Structures and Statistics of Citation Networks

Submitted in partial fulfillment of the requirements for

the degree of

Master of Science

in

Electrical and Computer Engineering

Miray Kas

B.S., Computer Engineering, Bilkent University, Ankara, TURKEY
M.S., Computer Engineering, Bilkent University, Ankara, TURKEY

Carnegie Mellon University
Pittsburgh, PA

May, 2011

Abstract

The growing of availability of electronic resources over the Internet enables rapid dissemination of the ideas and changes in the trends and the interaction patterns. In this work, we focus on dynamic, evolving social networks which exhibit numerous features that are also of interest to many researchers in non-social fields such as statistical physics, biology, applied mathematics, and computer science. We investigate how a specific research area (high-energy physics) changes over time, by building interlinked citation, publication, and co-publication networks that evolve and expand constantly through the emergence of new papers and authors. More specifically, following an interdisciplinary approach, we analyze the dataset in its full and reduced forms using techniques that are borrowed from social networks (key author/paper analysis), spatial analysis (relationship among involved countries), statistical physics (investigation of power laws in citation/authorship networks), and text mining (investigation of scientific breakthroughs). We also show how techniques such as Fourier analysis that are of particular interest to electrical engineers find their place in this interdisciplinary approach.

With this kind of comprehensive analysis, we aim to answer questions like :

•   Who are the key authors (researchers) in this field of research?

•   What do the topological features of citation/collaboration networks tell us? What can we learn from them?

•   What kinds of over-time trends are observed in these networks?

•   How often do authors publish?

•   Which organizations/cities/countries are involved in this research? How often do certain countries collaborate?

•   How can we use text contents to detect scientific breakthroughs?

Acknowledgements

Table of Contents

List of Figures

List of Tables

1. Introduction

The study of networks, including social networks, biological networks, information networks, and many others has been a major topic in scientific research. Traditionally, social networks have been studied in social sciences such as sociology, psychology, and business administration (1) (2). The general features of these classical studies are that they are often restricted to small networks, and often consider the networks as static graphs, whose nodes represent individuals and whose links represent the social interactions among these individuals. However, as more and more datasets become available online, the interest in social networks keeps growing. This is because many of the obtained results are not only solutions for problems in the field of social networks, but they are also applicable to other, even remote fields such as biological networks (3) or urban planning (4).

Citation networks, the principal focus of this study, are one kind of social networks that have been studied quantitatively almost from the moment citation databases first became available. In 1965, Derek J. de Solla Price described the inherent linking characteristic of the SCI in his seminal paper titled "Networks of Scientific Papers" (5). The links between citing and cited papers became dynamic when the SCI began to be published online. In 1973, Henry Small published his work on co-citation analysis (6) which became a self-organizing classification system that led to document clustering experiments and eventually what is called "Research Reviews" (7).

Autonomous citation indexing was introduced in 1998 by Giles, Lawrence and Bollacker (8), enabling automated extraction and grouping of citations for academic/scientific documents. While citation extraction was a manual process previously, citation measures can now be computed for any scholarly and scientific field with an online library.

Nowadays, many innovations and new research areas emerge from existing references. Creation of a specific research topic as well as its development can be traced by following paper citations. Citation networks can also help researchers identify topics that are related to a specific research topic and the subfields/communities structured around these topics. Along with the changes in the semantic network, changes of citation and collaboration networks help us understand how a research field evolves over time. To date, most studies that examine networks over time have focused on small binary networks, or large communication networks. The evolution of science is an area of interest and there is a growing body of work that shows that most people have few collaborators and that there is increasing scientific productivity outside the US.

1.1   Contributions

In this thesis, we present a comprehensive analysis of a complex, dynamic scientific citation/collaboration network following an interdisciplinary approach. With our approach, we are able to show how results of different approaches complement one another to give us a deeper understanding of the dataset. We also show how signal processing techniques that are particularly interesting for electrical and computer engineers can be used in spatial, dynamic, and evolving networks. Signal processing techniques enable us to answer questions about the periodicity of the activities in the research field such as 'How often do researchers publish?' and 'How often do researchers in different countries collaborate?'.

1.2   Thesis Organization

The rest of the thesis is organized as follows. In Section 2, we briefly review the related work in the literature. We first discuss dynamic complex networks, and then narrow our scope and move on to the citation/co-authorship networks, and the work that is specifically done on the high-energy physics dataset that we use in this work. In Section 3, we describe our dataset in more detail. In Section 4, we perform static analysis on the complete network, that is, including all the data we have available, without considering the time snapshots. In Section 5, we investigate degree distributions extracted from the networks, and discuss how the results obtained from power law distributions are in line with our results obtained in Section 4.

The remainder of the thesis is focused more on dynamic aspects. Section 6 starts with analyzing the over-time citation, publication trends, and the changes in network level parameters. Section 7 and Section 8 discuss using signal processing techniques for answering 'How often…?' questions embedded in our network data. Section 9 and Section 10 have more text oriented perspectives, demonstrating how different networks can be extracted from actual text contexts and what can be inferred from the text data. Finally, Section 11 summarizes the bigger picture while Section 12 concludes the paper.

2.   Related Work

In this section, we provide a brief overview of the related work available in the literature. We first provide background information about complex dynamic networks, statistical change detection in longitudinal analysis, and citation/co-authorship networks. Then, we briefly review the papers that use the same dataset.

2.1   Complex Dynamic Networks

The web based structure of nature, society, business, and science led many researchers to search for answers of questions such as 'How do networks emerge?', 'How do they evolve?', and 'What are the

underlying rules that govern their evolution?' (9). The processes that govern the growth and evolution of the networks have been studied in a number of papers (10), (11), (12). One of the most famous papers in this area is (10) which demonstrates the existence of scale power laws in dynamic, real life complex networks and explains the reason behind this phenomenon by the rule of 'preferential attachment'. Following the preferential attachment process, new nodes arriving to the network prefer getting attached to more popular nodes (i.e. connecting to the nodes with a higher number of connections). This leads to a power law degree distribution in the network.

(12) is another paper from the authors of (10) that proposes a continuum model for the evolution of dynamic complex networks. The model is primarily based on the observation that the networks grow by local events such as the addition of new nodes, new links and the rewiring of some of the existing links. The paper shows that depending on the frequency of these two processes (e.g. addition or rewiring) different networks can emerge with the connectivity distribution either following a power law or exponential distribution.

The model the authors propose for generating networks has three probability parameters (p, q, and *(1-p-q)*) whose sum is equal to 1. As initial input, m_0 isolated nodes are given. With probabilities p, q, and *1-p-q*, the following events happen: *(i)* addition of m new nodes, *(ii)* rewiring of m existing links, and *(iii)* addition of a new node with m links to existing nodes in the network. When nodes create links to the new nodes, they follow preferential attachment giving a higher probability to the nodes that already have a higher number of connections. Since the number of links a node receives changes over time and since the choice of new or rewired links depends on the current number of links potential destination nodes have, the growth/evolution process for the networks is a continuous process. To validate the accuracy of their model, authors use collaboration graph among movie actors. This model is also applicable to citation networks when $q=0$, which allows no rewiring of the links that have been established.

2.2   Longitudinal Analysis

There is another branch of research which primarily focuses on statistical change detection and longitudinal analysis on social networks. Techniques like Hamming distance (13) or Euclidean distance (1) have been long known to social network analysts. Hamming distance is often used in binary networks to measure the distance between two networks while Euclidean distance is primarily aimed for weighted networks. Although these methods can effectively quantify the differences in static networks, they overlook the underlying statistical distribution of the networks, preventing analysts from applying further statistical techniques.

The quadratic assignment procedure (QAP) (14) and its regression counterpart MRQAP (15) have been used to detect structural significance and compare networks in terms of their correlation (16). When applied over multiple snapshots of the same set of agents, these methods can be used for modeling the evolutionary change. Yet, such models are primarily designed for small, binary networks, and their computing time can be unfeasibly large for the large datasets similar to the ones that we study in this thesis. (16) focuses on a cumulative sum statistical process for rapid detection/identification of changes in social networks in real time; however, it shares the same characteristics as the other methods, given that the proposed methods are evaluated on datasets with less than 260 nodes/agents. Moreover, in our dataset, we observe the change of a scientific field by the emergence of new authors and papers which makes these methods inapplicable for our dataset unless we focus on a certain subset of authors or papers, as the methods discussed in this subsection require their node/agent sets to be fixed.

2.3   Citation / Co-Authorship Networks

There are also other papers on dynamic evolving networks that specifically investigate citation/co-authorship networks and evolution of these kinds of networks. The basic properties of citation/co-authorship networks can be briefly listed as follows:

- Citation networks are directed. The links go from one document to the other.
- Citation networks are acyclic because a paper can cite only existing papers.
- All edges in the citation networks point backwards in time.
- Links in co-authorship networks are reciprocal (symmetric).
- The link weights between two authors in co-authorship networks can increase over time if they have further collaboration.
- Vertices and edges added to the citation/co-authorship networks are permanent and cannot be removed at a later time.
- The already formed part of the network is mostly static; only the leading edge of the network changes.

An interesting problem in citation networks is to understand how topics evolve over time and how this can be detected using citation networks (17). For instance, (18) investigates the influence of marketing journals in subfields of marketing, analyzing which journals emerge as the most influential ones and how this changes over time. As another example, (11) studies the evolution of social networks in scientific collaboration (co-authorship) networks and present results that demonstrate the scale-free topology of co-authorship networks which are governed by preferential attachment rules in their growth.

Another paper, (19), investigates the graphical structure of the large-scale time evolving citation networks using three different techniques of analysis (*(i)* probabilistic mixture model using an expectation–maximization algorithm, *(ii)* modularity-maximization based network clustering method, and *(iii)* analysis of how eigenvector centrality scores vary over time). The paper uses a corpus of United States Supreme Court decisions as the dataset that enables revealing significant structural divisions in the network that have formed due to temporal reasons as well as the group divisions reflecting different logical ideologies.

