2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Six Degrees of Separation among US Researchers

Hakan Kardes, Abdullah Sevincer, Mehmet Hadi Gunes, and Murat Yuksel
Computer Science and Engineering Department
University of Nevada, Reno
1664 N. Virginia St. Reno, USA
{*hkardes, 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 National Science Foundation (NSF), a major source of research funding in the States, 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. We further analyze the directorates to identify the differences in collaboration trends between disciplines.**

***Keywords*-Complex networks; Complex network analysis; Research funding networks; Six degrees of separation; NSF.**

I. Introduction

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

In this paper, we analyze the collaboration of researchers when they obtain federal funding. For this study, we analyze the funding data of the National Science Foundation (NSF), an independent federal agency established in 1950. NSF has an annual budget of about $7.4 billion (FY 2011) [2], 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 [3]. The data released by NSF includes the Principle Investigator (PI), co-PIs (if any), organizations, directorate, grant amount and several others for the funded projects. In order to analyze the collaboration patterns within the NSF research funding network, we generate 3 types of networks from the provided dataset. First, we construct the *PI network* where we analyze the social interaction of researchers. The PI network shows the collaboration patterns 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 central organizations and collaboration trends. We also derive the *state network* to study the collaboration among the states in obtaining federal funding. We further analyze the funding network of each NSF directorate to find their distinct properties.

The main goal of this paper is to collect the NSF funding dataset, discover interesting complex network structures from the dataset and derive characteristics 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 NIH and NSF data sets using statistics [10] or visualization [4]. To best of our knowledge, however, this paper is the first study to analyze the funding data as a complex network.

II. Data Collection

NSF provides historic information on funded grants at its website. A search engine provides access to grant information. Each search query turns at most 3,000 grants at a time, and there is a rate limit of queries that a computer is allowed to perform. We implement a crawler using PlanetLab [5] infrastructure to download the data in parallel. Overall, we download a total of 279,862 entries for funded grants spanning from 1976 to December 2011.

Each NSF grant has a PI, organization, co-PI, directory and several other fields in the database. 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.

978-0-7695-4799-2 2012

DOI 10.1109/ASONAM.2012.109

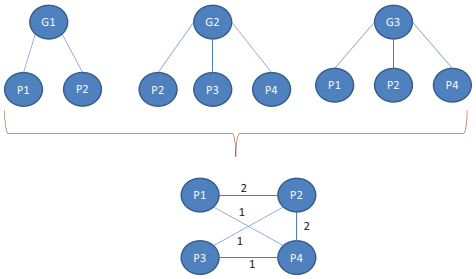


Figure 1. PI Network Construction

III. Networks Analysis of the NSF Funding

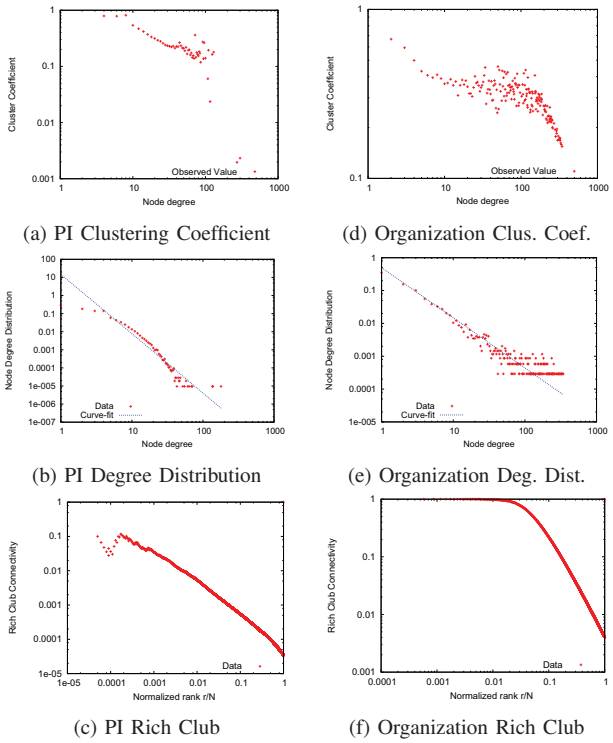
In order to analyze the collaboration patterns within the research funding network, we generate 3 types of networks from the dataset and visualize them with Gephi [6]. First network we explore is the *PI network*, i.e. the collaboration network between PIs of the grants to understand the relationships and characteristics of the collaboration between researchers. To construct the PI network, we connect 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 build the *organization network*, i.e. the collaboration network between the organizations of the PIs of the funded grants to observe the relations between institutions to receive grants from the NSF. Finally, we construct the *state network*, i.e. the collaboration network between the states of the PIs in order to analyze the patterns among the home states of researchers.

