικών της κάθε παρατήρησης για να αντιστοιχίσει την παρατήρηση στο πλησιέστερο κέντρο. Έπειτα υπολογίζει εκ νέου τα κέντρα της κάθε ομάδας, με τις μέσες αποστάσεις μεταξύ των τιμών των παρατηρήσεων που ανήκουν σε αυτήν την ομάδα. Η επανάληψη συνεχίζετε μέχρι τα κέντα να είναι σταθερά, δηλαδή δεν μεταβάλλετε η τιμή τους με τις επαναλήψεις. Η μετρική που χρησιμοποιήσαμε είναι η ευκλείδεια απόσταση, η οποία συνιστάται για συνεχείς τιμές.

3.5.3 ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Η διαδικασία επιλογής χαρακτηριστικών είναι μια συχνή και βασική διαδικασία στην εξόρυξη δεδομένων και στην ανάλυση στατιστικών στοιχείων για να παράγουν ένα υποσύνολο της αρχικής ομάδας δεδομένων, επισημαίνοντας ορισμένα χαρακτηριστικά που έχουν μεγαλύτερη σημασία ή επιρροή στο μοντέλο που εφαρμόζετε για την έρευνα που γίνετε. Πολλοί αλγόριθμοι και φόρμουλες προτείνονται, ανάλογα με το σκοπό και τη φύση της ανάλυσης. Τέτοιοι αλγόριθμοι συνήθως περιλαμβάνουν αναζήτηση υποσυνόλων στηλών και δείκτες αξιολόγησης για αυτά τα υποσύνολα. Όταν το μοντέλο είναι επιβλεπόμενο μέτρα όπως ο συντελεστής συσχέτισης του Pearson

(Pearson 1895) και η αμοιβαία πληροφορία (Manning, Raghavan και Schutze 2008) . Στην περίπτωση της μη ελεγχόμενης μάθησης η επιλογή βασίζεται στις τιμές και τις ιδιότητες των χαρακτηριστικών , όπως η απόκλιση ή η συσχέτιση μεταξύ τους.

3.5.3.1 ΣΥΜΠΙΕΣΗ ΔΕΔΟΜΕΝΩΝ

Για το τμήμα αυτό, προσπαθήσαμε να δημιουργήσουμε μια πιο συμπαγή ομάδα δεδομένων, με πιο εκφραστικές στήλες από τους δείκτες αλλαγών (3.2.2). Αυτό το πετύχαμε με τη συγκέντρωση αυτών των δεικτών, ώστε να αποκαλύψουμε τα πιο σημαντικά από τα αρχικά 13 συγγραφικά χαρακτηριστικά που αναφέρονται στα 3.1.3 και 3.2.1 .Επιλέξαμε αυτά τα χαρακτηριστικά γιατί είναι εύκολο να ερμηνευτούν, σε αντίθεση με τους δείκτες αλλαγών οι οποίοι είναι υπερβολικά περίπλοκοι για ένα τέτοιο έργο. Η συνάθροιση για κάθε χαρακτηριστικό επιτεύχθηκε με την ομαδοποίηση των δεικτών αλλαγής που προέρχονται από την ίδια λειτουργία και τις αλλαγές της. Για παράδειγμα οι 5 μέτρα που προήλθαν από την αρχική λειτουργία P_{now}(3.3.1.3) θα αθροίζονται έτσι:

$$aggrInd = \big(minC(P_{now}) + maxC(P_{now}) + lastC(P_{now}) + sumC(P_{now})\big) * featVal(P_{now}) \quad (22)$$

Αυτός ο τύπος μπορεί να εξηγηθεί, λαμβάνοντας υπόψη ότι όλες οι μετρικές αλλαγής μαζί είναι εξίσου σημαντικές με την αξία που αντιπροσωπεύει το χαρακτηριστικό.