Another approach, (20), investigates the over-time evolution of citation and author networks and discusses the two processes that mainly govern the growth/evolution of such networks: aging and growth. The authors propose a model called TARL which encompasses **T**opics, **A**ging, and **R**ecursive **L**inking. This model is shown to fit the systematic deviations from power-law distribution of citation networks well while accounting for the inter-related nature of the paper citation and co-authorship networks and importance of topic distributions. The model aims for integrating the following properties of citation networks the authors have observed:

- Authors have a bias to cite recent papers. Even highly cited papers stop receiving citations after a certain amount of time has passed. This feature works against the 'rich get richer' phenomenon enforced by aging, and frequently prevents a scale-free distribution of connectivity.
- Authors have a tendency to cite papers from the reference list of papers they have read, which is a recursive follow up of links in the network.

2.4 Work Done on High-Energy Physics Dataset

In the KDD Cup 2003 competition, there were mainly three tasks evaluated on the dataset:

- Predicting how the number of citations to each paper in the dataset will change over time

- Extracting useful data from a huge set of source/text files (*i.e.* data cleaning)

- Estimating the number of downloads a paper receives in its first two months after it is uploaded to arXiv.

For the citation prediction task, the method used by the winning team included the conversion of data into time series format and applying regression analysis on it (21). Another paper, (22), also focuses on time series conversion as its first step. The authors of (22) comment further on the factors that affect citations received by a paper, such as the reputation of the authors, publishing seasons (related to the academic year or conference schedules), and hot topics in the field. For download estimation tasks, the winning team focused on an extension of bag-of-words approach, using linear regression as the learning algorithm

(23). In a bag-of-words approach, the order of the words in a document is ignored, only the frequency (the number of occurrences of each word) is preserved.

3.   Data Description

The High Energy Physics (HEP) dataset that we use in this work (24) is a publicly available dataset compiled by arXiv (25) for the KDD Cup 2003 competition.

The dataset consists of 29,555 papers and the citations among them. The papers are in the field of high-energy physics, and they were added to the online library between 1992-2003. Each paper has a unique identifier. The paper IDs follow a certain convention, making it easy to create monthly, quarterly, and yearly snapshots. For instance, if the paper was published online in June 1998, the first four digits of its ID is 9806. Similarly, if the paper was published in December 2000, the paper's ID begins with 0012.

The dataset includes the following information: *(i)* LaTeX sources of each paper (classified by year), *(ii)* the abstract of each paper, *(iii)* SLAC dates of each paper in the form of (paper_id, SLAC_date), and *(iv)* citation graph data in the form of (citing_paper_id, cited_paper_id). SLAC date refers to the date the paper has been published online at the library of Stanford Linear Acceleration Center (SLAC) (26). The citation graph has 27,771 vertices and 352,807 edges. The citation graph does not contain any information about the citations to the papers that are not covered by this dataset.

In addition to the files provided in the high-energy physics dataset, in the arXiv website, there is structured metadata available for download in the form of XML files. Metadata XML files contain the following fields of information for each paper:

Header: identifier, datestamp, setSpec

Metadata: authors, title, categories, comments, and abstract.

SetSpec field shows the specific paper set a metadata file is generated for. The arXiv website includes papers from various fields such as physics:nuclear-experimental, physics:nuclear-theory, physics:high-energy, physics:astrophysics, mathematics, computer science, statistics and many others. The value of the setSpec field can be any of these. However, the value of the categories field might contain multiple set names listed. For instance, a paper can primarily be considered as a high-energy physics (HEP) paper, and it can be cross-listed in nuclear physics, and/or nuclear theory as well. Categories field includes all these set specification listings, which might be useful to detect to what extend these subfields overlap.

Inside the authors field, there are as many author fields as needed. Each author field has keyname (surname) and forenames as subfields. In some cases, an author field has affiliation as a subfield. However, since it is not mandatory, not all records include this information.

Using this dataset, we have extracted four different inter-linked networks. The respective sizes of the networks are listed in Table 1.

- Citation: Which paper cites which paper?

- Publication: Which author wrote which paper?

- Co-authorship: Which author writes papers with whom?

- Author-Citation: Which author cites papers from which author?

Entities		Networks			
Authors	Papers	Citations	Publication	Co-authorship	Author-Citation
7,962	27,802	352,768	55,432	38,524	371,547

Table 1- High Energy Physics Dataset Networks and Entities

3.1 Constructing Different Networks from High Energy Physics Dataset

The arXiv website provides metadata files that are available for download in the form of XML files. These metadata files list the author names, abstracts, and the SLAC dates for each unique paper. In some cases, the authors use '\affiliation{}' keyword in their LaTeX files so that their affiliations are listed along with their names. However, only 10-20% of the authors have the affiliation information readily available in the metadata files. To be able to perform spatial analysis, we have also constructed a subset of high-energy physics dataset restricted to only authors with known affiliations and the papers these authors are involved in. After analyzing the full dataset, we discuss our findings on this reduced form of the dataset.

3.1.1 Author Disambiguation

The first step in extracting the authors from the dataset is author disambiguation. It is a frequently observed problem that the same authors may use different names or abbreviations to identify themselves in different papers. In order to clean the data and decrease the redundancy in the dataset, we find these authors and revise their name to the same uniform variant. For instance, Alvarez_Gaume_Luis might use Alvarez_Gaume_Luis in some papers while he uses Alvarez_Gaume_L. or Alvarez_G._L. in other papers. If we directly use the names extracted from LaTeX files, the network will have some redundant nodes and the links between the author and papers will be inaccurate. In order to solve this problem, we change the author's name to a unified format: authors' full surname plus initials of other names. For instance, Alvarez_Gaume_Luis, Alvarez_Gaume_L., and Alvarez_G._L. will all be Alvarez_G._L. While standardization of the author names solved most of the problems, we have gone through additional manual processing to resolve other redundancy-causing issues such as misspelling of surnames.

Another problem in author disambiguation is to identify the authors that have the same name. For instance, there might be two different 'Chao Wang's in the field, and their publications might be combined under a single entity, inflating the importance of each of these separate individuals. Such authors can be distinguished from one another according to their affiliations.

3.1.2 Network Extraction via Matrix Multiplication

To be able to construct different networks using the same set of entities, we have heavily used matrix multiplication. For instance, to build co-authorship and author-citation networks, we have used the paper citation network and the author publication network.

Assume that paper citation network is denoted by the binary matrix *C*, where *C* (*i*, *j*) = 1 if paper *i* cites paper *j*. Similarly, assume that publication network is denoted by the binary matrix *P*, where *P* (*i*, *j*) = 1 if author *i* is an author of paper *j*. We calculate the author-to-author citation network *A* by *A* = *P* x *C* x *P'* while we use *W* = *P* x *P'* to get the co-authorship network *W*. We use matrix algebra in our reduced form of the dataset to get country-to-country, city-to-city, and organization-to-organization level citation and co-authorship networks.

4. Static Analysis on High-Energy Physics Dataset

In this section, we present some of the results we have obtained through the analysis of the complete dataset while we present our analysis on the reduced form of the dataset with spatial information in Section 8. We discuss different ways of extracting key papers and authors from the network, using centrality metrics that are primarily developed for social networks analysis and whose definitions are provided in Section 4.1.1. In Section 8, we present over-time publication and citation receiving trends of key author and papers and discuss their implications.

4.1.1 Overview of Centrality Metrics

4.1.1.1 Definitions

Total Degree Centrality of a node a_i is $Total\_Deg(a_i) = \frac{degree(a_i)}{n-1}$, where *n* is the number of nodes in the network.

Closeness Centrality (27) is defined as the inverse of the average of the distances between a given node and all other nodes in the network. The closeness centrality of a node a_i is:

$$Closeness(a_i) = \frac{n-1}{\sum_{j \neq i} distance(a_i, a_j)}$$

This metric describes the efficiency of information propagation from one node to all the others.

Betweenness Centrality (28) of node *v* in a network *N* is defined as the percentage of shortest paths across all possible pairs of nodes that pass through node *v*. Let *G = (V, E)* represent a square matrix where *V* is the network's nodes, and *E* represent the set of links. This is defined for directed networks. Let *n=|V|* and fix a node $v \in V$. For $(u, w) \in V \times V$, let $g_{u,w}$ be the number of shortest paths in *G* from *u* to *w* and $g_{u,w}(v)$ be the number of shortest paths from *u* to *w* that contain node *v*.

$$C_B = \sum_{u<w} \frac{g_{u,w}(v)}{g_{u,w}} \; where \; v \neq u, v \neq w$$

The value of C_B depends on the number of nodes in G. Therefore, this metric is normalized by the number of possible node pairs. Then,

For reciprocal (symmetric) graphs: $Betweenness(v) = \dfrac{C_B}{[\frac{(n-1)(n-2)}{2}]}$

For non-reciprocal graphs: $Betweenness(v) = \dfrac{C_B}{[(n-1)(n-2)]}$

4.1.1.2   *Implications and Usage*

Measure	Meaning	Usage
Degree Centrality	Node with most connections	Identifying sources for intel.
Closeness Centrality	Rapid access to information	Identifying source to transmit/acquire information.
Betweenness Centrality	Connects disconnected groups	Reducing activity by disconnecting groups.

Table 2 - Meaning and Usage of Centrality Metrics from Social Sciences

While degree centrality is useful for identifying the resources available to the nodes in the network, closeness centrality identifies nodes that have rapid access to information (which are close to many other nodes on average). As it is inversely proportional to the sum of the distances to all other nodes, closeness also provides an estimate of how long it will take information to spread from a node to others (29).

On the other hand, betweenness, focuses on another aspect of the topology: partitioning. For instance, nodes that are connected to many otherwise isolated nodes might appear to have high betweenness centrality because a node is high in betweenness if it resides on many best/shortest paths in the network. Similarly, if there is a clustered structure in the network, a node high in betweenness is likely to be a node that is connecting these two clusters, whose removal might result in partitioning of the network.

4.1.2 Key Authors

In this section, our aim is to figure out the most influential authors in the high-energy physics research area. In Table 3, we list the top 10 authors who appear as the central authors according to different metrics. For instance, 'publication out degree' column lists the authors that have published the highest number of papers while 'co-publication degree centrality[1]' ranks authors according to the total number of co-authors they have.

Although the ranking of authors are not exactly the same, the 'publication out-degree' column and 'co-publication degree centrality' column have many authors in common. This trend indicates the social network of an author increases his or her productivity in terms of the number of publications. The last column, 'author citation in-degree centrality' lists the names of highly cited authors, and this column presents names that are very different from the other columns, indicating a smaller effect of social network on the number of citations received. The social network effect on collaboration networks is discussed further in Section 4.1.4 where we present a visual snapshot of the core co-authorship network.

