Analyzing Stanford's Academic Network

Ashton Anderson
Stanford University
ashton@cs.stanford.edu

Stefan Krawczyk
Stanford University
stefank@cs.stanford.edu

ABSTRACT

1. INTRODUCTION

The university is a fundamental social institution. It is a haven for research and is where the next generation gets educated. Gaining knowledge of its properties and understanding how it functions are thus of inherent value to society. In our project, we analyzed Stanford's academic network, where faculty members are vertices and edges between faculty members represent academic relationships like publication co-authorship, being on the same dissertation committee, and being on the same grant.

Our motivation for studying this particular network, aside from its inherent social value, is that it has several interesting features from a network analysis standpoint. Firstly, it is the prototypical example of a hierarchical network. Secondly, there are many different relationships between people in the network; essentially there are many different networks that co-exist on the same underlying vertex set. Finally, our data has a strong temporal element since we have yearly snapshots of each network over the period from 1994-2007.

Our project is divided into three main goals. The first goal is to discover the community structure in the network. Its hierarchical structure is well-known, but how well does the activity of the network conform to the university's official divisions? Are there more coherent groups than the faculties? We applied graph clustering techniques to separate the network into natural clusters. With this partitioning, we quantify how isolated or connected certain fields of study are from the others, and how weak or strong interdisciplinary collaboration is between specific departments. The temporal nature of our data made this even more interesting; we quantify how collaboration varies over time. Our second goal is to verify if Burt's theory of structural holes apply to the university setting. Are faculty who fill structural holes more successful researchers? Finally, the third goal is to briefly explore if we can leverage the many different edge types we have to predict grant proposal success.

1.1 Related Work

1.1.1 Structural Holes and Good Ideas

The theory of structural holes was pioneered by Ron Burt [1, 4, 3]. He found that people who fill structural holes in networks derive social capital from their advantageous position. In his work, Burt proposes "network constraint measures" to detect hole signatures and thus identify structural holes. His studies were done in commercial settings (in [3], he analyzed a large electronics company), with information gathered from questionnaires. In this paper, he gave evidence that people who fill structural holes accrue various benefits by showing correlation between his constraint measures and higher compensation, promotions, and good ideas.

1.1.2 Academic Networks

The analysis of academic networks has been largely focused on citation graphs, but we are not aware of any previous work that focus on a single institution.

1.2 Data

Our networks are of the following form: faculty members are vertices and edges are academic interactions. We used five edge sets: `PubCoAuthor` is publication co-authorship, `DissCommCoMember` is being on the same dissertation committee, `GrantAwarded` is being on the same successful grant proposal, `GrantRejected` is proposing a grant which was rejected, and finally we call the union of the above edge sets `Combined`. The networks are yearly, from 1994 to 2007. We created graphs for each edge set per year and cumulatively.

The data was put together for the Mimir project, which studies the flow of knowledge/ideas in an academic network. We used the JUNG[7] network package in our work.

2. NETWORK CHARACTERISTICS

2.1 Basic statistics

The first phase of our project, like any network analysis project, was to analyze the basic statistics and properties of the overall network. As mentioned in Section 1.2, our "network" is actually a collection of networks over the same underlying vertex set and across many years. In this section, we'll recap the basic network characteristics that give a basic summary of the macroscopic properties of our networks.

As expected the degree distribution in our networks follow power laws. Since our data is small, the log-log plots of the

Network	Year	$\|V\|$	$\|E\|$	Density	Avg C.C.	Size of GC	α
`PubCoAuthor`	2005	1779	1338	4.23×10^{-4}	0.116	558	1.79
`PubCoAuthor`	All	2832	9804	1.23×10^{-3}	0.1744	1816	1.64
`DissCommCoMember`	2005	1779	1564	4.94×10^{-4}	0.248	662	1.61
`DissCommCoMember`	All	2832	10881	1.36×10^{-3}	0.269	1655	1.48
`GrantAwarded`	2005	1779	377	1.19×10^{-4}	0.0564	20	1.52
`GrantAwarded`	All	2832	4137	5.158×10^{-4}	0.2017	920	1.63
`GrantRejected`	2005	1779	592	1.87×10^{-4}	0.0794	134	1.55
`GrantRejected`	All	2832	5097	6.356×10^{-4}	0.198	1037	1.59
`Combined`	2005	1779	5281	1.67×10^{-3}	0.350	1262	1.62
`Combined`	All	2832	28022	3.5×10^{-3}	0.353	2521	1.44

Table 1: Basic Network Statistics

power laws are sparse. However we are confident that the degree distributions actually do represent power laws, because when we analyze the degree distribution of the networks over time (the union of networks of the same edge type over multiple years) the power laws become obvious and clear, and simply fill in the holes in the single year plots. This suggests that the single year plots are simply missing data to draw a totally compelling power law distribution. The clustering coefficients of our networks are much higher than in the corresponding rewired networks (using the configuration model), as expected. The diameter of the networks vary by edge type, but fit in with what we saw in class.