3.5.3.2 SINGULAR VALUE DECOMPOSITION

Το singular value decomposition είναι μια στατιστική τεχνική για την αποσυντίθεση μιας ενός πίνακα δεδομένων, της εύρεση των singular values που εξηγούν το μεγαλύτερο ποσοστό της διακύμανσης στα δεδομένα και έπειτα ανασυνθέτουν τον πίνακα των δεδομένων, κρατώντας μόνο αυτά τα singular values, για να επιτευχθεί μείωση διαστάσεων. Η διάσπαση ενός πίνακα δεδομένων X είναι:

$$X = U * D * V^T \quad (23)$$

Όπου V ο δεξιός singular vector, U ο αριστερός και D ένας διαγώνιος πίνακας με τις singular values. Κάθε singular value στο D εξηγεί ένα ποσοστό της διακύμανσης των δεδομένων. Κάθε ένα από τα δεξιά singular vectors, δείχνει τις στήλες που συνεισφέρουν στην διακύμανση των αντίστοιχων singular values. Με τον τρόπο αυτό,

μπορούμε να εξαγάγουμε στήλες που συμβάλλουν περισσότερο στην απόκλιση των δεδομένων, άρα και στην ομαδοποίηση. Αυτό μπορεί να επιτευχθεί κρατώντας τα singular values που εξηγούν το μεγαλύτερο ποσοστό της διακύμανσης από το D και ανακαλύπτοντας τις στήλες που συνεισφέρουν περισσότερο σε αυτά, από το V. Αθροίζοντας τις τιμές που έχει στο V κάθε στήλη, για τα αντίστοιχα πιο σημαντικά singular values, δημιουργούμε μια λίστα από μια τιμή για κάθε στήλη, που ουσιαστικά αναπαριστά την επίπτωση της στήλης στην διακύμανση του dataset.

3.5.4 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΜΕ ΧΡΟΝΟΣΕΙΡΕΣ

Έχοντας ένα σύνολο δεδομένων που σχετίζονται με τον χρόνο, μας δίνετε η δυνατότητα να αναπαραστήσουμε τις τιμές του κάθε χαρακτηριστικού σε κάθε χρονιά , ως διανύσματα στον χρόνο. Κάθε συγγραφέας έχει ένα διάνυσμα για κάθε χαρακτηριστικό του, και αυτά δείχνουν την εξέλιξη του συγγραφέα σε αυτόν τον τομέα. Για ένα μόνο χαρακτηριστικό, μπορούμε να κατασκευάσουμε ένα dataset με γραμμές τις τιμές του χαρακτηριστικού για κάθε author και στήλες τις χρονιές. Ύστερα μπορούμε να ομαδοποιήσουμε τους ερευνητές με κάποιον απλό αλγόριθμο που δέχεται χρονικό μέτρο απόστασης, όπως ο partitioning around medoids (PAM) (Kaufman και Rousseeuw 1987), η οποία στηρίζεται στην ομοιότητα του χρόνο διανύσματος με βάση την απόσταση dynamic time warping (DTW) (Berndt και Clifford 1994) ως απόσταση μέτρησης.

3.5.4.1 DYNAMIC TIME WARPING

Το DTW είναι ένας αλγόριθμος για μέτρηση ομοιότητας ανάμεσα σε δύο χρονικές ακολουθίες, ανεξαρτήτως διαφοράς μεγέθους ή ταχύτητας. Αυτό σημαίνει ότι μια χρονική σειρά μπορεί να συγκριθεί με μια άλλη σειρά ακόμα και αν τα μήκη τους διαφέρουν. Αυτό το καθιστά ιδανικό για δική μας υπόθεση, δεδομένου ότι θέλουμε να κατηγοριοποιήσουμε r συγγραφείς που εμφανίζονται σε διαφορετικά χρονικά σημεία κατά την τη χρονική περίοδο που εξετάζουμε, δημιουργώντας έτσι διανύσματα χαρακτηριστικών με διάφορα μήκη. Για να βρεί την ομοιότητα μεταξύ των δύο διανυσμάτων χαρακτηριστικών από δύο διαφορετικούς συγγραφείς, το DTW προβάλει τις δύο ακολουθίες στην χρονική διάσταση και μετράει την ομοιότητα τους, ξεχωριστά από μη γραμμικές διακυμάνσεις στην χρονική διάσταση.