Rank	Publication Out-Degree Centrality	Co-Publication Degree Centrality[2]	Author Citation In-Degree Centrality
1	Odintsov_S. (176)	Pope_C._N.	Witten_E.
2	Tseytlin_A._A. (131)	Lu_H.	Seiberg_N.
3	Cvetic_M. (128)	Odintsov_S.	Vafa_C.
4	Pope_C._N. (128)	Cvetic_M.	Maldacena_J.
5	Lu_H. (125)	Ferrara_S.	Sen_A.
6	Vafa_C. (115)	Bergshoeff_E.	Strominger_A.
7	Ferrara_S. (112)	Fre_P.	Douglas_M._R.
8	Witten_E. (112)	Ovrut_B.	Klebanov_I.
9	Nojiri_S. (107)	Vafa_C.	Polchinski_J.
10	Gibbons_G._W. (94)	Nojiri_S.	Susskind_L.

Table 3 – Key Authors Table

4.1.3 Key Papers

An analysis similar to the one presented in Section 4.1.2 can also be performed for the papers. The in-degree centrality shows the number of citations received by the top 10 papers, and the overall citation counts are listed in parenthesis in the 'In-Degree Centrality' column.

The patterns observed in paper rankings are different from the ones observed in author rankings. For instance, the total degree (in + out degrees) of a paper is dominated by its in-degree rankings because the number of citations a highly-cited paper receives is mostly larger than the number of papers it has cited

---

[1] Co-publication degree centrality refers to the number of links that start from an author *x* in the author-to-paper publication network defined in Section 3.

[2] Co-publication network is reciprocal (symmetric), implying that its in-degree and out-degree values are equal. Therefore, we only mention degree centrality.

itself. The only paper whose out-degree centrality is high enough to get it into the total degree centrality rankings is '9905111'. The papers with high out-degree centralities usually turn out to be survey papers, MS/PhD theses, or long technical reports. For instance, 9905111 is a long, 261-page technical report with 757 references, which has significantly more references than most other papers in this field. Therefore, it is not surprising that the paper with the highest betweenness is the paper with highest out-degree centrality as its collection of references is large enough to span multiple areas and connect sub-areas in the high-energy physics field. The papers that are high in betweenness tend to bring different sub-fields of the research area together given the citation clusters observed in the core citation network, as presented in Section 4.1.4.

Another interesting observation is that all the papers that are high in closeness are the most recent papers. This dataset covers papers from 1992 to April 2003, and all nodes that are high in closeness are from March and April 2003. Since the citation network is directed and acyclic, older papers have fewer edges initiated to the rest of the network; they don't have paths to access the more recent papers, they are only accessible by other, newer papers. Therefore, the most recent papers are the ones that have paths, hence rapid access to information in other papers.

Rank	Betweenness Centrality	Closeness Centrality	In-Degree Centrality	Out-Degree Centrality	Total Degree Centrality
1	9905111	0304271	9711200 (2414)	9905111	9711200
2	9810008	0304184	9802150 (1775)	9710046	9802150
3	0206223	0304138	9802109 (1641)	0110055	9802109
4	9509140	0304119	9407087 (1299)	0210157	9905111
5	9803001	0304251	9610043 (1199)	0101126	9407087
6	9912210	0304123	9510017 (1155)	0007170	9908142
7	9902121	0303268	9908142 (1144)	0204089	9610043
8	9607239	0303207	9503124 (1114)	0201253	9510017
9	0206182	0303144	9906064 (1032)	9809039	9503124
10	9907085	0303115	9408099 (1006)	9802067	9906064

Table 4 - Key Papers Table

4.1.4   Visual Inspection of Core Components of Networks

In this section, we present snapshots from the core of co-publication, author citation, and paper citation networks. To reduce them to a visually inspectable form, we have removed the weak links (e.g. removed the links between authors in the co-publication network if the link indicates that the authors wrote less than 15 papers together), and the nodes that became isolates after the removal of weak links.

The resulting topologies reveal different aspects of author/paper networks. For instance, in the co-publication network depicted in Figure 1, there are many closed triangles (i.e. cliques of 3) and cliques of 4. This also causes the co-authorship network to be much sparser than authors (see the network sizes

listed in Table 1). This triangulation is an effect of the social interaction among the authors since publishing papers together requires social acquaintance among authors, whereas it is possible for authors to receive citations from authors they do not know. The author-to-author citation network topology shown in Figure 2 resembles scale-free networks although it has closed triangles that are not connected with core giant component of the network. In this topology, there are authors that are clustered at the center of the network, who receive more citations than the others, following the 'rich get richer' phenomenon. However, this is not enough by itself to observe power law distributions, as there is also the impact of triangulations, explained in more detail in Section 7. Among these three networks, the paper-to-paper citation network (Figure 3) is the one that reveals clustering according to the subfields in the research area, as there are multiple highly-connected components.

Figure 1 - Co-publication network only showing authors that have more than 15 co-published papers

Figure 2 - Core of Author-to-Author Citation Network with link weights higher than 50.

Figure 3 - Core of Paper Citation Network (Snapshot from 2002 February)
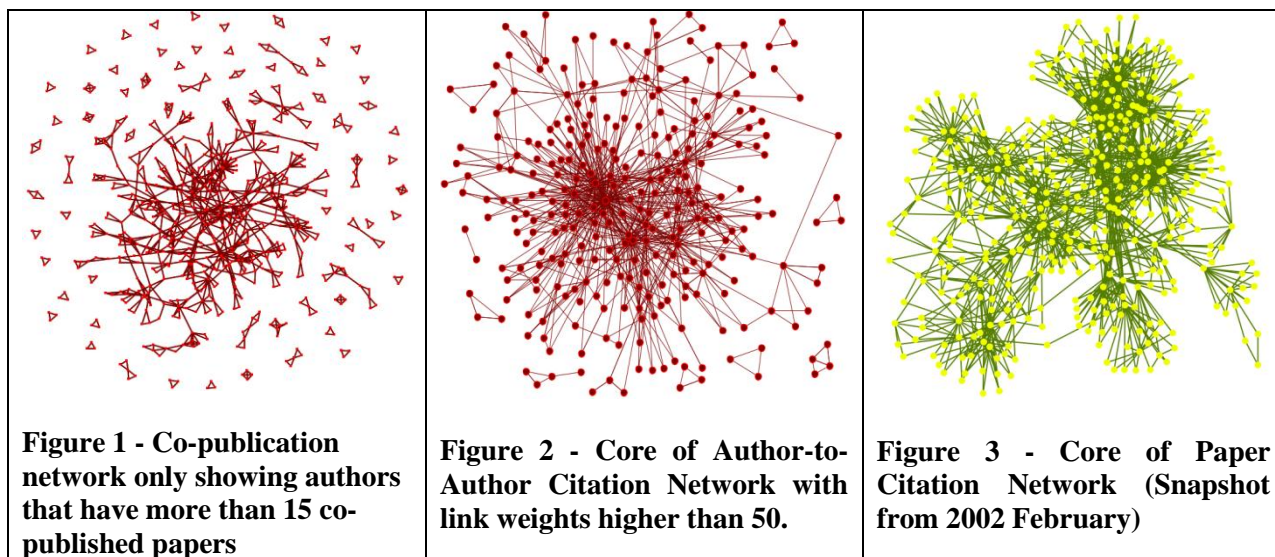
5. Investigation of Power-Law Distributions

Power laws are shown to exist in many natural and man-made phenomena such as social, biological, information, and technological networks (30) (10) (31). A few interesting examples include the frequency of terrorist attacks, the frequency of unique words in a novel, the number of calls received by AT&T customers, and the number of hyperlinks to websites (32). However, detection and characterization of power laws in empirical datasets are usually hard due to noise and the fluctuations in the tail of the distributions. Since they are hard to characterize, it is often assumed that there is a complex underlying process that is worthy of further exploration. These two points along with their special mathematical characteristics caused power laws to receive significant attention from researchers over the years.

After briefly reviewing the definition of power laws, we discuss the method we have used for estimating the power law parameters and we investigate the existence of power laws in networks extracted from our high-energy physics dataset.

5.1 Definitions

Mathematically, the distribution of a random variable x obeys power laws if its probability distribution satisfies $p(x) \propto x^{-\alpha}$ where α is the characterizing *scaling exponent* which typically lies in the range of $2 < \alpha < 3$ for power law data. More precisely, when the data is discrete, which is the case in our dataset, $p(x) = \Pr(X = x) = Cx^{-\alpha}$ where C is a constant. Various studies have shown that if an empirical dataset follows a power-law distribution, it usually only does so for values of x, where $x > x_{min}$ (32). In the cases where x_{min} is not known in advance, accurate estimation of x_{min} is very important for estimating α accurately. If the value chosen for x_{min} is too low, then we would try to fit power laws to a part of the dataset which does not necessarily follow power laws. Similarly, if the value chosen for x_{min} is too high, then we are effectively reducing the size of the dataset, making it prone to statistical errors.

5.2 Detection and Characterization of Power Laws

Estimating x_{min}: To estimate the lower bound of a power law distribution, we use the method proposed in (33). The basic idea is to choose the value of x_{min} such that the maximum absolute distance between the CDF functions of the original data and the pruned data (which contains only $x > x_{min}$) is minimized. The goal is to make the distribution of the original dataset and the best fitted power law as similar as possible.

Estimating α: For estimating the characterizing scaling exponent, we use the method described in (32), which essentially describes a maximum likelihood estimator that is equivalent to a discrete version of the Hill estimator (34). Mathematically, the estimated α, $\tilde{\alpha}$ is calculated as $\tilde{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}$.

5.3 Investigating the Existence of Power Laws in High-Energy Physics Dataset

We have extracted different distributions from the citation, publication, and co-authorship networks. In particular, we consider:

***Author Publication Weights*:** The distribution of the number of papers each author wrote from 1992 to 2003.

Co-authorship Degrees: The distribution of the number of co-authors each author has from 1992 to 2003.

Co-publication Link Weights: The distribution of the number of papers authors *X* and *Y* wrote together given that authors *X* and *Y* are co-authors.

