Application of Social Network Analysis to Collaborative Team Formation

Michelle Cheatham
Information Directorate
AFRL
WPAFB, OH 45433
michelle.cheatham@wpafb.af.mil

Kevin Cleereman
Information Directorate
AFRL
WPAFB, OH 45433
kevin.cleereman@wpafb.af.mil

ABSTRACT

Team formation is a challenging problem in many large organizations in which it is entirely possible for two individuals to work on similar projects without realizing it. By applying social network analysis to mappings of co-authors and to mappings of related research paper keywords, we are able to help generate teams of diverse individuals with similar interests and aptitudes.

KEYWORDS: Social network, concept map, collaboration, team formation.

1. INTRODUCTION

The first step in collaboration is knowing who to collaborate with. But today's organizations are so fast-paced and geographically distributed that there may be two employees working on the exact same project without realizing it. Worse, the company may be employing outside consultants unnecessarily. An effective way to insure that organizations are taking advantage of their human capital is needed. Typical approaches to this problem include reorganizing to put employees working on similar projects in close geographic proximity and the use of "yellow page" systems. Reorganizations can be a disruptive shock to employees and are impractical since they would need to be done frequently as employees' skill sets change over time. Yellow page systems generally ask an employee to fill out a questionnaire about his skill set which can then be searched by others. There are various problems with this approach, including keeping the directory current and relying on employees not to be too modest or boastful when listing their skills [6]. In this paper we consider using a modified version of social network analysis to construct a graph of employee interactions along with the subjects these employees are working on. This graph can then be used to suggest new collaborative teams.

Social Network Analysis (SNA) is a method of studying interactions among individuals or groups. SNA is best applied in situations where the data is inherently relational, meaning it is a property of the *interaction* of agents as opposed to individual agents [7]. Examples include computer networks, organizational relationships, and family trees. Much work has been done based on email or IM traffic [6], [8], [3], but this type of research raises privacy concerns. Newman points out the benefit of basing SNA on affiliation networks (a network in which people are related by membership in a common group or club): the data is readily available and does not rely on questionnaires or interviews [5] as is the case with yellow page systems. Plus, the information is often public, eliminating privacy concerns. For this project we have used coauthorship on published works as the basis for constructing the social network graph. Other types of data, such as a list of project titles and their teams or meeting minutes and the participants are also viable datasets. This paper first shows two different views of the paper-author data and considers the types of questions that can be answered with those views. We then take the analysis one step further and use the keywords of the papers to develop a graph of papers and the concepts they cover. Combining this information with the standard SNA allows us to suggest new collaborative teams.

2. RELATED WORK

Performing social network analysis based on coauthorship is not a new area of interest. In particular, Newman has specifically considered scientific collaboration networks based on publication information in several scientific databases [4]. His work found that these networks display the "small world" characteristic, are highly clustered, and obey a power-law distribution with an exponential cutoff. Newman did not go into detail on what these characteristics meant in the scientific collaboration context nor did he consider the implications for potential future collaborations, however.

Schwartz attempted to discover shared interests using graph analysis in the early nineties [6]. His method involved analyzing email traffic from 15 different organizations and running various algorithms on the resulting graph to determine which individuals shared his own interests. While the algorithms developed are very interesting, this work did not explicitly consider *which* subjects the individuals shared an interest in, only that they existed.

ReferralWeb, done by Kautz's group at AT&T Laboratories, is an interesting application in this area. It mines publicly available documents on the internet and constructs a graph of names that appear in close proximity [2]. The resulting system can then answer questions like "What is my relationship with Person A?" and "What people in my neighborhood know about topic x?" The method that ReferralWeb uses to answer this later type of question is not clear, however, and team formation is not mentioned. Khan et al. explicitly considered both the authors and subjects of papers, but the focus was on examining existing collaborations and predicting new ones [3]. The group did not explore the utility of their system for collaborative team formation.

3. TRADITIONAL SNA

The data we used consists of all the publications produced by a branch within the Air Force Research Laboratory between 2003 and 2005. To preserve employee privacy, the names of the authors have been replaced by numbers. The dataset consisted of 71 papers written by 80 different authors. The average number of papers written by an author was 2.4, and the average number of coauthors on a single paper was 2.6. The average author collaborated with 3.7 other people. Newman's corresponding numbers for the MEDLINE, Los Alamos, and NCSTRL databases containing scientific publications were 4.7, 2.8, and 10.5 respectively. The lower numbers for our dataset are likely due to the limited timespan considered and the omission of papers written by academics and contractors besides those done in conjunction with AFRL.