Furthermore, we analyze the funding network within each NSF directorate to find their distinct properties. We compared directorates to better understand the collaboration patterns within different research fields.

A. PI Network

The *PI network* has about 104K nodes and 204K 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. Average path length is a bit higher than similar other social networks. 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 PIs. Additionally, as the PI network is sparse, the number of interdisciplinary grants which would make the PI network better connected is low. As indicated in Directorates Networks Section, the PI network of each individual directorate is well-connected with a low diameter and average path length values but we do not observe this behavior when we consider all directorates together.

Figure 2-(a) presents the *clustering coefficient distribution* of the nodes in the PI network. The average clustering



(a) PI Clustering Coefficient (d) Organization Clus. Coef.

(b) PI Degree Distribution (e) Organization Deg. Dist.

(c) PI Rich Club (f) Organization Rich Club

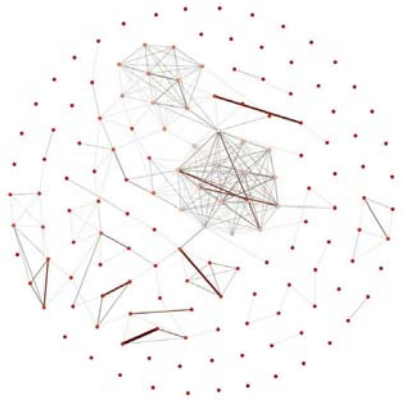
Figure 2. PI and Organization Network Metrics

coefficient of the graph is 0.46. This is considerably higher than a random network of similar size, which happens in *small world* [7] networks.

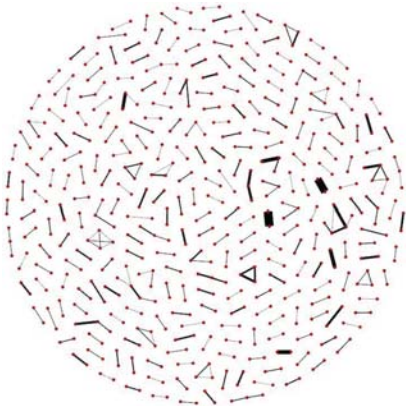
The *node degree distribution* in Figure 2-(b) does not exhibit a power-law distribution but rather results in a declining curve. We think this is mainly due to the fact that collaborations require considerable effort and researchers are limited in the number of projects they can contribute. The *average node degree* for the network is 3.94, while the *weighted node degree* is 4.8. 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 [8]. That is, PIs who have high collaborations slightly tend to work together rather than collaborating with PIs that have low collaborations. Moreover, Figure 2-(c) shows the *rich club connectivity* of the PI network [9]. There is not an obvious rich club that contains most of the collaborations even though such phenomenon has been observed in citation networks.

In order to better analyze highly collaborative PIs, we draw the network of the PIs with highest node degrees in Figure 3-(a). In this figure, the thickness of the edges illustrates the number of collaborations among PIs while the boldness of the color of each node represents the weighted node degree, i.e. total number of collaborative grants for that node. In this 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 collab-



(a) PIs with High Degrees



(b) PI Frequent Collaboration

Figure 3. PI Collaboration Networks from Different Perspectives

orations among researchers, we construct the PI network by only considering the highest weighted edges in Figure 3-(b). As seen in the figure, there are many distinct pairs of PIs while there are a few triangles and larger cliques in the network. This indicates most of the frequently funded research teams consist of two PIs.