Paper Citation Weights: The distribution of the number of citations each paper in the dataset has received. An interesting data point about paper citations is that a significant number of papers in the dataset have received no citations (16.6%) while the percentage of the papers that have 5 or fewer citations reaches 57.8%.

Among these four distributions, only the paper-to-paper citation distribution resembles power laws (Figure 10). The visual inspection of power laws involves observing a straight line on a log-log scale for the complementary cumulative distribution function. Mathematically, α should satisfy *2 < α < 3* for power law data. In our dataset, this condition holds only for paper-to-paper citation network with $\alpha = 2.7$. In all author related distributions, an exponential decay is observed. The exponential decay suggests that the creation of a large fraction of links arises from local triangulations as observed in Figure 1 which is an indication of authors who are close in the network (e.g., have a common co-author) are likely to become co-authors themselves (35). This locality property works against the emergence of power laws since preferential attachment is inherently non-local (i.e. does not have to stay local), as can be observed in the paper citation network. This matches intuition as well, as people usually find co-authors through their social networks while the popularity of a paper influences the chances that it will show up as a hit for a search query.
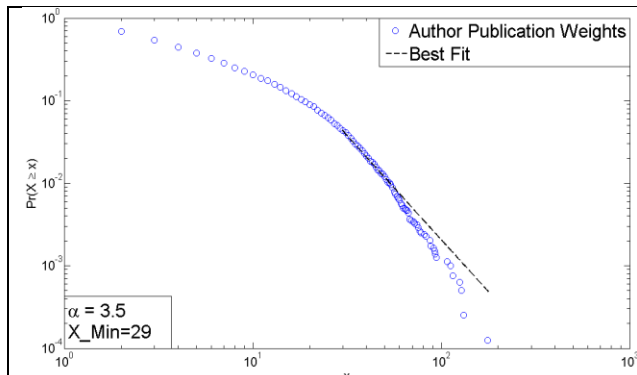

Figure 4 - Complementary Cumulative Distribution Function of Author Publication Counts and the Best Maximum Likelihood Power-Law Fit.
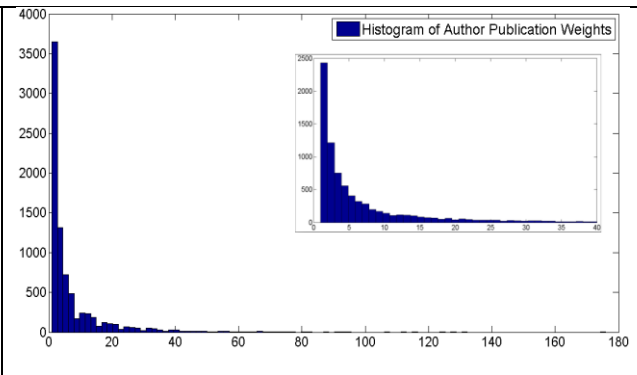

Figure 5 - Histogram of Author Publication Counts. The smaller figure covers the number of author publication counts less than or equal to 40.
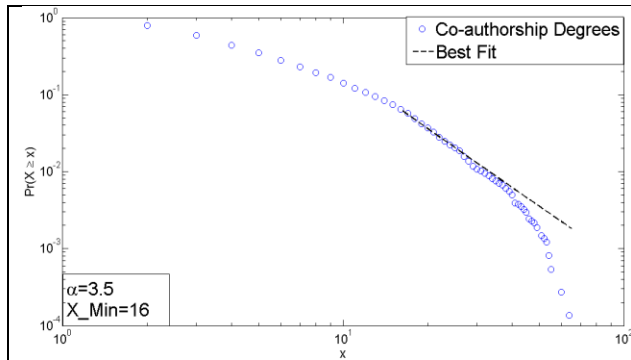
Figure 6 - Complementary Cumulative Distribution Function of Co-authorship Degrees and the Best Maximum Likelihood Power-Law Fit.
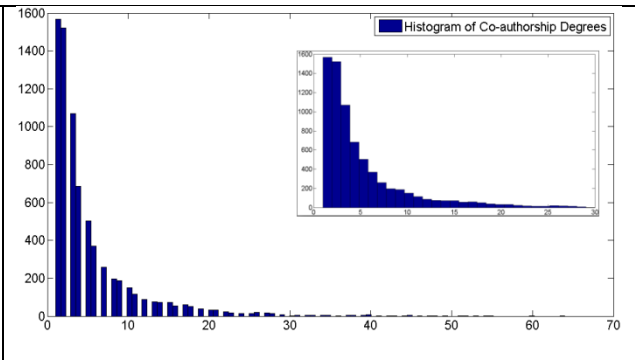

Figure 7 - Histogram of Co-authorship Degrees. The smaller figure covers the co-authorship degrees less than or equal to 30.
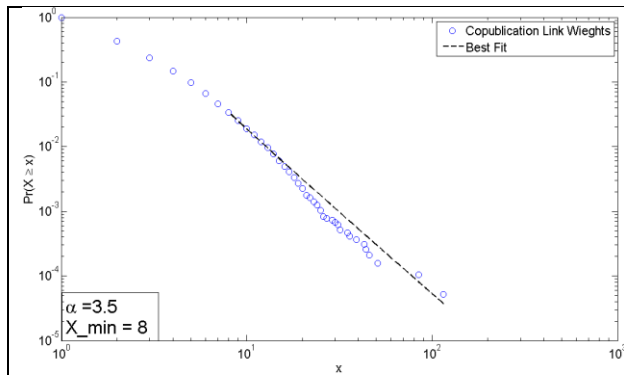

Figure 8 - Complementary Cumulative Distribution Function of Co-publication Link Weights and the Best Maximum Likelihood Power-Law Fit.

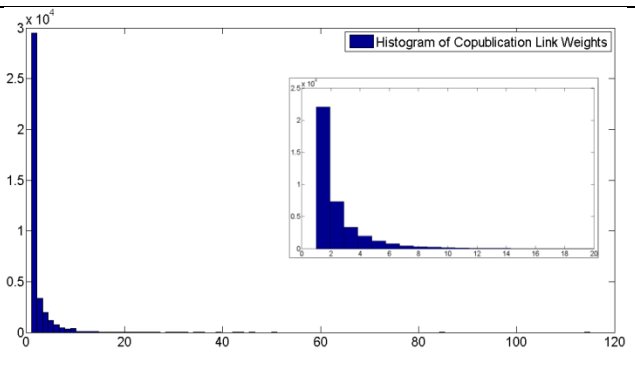**Figure 9 - Histogram of Co-publication Link Weights. The smaller figure covers the co-authorship degrees less than or equal to 20.**
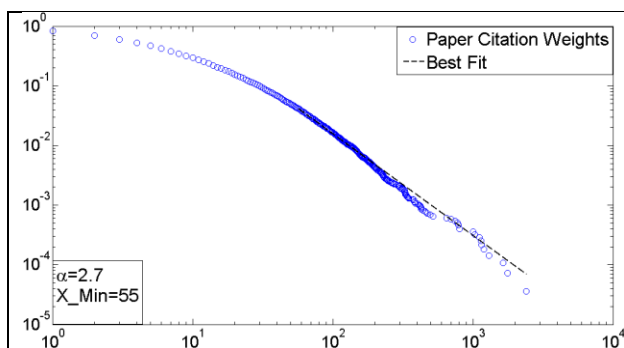

Figure 10 - Complementary Cumulative Distribution Function of Paper Citation Weights and the Best Maximum Likelihood Power-Law Fit.
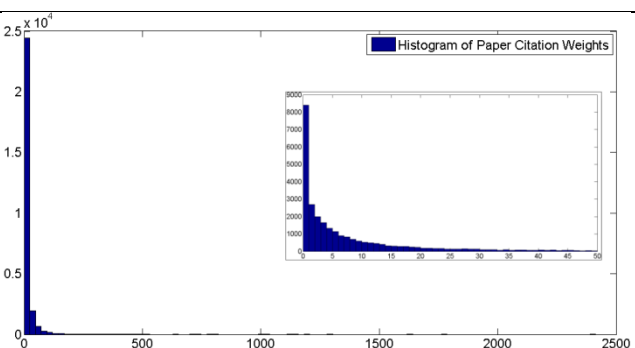

Figure 11 - Histogram of Paper Citation Weights. The smaller figure covers the paper citation weights less than or equal to 50.

21

6.   Dynamic Analysis of High-Energy Physics Dataset - Trends over Time

In this section, we present results from our over-time analysis on the paper-to-paper citation and publication networks. In Figure 12, we present the in-degree centrality of the top three (i.e. most cited) papers per snapshot. The in-degree centrality of each paper is the number of its received citations normalized by the number of nodes in the network. Therefore, it shows the relative importance of each paper across the entire snapshot. Looking at the citation trends of these three papers, one can observe that there are three distinct phases of receiving new citations, each with significant fluctuations: *(i)* generally increasing (roughly up to the first half of 1999), *(ii)* generally decreasing (roughly up to July 2001), and *(iii)* flattening out (after July 2001).
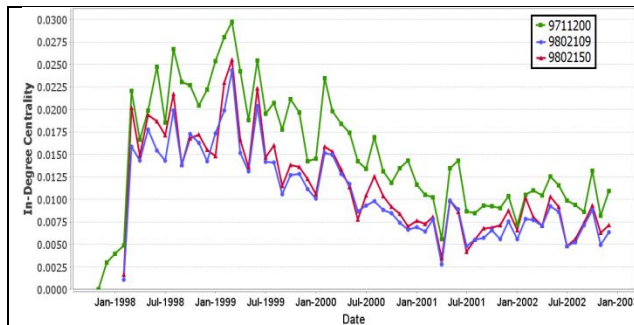
Figure 12 – New Citations Received by Most Cited Papers per Snapshot
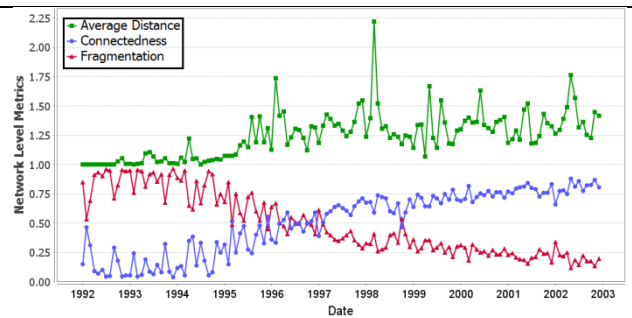
Figure 13 - Network Level Metrics for Paper-to-Paper Citation Network
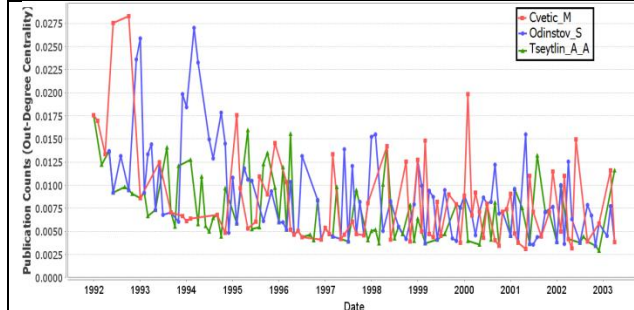
Figure 14 - New Papers Published by Most Prolific per Snapshot
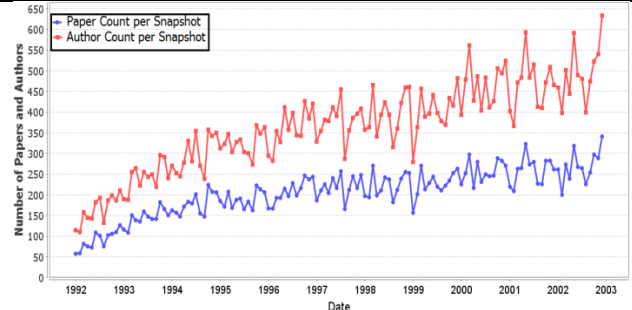
Figure 15 - Number of Active Papers and Authors per Snapshot