Figures 1 and 2 show two traditional views of the social network graph for this data. In Figure 1, the nodes of the graph are authors, and the lines between them indicate that the connected authors have collaborated on a paper. The thicker the line, the more papers they have worked together on. This view makes it easy to see the different workgroups that exist, as well as which individuals primarily work alone and which serve as "bridges" between different workgroups. In the dataset examined here, there are seven major clusters and several smaller groups. It is readily observed that authors 63 and 13 have collaborated extensively (with one another and with a group of other colleagues) while 29 has only worked individually on publications. 13 also acts as the only bridge between two large groups of researchers. This graph also shows that 52 acts as a hub for a group of eight researchers. This gives a coarse indication of what individuals would work well together in a group setting. For example, creating a group that consisted entirely of employees who had only worked alone on papers in the past, without any of the potentially more outgoing "hub" employees, may not be productive. In addition, if employees have already worked together numerous times in the past they will likely be able to function smoothly in a group from the start, whereas employees from different clusters will bring a more diverse set of viewpoints to bear on a problem at the expense of easy communication [1].

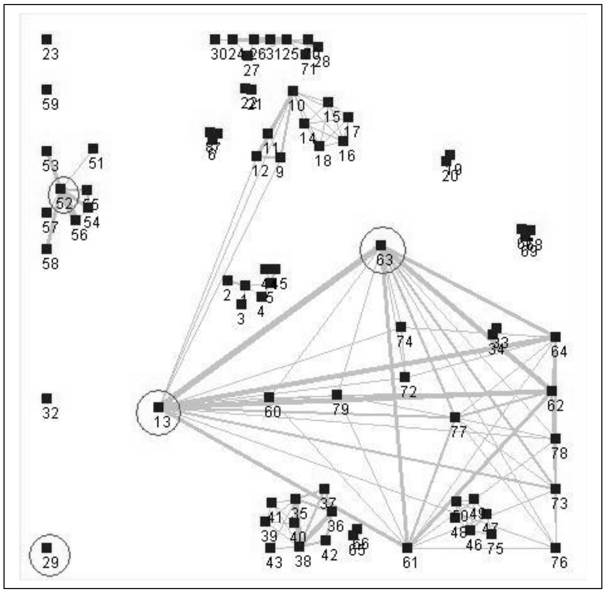

Figure 1: Traditional View

Newman suggests that affiliation networks are fundamentally bipartite graphs with one type of node representing individuals and the other representing the groups they belong to and edges can only connect vertices of unlike type [5]. In Figure 2, the same dataset is shown as a bipartite graph where one class is the authors (represented as squares) and the other is the papers (represented as circles). The size of the node is based on its degree, so that more prolific authors and papers with more coauthors are larger. This view still allows us to see the different workgroups that exist, but now it is easier to see what those groups are working on. This graph also makes it easy to determine who has been publishing the most and which papers have been the focus of attention. In the graph of our data, it is easy to see that 52 has written the most papers while "Real Time Streaming Data Grid Applications" had the most contributors (shown in black). We can also use this graph (or the one shown in Figure 1) to answer some common queries in social network analysis, such as "How are 3 and 44 connected?" (Figure 3) and "What neighborhood of people can be reached in a given number of levels, starting from 51?" (Figure 4). This type of analysis has implications for team formation. Kautz reasons in [2] that an individual is more likely to trust a person recommended to them if they can see the chain of known acquaintances between themselves and the recom-

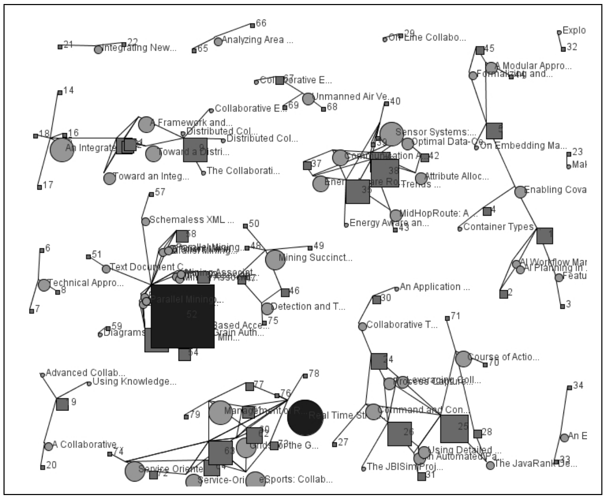

Figure 2: Bipartite View

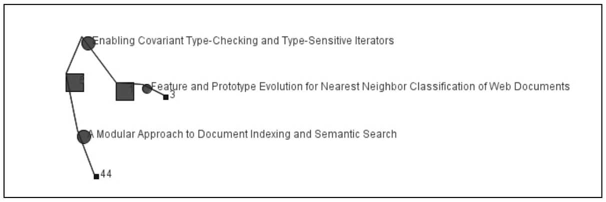
Enabling Covariant Type-Checking and Type-Sensitive Iterators
Feature and Prototype Evolution for Nearest Neighbor Classification of Web Documents
3
A Modular Approach to Document Indexing and Semantic Search
44