However, we did notice one major difference from most of the networks that we've seen in class: the size of the giant component is consistently less than the giant component in the rewired network. This is consistent with the observation that academic networks are hierarchical, and hence more "cliquish" than other networks. In such cases, one would expect a smaller giant component.

One peculiarity of the Stanford academic network that influenced our results is the following fact: of all Stanford professors, a full 45% are in the Medicine faculty! In particular, the giant component in many of our networks is almost exclusively made up of the School of Medicine.

The basic statistics of our networks are shown in Table 2. $|V|$ is the number of vertices, $|E|$ is the number of edges, density $= |E|/|V|^2$, Avg C.C. is the average clustering coefficient, Size of GC is the size of the giant component, and α is the power law coefficient for the degree distribution.

2.2 Network similarity

Since our project is an analysis of many related networks, we calculated various measures to quantify how the networks are related to each other. We calculated vertex and edge overlap between networks of difference edge type and for networks with the same edge type over time. The vertex overlap from one year to the next is consistently between 90-95%, meaning the vertex overlap between two networks separated by n years is around 0.93^n. The edge overlap over time for a particular network follows an intuitive decay pattern, as shown in Figure FIG. The number of common edges seems to fall off exponentially with time until it settles at a relatively low stable state.

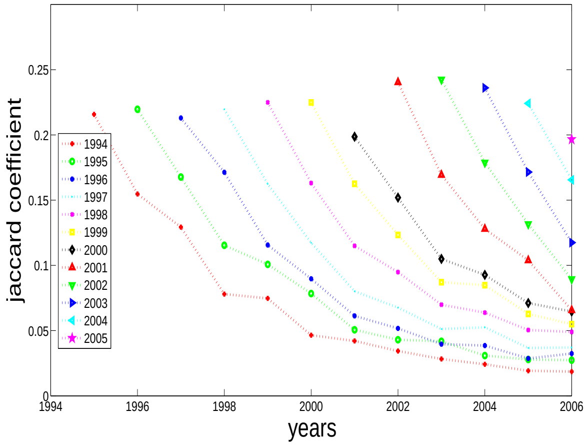

Figure 1: Edge overlap decays as a function of time

2.3 Nested core periphery

Following Leskovec et al.'s work on detecting community structure [5], we tried to determine whether our networks follow the nested core periphery structure that they discovered in many real-world networks. Our networks' small size prohibited us from exhibiting the downward-then-upward-sloping Network Community Profile plot characteristic of the nested core periphery structure (the upward-sloping part wouldn't show up, since it only happens beyond a certain threshold size). However, "whiskers" are crucial to the definition of the nested core periphery structure and they can be found even at the small network sizes we are dealing with, so we focus on them.

Leskovec et al. observe that whiskers have a surprisingly significant effect on the community structure of real networks, and the same is true for our networks. In particular, our networks all (except for `GrantAwarded`, which is too small to be meaningful) consistently have much larger 1-whiskers than their rewired counterparts. We also empirically observe that whiskers were very often the best clusters.

3. CLUSTERING

3.1 Identifying Community Structure

Our first major goal was to discover the community structure of the network. It is common knowledge that univer-

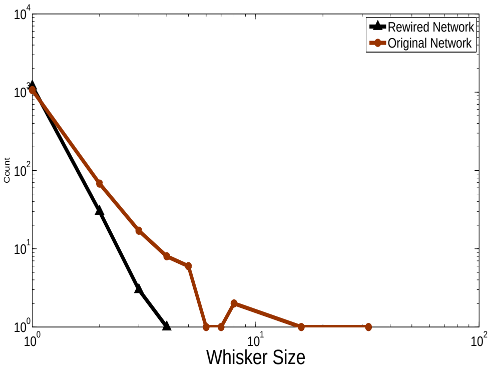

Figure 2: Our networks have larger 1-whiskers than their rewired counterparts

sities are hierarchically structured – indeed, they are the prototypical example of hierarchically-structured networks. Faculties form the top-level breakdown of nodes into clusters, then these faculties further break down into departments, and different groups often make up the departments. The main question we wish to address in this section is: do the interactions exhibited in the Stanford academic network follow the intuitive pattern described above? In other words, do the "natural" clusters in the network correspond to the different faculties and/or departments of the university? If so, this would be a rigorous quantitative confirmation that the structure of the university is well-described by the university's official divisions; if not, then depending on the extent to which the network clusters differ from the official divisions, there exist more coherent communities than those explicitly imposed by the university system.

