Complex Network Analysis of Research Funding: A Case Study of NSF Grants

Hakan Kardes[1], Abdullah Sevincer[2], Mehmet Hadi Gunes[2], and Murat Yuksel[2]

[1] inome Inc,
500 108th Ave NE Bellevue, WA 98005 USA
hkardes@cse.unr.edu
[2] University of Nevada, Reno,
1664 N Virginia St Reno, NV 89557 USA
{asev, mgunes, yuksem}@cse.unr.edu

Abstract. Funding from the government agencies has been the driving force for the research and educational institutions particularly in the United States. The government funds billions of dollars every year to lead research initiatives that will shape the future. In this paper, we analyze the funds distributed by the United States National Science Foundation (NSF), a major source of academic research funding, to understand the collaboration patterns among researchers and institutions. Using complex network analysis, we interpret the collaboration patterns at researcher, institution, and state levels by constructing the corresponding networks based on the number of grants collaborated at different time frames. Additionally, we analyze these networks for small, medium, and large projects in order to observe collaboration at different funding levels. We further analyze the directorates to identify the differences in collaboration trends between disciplines. Sample networks can be found at http://www.cse.unr.edu/∼mgunes/NSFCollaborationNetworks/.

Keywords: Complex networks, Complex network analysis, Research funding networks, Six degrees of separation, NSF.

1 Introduction

As data about social networks has grown vastly in size and heterogeneity, complex network analysis of such networks have become more popular. Many researchers are modeling the growth and the structure of the networks from different fields including biology, chemistry, geography, mathematics and physics. Complex network analysis helps to capture the small-scale and the large-scale features of these networks that are not evident. Such analysis may uncover the underlying dynamics of the network patterns. In this direction, researchers have investigated interactions of different systems in various fields as a complex network [1].

Many studies [2–5] look into popular social networks such as Facebook, Twitter and YouTube. Newman provided the first study on co-authorship networks by analyzing the macroscopic properties of different domains [6, 7]. Similarly, researchers have studied academic ties [8], air transport [9], authors network [10], citation networks [11, 12], friend recommendation [13], influenza spread [14, 15], Internet topology [16–18], news networks [19, 20], patent networks [21, 22], protein interactions [23], software collaborations [24, 25], and video industry [26] as complex networks.

In this paper, we analyze the collaboration of researchers when they obtain federal funding[3]. For this study, we obtain the funding data of the National Science Foundation (NSF), an independent federal agency established by the U.S. Congress in 1950 to promote the progress of science; to advance the national health, prosperity, and welfare; and to secure the national defense. NSF has an annual budget of about $7.4 billion (FY 2011) [28], and funds research and educational activities at various institutions including universities, research institutes, foundations and industry.

As a public institution, NSF shares its funding information [29]. The data released by NSF includes the Principle Investigator (PI), i.e., the researcher responsible for leading the project, co-PIs (if any), organizations, directorate, grant amount and several other fields for the funded projects. In order to analyze the collaboration structures within the NSF research funding network, we generate three types of networks from the provided dataset based on the number of collaborations for different time frames. First, we construct the PI collaboration network where we analyze the social interaction of researchers. The PI network shows the structure of the collaboration and different characteristics of the NSF grants among PIs. Moreover, from the institution information of co-PIs, we build an organization network where we inspect the collaboration among research institutions. This analysis reveals the most central organizations and collaboration trends. We also derive the state collaboration network to study the collaboration among the states in obtaining federal funding.

Since, we construct these networks both for different time frames and as a whole; we compare the network characteristics of these networks for different time frames and capture the changes in the NSF collaboration network over the time. Additionally, we analyze these networks for small, medium, and large projects in order to observe collaboration patterns at different funding levels. We further analyze the funding networks within each NSF directorate and find their distinct properties. We compare each directorate with the other directorates to better understand the collaboration in the NSF funding data.

The main goal of this paper is to collect the NSF funding dataset, discover interesting complex network structures from the dataset, and derive new insights from it. The newly discovered properties from the dataset will give an idea of the collaboration among researchers in obtaining federal funding. Researchers have studied National Institutes of Health (NIH) and NSF data sets for visualization. For instance, Herr et al. presents an interactive two dimensional visualization of the 60,568 grants funded by NIH in 2007 [30]. However, this paper is, to best of our knowledge, the first study to analyze the funding data as a complex network.

In the rest of the paper, first we clarify the metrics that we use during our analysis and we describe data collection and network construction procedures. We then present analysis of research funding networks derived from the NSF data at different levels. Finally, we conclude and provide future directions.

[3] An earlier version of this study appeared in [27].

2 Preliminaries and Definitions

There are several well known metrics which are widely utilized in complex network analysis. In this section, we briefly provide an overview of the metrics that we use in our analysis.

Size is one of the most basic properties of a network, and is quantified by the number of nodes n and the number of edges e.

The basic characteristic to infer a network's connectivity is **average node degree** $\bar{k} = 2n/e$. The **degree** k of a node is the number of edges that are adjacent to the node. A node with degree k is called as $k - degree$ node, and $n(k)$ is the set of all k-degree nodes in a network. The average node degree can also be calculated by taking the mean of the degree of all nodes in the network. **Weighted Degree** of a node is the sum of the weights of all of the edges that this node has. **Node degree distribution** is the probability distribution of the node degrees where the probability of having a $k - degree$ node in the network is expressed as $P(k) = n(k)/n$.

Distance is the shortest path length between a pair of nodes in the network. **Average Path Length** stands for the average distance between all pairs of nodes in the network. **Diameter** is the maximal shortest distance between all pairs of nodes in the graph, and gives an idea of how far apart are the two most distant nodes.

Assortativity illustrates the link behavior of nodes, and measures whether similar degree nodes are more likely to be connected to each other. **Rich Club** measures how well the highest degree nodes in the network are connected.

Clustering coefficient is the measure of how well the adjacency (i.e., neighbors) of a node are connected. The neighbor set ns of a node a is the set of nodes that are connected to a. If every node in the ns is connected to each other, then the ns of a is complete and will have a clustering coefficient of 1. If no nodes in the ns of a are connected, then the clustering coefficient of a will be 0. High clustering coefficient is the indicator of **small-world** effect along with small average shortest path.

There are several measures for the *centrality* of a node within the network. Such centrality measures are important in analyzing the funding network since they may determine the relative importance of a node within the network. **Betweenness Centrality** of a node is the measure of how often this node appears on the shortest paths between any node pair in the network. **Closeness Centrality** of a node is the average distance of this node to all other nodes in the network. **Eigenvector Centrality** measures the importance of a given node based on its connections.

3 Data Collection

NSF provides historic information on funded grants at its website. A search engine provides access to the grant information. Each search query turns at most 3,000 grants at a time, and there is a rate limit for queries from a computer. This rate limiting of NSF website necessitates using multiple computers if one wants to download the grant data faster. We implemented a crawler using the PlanetLab [31] infrastructure to download the NSF grants database in order to parallelize the download process. Overall, we downloaded a total of 279,862 funded grant data spanning from 1976 to December 2011.

Each NSF grant has a Principal Investigator (PI), organization, co-PIs, directory and several other fields in the database. We ignored some of these fields since our aim is to analyze the network of collaborations among the NSF grants. The individual grants such as fellowships or presidential awards are not included in the dataset as they are not collaborative works. A collaborative research grant with co-PIs from the same institution has a single entity in the NSF database. However, if the co-PIs are from different organizations, there may be multiple entities in the database for this grant. If it appears in multiple entities, the title of the grant should be the same and begin with ‘Collaborative Research’. We filter the dataset considering these rules and similar naming conventions of the NSF.

4 Networks Analysis of the NSF Funding

In order to analyze the collaboration patterns within the research funding network, we generated three types of networks from the dataset and visualized them with Gephi [32]. First network we explore is the *PI network*, i.e., the collaboration network between Principal Investigators (PIs) of the grants. By constructing this network, we aim to understand the relationships and characteristics of the collaboration between researchers. To construct the PI network, we connected co-PIs of each grant as in Figure 1. In this network, each node $P_i \in PIs$ represents a PI and each edge between P_i and P_j indicates that these two PIs have a collaborative grant. This network is weighted and the weight of the edges represents the number of grants collaborated among the two PIs. Moreover, we built the *organization network*, i.e., the collaboration network between the organizations of the PIs of the funded grants to observe the relations between institutions in receiving grants from the NSF. Finally, we constructed the *state network*, i.e., the collaboration network between the states of the PIs in order to analyze the patterns among the home state of researchers.

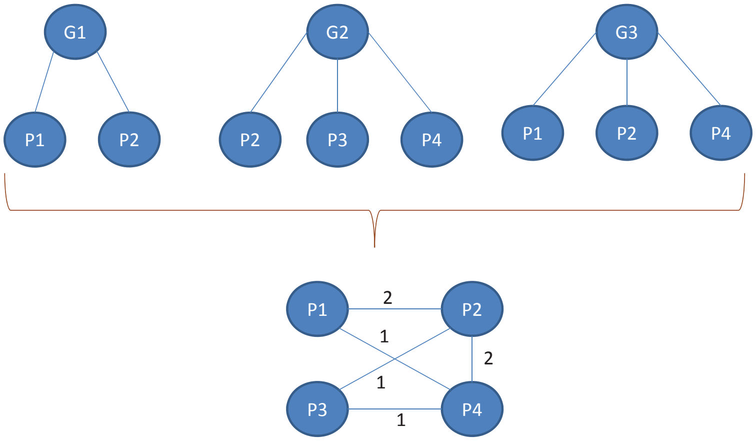


Fig. 1. PI Network Construction

Furthermore, we drew the same networks for different time frames, namely 80s (i.e., 1980-1989), 90s (i.e., 1990-1999), and 2000s (i.e., 2000-2009). Although NSF was established in 1950, it has begun to gain more importance since 1980s as the country realized the significance of research in science, technology, and education. In 1983, NSF budget exceeded $1 billion for the first time. Later, it passed $2 billion in 1990, $4 billion in 2001 and became $6.9 billion in 2010. Therefore, in this analysis, we examine the evolution of the collaboration networks and the effect of the growth of the budget to the collaboration of the researchers and organizations.