Figure 3: Shortest path between two authors

mended person. In addition, if there are many people in the social network who cannot reach one another (i.e. many people whose neighborhood is a small subset of the overall graph), then knowledge cannot effectively spread throughout the branch or organization. When forming new teams, individuals can be chosen such that the average neighborhood size increases by creating new links that will allow information to flow more easily to all individuals. The easiest way to do this would be to put the hub individuals from two different groups on the new team.

As with Newman's work on scientific networks, we see in this case that the network is highly clustered (two authors are much more likely to have collaborated if they have both worked with a third person). There are seven large clusters and several smaller groups in our dataset. Unlike his results, however, this graph does not display the "small world" phenomena, in which the path between any two randomly chosen nodes is a small value, typically six or less [4]. If the hubs of the various clusters (i.e. 1, 9, 13, 25, 38, 52) are asked to collaborate on a new project, then the graph becomes much more connected. Of course, ideally this collaboration would be centered on a common subject of interest. This is the goal of our proposed concept view.

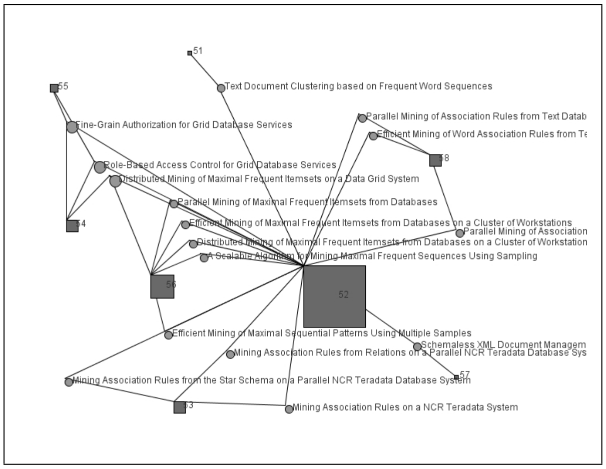
51
Text Document Clustering based on Frequent Word Sequences
55
Fine-Grain Authorization for Grid Database Services
Parallel Mining of Association Rules from Text Datab
Efficient Mining of Word Association Rules from Te
58
Role-Based Access Control for Grid Database Services
Distributed Mining of Maximal Frequent Itemsets on a Data Grid System
Parallel Mining of Maximal Frequent Itemsets from Databases
4
Efficient Mining of Maximal Frequent Itemsets from Databases on a Cluster of Workstations
Parallel Mining of Association
Distributed Mining of Maximal Frequent Itemsets from Databases on a Cluster of Workstation
A Scalable Algorithm for Mining Maximal Frequent Sequences Using Sampling
56
52
Efficient Mining of Maximal Sequential Patterns Using Multiple Samples
Schemaless XML Document Managem
Mining Association Rules from Relations on a Parallel NCR Teradata Database Sys
57
Mining Association Rules from the Star Schema on a Parallel NCR Teradata Database System
3
Mining Association Rules on a NCR Teradata System

Figure 4: Neighborhood of the most well-connected author

4. CONCEPT VIEW

As discussed in the previous section, traditional social network analysis can provide valuable information about the informal structure of the organization that is useful when creating new teams. However, the problem remains that two individuals may be working on very similar topics without realizing it. To address this issue, we suggest combining social network analysis with a concept map.

To create the concept map (shown in Figure 5), we came up with a list of keywords for each paper based on its title. Ideally, we would use the author-specified keywords included in the paper, but only bibliographic information was available for this study. In the future we would like to incorporate natural language processing techniques to extract the keywords automatically. Similar to Figure 2, the concept map allows us to quickly see (based on the size of the nodes) what subjects the branch is focusing on. In this particular branch, the primary subject being researched is collaboration. It is also possible to see the path between two different concepts or the neighborhood of an individual concept. The neighborhood surrounding the concept *collaboration* is significantly larger than the neighborhood surrounding the most well-connected individual in the social network. In fact, the concept map is generally more interconnected than the corresponding graph of the social network (the neighborhood of the most well-connected concept is almost the size of the entire network). This indicates that researchers are working on a coherent and interrelated set of subjects, but they are working on these subjects in isolated groups.