As we learned in class, the Girvan-Newman [6] algorithm is naturally suited to hierarchical networks; therefore we used it to cluster our networks. Since Stanford has approximately 80 different departments, the departmental-level clustering is too fine-grained to give meaningful clusters. Therefore, we compared the faculty-level clustering (there are under 10 faculties) with the clustering found by the algorithm. The number of clusters found by the Girvan-Newman algorithm depends on the number of edges removed, therefore by varying this parameter we got clusterings at varying levels of granularity. We selected clusterings that contained around the same number clusters as there are faculties for a fair comparison.

We clustered only the giant component to avoid having clusters that are groups of many small connected components, since these have little value in reflecting the community structure of the network. The giant component typically comprised around 40% of the network.

We used the standard measure of conductance to evaluate the clusters found. Although we were aiming to find around 5-10 clusters with the Girvan-Newman algorithm, we still usually had a few candidate clusterings which satisfied this constraint to choose from. Aggregating the conductance scores into a single statistic reflecting the quality of the clustering is a non-trivial problem, so to select the final clustering for a given network we ended using a blend of heuristics and manual decisions. The heuristics we used were as follows. Let $\mathcal{C}$ be the set of clusters (which are themselves sets of nodes), and let $\alpha(C)$ denote the conductance of cluster C. Then let $\mathcal{M} = \max_{C \in \mathcal{C}} \alpha(C)$ be the largest conductance over all clusters in the clustering. We empirically identified a low threshold T, such that if the largest conductance was below this threshold ($\mathcal{M} < T$), the clustering was usually very good. We always ran the algorithm for as long as this max value stayed below T. The second heuristic was that after many trial runs, we noticed that there was usually a sharp threshold between good clusterings and quite bad clusterings; the decline was quick and steep as opposed to slow and gradual. Therefore, to save time we stopped the algorithm if the average conductance score of the clusters grew by 40% (determined empirically) in any one step (where one step corresponds to removing 5 edges in the Girvan-Newman algorithm). The table above compares the average conductance scores of the faculty clustering versus those of the graph clustering ("faculty" means the explicit faculty clustering and "graph" means the clustering found by the algorithm).

Table 2: Average Conductance Scores

Network	Faculty	Graph
`DissCommCoMember 1995`	1.91	0.13
`DissCommCoMember 2000`	1.29	0.05
`DissCommCoMember 2005`	1.13	0.127
`Combined 1995`	2.55	0.19
`Combined 2000`	2.39	0.22
`Combined 2005`	1.903	0.191

3.2 Quantifying Interdisciplinary Work

Having identified, analyzed, and visualized both the official academic clusterings and our computed graph clusterings, our next step was to use this information to quantify how much interdisciplinary work is happening at the University, and between which departments. Interdisciplinary work has been strongly emphasized in the past few years, especially at Stanford. Does the data support the rhetoric? In this section, we use our work in clustering described in the previous section to quantify the interdisciplinary work being conducted at Stanford. A main point of this section is to leverage the temporal nature of our data: we are lucky that our data provides us with not only a snapshot of the Stanford academic network, which is interesting in itself, but it also provides us with several snapshots over time so we can analyze the evolution of the network.

Concretely, for every pair of clusters, we calculate the number of links between them divided by the total number of possible links. This ratio is an approximation of how much the two clusters collaborate. To aggregate these numbers into an overall statistic for a network, we simply calculate the total number of links crossing any cut between clusters and divide by the total number of links that could have crossed a cut. This is a useful summary statistic to capture how much the academic groups collaborate with each other. We calculate these statistics using both the faculty clustering and the discovered graph clustering. We focus mainly on interdisciplinary coefficients derived from the faculty clus-