Moreover, we analyzed these networks for small, medium, and large projects in order to observe the collaboration at different funding levels. Similarly, we analyzed the funding network within each NSF directorate to find their distinct properties. We compared each directorate with the other directorates to better understand the collaboration patterns within different research fields.

4.1 PI Network

The PI network constructed from the dataset is shown in Figure 2. In this network, there are about 106K nodes and 197K edges which makes it hard to visualize. The diameter of the PI network, which is constructed from all PIs with a collaboration, is 29 and the average path length is 24.4. The average path length is higher than other similar social networks. There are several directorates such as Biological Sciences (BIO), Computer and Information Sciences (CSE), Education and Human Resources (EHR), Geosciences (GEO), Mathematical and Physical Sciences (MPS), Office of Polar Programs (OPP), Social Behavioral and Economic Sciences (SBE) in NSF grants. Thus, in our opinion, the main reason for having high diameter and average path length values for the PI network is due to the diverse fields of studies of the PIs. Additionally, as the PI network is sparse, the number of interdisciplinary grants which would make the PI network more connected is low. As indicated in the Directorates Networks Section, the PI network of each individual directorate is well-connected with low diameter and average path length values but we do not observe this behavior when we consider all of the directorates together.

Figure 3-(a) presents the *clustering coefficient distribution* of the nodes in the PI network. The average clustering coefficient of the graph is 0.46. This is considerably higher than a random network of similar size, which happens in *small world* [33] networks.

The *node degree distribution* in Figure 3-(b) does not exhibit a power-law distribution as observed in many social networks but rather results in a declining curve. We think this is mainly due to the fact that funding collaborations require considerable effort and researchers are limited in the number of collaborations they can form. The *average node degree* for the network is 3.71, while the *weighted node degree* is 4.5. The number of collaborations, if any, among PIs is 1.22 on average.

The *assortativity* of the graph is 0.18, which means the network is non-assortative [34]. That is, PIs who have high collaborations slightly tend to work together rather than collaborating with PIs that have low collaborations.

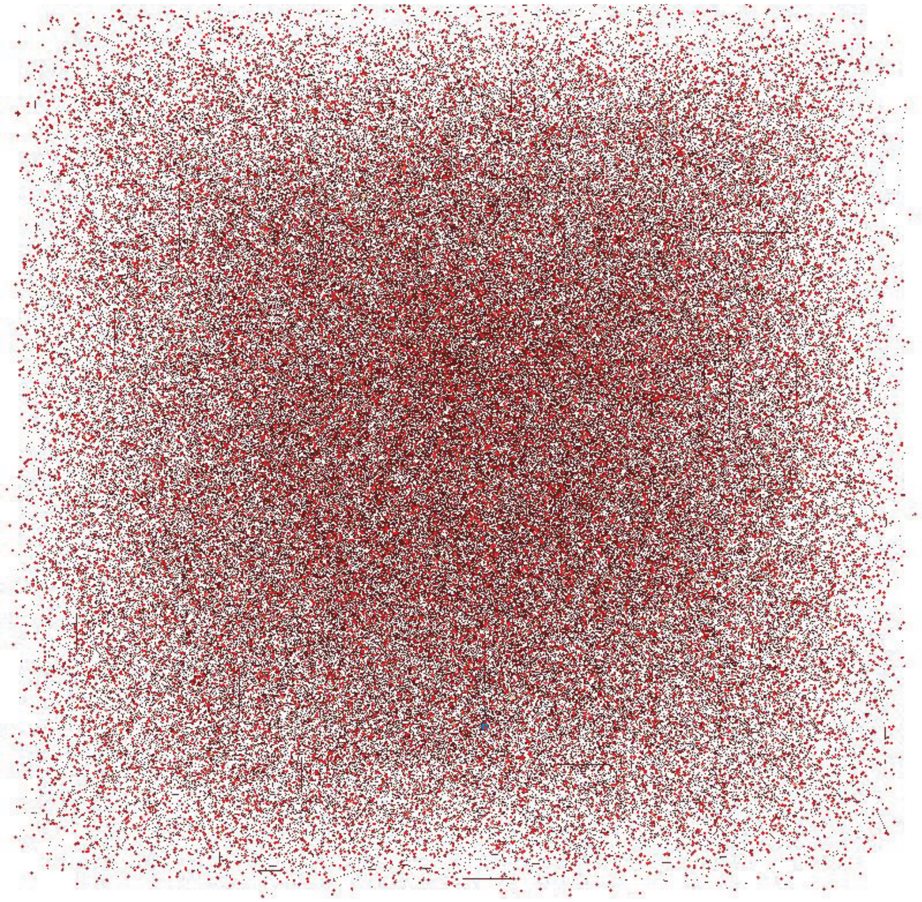
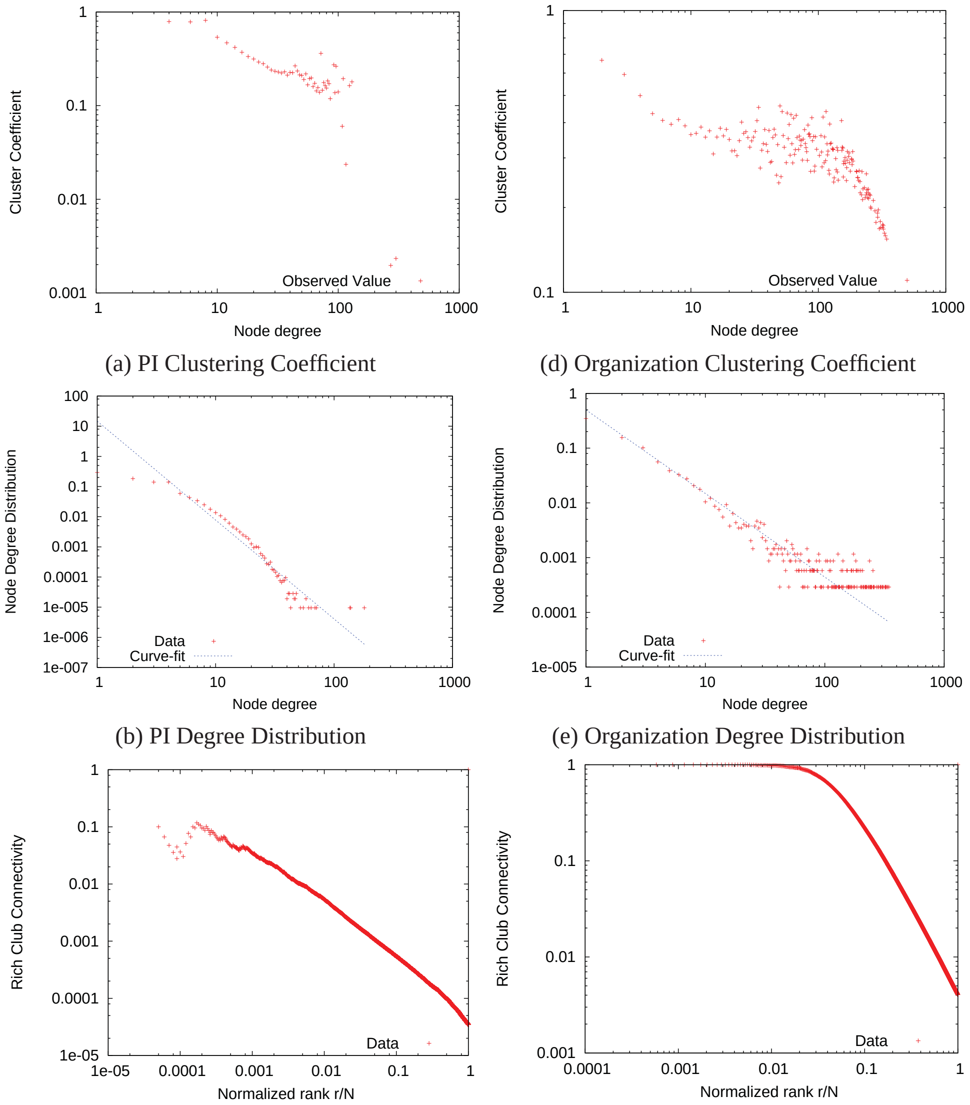


Fig. 2. PI Collaboration Network

Moreover, Figure 3-(c) shows the *rich club connectivity* of the PI network. According to this graph, there is not an obvious rich club that contains most of the collaborations even though such phenomenon has been observed in citation networks [35].

In order to better analyze highly collaborative PIs, we draw the network of the PIs with the highest node degrees in Figure 4. In this figure, the thickness of the edges illustrate the number of collaborations among PIs while the boldness of the color of each node represents the weighted node degree, i.e., the total number of collaborative grants for that node. In the figure, we observe few cliques indicating a highly collaborative group of researchers and some isolated nodes indicating researchers with a large number of distinct collaborations.