3.5.4.2 PARTITIONING AROUND MEDOIDS

Ο PAM είναι μια προσέγγιση συσταδοποίησης παρόμοια με τον k-means στο σημείο 3.3.1 . Και οι δύο στοχεύουν στη μείωση της απόστασης μεταξύ των σημείων σε ένα cluster. Και οι δύο από αυτούς είναι επαναληπτικοί και επαναυπολογίζουν τα cluster σε κάθε επανάληψη μέχρι τα σημεία να συγκλίνουν, δηλαδή να μην αλλάζουν. Βασική τους διαφορά είναι ότι ο PAM χρησιμοποιεί ένα από τα υπάρχοντα σημεία ως κέντρο της ομάδας ενώ ο k-means χρησιμοποιεί σημεία που αντιπροσωπεύουν τις μέσες τιμές των σημείων του. Επιπλέον, ο PAM λειτουργεί με διάφορα μέτρα απόστασης, ο κύριων λόγος για τον οποίο τον επιλέξαμε για την DTW απόσταση.

3.5.5 ΧΑΡΑΚΤΗΡΙΣΜΟΣ ΤΩΝ ΟΜΑΔΩΝ

Οι συστάδες που παράγονται από τη διαδικασία της ομαδοποίησης, σχηματίζονται λόγω ορισμένων μοτίβων στα δεδομένα. Δηλαδή στην περίπτωση μας, οι ερευνητές κατηγοριοποιούνται με βάση την συμπεριφορά κάποιον χαρακτηριστικών τους. Αυτό έχει ως αποτέλεσμα ορισμένες συστάδες να δείχνουν ιδιαίτερα ξεχωριστές τιμές σε κάποια χαρακτηριστικά σε σχέση με τις υπόλοιπες. Αυτή η ιδιότητα αξιοποιείται, ώστε να ονοματίσουμε το κάθε cluster βασιζόμενη στα χαρακτηριστικά στα οποία εξέχει. Η τιμή κάθε χαρακτηριστικού ενός cluster είναι η τιμή του κέντρου του.

Το κείμενο που περιγράφει τις τεχνικές λεπτομέρειες της υλοποίησης παρατίθεται στο Παράρτημα.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ & ΜΕΛΛΟΝΤΙΚΟΙ ΣΤΟΧΟΙ

Στην παρούσα εργασία προσπαθήσουμε να πετύχουμε μια κατάταξη ερευνητών σε ομάδες και γενική εξαγωγή γνώσης, αξιοποιώντας διάφορες πτυχές των πληροφοριών που είχαμε στην διάθεση μας για τους συγγραφείς και τις συνεργασίες τους. Μαζέψαμε βιβλιογραφικά δεδομένα που αναφέρονται σε συγκεκριμένο χρονικό διάστημα, από την ψηφιακή βιβλιοθήκη Scopus καλύπτοντας τυχόν ασυνέπειες των δεδομένων, εφαρμόζοντας μια τεχνική παρόμοια με συνεργατικό φιλτράρισμα. Οι αναπαράσταση του δικτύου των ερευνητών σε γράφους και οι τεχνικές εξόρυξης γνώσης από αυτούς, μας επέτρεψαν να συλλάβουμε την κοινωνική, ατομική και χρονική πληροφορία του κάθε συγγραφέα, ταυτόχρονα εξάγοντας ποικίλες πληροφορίες σε σχέση με τον ρυθμό παραγωγικότητας, την λήψη αναφορών, τους πιο επιτυχημένους ερευνητές, την αντοχή τους στον χρόνο καθώς και την επιτυχία των ομάδων στις οποίες ανήκουν. Κατά τη διάρκεια αυτής της διαδικασίας γεννήθηκαν νέες μετρικές για τον χαρακτηρισμό των συγγραφέων, εισάγοντας χρονικές μειώσεις σε βιβλιογραφικές μετρικές, power graph και ανάλυση κοινωνικών δικτύων για να πετύχουμε την αναπαράσταση του τρέχοντος χαρακτήρα μια επιτυχίας η μιας συνεργασίας. Τα χαρακτηριστικά αυτά αξιοποιήθηκαν στην κατασκευή χρονικών dataset και έπειτα δημιουργήσαμε δείκτες αλλαγών σε αυτά τα χαρακτηριστικά. Με αυτούς τους δείκτες συμπεριλάβαμε την εξέλιξη του κάθε χαρακτηριστικού για κάθε συγγραφέα, σε ένα dataset που χρησιμοποιήθηκε για να κατηγοριοποιήσουμε τους συγγραφείς με τον αλγόριθμο K-means. Ο αριθμός των cluster ορίστηκε από πειραματισμούς και χρησιμοποιώντας καθιερωμένες μετρικές εγκυρότητας clustering. Η δημιουργία ετικετών για τις 7 κατηγορίες συγγραφέων που παρήχθησαν, βασίστηκε στα χαρακτηριστικά που ξεχώριζε το κάθε cluster, τα οποία διαφοροποιούνται κυρίως στην δύναμη της κοινότητας, στον ρυθμό λήψης αναφορών και στο κοινωνικό κύρος. Επίσης έγινε μια προσπάθεια να δείξουμε το πιο σημαντικό από τα χαρακτηριστικά των συγγραφέων, ως προς την ομαδοποίηση, εφαρμόζοντας singular value decomposition. Τα αποτελέσματα δείχνουν πως τα πιο σημαντικά χαρακτηριστικά είναι οι αναφορές που δέχεται ανά χρόνο ο συγγραφέας μαζί με την τάση που ακολουθούν, ο αριθμός των άρθρων που γράφει μειωμένος με βάση τον δείκτη παλαιότητας και ο μέσος όρος του αριθμού των άρθρων που γράφει ανά χρόνο. Κάθε συγγραφέας έχει μια χρονοσειρά από τιμές για το κορυφαίο χαρακτηριστικό που εκθέσαμε προηγουμένως. Αυτές οι χρονοσείρες χρησιμοποιήθηκαν για ομαδοποίηση χρονοσειρών με dynamic time warping σαν μέτρο απόστασης χρονοσειρών. Οι ομάδες