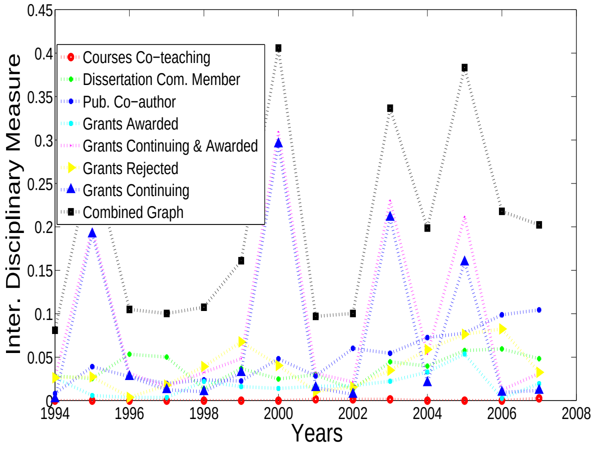

Figure 3: Interdisciplinary coefficient for all networks over all years

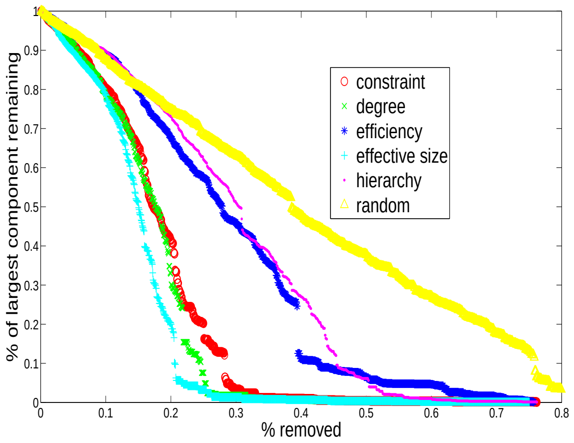

Figure 4: Graph Disintegration on combined network 2007

tering, since the clusters there all have intuitive meaning. However, the interdisciplinary coefficients derived from the discovered graph clusterings, although less intuitive, are still relevant: the amount of "interdisciplinary" work between these clusters is a lower bound on how much different groups collaborate. If even the most "cliquish" clusters we can find are still collaborating with each other a lot, then there exists no way to cut the graph so that few edges cross the cut, and so work in Stanford is "well-mixed".

In Figure 3 we plot our interdisciplinary coefficient for all networks over all years for the School of Medicine and School of Engineering. Interestingly, the `PubCoAuthor` coefficient quadruples from 2001 to 2007, meaning that collaboration between these two departments quadrupled in 6 years!

4. STRUCTURAL HOLES

4.1 Structural Holes

As we saw in class, structural holes are *"the 'empty space' in the network between two sets of nodes that do not otherwise interact closely."*. Ron Burt[1, 4] investigated the social capital that accrue to people who span structural holes, and showed that they tend to receive increased benefits through higher salaries, promotions, and production of good ideas. The main question we seek to answer in this section is: do the more successful Stanford faculty occupy these structural holes within Stanford?

4.2 Faculty Success

Burt's work correlated his structural hole measures with faster promotions, salaries, etc., which are reasonable measures of success in the company context. But what kind of success measure is applicable to the academic setting? H-index, tenure, grant success, salary, promotions, number of publications?

The H-index would be ideal, but unfortunately individual publication and citation data is not available. Neither are salary data and promotions. Tenure is also reasonable, but we don't have data on why people leave the network and assuming everybody who left the network did so because they didn't get tenure is susceptible to false negatives, as faculty might leave to go somewhere else for any number of reasons. We elected to use grant success rate as a proxy for faculty success because we have ample data (over 22K unique grant proposals), which in any case encompasses a lot more of the faculty than tenure does, and varies interestingly over time, whereas tenure is a once-in-a-lifetime decision.

4.3 Measures

Burt describes five measures: constraint, aggregate constraint, effective size, efficiency and hierarchy. Constraint measures the extent to which your access to information is constrained by your neighbors; lower constraint means more structural holes around you. Aggregate constraint [2] takes into account size (Herfindal index), density and hierarchy; again, lower values indicate more nearby structural holes. Effective size is the amount of non-redundant contacts you have; large effective size means more opportunities for structural holes. Efficiency is the ratio of effective size to degree. Hierarchy measures how much constraint is concentrated in a single relationship.

4.4 Approach

Our approach to answering the main question of the section was two-fold. Firstly, we disintegrated the giant component by removing faculty in descending "structural hole-ness" using the measures described above. Intuitively, if this breaks the graph apart faster than other methods then significant structural holes exist in the network.

Secondly, we plotted each structural hole measure versus grant success rate for each network and produce a line of best fit to see if there was any correlation.