Moreover, in order to study frequent collaborations among researchers, we construct the PI network by only considering the highest weighted edges in Figure 5. As seen in the figure, there are many distinct pairs of PIs while there are a few triangles and larger



(a) PI Clustering Coefficient

(d) Organization Clustering Coefficient

(b) PI Degree Distribution

(e) Organization Degree Distribution

(c) PI Rich Club

(f) Organization Rich Club

Fig. 3. PI and Organization Network Metrics

cliques in this network. This indicates most of the frequently funded research teams consist of two PIs. Though more statistical evidence is needed, one may concur that frequent collaboration with another PI is more likely to increase chances of a successful project compared to new collaborations that might be more fruitful while being more risky.

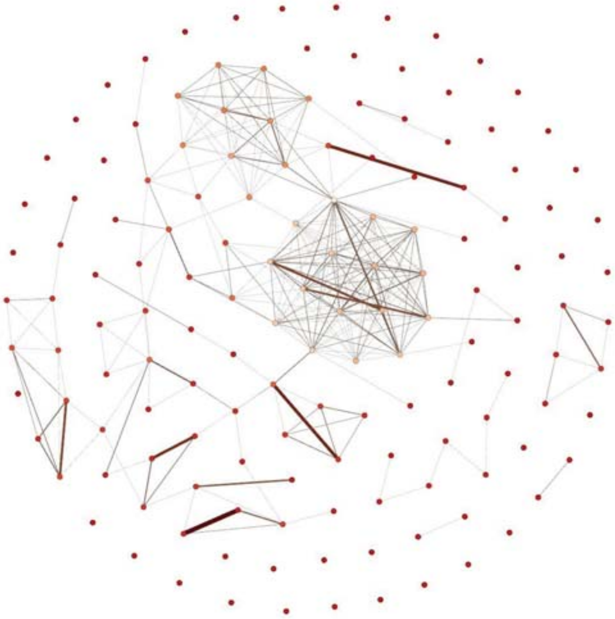


Fig. 4. PI Collaboration Network for PIs with High Degrees

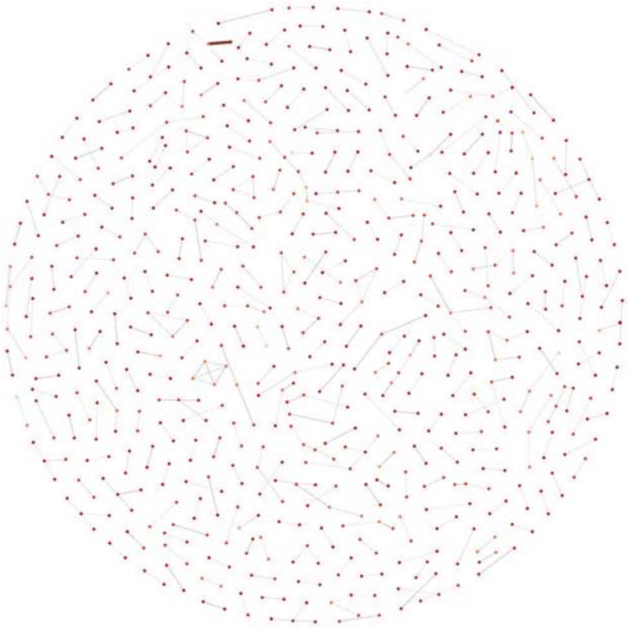


Fig. 5. PI Frequent Collaboration Network

4.2 Organization Network

To observe the relations between institutions receiving grants from the NSF, we build the *organization network*, i.e., the collaboration network between the organizations of the PIs of the funded grants. The constructed network of 3,450 nodes and around 27K edges is visualized in Figure 6. In this visualization, the nodes with high node degrees are located at the core of the network. The edge weights of these core nodes are usually high as well. This network is also weighted and the weight of the edges represents the number of grants collaborated among the two organizations. As seen in the figure, there is a group of nodes that are highly collaborative at the center of the figure.

The diameter of the organization network is 6.5 and the average path length is 3.07. However, we observed that there are many organizations that collaborate just once or twice. Many of these organizations are some short-run companies which were in busi-

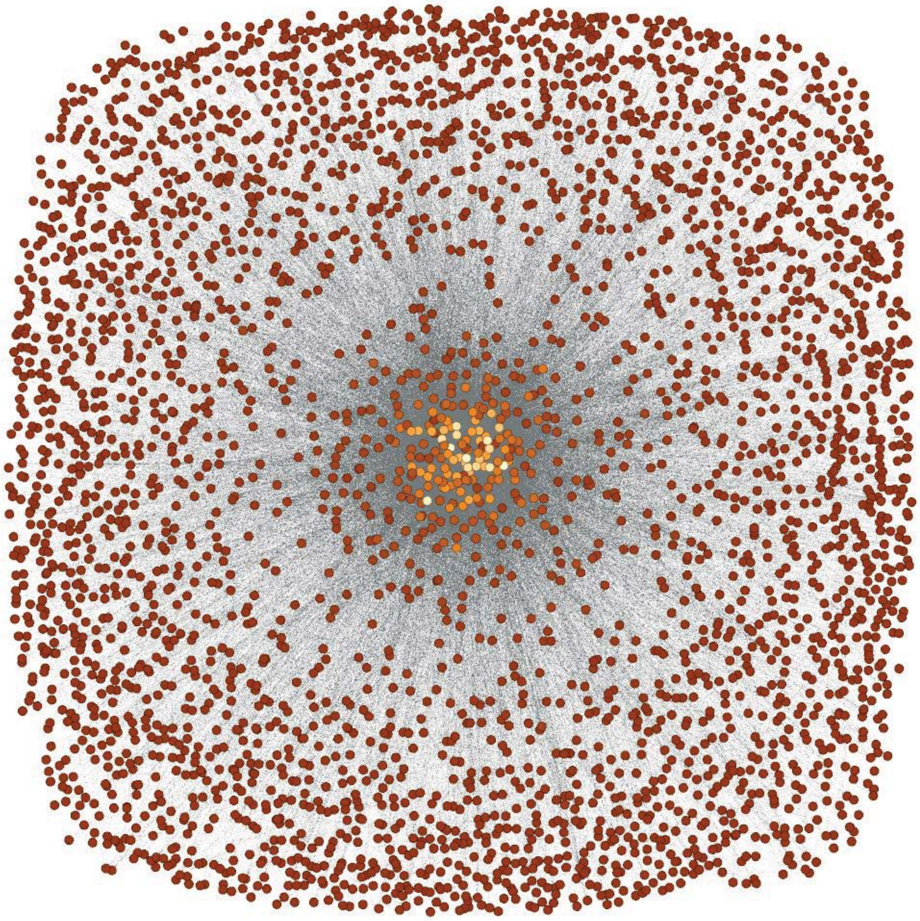


Fig. 6. Organization Collaboration Network

ness for a limited time. When we exclude such organizations from the network, the diameter of the network becomes 6.0 and the average shortest path becomes 2.75. Therefore, it can be concluded that the *six degrees of separation* is also observed in this network.

Figure 3-(d) presents the *clustering coefficient distribution* of the nodes in the organization network. The average clustering coefficient of the network is 0.34. The top clique size is 20 indicating that there are 20 organizations that have pairwise collaborated with each other. Along with the small average path length, the very high clustering coefficient compared to a random network of similar size indicates the *small world* characteristics for the collaboration network of organizations.

The *node degree distribution* of the organizations network is shown in Figure 3-(e). The degree distribution follows a power-law distribution with a fat tail. The average node degree for the network is 15.85, while the average weighted degree is 33.36. This indicates that on average each organization collaborated with its peers twice.

According to the Figure 3-(f) which presents the *rich club connectivity*, there is a rich club among organizations that receive federal funding. As observed as a highly connected core in the Figure 6, a group of organizations participate in most of the collaborations. To further investigate the rich club, we calculate the betweenness centrality, node degree, and weighted node degree for each node. Table 1 shows the rankings of the top 10 organizations based on the betweenness centrality and node degree values. Essentially, these top 10 organizations are part of the rich club in the network. As indicated above, for an organization, node degree expresses the number of distinct organizations which a collaboration was made while weighted node degree represents the total number of grants collaborated with the other institutions. According to the table, University of Colorado at Boulder is ranked 1^{st} both according to the betweennes centrality and node degree, while ranked 5^{th} based on weighted degree. This illustrates that even though University of Colorado at Boulder has collaborated with the highest number of organizations, it is not the highest according to the total number of grants collaborated. Another interesting result is that even though MIT is not one of the top ten organizations based on the node degree, it is the 4^{th} institution according to weighted node degree.

The *assortativity* value of this network is -0.09, which indicates that the organizations equally prefer to collaborate with high or low degree organizations. That is, different from the PI network where highly collaborating researchers slightly prefer to collaborate with researchers that also have high degrees, organizations are indifferent to the degree or popularity of the collaborators.

In order to illustrate the collaboration of organizations with the highest number of collaborative grants, we draw the network of the top 10 organizations in Figure 7. This network forms a clique, i.e., all organizations collaborated in grants with the others. The thickness of the edges presents the number of collaborations among these organizations. The boldness of the color of each node represents the weighted node degree for that node. The highest number of collaborations is between the University of Washington and the Arizona State University with 27 grants. The lowest collaboration among this group is between the Arizona State University and the Columbia University with 5 grants.