In Figure 13, we present three network-level metrics from paper-to-paper citation snapshots. Fragmentation and connectedness are the exact opposite of one another, and the papers become better connected over time. The average distance has a slightly increasing trend because some of the added papers are only in the citing position; they never receive citations. Therefore, there are no paths in the network to reach such papers, resulting in a slight increase in the average distance between papers.

In Figure 13 and Figure 14, we present over-time trends from the author-to-paper publication networks. Figure 13 presents the activity of the three most prolific authors over 11 years, which shows periodic

spikes for each author. Finally, Figure 15 presents the number of published papers per snapshot and the corresponding number of authors. Despite fluctuations, the general trend is an increase in the number of published papers and published authors following the 1.99 author-to paper ratio closely (55432/27802 =1.99, see Table 1).

7. Mining Periodic Activities: Fourier Analysis

Large datasets with monthly/quarterly/yearly snapshots often involve discrete signals that have time-domain frequency-domain characteristics that are hard to recognize using traditional social network analysis. Another important characteristic of such signals is that they do not have a particular defining equation that we can work with (36). We refer to signals that are in the discrete-time and discrete-frequency domains as 'discrete signals'. However, the amplitude values are continuous.

Since we have monthly/quarterly/yearly snapshots of the high-energy physics dataset, it is possible to apply digital signal processing techniques such as Discrete Fourier Transform (DFT). Discrete Fourier Transform (DFT) takes a discrete time-domain signal (time/amplitude function) and converts it into frequency domain (frequency/magnitude) signal. It takes a time series, a signal, as its input, and outputs the dominant frequencies of its input signal. Therefore, DFT is appropriate for revealing periodicities of recurring activities in social networks (16). If we treat the sequence of number of papers an author published during each interval (month, quarter or year) as a signal, then it becomes applicable to publication networks as well. A basic assumption here is that the discrete signal sequence we have is just one segment of an infinitely repeating steady-state sequence (36).

In Figure 16, we illustrate how DFT can be used to analyze publication signals at an abstract level. Assume that we have two authors of interest. The first author (*Author1*) periodically publishes a research review every other month while the second author (*Author2*) writes quarterly reviews. When the activity periods of these two authors are combined together as time series (i.e. within the time domain), we may not be able to detect these two signals because they are mixed (e.g. top right subfigure of Figure 16). However, using a Fourier transform, we can observe both signals and detect the dominant activity frequencies as *(i)* every two months and *(ii)* every three months as shown in bottom right subfigure of Figure 16.
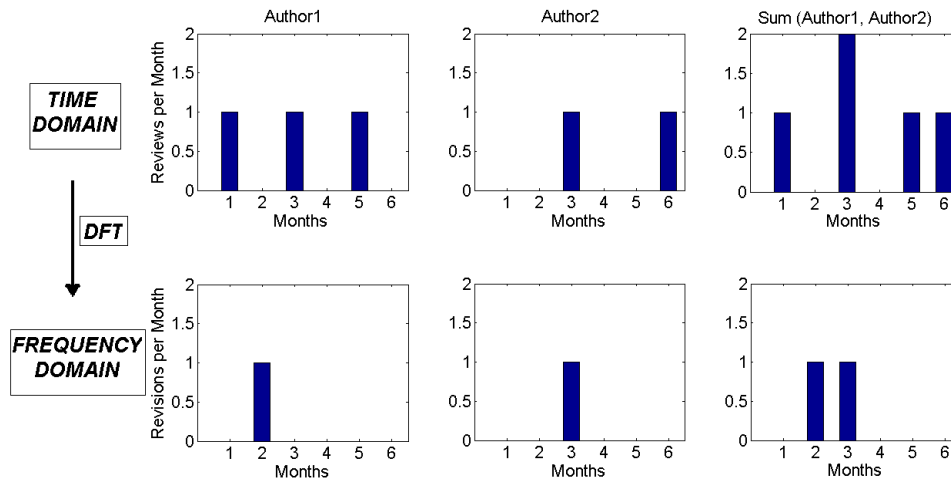
Figure 16 - DFT Analysis Example

7.1   FFT: How it works?

An efficient algorithm to compute the DFT of a signal is the FFT algorithm. The main strategy behind most FFT algorithms is to factor a DFT of length *N* into a number of shorter-length DFTs whose outputs are reused multiple times to compute the final results (37).
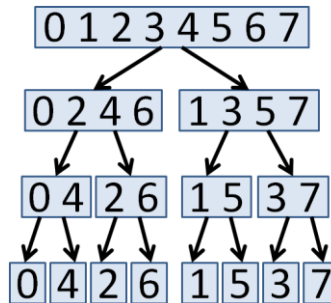
Figure 17- FFT Decomposition

Basically, FFT decomposes a length-*N* time domain signal into *N* length-1 time domain signals (i.e. top-down). This decomposition is performed over $log_2 N$ phases, and the result sequence is a reordering of the original sequence, which is usually carried out by a bit reversal sorting algorithm. Figure 17 shows the time-domain decomposition on a signal of length 8. The next step is to find the frequency spectra of the length-1 time domain signals. The frequency spectrum of a single-point signal is equal to itself. Therefore, there is nothing that needs to be done for this step. However, the values are now in frequency domain rather than time domain. The final step is the synthesis of these frequency spectra. Frequency spectra are combined in the reverse order of time domain decomposition (i.e. bottom-up). The last stage of this synthesis results in the output of the FFT, an *N* point frequency spectrum (38).

24

7.2 Fourier Analysis on High-Energy Physics Dataset

Figure 18- DFT of Publication Activities of Most Prolific Authors (Top 10 authors).

Figure 19 - DFT of Publication Activities of all Authors (only authors with at least 2 papers).
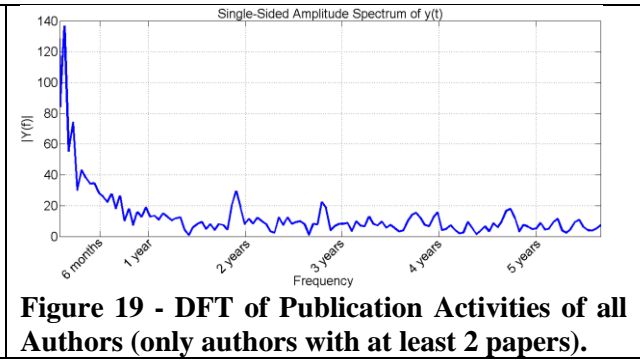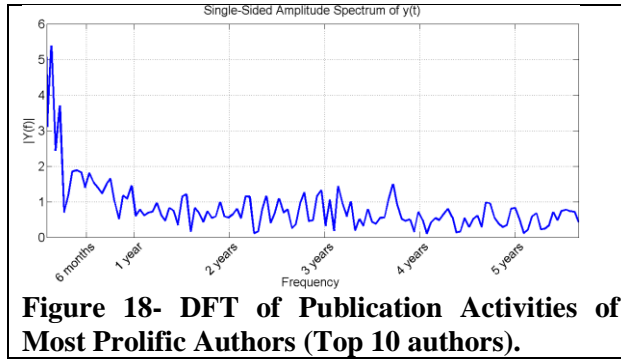
We have performed Fourier transform (DFT) on the publication activities of authors extracted from monthly snapshots of the author-to-paper publication network. As discussed in Section 7.1, the most commonly used and the most efficient FFT algorithms require $N = 2^K$ to be a power of two. From 1992 January to 2003 April, we have 136 snapshots in total. The FFT algorithms that require the number of samples to be a power of two, usually pad the end of the sequence with 0s to until the number of samples meet the next power of two. Instead of padding the publication sequence with 0s we have excluded the first 8 snapshots, which are from the immature phases of the network and have fewer papers. In addition, in its most generic form, FFT is computed on the real and imaginary parts of the signal. However, most real life (experimental) signals come with real parts only, where imaginary parts are set to 0. Yet, any real signal that is not antisymmetric around center point will still have both real and imaginary FFTs.

In Figure 18 and Figure 19, we present DFT results we have obtained using the publication signals of most prolific ten authors and of all authors who published at least two papers. These DFT results try to answer the following question: "How often do authors publish?" In Figure 18 and Figure 19, the y-axis shows $\sqrt{(Real^2 + Imaginary^2)}$. In Figure 19, in addition to the high frequency components around one and two months which dominantly come from the most prolific authors shown in Figure 18, there are spikes in periods close to two years, and three years. However, there are many more high frequency components in Figure 18 than in Figure 19. This is intuitive in the sense that the authors with many publications will publish more frequently than the community average, and are more likely to publish any time. For the most prolific authors, the magnitudes of spikes around 1 year, 1.5 years, 26 months, and 3 years are very close to one another, which essentially states that the possibility of a prolific author's publishing a paper every year is approximately the same as his publishing a paper every 26 months, although their highest frequencies lie around 1-2 months.