B. Organization Network

To observe the relations between institutions to receive funding 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 4-(a) where the nodes with high degrees are located at the center. The edge weights of these core nodes are usually high as well. The edge weights represent the number of grants collaborated among the two organizations. As seen in the figure, there is a group of nodes that are highly collaborative.

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 business 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 observed in this network.

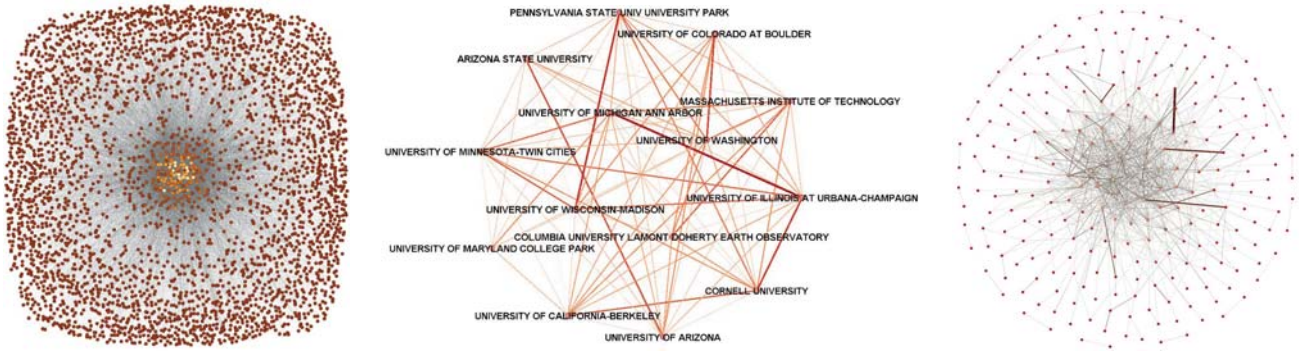
Table I
TOP 10 ORGANIZATIONS

Metric	Organization	Value
Betweenness Centrality	Univ. of Colorado at Boulder	213,721
	Arizona State Univ.	192,345
	Univ. of Michigan Ann Arbor	183,380
	Univ. of Wisconsin Madison	182,452
	Pennsylvania State Univ Univ. Park	180,111
	Univ. of Illinois at Urbana-Campaign	179,725
	Univ. of Washington	175,303
	Colombia Univ.	163187
	Massachusetts Institute of Technology	153,406
	Cornell Univ.	151,373
Node Degree	Univ. of Colorado at Boulder	344
	Univ. of Washington	336
	Univ. of Wisconsin-Madison	330
	Colombia Univ.	324
	Pennsylvania State Univ Univ. Park	323
	Univ. of Illinois at Urbana-Campaign	320
	Univ. of Michigan Ann Arbor	319
	Arizona State Univ.	308
	Cornell Univ.	306
	Univ. of California-Berkeley	301
Weighted Node Degree	Colombia Univ.	1197
	Univ. of Illinois at Urbana-Campaign	1183
	Univ. of Washington	1152
	Massachusetts Institute of Technology	1136
	Univ. of Colorado at Boulder	1120
	Univ. of Michigan Ann Arbor	1050
	Pennsylvania State Univ Univ. Park	1040
	Cornell Univ.	1035
	Univ. of California-Berkeley	1107
	Univ. of Wisconsin-Madison	992

Figure 2-(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 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.

The *node degree distribution* follows a power-law distribution with a fat tail as shown in Figure 2-(e). 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 twice with its peers.