Table 1. Top 10 Organizations

Organization Rankings								
	1980s		1990s		2000s		Overall	
Metric	Organization	Value	Organization	Value	Organization	Value	Organization	Value
Betweenness Centrality	U. of Washington	11,540	U. of Washington	60,317	U. of Washington	47,912	U. of Colorado at Boulder	213,721
	U. of California-San Diego	11,147	U. of Minnesota-Twin C.	47,824	Arizona State U.	44,406	Arizona State U.	192,345
	U. of Michigan Ann Arbor	9,200	Pennsylvania State U.	43,756	U. of Colorado at Boulder	41,063	U. of Michigan Ann Arbor	183,380
	MIT	7,737	U. of Michigan Ann Arbor	37,827	Pennsylvania State U.	39,131	U. of Wisconsin Madison	182,452
	U. of Illinois at U-C	6,493	U. of Colorado at Boulder	37,214	GA Inst. of Technology	38,971	Pennsylvania State U.	180,111
	U. of California-Berkeley	6,358	U. of Wisconsin-Madison	36,730	U. of Michigan Ann Arbor	36,798	U. of Illinois at U-C	179,725
	Cornell U.	6,299	U. of California-Berkeley	35,989	Virginia Poly. Ins.	36,119	U. of Washington	175,303
	Colombia U.	6,275	Colombia U.	35,595	Ohio State U.	35,710	Colombia U.	163187
	Carnegie-Mellon U.	5,847	U. of Illinois at U-C	34,931	Colombia U.	34,570	MIT	153,406
	Indiana U. Of Pennsylvania	4,595	Suny at Stony Brook	33,796	U. of Minnesota-Twin C.	33,117	Cornell U.	151,373
Node Degree	U. of Washington	52	U. of Washington	121	U. of Washington	225	U. of Colorado at Boulder	344
	U. of California-San Diego	49	U. of Minnesota-Twin C.	112	Pennsylvania State U.	212	U. of Washington	336
	U. of Michigan Ann Arbor	44	Pennsylvania State U.	101	U. of Colorado at Boulder	205	U. of Wisconsin-Madison	330
	MIT	43	U. of Illinois at U-C	101	U. of Illinois at U-C	200	Colombia U.	324
	U. of Illinois at U-C	43	U. of Colorado at Boulder	100	U. of Arizona	197	Pennsylvania State U.	323
	U. of California-Berkeley	41	U. of Wisconsin-Madison	100	U. of Wisconsin-Madison	195	U. of Illinois at U-C	320
	Colombia U.	38	Colombia U.	100	U. of Michigan Ann Arbor	193	U. of Michigan Ann Arbor	319
	Cornell U.	36	U. of Michigan Ann Arbor	96	U. of California-Berkeley	191	Arizona State U.	308
	Carnegie-Mellon U.	34	U. of Arizona	96	Arizona State U.	187	Cornell U.	306
	U. of Wisconsin-Madison	34	MIT	92	Purdue U.	187	U. of California-Berkeley	301
Weighted Node Degree	U. of California-San Diego	90	U. of Minnesota-Twin C.	256	U. of Washington	706	Colombia U.	1197
	U. of Washington	73	U. of Washington	213	U. of Colorado at Boulder	606	U. of Illinois at U-C	1183
	U. of Illinois at U-C	65	U. of Illinois at U-C	197	U. of Illinois at U-C	604	U. of Washington	1152
	MIT	63	MIT	185	Pennsylvania State U.	599	MIT	1136
	U. of Michigan Ann Arbor	60	Colombia U.	182	Colombia U.	562	U. of Colorado at Boulder	1120
	U. of Wisconsin-Madison	60	U. of California-Berkeley	173	U. of Wisconsin-Madison	558	U. of Michigan Ann Arbor	1050
	Colombia U.	55	U. of Michigan Ann Arbor	169	U. of California-Berkeley	549	Pennsylvania State U.	1040
	Cornell U.	55	U. of Arizona	163	U. of Michigan Ann Arbor	537	Cornell U.	1035
	U. of California-Berkeley	55	U. of Colorado at Boulder	162	MIT	529	U. of California-Berkeley	1107
	U. of Georgia	50	U. of Wisconsin-Madison	155	U. of Arizona	520	U. of Wisconsin-Madison	992

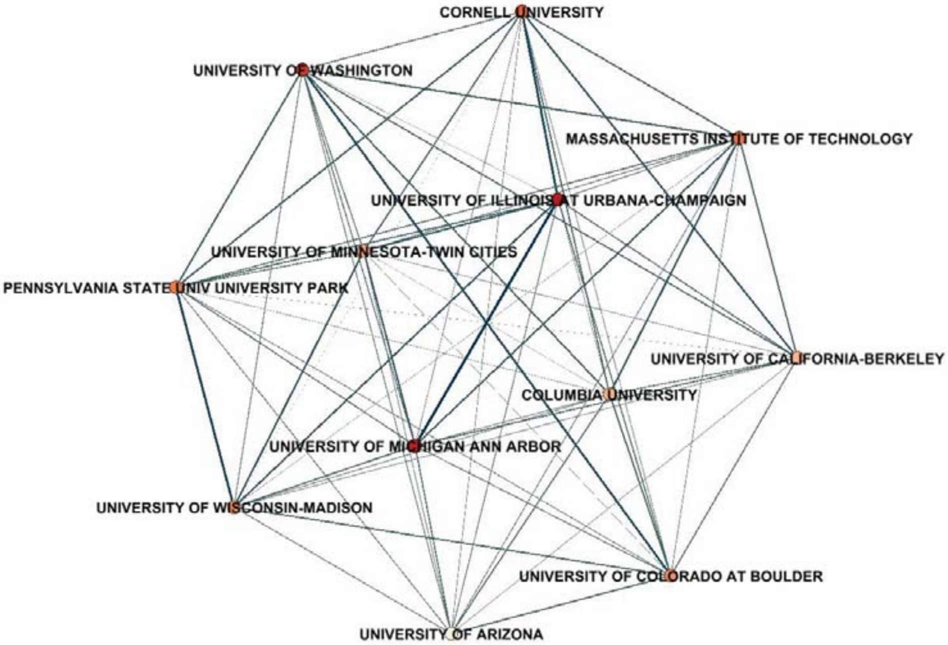

CORNELL UNIVERSITY
UNIVERSITY OF WASHINGTON
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
UNIVERSITY OF MINNESOTA-TWIN CITIES
PENNSYLVANIA STATE UNIV UNIVERSITY PARK
UNIVERSITY OF CALIFORNIA-BERKELEY
COLUMBIA UNIVERSITY
UNIVERSITY OF MICHIGAN ANN ARBOR
UNIVERSITY OF WISCONSIN-MADISON
UNIVERSITY OF COLORADO AT BOULDER
UNIVERSITY OF ARIZONA

Fig. 7. High Node Degree Organizations' Collaboration Network

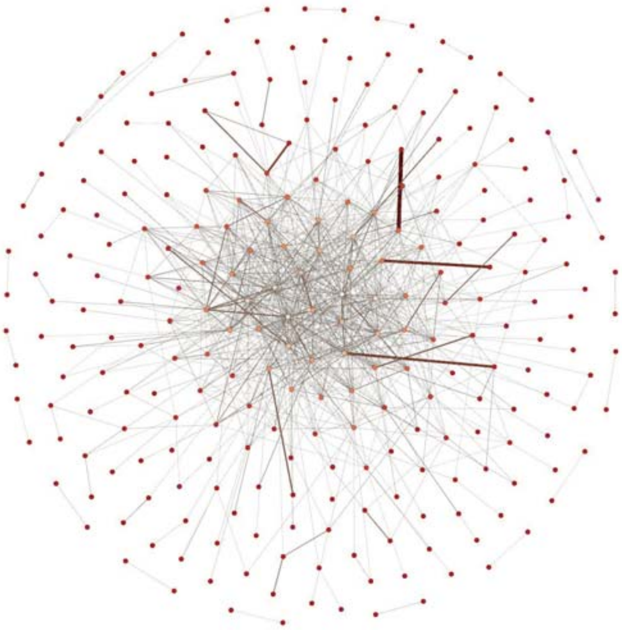


Fig. 8. Organization Frequent Collaboration Network

Moreover, in order to study frequent collaborations, we only consider edges where there are more than 10 collaborations in Figure 8. As seen in the figure, the ratio of the distinct pairs is lower than that of the PIs' frequent collaboration network in Figure 5. There are more triangles and even larger cliques in this network indicating frequent collaboration among those organizations.

Historical Perspective Above, we analyze organization collaboration network of the NSF grants from 1976 to 2011. However, in order to capture the changes within this network and to analyze the evolution of the network better, one needs to analyze the network at different time frames. Therefore, in this section, we generate the organization collaboration network for 1980s, 1990s and 2000s. For 2000s, we just analyze the grants awarded from 2000 to 2009 in order to have the same time frame length in each network.