8. Spatial Analysis on Reduced Dataset (Only with Location Stamps)

In this section, we perform analysis on a subset of the high-energy physics dataset, focusing on the authors that have affiliation stamps. For each organization, we have also extracted the city and country information. From this analysis, we can find out organization-to-organization, city-to-city, and country-to-country relationships. In Figure 20 and Figure 21, we show the country-by-country co-publication and citation networks. Similar to the author relations, the collaboration (i.e. co-publication) network is sparser than the citation network. Investigating the link weights in Figure 20, we found that 'Germany-USA', 'Germany-Japan', and 'Japan-USA' have more collaboration.
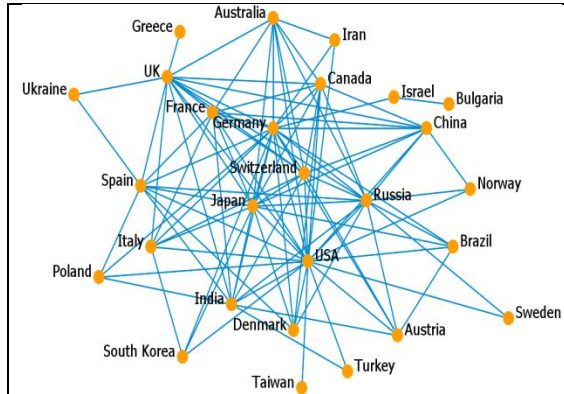
Figure 20 - Country-by-Country Co-publication Network
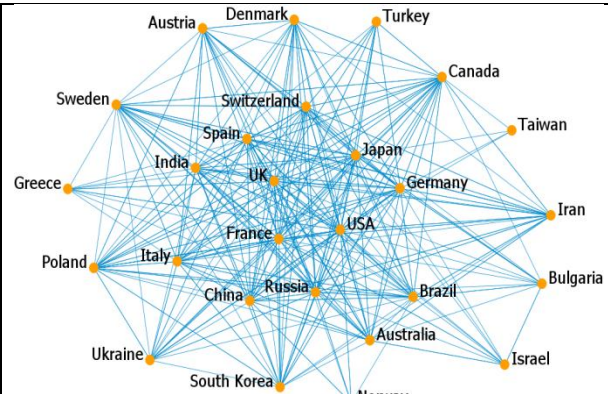
Figure 21 - Country-by-Country Citation Network
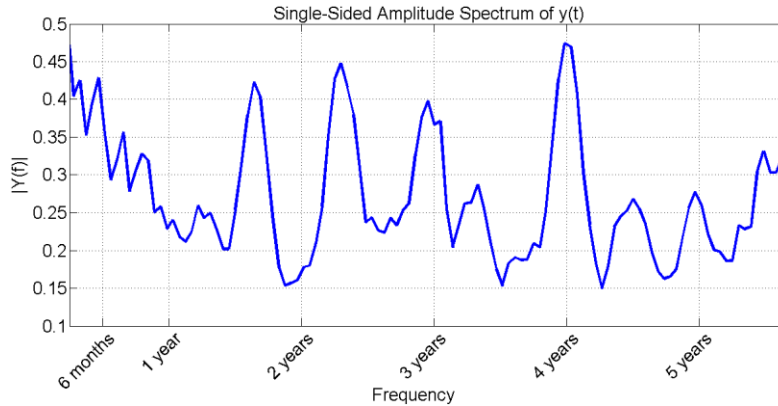
Figure 22- DFT of Country Collaboration

In Figure 22, we present the DFT results for the collaboration among USA, Germany, and Japan. We have formed three different time series from monthly snapshots of the co-publication network; one for each country pair. The values in these time series represent the number of papers published by these countries together during a certain time period (e.g. month). DFT is performed over the sum of these three signals. According to the results presented in Figure 22, 4 years is the strongest collaboration frequency,

26

followed by approximately 28 months, 20 months, and 3 years. This conclusion is in line with our conclusions from Section 7. Since average authors write papers every 2-3 years, and since most of the co-authorships stay within the same country, it is reasonable for the country-to-country relationships' to be less frequent.

9.   Extracting Social Networks from Texts: Data-to-Model

For constructing the semantic networks, we use Automap (39) as our main processing tool and iterate over the steps of the Data-To-Model (D2M) process until we get our dataset in a form that is appropriate for performing analysis. Data-To-Model is a computerized data mining procedure for extracting social networks from text files.

Step-1: Most of the text files we have downloaded are in LaTeX format, while some others are Word or PDF documents. LaTeX files include many keywords that are used for structuring the document and the initial cleaning of data includes removal of those keywords. In some cases, duplication is also an issue (i.e. repeated articles) which might amplify the relative importance of certain terms, calling for deduplication. Our main data source, arXiv, is an online library where authors upload their papers on a voluntary basis. Hence, duplication is possible. However, we have noticed many files that have been removed by the website admins upon detection of duplicate entries. Therefore, we do not need to perform our own deduplication in practice.

Step-2: In this step, we perform more detailed cleaning on the dataset, such as removing extra space, blank lines, numbers and individual letters. For the papers from the field of high energy/nuclear physics, this becomes a major problem because such papers use advanced mathematical formulations with many single letter, super/sub-scripted variables and numbers. This step is important for forming n-grams and identifying proper nouns as those characters would otherwise appear as valid characters interfering with the named entity extraction.

Step-3: The next step involves text refinement. We create stemmed/non-stemmed versions of the nouns and verbs (e.g. detensing/depluralization). Then, we delete noise words such as prepositions and helping verbs. Within this step, pronoun resolution is performed as well. This step is completely automated.

Step-4: In this step, we identify entities and n-grams that are listed as named entities. The result is a thesaurus of named entities. However, the initial thesaurus can contain invalid information which requires additional semi-automated cleanup. This semi-automated cleanup is a major bottleneck in the D2M process as it involves manual processing.

Step-5: In this step, we form a thesaurus for ontological cross-classification. The identified entities are classified into following ontological categories:

Agent – Resource – Organization – Task – Location – Knowledge – Time – Belief – Event

Ontological classification enables us to construct social networks such as (Agent x Agent), (Knowledge x Location), and (Time x Event). Unclassified entries appear under the Concept class, and may contain a significant number of invalid entries, which may require manual/semi-automated cleaning according to the focus of the project.

Step-6: We have formed a high-energy/nuclear physics specific thesaurus file using the information available in the International Nuclear Information System (INIS) (40). In this step, a merged project thesaurus is formed which is a combination of the named entity and ontological cross-classification thesauri that are generated and cleaned up in the earlier steps of D2M process.

Several iterations of these above-mentioned steps may be required. At each iteration, the text is first automatically scanned for named entities. Then, manual inspection of the list of named entities is required to remove spurious entries and tag legitimate entities by their ontological category (e.g. "Harvard" is an *organization*). The entries that are identified to have legitimate ontological tags are merged to the master thesaurus file and therefore do not appear in the list of entities to be cleaned or tagged in the following iterations. This way, the number of entries requiring manual attention decreases progressively. This gives a feel of the completeness of the dataset, signaling when we can finalize the text processing and cleaning.

The next and final step of the D2M process is to form the networks extracted from the text files and perform network analysis. In the following figures (Figure 23, Figure 24, and Figure 25) we show sample networks extracted from the text files to give a feel of how these networks look like.
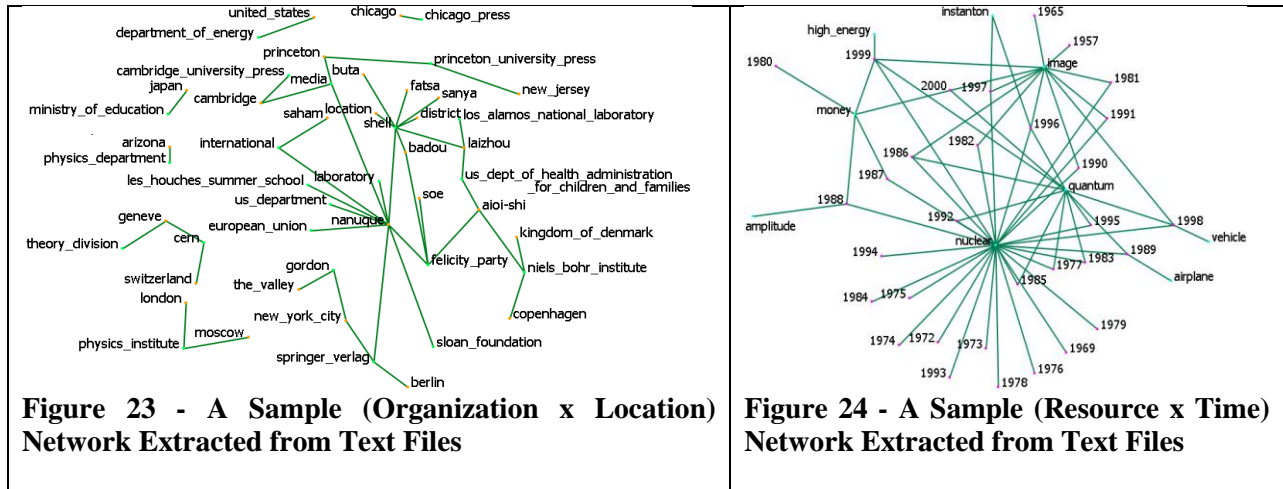
Figure 23 - A Sample (Organization x Location) Network Extracted from Text Files

Figure 24 - A Sample (Resource x Time) Network Extracted from Text Files
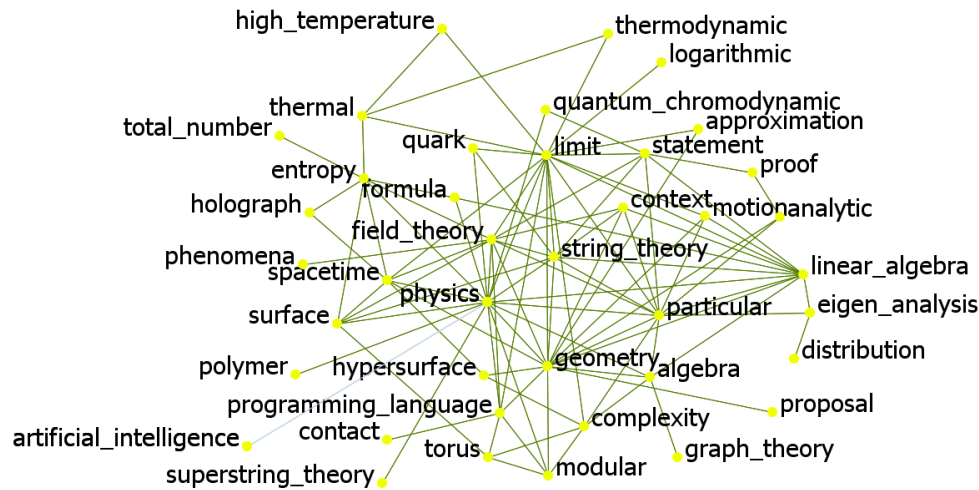
Figure 25 - A Sample (Knowledge x Knowledge) Network Extracted from Text Files

10. Breakthrough Investigation

Breakthroughs in technological advancement, science and public opinions have been investigated in various studies (41), (42), (43). For instance, (42) focuses on RNA interference as it can be considered a scientific breakthrough due to its therapeutic and scientific potential, and puts together a framework for identifying breakthroughs and deriving interesting conclusions from them.