According to the Figure 2-(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 4-(a), 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 I shows the rankings of the top 10 organizations based on betweenness centrality and node degree values. These top 10 organizations are part of the rich club in the network. For an organization, node degree values represent the number of distinct organizations which a collaboration was made while weighted node degree represents the total number



(a) Organization Col. Network (b) Organizations with High Degrees (c) Organization Frequent Col. Network

Figure 4. Organization Collaboration Networks from Different Perspectives

of grants collaborated with 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 4^{th} based on weighted degree. This illustrates that even University of Colorado at Boulder has collaborated with highest number of organizations, it is not the first according to the total number of grants collaborated. Another interesting result is that even 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.

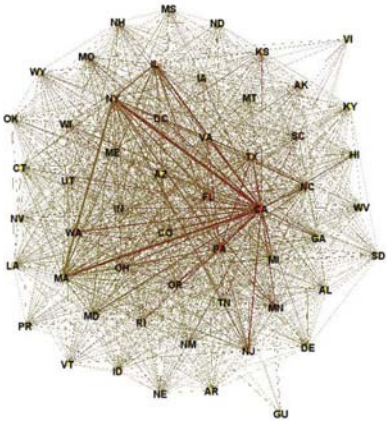
The *assortativity* value of this network is -0.09, which indicates that the organizations equally prefer to collaborate with high or low degree organizations. Different from PI network where highly collaborating researchers slightly prefer to collaborate with researchers that also have high degrees,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

In order to illustrate the collaboration of organizations with the highest number of collaborative grants, we draw the network of top 10 organizations in Figure 4-(b). 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 highest number of collaborations is between the University of Illinois at Urbana-Campaign and University of Michigan Ann Arbor 31 grants. The lowest collaboration among this group is between the Arizona State University and the Columbia University with 4 grants. The boldness of the color of each node represents the weighted node degree for that node.

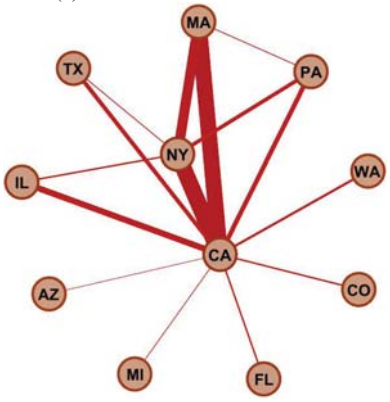
To study frequent collaborations, we only consider edges where there are more than 10 collaborations in Figure 4-(c). As seen in the figure, the ratio of distinct pairs is lower than that of PI's frequent collaboration network in Figure 3-(b). There are more triangles and even larger cliques in this network indicating frequent collaboration of those organization groups.

C. 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 5-(a) illustrates the state network where the nodes with higher betweenness centrality are located at the center. In this network, there are 54 nodes and 1,289 edges. This network is highly clustered as the maximal clique size is 35 indicating that 35 states pairwise collaborate with each other. The diameter of the network is 2 and average path length is 1.1. The average node degree of the network is 47.7 and the clustering coefficient is 0.95. All these metrics indicate a highly connected network. The assortativity coefficient is -0.13 for this network.



(a) State Collaboration Network



(b) State Frequent Collaboration Network

Figure 5. State Collaboration Networks from Different Perspectives

Table II
TOP 10 STATES

Betweenness Centrality	State	Weighted Node Degree	State
7.78	CA	8064	CA
7.78	NC	6341	NY
7.78	OH	5147	MA
7.78	PA	3878	PA
7.78	TX	3372	IL
4.82	DC	3202	TX
4.82	IL	2676	CO
4.82	NJ	2381	MI
4.82	NY	2369	FL
4.52	FL	2364	NC

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 between the states of New York (NY), California (CA) and Massachusetts (MA), which points to a high number of collaborations.

Furthermore, we tabulate the betweenness centrality, and weighted node degree for each node in Table II. According to the table, betweenness centrality values are very close to each other for the top 5. However, average weighted node degree results indicate some differences where California (CA) is the most collaborative state with 8,064 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 5-(b), we draw the network for 11 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.