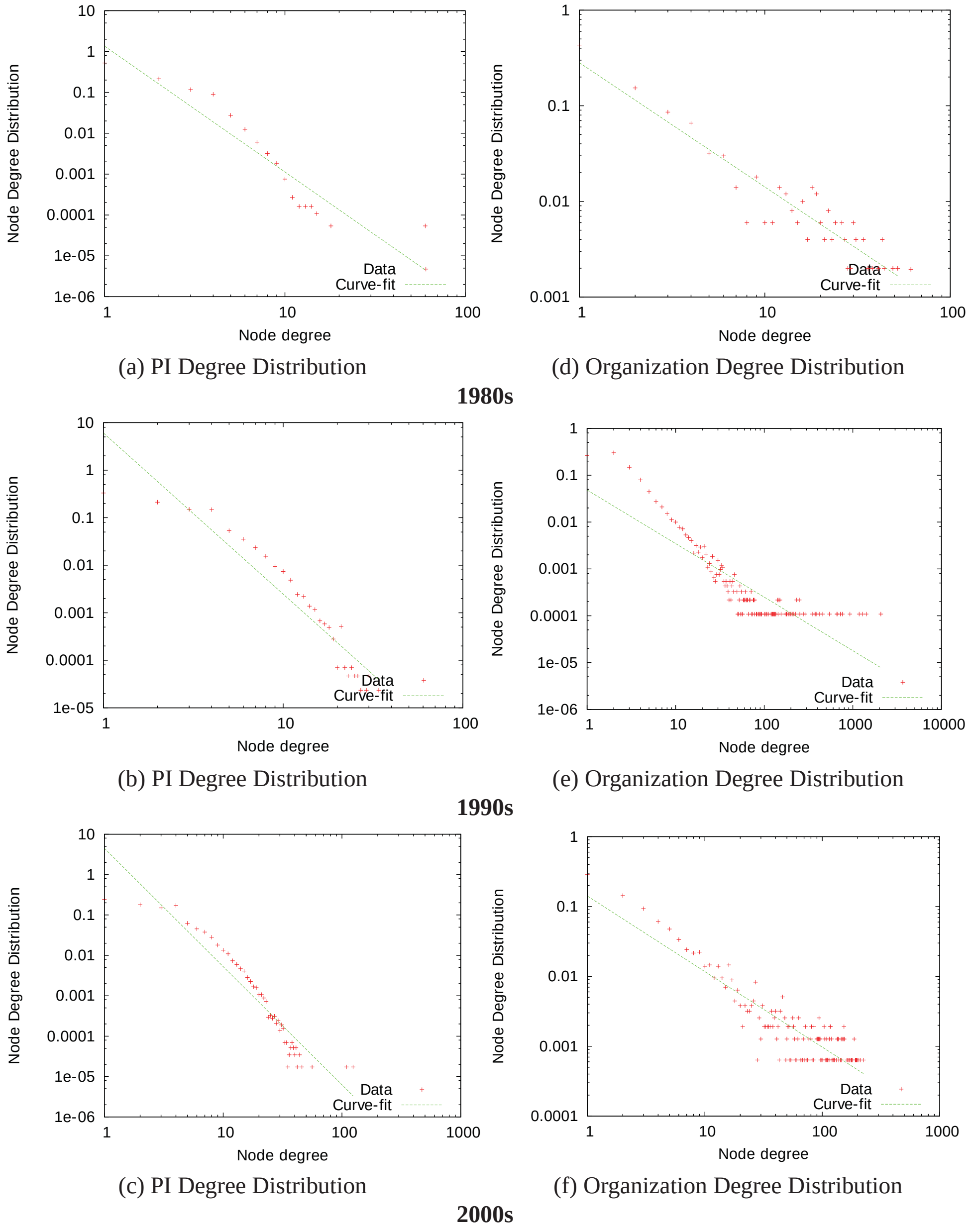
Table 2 represents the characteristics of organization collaboration networks of 1980s, 1990s and 2000s. According to this table, there is a steep increase in the average node degree and the average weighted node degree. The average node degrees of the networks are 5.39, 7.36 and 15.3, respectively, while the average weighted degrees of the networks are 7.5, 11.03, and 27.9, respectively. These values clearly illustrate that both the average number of collaborators and the total number of collaborations with other peer institutions have increased considerably. Additionally, the average number of collaborations made by an organization with its peers has become 1.8, while it was 1.4 in 1980s.

Parallel to the increase in the node degree, the organization network has become denser over the years. The diameter of the network is 9, 9, and 7, respectively for 1980s, 1990s, and 2000s. However, when we look at the overall network of the organization collaborations, the diameter is 6. Thus, the *six degrees of separation* has persisted in the organization collaboration network though the past three decades. Moreover, the average path length of the network decreases over the years, while the average clustering coefficient rises. In addition to the *small-world* characteristic of the organization collaboration network, it has become denser over the years as observed in typical social networks.

Table 1 shows the rankings of the top 10 organizations based on the betweenness centrality and node degree values for the 1980s, 1990s, 2000s, and overall (i.e., including all three periods in the network formation) networks. These top 10 organizations are part of the rich club in the network. According to the table, we can conclude that the *rich-get-richer* phenomenon is observed in the organization collaborations networks. Finally, Figure 9 and Figure 10 present several network characteristics of the Organization and PI collaboration networks for different time frames.

Table 2. Organization Network Characteristics Over Years

	80s	90s	00s	Overall
Avg. Degree	5.39	7.36	15.3	15.85
Avg. W. Degree	7.50	11.03	27.9	33.36
Diameter	9	9	7	6
Avg. Path Length	3.54	3.53	3.1	3.07
Avg. Clustering Coef.	0.15	0.19	0.32	0.34



(a) PI Degree Distribution

(d) Organization Degree Distribution

1980s

(b) PI Degree Distribution

(e) Organization Degree Distribution

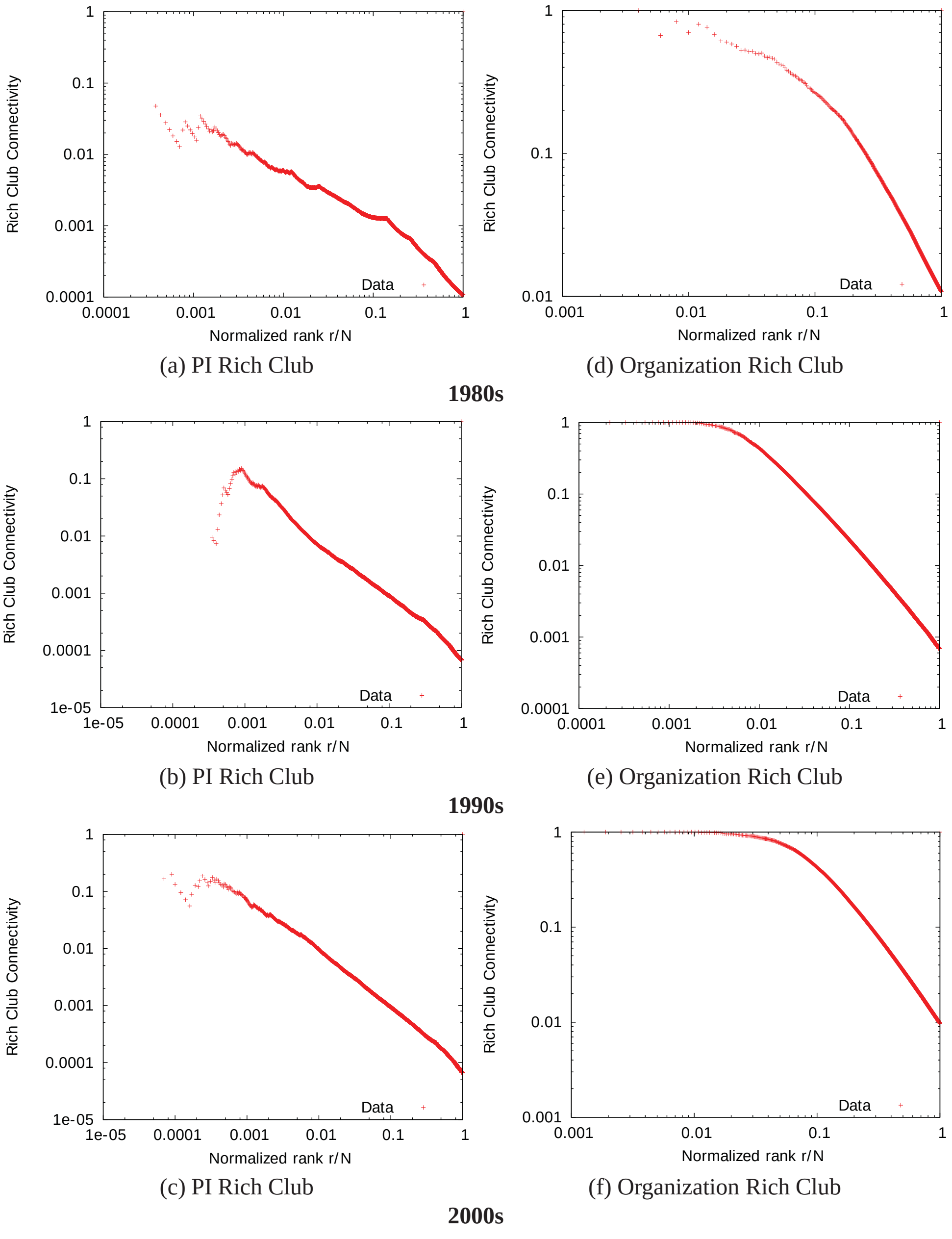
1990s

(c) PI Degree Distribution

(f) Organization Degree Distribution

2000s

Fig. 9. PI and Organization Degree Distributions (Historic)



(a) PI Rich Club

(d) Organization Rich Club

1980s

(b) PI Rich Club

(e) Organization Rich Club

1990s

(c) PI Rich Club

(f) Organization Rich Club

2000s

Fig. 10. PI and Organization Network Rich Club Metric (Historic)

4.3 State Network

In order to analyze the patterns among the home state of researchers, we construct the *state network*, i.e., the collaboration network between the states of the PIs. Figure 11 illustrates the state network constructed from the dataset where the nodes with higher betweenness centrality are located towards the center. In this network, there are 54 nodes and 1,285 edges. This network is highly clustered as the maximal clique size is 35 indicating that 35 states pairwise collaborate with each other. The assortativity coefficient is -0.13 for this network. The diameter of the network is 2 and average path length is 1.1. The average node degree of the network is 47.6 and the clustering coefficient is 0.95. All these metrics indicate a highly connected network.