The authors of (42) classify scientists as 'specialists' and 'generalists' according to the cohesiveness of their publications and discuss that generalists are more inclined to create breakthroughs as their thinking is not bound by any special topics. However, they also note that if the breakthrough occurs due to the solution of a long-unsolved problem, then the solution is very likely to come from a specialist since the specialists have deep knowledge of their fields.

29

Another aspect considered in (42) is core/periphery analysis which argues that the breakthroughs emerge either from the core or the periphery of the network, but are not as likely to emerge from the middle-status scientists. To elaborate, the scientists at the core of the network can afford a breakthrough because they have the major investments and required tools. The scientists at the peripheral parts of the network can also afford breakthroughs because they can significantly deviate from the mainstream and do not have too much to lose. However, breakthroughs are not that affordable for middle-class researchers as they tend to be more concerned about visibility and status and hence, have less opportunity to experiment.

Breakthroughs are not only considered for research areas in science. As an example, (41) tries to identify breakthroughs in public opinion trends by analyzing daily tweets from the Twitter website (44). In (41), the authors describe a set-space model based on Jaccard similarity (45) which we also discuss and evaluate for the reduced form of the high-energy physics dataset.

10.1   Set-Space Model

In the set space model, all documents published during a certain interval are combined into a single set by taking their union. The Jaccard similarity of two intervals T_1 and T_2 is calculated as follows:

$$Sim(T_1, T_2) = \frac{|Set(T_1) \cap Set(T_2)|}{|Set(T_1) \cup Set(T_2)|}$$

Looking at the evolution of Jaccard similarities over multiple time intervals, interval T_n is identified as a break(through) point if both of the following conditions are satisfied:

$$Sim(T_{n-1}, T_n) < Sim(T_{n-2}, T_{n-1}) \text{ and } Sim(T_{n-1}, T_n) < Sim(T_n, T_{n+1})$$

Basically, the abovementioned conditions look for dip points across the entire interval of interest. In (41), this idea has been shown to work with daily tweets. However, in both (41) and (42), the authors target periods of time for which they are already aware of the breakthroughs that happened.

10.2   Evaluation of the Set-Space Model on High-Energy Physics Dataset

To evaluate the set-space model on our dataset, we have created monthly and quarterly snapshots (unions) of the per-text semantic and social networks obtained via applying the Data-to-Model data mining procedure and calculate the similarities of the per-month and per-quarter snapshots. Different from (41) and (42), we initially had no idea whether we should expect a breakthrough.

Figure 26, Figure 27, and Figure 28 show the Jaccard similarity over multiple snapshots. Figure 26 shows the Jaccard similarity calculated over 136 months (Jan 1992 – Apr 2003) using the semantic networks extracted from Automap output files directly. In Figure 27, the Jaccard similarity is calculated for the

same time period, covering 45 quarter snapshots. However, neither Figure 26 nor Figure 27 has major dips and the slope of the linear trend fitted across the entire time period is very close to zero. Since there were not any significant enough dips in the Jaccard similarities, we also tried observing the Jaccard similarities after cleaning Automap result files. Figure 28 compares original Automap output files against 'Concepts-Cleaned' files *(i)* that have the Concept class cleaned from invalid entries, and against 'Concepts-Deleted' files *(ii)* that have the Concept, Agent, Location, and Time classes deleted. The cleaner the files, the higher the Jaccard similarities are. The lines depicted in Figure 28 turned out to be quite smooth as well.

Before concluding that there were not any significant breakthroughs during 1992-2003 periods, we have also calculated Jaccard similarity for the first (Jan-Mar 1992) and last (Jan-Mar 2003) snapshots on the 'Concepts-Deleted' files, which turned out to be 0.585, as compared to 0.67-0.68 range observed among the five snapshots shown in in Figure 28. Having a more than 50% similarity between the first and last snapshots indicates that there are not any obvious breakthroughs in the high-energy physics field that changed the interests in the field significantly.
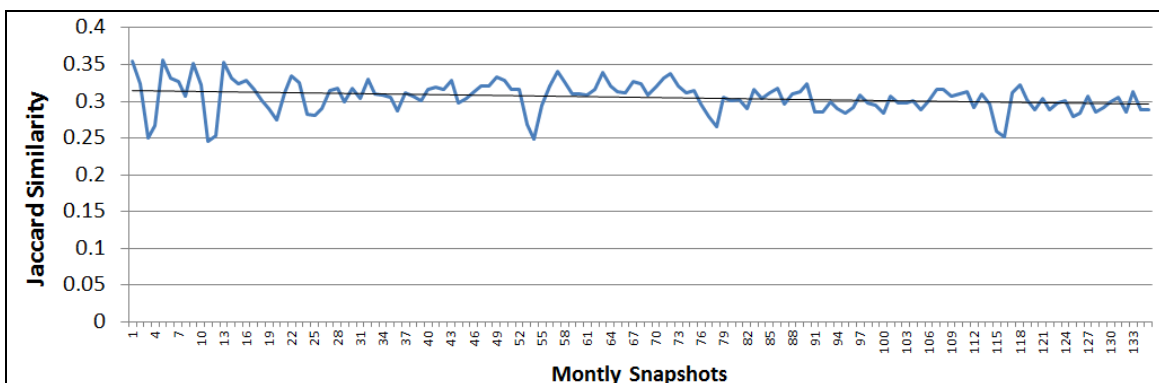
Figure 26- Jaccard Similarity of Monthly Snapshots
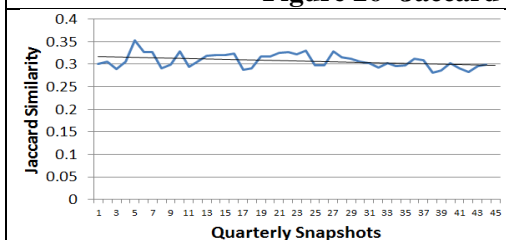
Figure 27- Jaccard Similarity of Quarterly Snapshots
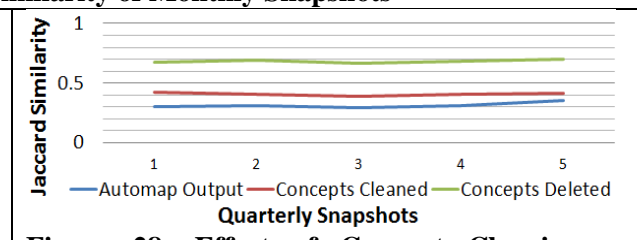
Figure 28- Effect of Concept Cleaning on Jaccard Similarity

11.   Big Picture and Ongoing Work

The study of those working on Weapons of Mass Destruction (WMD) is a very diverse and interdisciplinary area which requires input from various science and engineering fields as well as various regions around the world. The ability to remotely assess the capabilities of certain regions, countries, and

researchers requires a wide set of data such as citation networks, publication/collaboration networks among authors, affiliation data, and location data for the key set of people, along with the topics that are extracted from the relevant texts.

The questions that can be considered within the scope of remote capabilities assessment are questions as follows:

1- Is there anyone doing XXX in the region YYY? If yes, who should I talk to?
2- Is this region close to doing ZZZ (for example building a reactor)?

Remote capabilities assessment is particularly important for providing early signals on what countries, groups, or key individuals are attempting to acquire scientific knowledge to build Weapons of Mass Destruction (WMD) by looking at patterns over time and space. In this perspective, our long-term plans for this project cover proposing techniques for analyzing and predicting the moving edge of the science networks of nuclear, chemical, and biological areas.

In addition to the 29K papers from the high energy physics dataset, we have downloaded 4.6K nuclear experiment and 16.4K nuclear theory physics papers in order to construct a more comprehensive nuclear dataset. Currently, we are in the process of building a nuclear thesaurus and constructing semantic/social networks from these datasets which will eventually merged with the citation, co-publication, affiliation, and location networks that are presented in part in this thesis. Our primary goal will be to achieve over-time prediction ability, which can be performed using time series analysis and/or signal detection.

Another direction we are planning to focus on is to speed up the D2M process by reducing the amount of manual processing required. Our ongoing efforts in this direction are focused on integrating the following steps into the D2M process, mainly using a set of keywords for different ontological categories:

Step-1. Converting to time

Concept_from,	**Concept_to,**	**Metaontology**
May-99,	May_1999,	time
Jun-1999,	June_1999,	time
Jul-99,	July_1999,	time

The examples shown above are obviously time entries. However, initially, they were not listed as time entries. Using 3-letter short form or the full form of months as keywords, the entries that contain one of these keywords can be flagged as 'time' entries. This list of keywords has also been extended for days of the week as well.

```
months = { "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep",
"Oct", "Nov", "Dec","January", "February", "March", "April", "June", "July",
"August", "September","October", "November", "December"}

days = {"Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun","Monday", "Tuesday",
"Wednesday", "Thursday", "Friday", "Saturday", "Sunday"}
```

Step-2. Postal codes in locations

Concept_from,	Concept_to,	Metaontology
"California 91125",	"California_91125"	"agent"

After removing time entries as above, there are still many entries that contain location+postal code. So, if an entry contains number and word, we can check the word against the locations thesaurus. If a match is found, the ontological category can be changed to location. While preparing the entry's concept_to field the postal code that appears in the concept_from field can also be removed.