που παρήχθησαν, αντιστοιχήθηκαν με αυτές από το k-means clustering και καταγράφηκε ποσοστό ομοιότητας σχεδόν στο 55%. Αυτό σημαίνει ότι η ομαδοποίηση χρονοσειρών αποτυπώνει ένα μεγάλο ποσοστό της πληροφορίας και μπορεί να αξιοποιηθεί στο μέλλον για πιο αποτελεσματικές ομαδοποιήσεις με λιγότερες απαιτήσεις δεδομένων.

Τα σχέδια μας για το μέλλον επικεντρώνεται στην κατασκευή ενός μηχανισμού κατάταξης, που θα κατατάσσει έναν συγγραφέα σε μια αντίστοιχη ομάδα, δεδομένων των απαιτούμενων χαρακτηριστικών. Επιπλέον, είναι σημαντικό να εργαστούμε με ολόκληρο το σύνολο δεδομένων που έχουμε στην διάθεση μας, πράγμα που δεν έγινε λόγω χρονικών και υπολογιστικών περιορισμών , γιατί περικλείει ένα σημαντικά περισσότερες πληροφορίες, όσον αφορά τα κοινωνικά και βιβλιογραφικά στοιχεία των συγγραφέων. Τέλος, να διερευνήσουμε πιο σχολαστικά τις δυνατότητες της ομαδοποίησης με χρονοσειρές, δημιουργώντας συνισταμένα διανύσματα από πολλά χαρακτηριστικά για την καλύτερη επίδοση την συσταδοποίησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

An, Y., J. Janssen, και E. Milios. 2004. «Characterizing and mining the citation graph of the computer science literature.» *nowledge and Information Systems, 6(6)* 664-678.

Balakrishnan, R, και K Ranganathan. 2012. *A textbook of graph theory.*

Berndt, D., και J. Clifford. 1994. «Using Dynamic Time Warping to Find Patterns in Time Series.» *KDD workshop Vol. 10. No. 16.*

Bonacich, P. 1972. «Factoring and weighting approaches to status scores and clique identification.» *Journal of Mathematical Sociology 2.1* 113-120.

Bonacich, P. 2007. «Some unique properties of eigenvector centrality.» *Social Networks.*

Borgatti, SP, και MG Everett. 1999. «The centrality of groups and classes.» *The Journal of Mathematical Sociology 23(3)* 181-201.

Davies, D.L., και D.W. Bouldin. 1979. «A Cluster Separation Measure.» *IEEE Transactions on Pattern Analysis and Machine Intelligence (2)* 224-227.

Dunn, J.C. 1973. «A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.» 32-57.

Erten, C, PJ Harding, SG Kobourov, K Wampler, και G Yee. 2004. «GraphAEL: Graph animations with evolving layouts.» *Graph Drawing.*

Falagas, ME, EI Pitsouni, GA Malietzis, και G Pappas. 2008. «Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses.» *FASEB Journal 22 (2)* 338–42.