D. Directorates Networks

In the previous subsections, we construct three kinds of networks based on the complete 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), and Social Behavioral and Economic Sciences (SBE).

By considering each directorate we calculate node degree distribution of the PI, organization, and state networks as shown in Figure 6. When considering each directorate individually, the corresponding networks do not have a rich club. 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 among 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 collaboration within these directorates are much more emphasized than the other directorates.

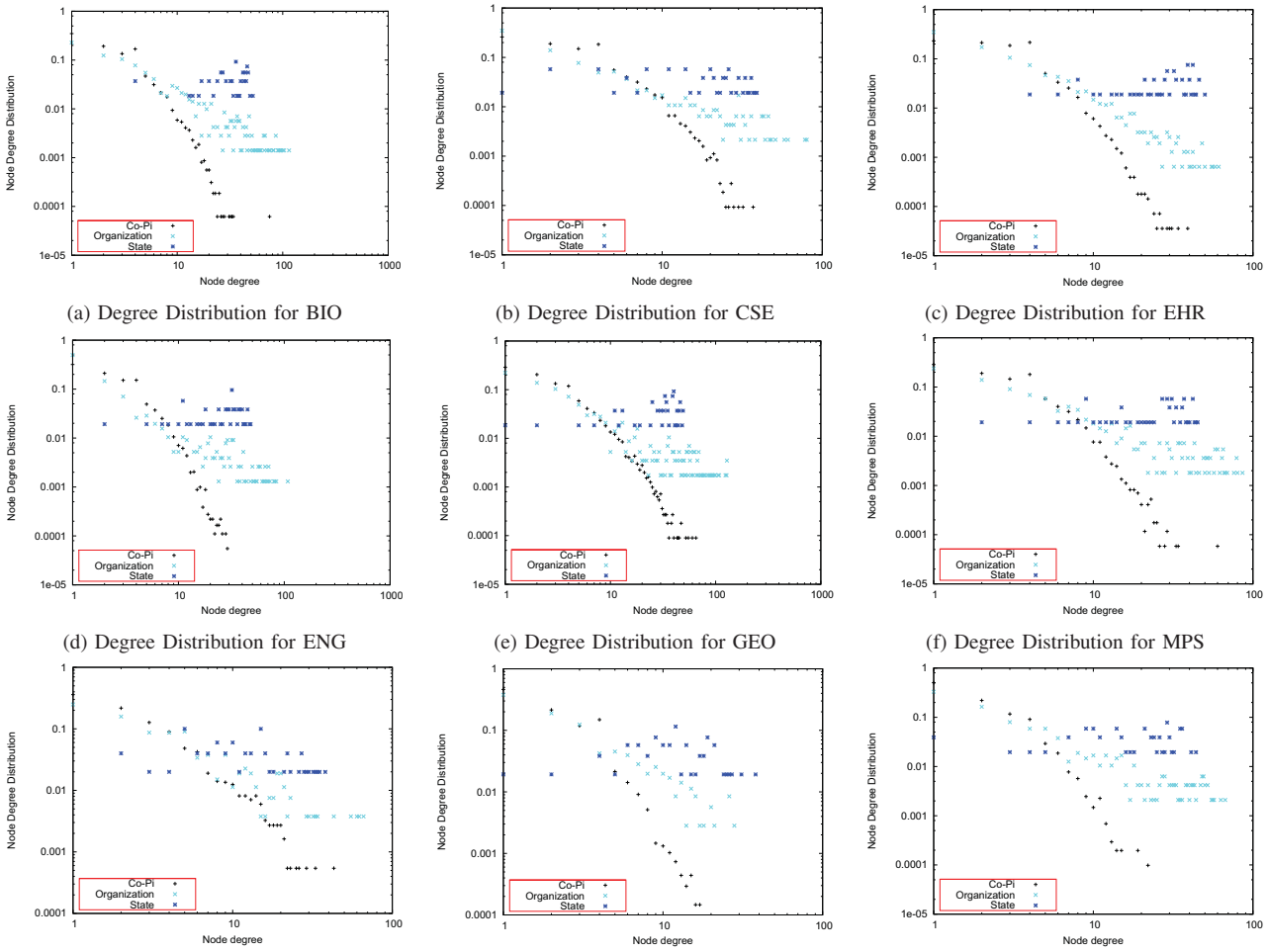
Additionally, as expected, the PI networks of directorates are better clustered than the overall PI network. Their diameter and average shortest path values are much smaller than those of the overall PI network as well.