There is no rich club in this network as almost all nodes are well connected. However, we can see the states that have many connections with higher degrees and weights represented with thick lines in the network. For instance, there is a frequent collaboration triangle between the states of New York (NY), California (CA) and Massachusetts (MA), which points to a large number of collaboration among these three states.

Furthermore, we tabulate the betweenness centrality, and weighted node degree for each node in Table 3. According to the table, betweenness centrality values are very close to each other for the top 5 collaborative states. However, average weighted node

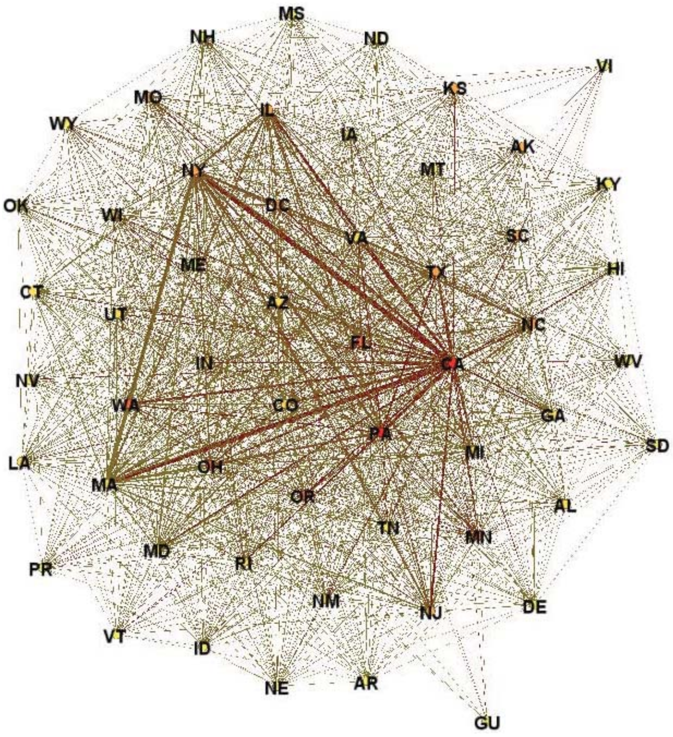


Fig. 11. State Collaboration Network based on betweenness centrality

Table 3. Top 10 States

State Rankings								
	1980s		1990s		2000s		Overall	
Metric	State	Value	State	Value	State	Value	State	Value
Weighted Node Degree	CA	379	CA	1,270	CA	3,982	CA	7,730
	NY	284	NY	1,024	NY	2,716	NY	5,946
	PA	187	MA	814	MA	2,384	MA	4,748
	MA	179	PA	605	PA	1,764	PA	3,774
	IL	166	IL	512	IL	1,594	IL	3,289
	TX	137	TX	465	TX	1,563	TX	3,178
	WA	101	MI	402	CO	1,330	CO	2,520
	MI	94	MD	399	FL	1,166	MI	2,333
	FL	93	NC	363	VA	1,105	FL	2,309
	CO	90	CO	333	MI	1,099	NC	2,146
Betweenness Centrality	CA	204.7	PA	75.8	FL	21.8	CA	8.2
	VA	77.2	HI	54.6	OR	20.6	NC	8.2
	NY	69.9	CA	26.1	WA	20.5	NY	8.2
	IL	59.3	NY	26.1	CA	12.8	OH	8.2
	FL	48.8	MA	23.2	IL	12.8	TX	8.2
	MA	36.6	NC	20.2	NC	12.8	FL	4.9
	PA	34.3	CO	19.5	NY	12.8	TN	4.9
	CO	34.1	TX	19.2	SC	9.6	MI	4.7
	NC	33.5	MI	16.3	AK	8.9	PA	4.6
	TX	32.1	WA	15.3	KS	7.1	IL	4.6

degree results indicate some differences among the top collaborative states. California (CA) is the most collaborative state with 7,730 inter-state collaborations. Since the node degrees are very close to each other we don't tabulate them. California (CA), North Carolina (NC), Ohio (OH), Pennsylvania (PA) and Texas (TX) have a node degree value of 53; which indicates that they have collaborated with all other states in at least one grant. On the other hand, Virgin Islands (VI), Guam (GU), Puerto Rico (PR), Wyoming (WY), South Dakota (SD), and Mississippi (MS) has collaborated with 13, 14, 35, 40, 41, 42, and 43 states, respectively, and are the states with the smallest node degrees.

Moreover, we analyze frequent collaborations among the states. In Figure 12, we draw the state collaboration network when the number of collaborations is greater than 250. There are 10 states which collaborated in more than 250 grants. As seen in the figure, California (CA) collaborated at least 250 times with all the other states in this network. The high collaboration among NY, CA and MA is more visible in this figure.

Historical Perspective Table 4 represents the network characteristics of state collaboration networks of 1980s, 1990s and 2000s, respectively. According to this table, there is a considerable increase in the average node degree and the average weighted node degree values. The average node degrees of the networks are 17.9, 34.2 and 43.3, respectively while the average weighted degrees of the networks are 57.5, 229.6, and 695.1, respectively. These values clearly illustrate that inter-state research collabora-

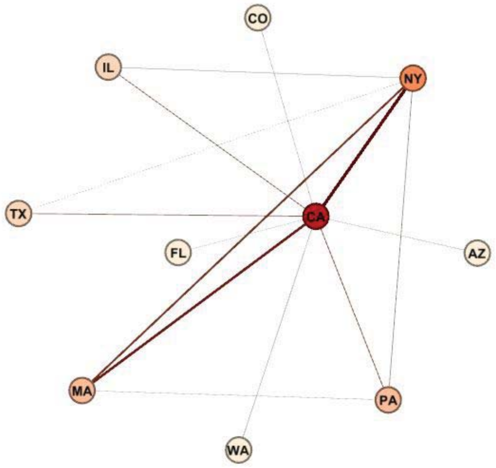


Fig. 12. State Frequent Collaboration Network

tion has increased over the years. Additionally, the average number of collaborations made by a state with its peers has become 16 in 2000s, while it was around 3 in 1980s. Thanks to this increase in the node degree, the overall state collaboration network has become almost a *clique*, i.e., full mesh. The diameters of the networks are 3, 3, and 3, respectively over the three decades. This is mainly due to two states, namely Virgin Islands (VI) and Guam (GU), which have very low collaborations. They don't have a research collaboration and a common collaborator in given time frames. However, when we look at the overall network of the organization collaborations, the diameter reduces to 2. Moreover, average path length of the network decreases over the years and has become 1.09 while the average clustering coefficient rises and has become 0.95 in the overall network.

Table 3 shows the rankings of the top 10 states based on the weighted node degree and the betweenness centrality values for the 1980s, 1990s, 2000s, and overall networks. The top 10 states have slightly changed over the years. Additionally, according to this table, we can conclude that *rich-get-richer* phenomenon applies to the state collaborations network, as well.

Table 4. State Network Characteristics Over Years

	80s	90s	00s	Overall
Avg. Degree	17.9	34.15	43.3	47.7
Avg. W. Degree	57.5	229.6	695.1	1405.0
Diameter	3	3	3	2
Avg. Path Length	1.69	1.37	1.18	1.09
Avg. Clustering Coef.	0.63	0.78	0.81	0.95

4.4 Directorates Networks

In the previous subsections, we construct three kinds of networks based on the whole NSF funding data. In this section, we construct these networks for each directorate separately to analyze the funding structures within each NSF directorate. The dataset contains 9 different NSF directorates, namely: Biological Sciences (BIO), Computer and Information Sciences (CSE), Education and Human Resources (EHR), Engineering (ENG), Geosciences (GEO), Mathematical and Physical Sciences (MPS), Office of Polar Programs (OPP), Social Behavioral and Economic Sciences (SBE), and Office of the Director (OD).

By considering each directorate we calculate node degree values of the PI, the organization, and the state networks. The graphs for node degree distributions of each directorate are shown in Figure 14 and Figure 13. When considering each directorate individually, the corresponding networks do not have a rich club similar to the whole network. Additionally, the assortativity value of each individual directorate network is close to zero indicating indifference to the popularity of the peers.

According to the clustering coefficient values of the directorate networks, GEO directorate has the highest clustering coefficient in the state network followed by BIO and ENG. These three directorates have the highest clustering coefficient values in the PI and the organization networks, as well, which indicates that the collaboration within these directorates are much more emphasized than the other NSF directorates. It also indicated, however, that researchers whose home directorate is one of these three directorates have a lower likelihood of collaborating with researchers from other directorates.

Additionally, as expected, the PI networks of directorates are better clustered than the overall PI network. Their diameter and average shortest path values are much smaller than those of the overall PI network, as well.