Freeman, L. 1977. «A set of measures of centrality based on betweenness.» *Sociometry* 35-41.

Halkidi, M, Y Batistakis, και M. Vazirgiannis. 2001. «On clustering validation techniques.» *Journal of Intelligent Information Systems 17(2-3)* 107-145.

Harth, A., J. Umbrich, και S. Decker. 2006. «Multicrawler: A pipelined architecture for crawling and indexing semantic web data.» *The Semantic Web-ISWC 2006* 258-271.

Hirsch, J.E. 2005. «An index to quantify an individual's scientific research output.» *Proceedings of the National academy of Sciences of the United States of America 102.46.* 16569-16572.

Johnson, SC. 1967. «Hierarchical clustering schemes.» *Psychometrika 32.3* 241-254.

Kaufman, L., και P.J. Rousseeuw. 1987. «Clustering by means of Medoids.» *Statistical Data Analysis Based on the L1 Norm*, 405-416.

Kulkarni, A. V., B. Aziz, I. Shams, και J. W. Busse. 2009. «Comparisons of Citations in Web of Science, Scopus, and Google Scholar for Articles Published in General Medical Journals.» *JAMA 302 (10)* 1092–6.

Levenshtein, Vladimir I. 1966. «Binary codes capable of correcting deletions, insertions and reversals.» *Soviet physics doklady Vol. 10.*

MacQueen, J.B. 1967. «Some Methods for classification and Analysis of Multivariate Observations.» *Berkeley Symposium on Mathematical Statistics and Probability Vol. 1. No. 281-297.* University of California Press.

Manning, C., P. Raghavan, και H. Schutze. 2008. *An Introduction to Information Retrieval.* Cambridge University Press.

Martin, E., H. Kriegel, J. Sander, και X. Xu. 1996. «A density-based algorithm for discovering clusters in large spatial databases with noise.» *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining Vol. 96.* 226-331.

Martyn, J. 1964. «Bibliographic coupling.» *Journal of Documentation 20.4* 236.

Odda, Tom. 1979. «On properties of a well-known graph or what is your Ramsey number? Topics in graph theory.» *Annals of the New York Academy of Sciences 328.1* 166–172.

O'Madadhain, J., J. Hutchins, και P. Smyth. 2005. «Prediction and ranking algorithms for event-based network data.» *CM SIGKDD Explorations Newsletter 7.2* 23-30.

Pearson, K. 1895. «Notes on regression and inheritance in the case of two parents.» *Proceedings of the Royal Society of London 58(347-352)* 240–242.

Rand, W.M. 1971. «Objective criteria for the evaluation of clustering methods.» *Journal of the American Statistical Association 66(336)* 846–850.

Rousseeuw, P.J. 1987. «Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis.» *Computational and Applied Mathematics 20* 53-65.

Royer, Loïc, Matthias Reimann, Bill Andreopoulos, και Michael Schroeder. 2008. «Unraveling Protein Networks with Power Graph Analysis.» *PLoS Computational Biology 4(7).*

Tsatsaronis, G., I. Varlamis, S. Torge, M. Reimann, K. Norvag, M. Schroeder, και M. Zschunke. 2011. «How to Become a Group Leader? or Modeling Author Types Based on Graph Mining.» *Research and Advanced Technology for Digital Libraries* 15-26.

Varlamis, I., και G. Tsatsaronis. 2012. «Mining Potential Research Synergies from Co-Authorship Graphs using Power Graph Analysis.» *International Journal of Web Engineering and Technology 7(3)* 250 - 272.

Wall, M.E., A. Rechtsteiner, και L.M. Rocha. 2003. «Singular value decomposition and principal component analysis.» *A practical approach to microarray data analysis* 91-109.

West, J.D., M.C. Jensen, R.J. Dandrea, Gregg Gordon, και C.T. Bergstrom. 2013. «Author-Level Eigenfactor Metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community.» *Journal of the American Society for Information Science and Technology 64(4)* 787-801.

West, J.D., και M.A. Wiseman. 2008. «The Eigenfactor Metrics.» *Journal of Neuroscience 28 (45)* 11433–11434.

Xiaofei, He, Cai Deng, Shao Yuanlong, Bao Hujun, και Han Jiawei. 2011. «Laplacian regularized gaussian mixture model for data clustering.» *Knowledge and Data Engineering, IEEE Transactions 23(9)* 1406-1418.