Concept_from,	Concept_to,	Metaontology
"California 91125",	"California"	"location"

Step-3. Identifying organizations

We use the list of keywords provided below to identify the organizations.

```
organization_key_words = {"ministry", "university", "universtat", "institute",
"institut", "college", " press","universite", "laboratory", "foundation",
"school","council", "community", "society"}
```

Step-4. Identifying events

We use the list of keywords provided below to identify the events.

```
events ={ "conference", "meeting", "symposium", "workshop", "anniversary" }
```

12.   Conclusion

In this work, we have focused on providing a thorough analysis of our high-energy physics dataset. Considering all the results presented in this thesis, we can describe our findings about the high-energy physics dataset as follows.

The overall network is in its early phases of development with growing number of papers and authors added every month. The major networks that can be extracted from the KDD Cup 2003 dataset are author-to-author collaboration, author-to-author citation, author-to-paper publication, and paper-to-paper citation networks. These network show different topologies and follow different evolution (growth) processes. For instance, in author-to-author collaboration, the impact of social networks is huge as most of the authorships stay local. However, when it is about citations, this localization effect diminishes,

allowing us to see an impact of preferential attachment, as also supported by power law investigation presented in Section 5. The existence of significant deviations from power laws and the fact that the top-cited papers are the oldest papers are indicators of evolution processes (e.g. aging, tendency to cite more recent papers) that work against the emergence of power laws.

Our Fourier analysis results reveal that most authors have a 2-3 year publication cycle while the prolific authors are equally likely to publish at almost any frequency. Fourier analysis on our spatial data shows that the players (countries) in this field of research mostly overlap with countries with nuclear power, and the most active countries are the USA, Japan, and Germany with 2 to 4 years of collaboration cycles. Finally, high similarities between the monthly and quarterly snapshots of the papers indicate that there were not any significant breakthroughs in the field of high-energy physics between 1992 and 2003.

13. Bibliography

1. **S.Wasserman, K.Faust.** *Social Network Analysis.* Cambridge : Cambridge University Press, 1994.

2. *The Small World.* **(Ed.), M.Kochen.** NJ : Ablex, Norwood, 1989.

3. *Ecosystems and the Biosphere as Complex Adaptive Systems.* **Levin, S. A.** s.l. : Ecosystems, 1998.

4. *The Size, Scale and Shape of Cities.* **Batty, M.** s.l. : Science, 2008, Vol. 319.

5. *Networks of Scientific Paper.* **Price, Derek J. de Solla.** s.l. : Science, 1965, Vol. 149.

6. *Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents.* **Small, Henry.** s.l. : Journal of the American Society for Information Science, 1973, Vol. 24.

7. Wikipedia. [Online] 2011. http://en.wikipedia.org/wiki/Citation_index.

8. *CiteSeer: An Automatic Citation Indexing System.* **Giles, C.L. and Bollacker, K.D. and Lawrence, S.** 1998 : Proceedings of the 3rd ACM Conference on Digital Libraries.

9. *LINKED: The New Science of Networks.* **Barabasi, Albert-Laszlo.** Cambridge, MA : Perseus Publishing, 2002.

10. *Emergence of Scaling in Random Networks.* **Barabasi, A.L. and Albert, R.** s.l. : Science, 1999, Vol. 286.

11. *Evolution of the Social Network of Scientific Collaborations.* **Barabasi, A.L. and Jeong, H. and Neda, Z. and Ravasz, E. and Schubert, A. and Vicsek, T.** 3-4, s.l. : Physica A: Statistical Mechanics and its Applications, 2002, Vol. 311.

12. **Barabási, Réka Albert and Albert-László.** Topology of Evolving Networks: Local Events and Universality. *PHYSICAL REVIEW LETTERS.* 1999, Vol. 85.

13. *Error Detecting and Error Correcting Codes.* **Hamming, R.W.** 2, s.l. : Bell System Technical Journal, 1950, Vol. 29, pp. 147--160.

14. *QAP Partialling as a Test of Spuriousness* 1.* **Krackardt, D.** 2, s.l. : Social Networks, 1987, Vol. 9.

15. *A Caveat on the Use of the Quadratic Assignment Procedure.* **Krackhardt, D.** s.l. : Journal of Quantitative Anthropology, 1992, Vol. 3.

16. **McCulloh, Ian.** *Detecting Changes in Dynamic Social Networks.* Pittsburgh : COS PhD Program, CMU, 2008.

17. *Detecting Topic Evolution in Scientific Literature: How Can Citations Help?* **Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, C. Lee Giles.** Hong Kong, China : CIKM, 2009.

18. *The structural influence of marketing journals: A citation analysis of the discipline and its subareas over time.* **Baumgartner, H. and Pieters, R.** s.l. : Journal of Marketing, 2003, Vols. 123--139.

19. *Large-scale Structure of Time Evolving Citation Networks.* **Leicht, EA and Clarkson, G. and Shedden, K. and Newman, M.E.J.** 1, s.l. : The European Physical Journal B-Condensed Matter and Complex Systems, 2007, Vol. 59.

20. *The simultaneous evolution of author and paper networks.* **Borner, K. Maru J. Goldstone R.** s.l. : PNAS, 2004.

21. *Citation Prediction Using Time Series Approach KDD Cup 2003 (Task 1).* **Manjunatha, J. N. and Pandey, Raghavendra and Sivaramakrishnan, R. and Murty, Narasimha.** Washington, DC : SIGKDD, 2003.

22. *Predicting Citation Rates for Physics Papers: Constructing Features for an Ordered Probit Model.* **Mackassy, Claudia Perlich and Foster Provost and Sofus.** Washington, DC. : SIGKDD, 2003.

23. *The Download Estimation Task on KDD Cup 2003.* **Leskovic, Janez Brank and Jure.** Washington, DC. : SIGKDD, 2003.

24. SIGKDD CUP. [Online] 2003. \url{http://www.sigkdd.org/kdd2003/kddcup.html}.

25. *Cornell University Library (arXiv).* [Online] http://arxiv.org/.

26. Stanford Linear Acceleration Center (SLAC). [Online] www.slac.stanford.edu.

27. *The centrality index of a graph.* **Sabidussi, G.** s.l. : Psychometrika, 1966.

28. *A Set of Measures of Centrality based on Betweenness.* **Freeman, L.C.** s.l. : Sociometry, 1977.

29. *Social Network Analysis Plugin (SNAP) for Mesh Networks.* **Nanda S. and Kotz, D.** s.l. : IEEE Wireless Communications and Networking Conference (WCNC), 2011.

30. *The Structure and Function of Complex Networks.* **Newman, M. E. J.** 2, s.l. : SIAM Review, 2003, Vol. 45.

31. *Power Laws, Pareto Distributions, and Zipf's Law.* **Newman, M.E.J.** s.l. : Contemporary Physics, 2005, Vol. 46.

32. *Power-Law Distributions in Empirical Data.* **Aaron Clauset, Cosma Rohilla Shalizi, M.E.J. Newman.** s.l. : SIAM Review, 2009.

33. *On the Frequency of Severe Terrorist Attacks.* **A. Clauset, C. Moore and M. E. J. Newman.** 58, s.l. : Journal of Conflict Resolution, 2007, Vol. 51.

34. *A Simple General Approach to Inference about the Tail of a Distribution.* **Hill, B. M.** 5, s.l. : The Annals of Statistics, 1975, Vol. 3.

35. *Microscopic Evolution of Social Networks.* **Jure Leskovec, Lars Backstrom, Ravi Kumar, Andrew Tomkins.** Las Vegas, Nevada, USA : SIGKDD 2008, 2008.

36. **Sabin, William E.** *Discrete-Signal Analysis and Design.* s.l. : Wiley-Interscience, 2008.

37. Overview of FFT Algorithms. [Online] http://cnx.org/content/m12026/latest/?collection=col10281/latest.

38. How the FFT works? [Online] http://www.dspguide.com/ch12/2.htm.

39. **Carley, K.M. and Columbus, D. and Bigrigg, M. and Kunkel, F.** *AutoMap User's Guide 2010.* Pittsburgh : Carnegie Mellon University, 2010.

40. **INIS.** International Nuclear Information System. [Online] www.iaea.org/inisnkm.

41. *Identifying Breakpoints in Public Opinion.* **Cuneyt Akcora, Murat Bayir, Murat Demirbas, Hakan Ferhatosmanoglu.** Washington, DC. : 1st Workshop on Social Media Analysis (SOMA '10)., 2010.

42. *Emergence of Breakthroughs: a Case Study of RNA Interference.* **Sen Chai, Wyss House, Lee Fleming.** Boston, MA : Harvard Business School, 2011.

43. *Reaping the Fruits of Science: Comparing Exploitations of a Scientific Breakthrough in European Innovation Systems.* **Valentin, F. and Jensen, R.L.** s.l. : Economic Systems Research, 2002.

44. *Twitter.* [Online] 2011. http://twitter.com/.

45. *Etude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura.* **Jaccard, Paul.** s.l. : Bulletin de la Société Vaudoise des Sciences Naturelles , 1901, Vol. 37.

46. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Citation_index#cite_note-2.

47. *Evolution of the social network of scientific collaborations.* **A.L.Barabasi, H.Jeong, Z.Neda, E.Ravasz.** 690-614, s.l. : Physica, 2002, Vol. 311.

48. **Carley, K.M. and Reminga, J. and Storrick, J. and Columbus, D.** *CMU-ISR-10-120 ORA user's Guide 2010.* Pittsburgh : Carnegie Mellon University. , 2010.

49. **Batagelj, Vladimir.** *Efficient Algorithms for Citation Network Analysis.* Ljubljana, Slovenia : University of Ljubljana, Department of Mathematics, 2003.

50. **R.Albert, H.jeong, A.L. Barabasi.** s.l. : Nature, 1999, Vol. 400.

51. *Giles.* **S.Lawrence, C.L.** s.l. : Science, 1998, Vol. 280.

52. *Efficient Similarity Search In Sequence.* **Faloutsos, Rakesh Agrawal and Christos.** San Jose : s.n., 1993.