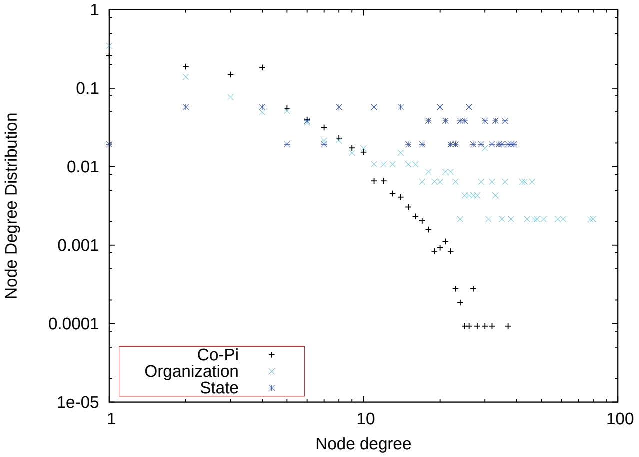
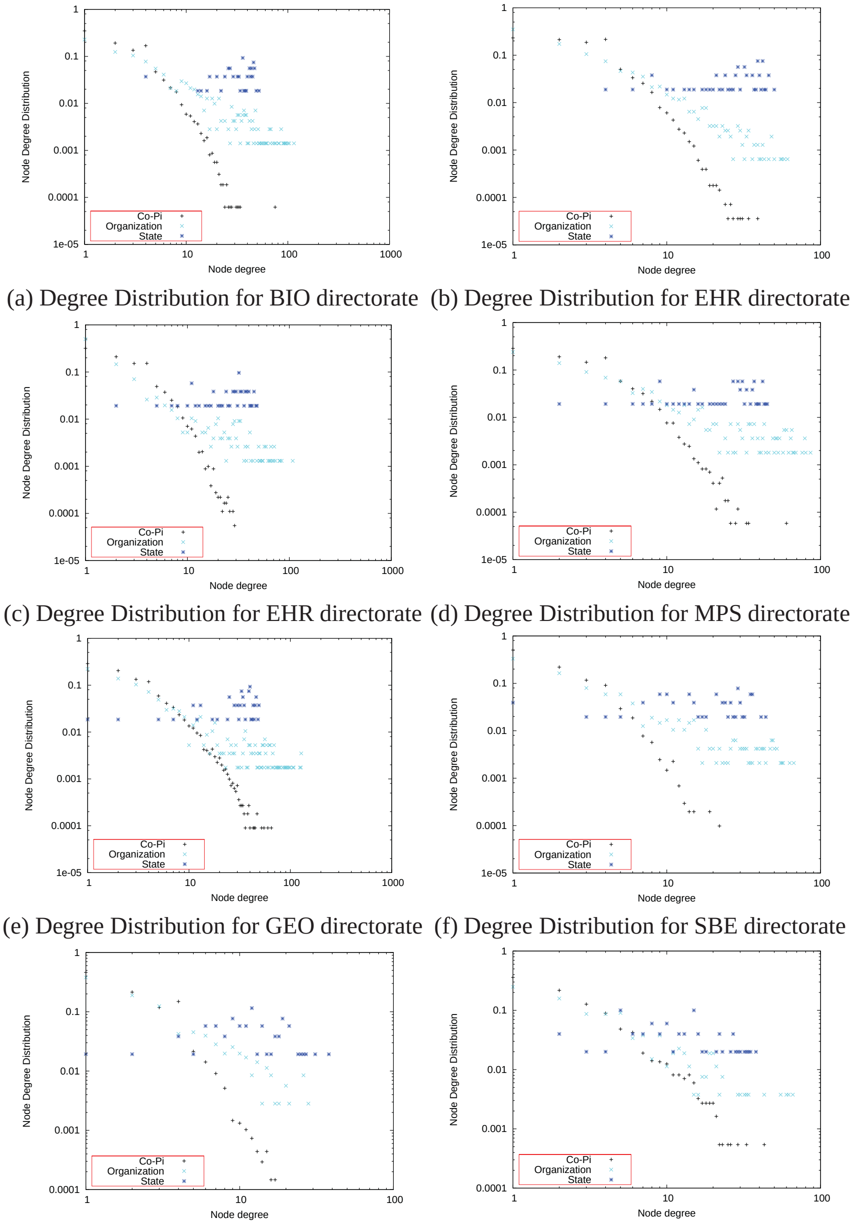


Fig. 13. Degree Distribution for CSE directorate



(a) Degree Distribution for BIO directorate (b) Degree Distribution for EHR directorate

(c) Degree Distribution for EHR directorate (d) Degree Distribution for MPS directorate

(e) Degree Distribution for GEO directorate (f) Degree Distribution for SBE directorate

(g) Degree Distribution for OD directorate (h) Degree Distribution for OPP directorate

Fig. 14. Metrics for Directorates Networks

4.5 Project Size

NSF categorizes the research projects based on funding levels. There are mainly three types of projects: small projects (typically, $<$500K), medium projects (typically, 500K-2M), and large projects (typically, $>$2M). In order to analyze the collaboration patterns within different project sizes, we generate the organization networks to investigate the collaboration among organizations at different funding levels.

Table 5 represents the network characteristics of organization collaboration networks of small, medium and large projects. According to the table, the average node degrees of the networks are 9.46, 6.42 and 14.6, respectively. Interestingly, organizations collaborated with more different peers in smaller projects than the medium projects. The average weighted degrees of the networks are 13.3, 16.2, and 21.1, respectively. Accordingly, the average number of collaborations made by an organization with its peers is 1.4 in small and large projects while it is 2.5 in medium projects. This also indicates that organizations collaborate with more peers in small and large projects. However, the average number of collaborations made by an organization with its peers is higher in medium projects, indicating more persistent collaborations at medium level of funding.

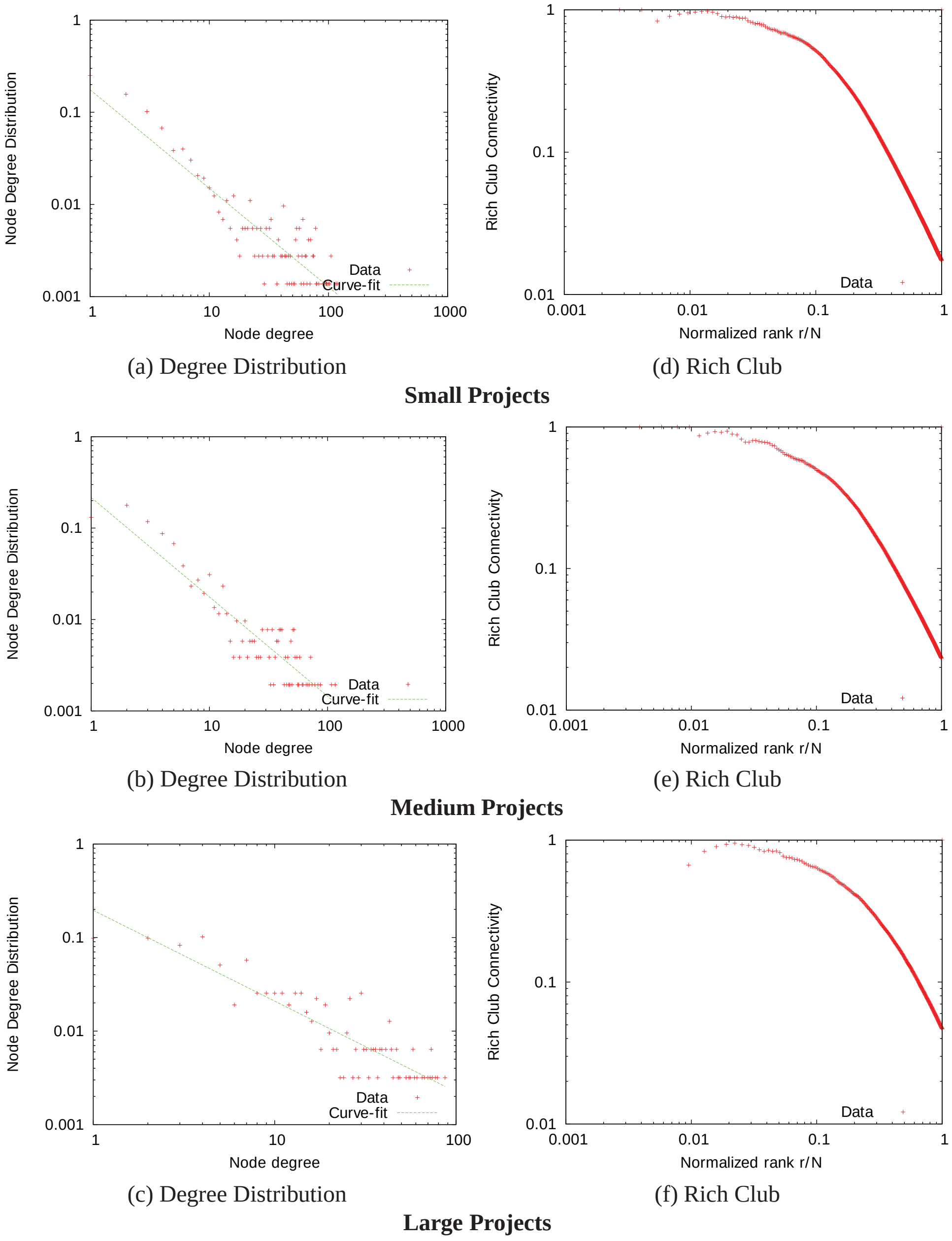
The diameters of the networks are 6, 8, and 6, respectively. Since the average number of collaborators of an organization is the lowest in medium project network, this network has the highest diameter. Moreover, in the large project collaboration network, we have the lowest average path length and the highest average clustering coefficient values. Thus, while all networks are *small-worlds*, the large project collaboration network exhibits the *small-world* characteristics more than the other funding levels.