Combining the graphs shown in Figures 2 and 5 into a single social-concept network graph makes things a little cluttered (see Figure 6), but we are now able to simultaneously see who has been working together and what they were working on. This has several uses. In particular, if an individual has published extensively on a given subject through collabora-

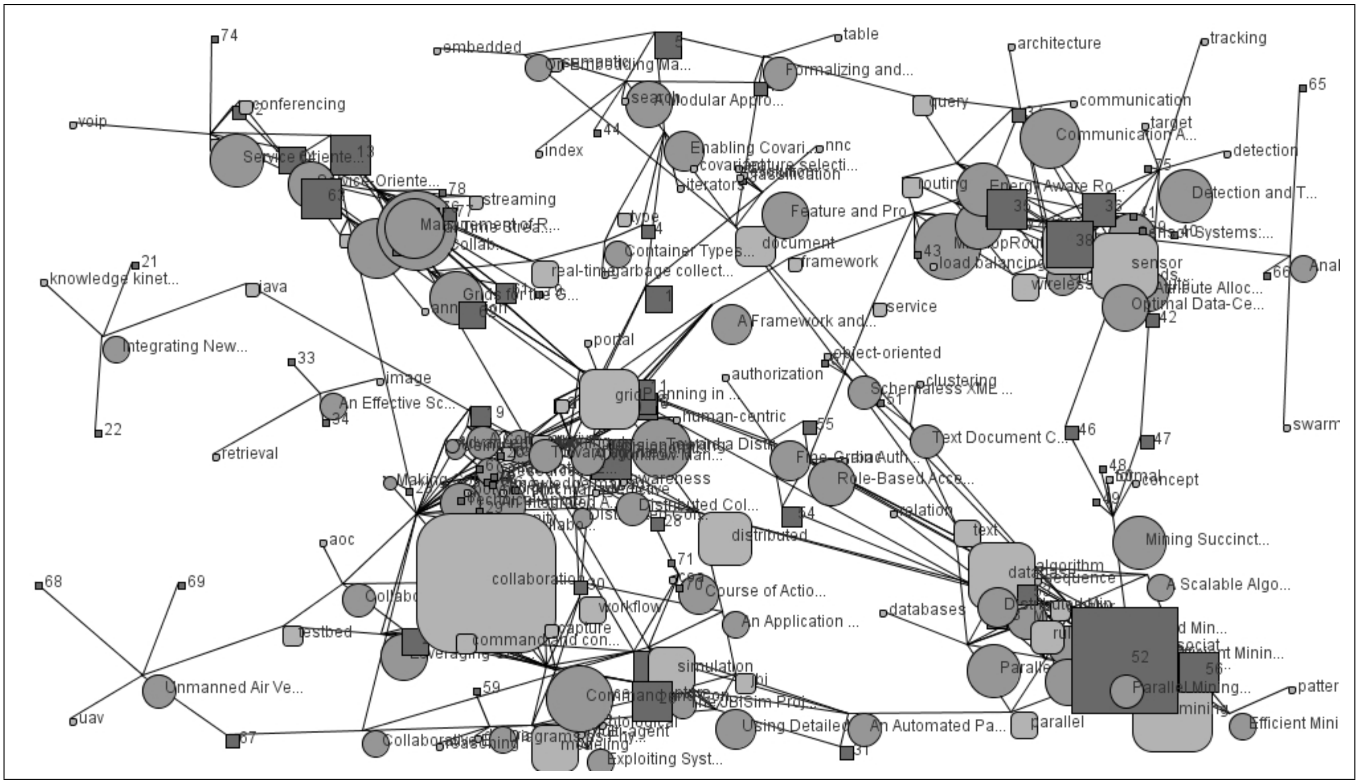

Figure 6: Enhanced Social Network

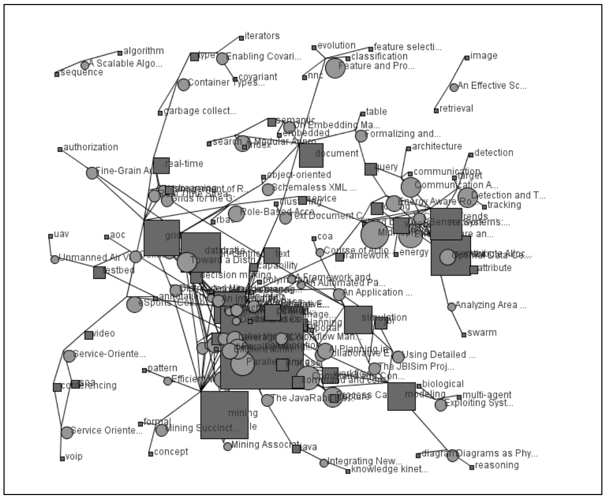

Figure 5: Concept Map

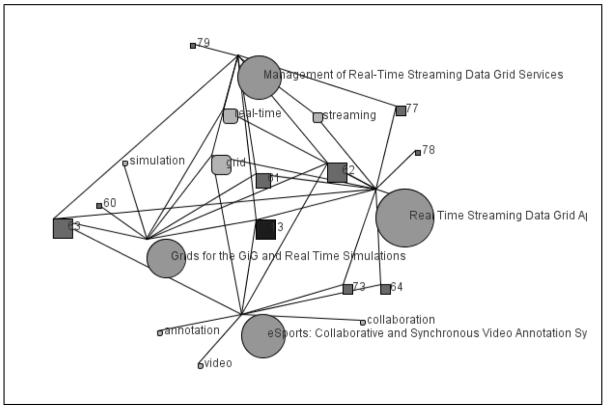
79
Management of Real-Time Streaming Data Grid Services
real-time
streaming
77
78
simulation
grid
61
62
60
13
Real Time Streaming Data Grid A
3
Grids for the GiG and Real Time Simulations
73
64
collaboration
annotation
eSports: Collaborative and Synchronous Video Annotation Sy
video

Figure 7: Identification of Subject Matter Expert

tions with many other scientists, that person may be deserving of a higher degree of trust than someone for whom this is not the case. For example, 13 has written papers on grid computing with ten other people and may therefore be considered a potential authority on the subject (Figure 7). By querying over the same graph it is also possible to see who is working on similar subjects and yet not currently collaborating. For instance, 51 has written a paper titled *Text Document Clustering Based on Frequent Word Sequences*, which has *document* as one of its keywords. 44, 45, and 5 have produced a related paper titled *A Modular Approach to Document Indexing and Search*. Figure 8 shows that our modified social network graph can be used to suggest a collaboration among these researchers. This is significant because, as the traditional social network in Figure 2 shows, 44, 45, and 5 are not even part of 51's neighborhood (there is no path of acquaintances between them).