4.5 Results

4.5.1 Graph Disintegration

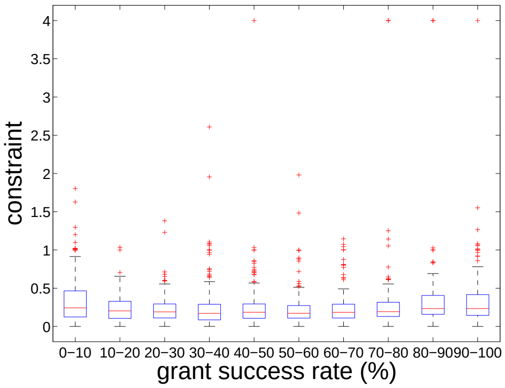

Figure 5: Constraint distributions for each grant success bins (combined network, all years)

Figure 4 shows the measures[1], along with random node removal and removing by degree. This graph was representative of the results as a whole.

What is noticeable is that effective size produces the fastest network disintegration, followed by disintegration by degree, then the constraint measure. This indicates that the structural holes do exist in the graph, since removing those who bridge them first breaks the graph apart the fastest. Notably, removing by effective size even beats removing nodes by degree.

Our plots show that there is basically no correlation between grant success rate the structural hole measures. To see how the constraint measures were distributed, we binned nodes according to their grant success rate in 10% intervals, and produced a box plot. As shown in Figure 5 the distribution of each bin is similar, thus showing that there is no correlation between grant success rate and constraint. Figure 5 is representative of most graphs.

There are two main explanations for these results: (1) grant success rate is not a good proxy for faculty quality, and (2) the structural hole argument for social capital does not apply to academic networks.

Burt [3] has argued that the main mechanism by which people gain social capital from bridging social holes is *brokerage*. These people control the flow of information between two far-apart parties and thus can use this as leverage for potential benefit. This explanation is more plausible in a company context than in a university context in our opinion. However, Burt has also argued and shown that people who bridge structural holes are “more at risk for being having good ideas”, which seems applicable to the university setting.

The first explanation, that grant success rate is not a good proxy for faculty quality, is also possible. We assume that faculty mostly only apply for grants they wish to receive, but there is anecdotal evidence that this not always the case (e.g. applying for multiple grants and wishing to receive some of them, but not all). Grant success rate is also affected by many things, and can vary widely depending on faculty, funding agency, etc. However it still seems that grant success rate should be some indicator of the quality of a researcher.

It is thus unclear which of these two explanations is correct and more experiments will need to be run to conclude that bridging structural holes is not beneficial in academic networks. In particular, other measures of faculty quality should be used to obtain a compelling answer.

5. GRANT PREDICTION

Since the Mimir publication data is incomplete, we were unable to pursue our original idea of predicting edges in the grant network from edges in the paper network. In its place, we tried to predict whether grants would be accepted or rejected. We have 22,000 grant proposals with a roughly an even split of approvals and rejections.

5.1 Baseline Features – Grant Signature

Our baseline feature set is a collection of grant features that we extracted entirely from the proposal. It is comprised of: a Bernoulli bag of word model on the project title, a Bernoulli bag of word model on the sponsoring organization, the proposed amount, the proposer’s faculty and department, and the year.

5.2 Network Faculty Features

The following was extracted for each faculty member: five structural hole signature measures[1]: aggregate constraint, constraint, effective size, efficiency, hierarchy; barycenter value[2];random walk betweenness value[3]; betweeness centrality[4]; closeness centrality[5]; eigenvector centrality[6]; clustering coefficient; degree; radius one and two features: number of tenured and untenured links, number of awarded and continuing grants, number of rejected grants, number of grant and publication edges that overlapped, number of publications, neighbour edge incidence; and cumulative weighted, unweigted and average weight edge incidence.

Note: Our edge weights represent the number of interactions that the faculty had together (e.g. an edge weight of 3 in `PubCoAuthor` 2007 means that the two endpoints wrote three papers together in 2007). Weighted and unweighted versions were produced where possible.

5.3 Results

We ran two types of test: (1) training and testing on random years (not using any cumulative features), and (2) testing on a particular year and training on all previous years.

We expected tests of the second type to get better with time (since more training data is being used).

[1] Aggregate constraint was extremely similar to constraint and was thus omitted.