Table 5. Network Characteristics for Different Project Sizes

	Small P.	Medium P.	Large P.	Overall
Avg. Degree	9.46	6.42	14.6	15.9
Avg. W. Degree	13.3	16.2	21.1	33.3
Diameter	6	8	6	6
Avg. Path Length	2.97	2.88	2.60	3.07
Avg. Clustering Coef.	0.36	0.50	0.59	0.34

5 Conclusion and Future Work

In this paper, we analyzed publicly available data on NSF funded grants to reveal the collaboration patterns among researchers. We derived three different kinds of networks to analyze the trends within the funding of the PI network, the organization network, and the state network. The PI network reveals a small-world characteristic but does not exhibit a power-law degree distribution. However, organization network exhibits a power-law degree distribution with a rich club that has majority of the collaborations. The state network is highly clustered but we identified the most central states in terms of collaborations. We construct these networks both for different time frames and as a



(a) Degree Distribution

(d) Rich Club

Small Projects

(b) Degree Distribution

(e) Rich Club

Medium Projects

(c) Degree Distribution

(f) Rich Club

Large Projects

Fig. 15. Organization Network Characteristics of Different Project Sizes

whole in order to compare the network characteristics of these networks for different time frames and capture the evolution of the NSF collaboration network over the time. We further analyze the funding network within each NSF directorate and find that some research fields are more collaborative than others in obtaining federal funding. Finally, we analyze these networks for small, medium, and large project sizes in order to observe the collaboration at different funding levels.

Our study revealed several interesting findings while reaffirming some of the anticipated characteristics of the funding network. We clearly observed a six degrees of separation in the state and the organization collaboration networks, while the degree of separation in the PI network is much higher. Another observation was that most of the funded collaborative projects had only two PIs.

Several extensions to the grant network analysis is of interest. In our study, we focussed on the successful grant proposals. To obtain a better picture of the collaboration patterns in the research funding, it would be very helpful to consider unsuccessful proposals. Furthermore, NSF uses different recommendation levels to rank grant proposals, e.g., Highly Recommended, Recommended, or Low Recommended. Consideration of these recommendation levels while constructing the collaboration networks would reveal more refined patterns. However, the challenge is to obtain such data without violating the privacy of PIs. Lastly, it would be interesting to observe the collaboration patterns in agencies other than the NSF and the United States.

References

1. M. Newman, A.-L. Barabasi, and D. J. Watts, *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton, NJ, USA: Princeton University Press, 2006.
2. M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, “Walking in Facebook: A Case Study of Unbiased Sampling of OSNs,” in *Proceedings of IEEE INFOCOM ’10*, San Diego, CA, March 2010.
3. M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, “I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system,” in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurements*, ser. IMC ’07. New York, NY, USA: ACM, 2007, pp. 1–14.
4. D. Ediger, K. Jiang, E. J. Riedy, D. A. Bader, C. Corley, R. Farber, and W. N. Reynolds, “Massive social network analysis: Mining twitter for social good,” in *39th International Conference on Parallel Processing (ICPP)*, San Diego, CA, Sep. 2010.
5. A. Java, X. Song, T. Finin, and B. Tseng, “Why we twitter: Understanding microblogging usage and communities,” in *WebKDD/SNA-KDD ’07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007, pp. 56–65.
6. M. E. J. Newman, “Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality,” *Phys. Rev. E*, vol. 64, no. 1, p. 016132, Jun 2001.
7. “Igraph,” http://igraph.sourceforge.net/.
8. E. Arslan, M. H. Gunes, and M. Yuksel, “Analysis of Academic Ties: A Case Study of Mathematics Genealogy,” in *IEEE Globecom Workshop on Complex and Communication Networks*, Houston, TX, December 9, 2011.
9. D. Cheung and M. H. Gunes, “A Complex Network Analysis of the United States Air Transportation,” in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Istanbul, Turkey, August 26–29, 2012.

10. C. Cotta and J.-J. Merelo, “The complex network of ec authors,” in *ACM SIGEVOlution*, vol. 1, pp. 2–9, June 2006.
11. S. Nerur, R. Sikora, G. Mangalaraj, and V. Balijepally, “Assessing the relative influence of journals in a citation network,” *Communications of the ACM*, vol. 48, pp. 71–74, November 2005.
12. P. Zhang and L. Koppaka, “Semantics-based legal citation network,” in *ACM Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pp. 123–130, Palo Alto, CA, June 4–8, 2007.
13. J. Naruchitparames, M. H. Gunes and S. Louis, “Friend Recommendations in Social Networks using Genetic Algorithms and Network Topology,” in *IEEE Congress on Evolutionary Computation*, New Orleans, LA, June 5-8, 2011.
14. A. E. Breland, M. H. Gunes, K. A. Schlauch, and F. C. Harris, in “Mixing Patterns in a Global Influenza A Virus Network Using Whole Genome Comparisons,” in *IEEE Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Montreal, Canada, May 2–5, 2010.
15. A. E. Breland, K. A. Schlauch, M. H. Gunes, and F. C. Harris, in “Fast Graph Approaches to Measure Influenza Transmission Across Geographically Distributed Host Types,” in *ACM International Workshop on Graph Theoretic Approaches for Biological Network Analysis (IWBNA 2010)*, Niagara Falls, NY, Aug 2-4, 2010.
16. M. B. Akgun and M. H. Gunes, “Link-level Network Topology Generation,” in *ICDCS Workshop on Simplifying Complex Networks for Practitioners*, Minneapolis, MN, June 24, 2011.
17. H. Kardes, M. Gunes, and T. Oz. Cheleby: A subnet-level internet topology mapping system. In *Fourth International Conference on Communication Systems and Networks (COMSNETS)*, pp. 1–10, Jan. 2012.
18. M. B. Akgun and M. H. Gunes, “Bipartite Internet Topology at the Subnet-level,” in *IEEE International Workshop on Network Science (NSW 2013)*, West Point, NY, 29 Apr – 1 May 2013.
19. D. M. Ryfe, D. Mensing, H. Ceker, and M. H. Gunes, “Popularity is Not the Same Thing as Influence: A Study of the Bay Area News System,” in *Journal of the International Symposium of Online Journalism (#ISOJ)*, vol. 2(2), Fall 2012.
20. D. Ramos, M. H. Gunes, D. Mensing and D. M. Ryfe, “Mapping Emerging News Networks: A Case Study of the San Francisco Bay Area,” in *Studies in Computational Intelligence: Complex Networks*, vol. 424, pp. 237–244, 2013.
21. X. Li, H. Chen, Z. Zhang, and J. Li, “Automatic patent classification using citation network information: an experimental study in nanotechnology,” in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 419–427, Vancouver, BC, Canada, June 18–23, 2007.
22. B. H. Hall, A. B. Jaffe, and M. Trajtenberg, “The nber patent citations data file: Lessons, insights and methodological tools,” in *NBER Working Papers 8498, National Bureau of Economic Research, Inc*, 2001.
23. K. Komurov, M. H. Gunes, and M. A. White, “Fine-scale dissection of functional protein network organization by statistical network analysis,” in *PLoS ONE*, vol. 4(6), June 2009.
24. A. Dittrich, M. H. Gunes and S. Dascalu, “Network Analysis of Software Repositories: Identifying Subject Matter Experts,” in *Studies in Computational Intelligence: Complex Networks*, vol. 424, pp. 187–198, 2013.
25. C. Zachor and M. H. Gunes, “Software Collaboration Networks,” in *Studies in Computational Intelligence: Complex Networks*, vol. 424, pp. 257–264, 2013.
26. T. Morelli and M. H. Gunes, “Video Game Industry as a Complex Network,” in *2nd Workshop on Social Network Analysis and Applications (SNAA 2012)*, Istanbul, Turkey, 26 Aug, 2012.

27. H. Kardes, A. Sevincer, M. H. Gunes and M. Yuksel, “Six Degrees of Separation among US Researchers,” in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, Istanbul, Turkey, 26-29 Aug, 2012.
28. “National science foundation,” http://www.nsf.gov/about/.
29. “National science foundation award search,” http://www.nsf.gov/awardsearch/.
30. Herr II, Bruce W., Talley, Edmund M, Burns, Gully APC, Newman, David La Rowe, Gavin., “Interactive science map of nih funding,” http://scimaps.org/maps/nih/2007/, 2009.
31. B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, “Planetlab: an overlay testbed for broad-coverage services,” *SIGCOMM Comput. Commun. Rev.*, vol. 33, pp. 3–12, July 2003.
32. M. Bastian, S. Heymann, and M. Jacomy, “Gephi: An open source software for exploring and manipulating networks,” 2009.
33. D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.
34. M. E. J. Newman, “Assortative Mixing in Networks,” *Physical Review Letters*, vol. 89, no. 20, pp. 208 701+, Oct. 2002.
35. J. J. McAuley, Costa, and T. S. Caetano, “Rich-club phenomenon across complex network hierarchies,” *Appl. Phys. Lett.*, vol. 91, p. 084103, 2007.