IV. CONCLUSION AND FUTURE WORK

In this paper, we analyzed publicly available data on NSF funded grants to reveal the collaboration among researchers. We derived three different kinds of networks to analyze the trends within the funding of PI, organization and state networks. The PI network reveals small world characteristics but does not exhibit a power-law degree distribution. However, organization network exhibits a power-law degree distribution with a rich club of organizations that has most of the collaborations. The state network is highly clustered. We further analyzed the funding network within each NSF directorate and found that some research fields are more collaborative than others in obtaining federal funding. Even though researchers have studied NIH and NSF data using statistics or visualization, this paper is, to best of our knowledge, the first study to analyze the funding data as a complex network.

Our study revealed several interesting findings while reaffirming some of the anticipated properties of the funding network. We clearly observed a six degrees of separation in the state and 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 successful grant proposals. To obtain a better picture of collaborative patterns in the funded research network, consideration of unsuccessful proposals would be very helpful. Further, NSF uses different recommendation levels to rank grant proposals, e.g., Highly Recommended, Recommended, or Low Recommended. Consideration of these recommendation levels of each grant while constructing the collaboration networks would surely reveal more refined patterns. However, the challenge is to obtain such data with privacy restrictions.



(a) Degree Distribution for BIO

(b) Degree Distribution for CSE

(c) Degree Distribution for EHR

(d) Degree Distribution for ENG

(e) Degree Distribution for GEO

(f) Degree Distribution for MPS

(g) Degree Distribution for OPP

(h) Degree Distribution for OD

(i) Degree Distribution for SBE

Figure 6. Metrics for Directorates Networks

Additionally, we may consider grant amount in dollars as a metric while constructing the networks. Furthermore, in order to analyze the collaboration patterns within different project sizes, these networks might be generated and analyzed for different funding levels. Moreover, the grant networks may be generated for certain time intervals in order to analyze the networks for different times. This will allow us to capture the evolution of the collaboration networks over the time. Lastly, it would be interesting to observe the collaboration network patterns in agencies other than NSF and the U.S.

REFERENCES

[1] Mark Newman, Albert-Laszlo Barabasi, and Duncan J. Watts. 2006. The Structure and Dynamics of Networks: (Princeton Studies in Complexity). Princeton University Press, Princeton, NJ, USA.

[2] "National science foundation," http://www.nsf.gov/about/.

[3] "National science foundation award search," http://www.nsf.gov/awardsearch/.

[4] 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.

[5] Brent Chun, David Culler, Timothy Roscoe, Andy Bavier, Larry Peterson, Mike Wawrzoniak, and Mic Bowman. 2003. PlanetLab: an overlay testbed for broad-coverage services. SIGCOMM Comput. Commun. Rev. 33, 3 (July 2003), 3-12.

[6] Bastian M., Heymann S., Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[7] Watts, Duncan J.; Strogatz, Steven H. 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684): 440442.

[8] M. E. J. Newman. 2002. Assortative mixing in networks. Physical Review Letters 89 (20): 208701

[9] J. J. McAuley, Costa, and T. S. Caetano, "Rich-club phenomenon across complex network hierarchies," *Appl. Phys. Lett.*, vol. 91, p. 084103, 2007.

[10] Hather GJ, Haynes W, Higdon R, Kolker N, Stewart EA, et al. (2010) The United States of America and Scientific Research. PLoS ONE 5(8): e12203. doi:10.1371/journal.pone.0012203