[2] sum of distances to each vertex

[3] measures the expected number of times a node is traversed by a random walk averaged over all pairs of nodes

[4] how many shortest paths go through a vertex

[5] based on average distance to each vertex

[6] the fraction of time that a random walk will spend at that vertex over an infinite time horizon

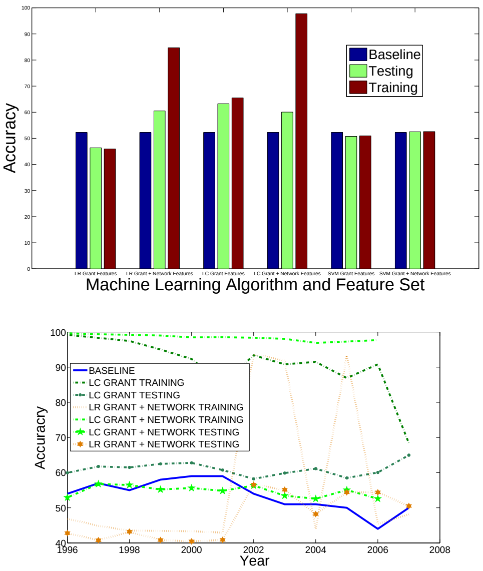

Figure 6: Top graph shows results of testing on randomly intermingled years. The bottom shows the testing on all years, training on the previous years.

5.3.1 Linear Regression

As seen in the top of figure 6 for the randomly chosen years test, the addition of the network features increases the classifier accuracy. However the training accuracy with the network features indicates that it was beginning to overfit. Playing with the convergence parameters did not yield better results.

Both feature sets do poorly on the year to year test, getting below baseline until enough training data is available.

5.3.2 Linear classifier

The linear classifier using the Quasi-Newton minimizer for gradient calculation was chosen next. As can be seen in the middle two sets of columns at the top of figure 6 the linear classifier achieves the best results. Interestingly the baseline feature set performs the best, with the network features actually hurting testing accuracy. Also the training accuracy including the network features leads to big overfitting.

The year to year results show that the baseline features perform quite well from the start, staying above baseline at the end. The curve that exhibits a great example of more training data decreasing training accuracy and increasing testing accuracy. As for the including the network features, one can observe that more training data slowly decreases the training accuracy, but it is still overfitting, and thus the testing accuracy is not that great. It does come above baseline, but at no point does it increase the testing accuracy over just using the baseline features.

5.3.3 Top Features

The top features for all four models were bag of word features, except for logistic regression on grants which chose the proposed total as the most influential feature. Logistic regression was not able to discern betters weights for the features than the linear classifier. However, with the addition of network features, the logistic classifier is then able to find better weights for these features.

Overall the addition of the network features changes the top features to be more title oriented. No network features appear in the top 20 weighted features in either logistic regression or the linear classifier.

5.3.4 Summary

The bag of word model on the grant title and sponsoring organization does quite well overall. But by just looking at the logistic regression results one could hypothesize that the network features can help, but then looking at the linear classifier the network features actually hurt performance. This is probably due the overfitting that is happening, not allowing the model to generalize itself as well as just using the baseline features. More data or removing features would be the next approach to see whether the overfitting can be reduced.

6. CONCLUSIONS AND FUTURE WORK

In this project, we analyzed Stanford's academic network and found several interesting results and further avenues we'd like to explore. We found that there are more coherent groups in the network than the explicit faculty clustering. We quantified the amount of interdisciplinary work as a function of time, and found that some departments (for example the Medicine and Engineering) are much more connected than they used to be. We then gave evidence that structural holes exist by decomposing the network with structural hole measures, but didn't find any correlation between these measures and grant success rate. This suggests that the theory of structural holes might not apply in the academic work setting, although this is something we'd like to further explore. Finally, we tried to use network features to predict grant success or failure but were unable to leverage them to gain any predictive power over baseline. This is probably due to the fact that individual grants rely more on non-network factors, and in the future we'd like to try predicting the grant success rate of faculty members rather than predicting individual grants.

7. REFERENCES

[1] R. Burt. *Structural holes: The social structure of competition*. Belknap Pr, 1995.

[2] R. Burt. The gender of social capital. *Rationality and society*, 10:5–46, 1998.

[3] R. Burt. Structural Holes and Good Ideas 1. *American journal of sociology*, 110(2):349–399, 2004.

[4] R. Burt. Structural Holes and Good Ideas 1. *American journal of sociology*, 110(2):349–399, 2004.

[5] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. 2008.

[6] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):26113, 2004.

[7] J. OŠMadadhain, D. Fisher, S. White, and Y. Boey. The jung (java universal network/graph) framework. *University of California, Irvine, California*, 2003.