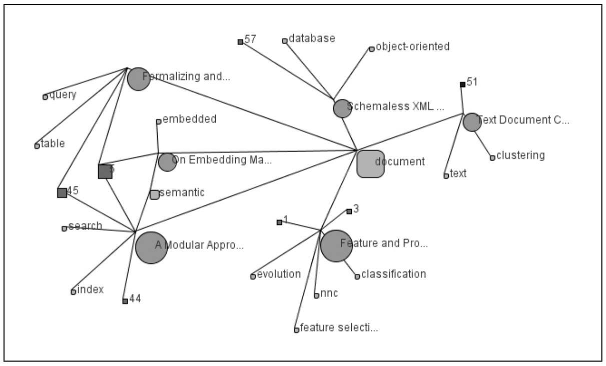

Figure 8: Recommendation for Future Collaboration

5. FUTURE WORK

The work we have done thus far allows us to build teams in an ad hoc manner. In the future, we would like to expand on this work by creating a tool to enable more systematic collaborative team formation. We envision this tool allowing a user to specify what concepts the team needs to be familiar with and to what degree. In addition, users should be able to specify how important it is that team members are trusted experts, have worked together in the past, etc. This type of information can be gathered from the social-concept network and fed to a genetic algorithm that will use a fitness function based on the user's parameters to suggest an appropriate team.

6. CONCLUSION

Often the hardest part of collaborating is knowing whom to collaborate with. We have proposed using techniques from social network analysis to help identify colleagues that would be helpful in a newly formed team. Traditional SNA allows us to see which employees are already collaborating, how often they are working together, and how many others in the organization a given employee can reach through intermediaries. This facilitates creating new teams consisting of people who have already been working together so that the team members will already be comfortable with one another and can begin work quickly, or of people from completely different workgroups who can bring a wide variety of perspectives to the table. In addition, the employees that act as hubs of the current workgroups can be introduced to one another to more fully connect the various clusters and ease the flow of knowledge throughout the organization. We also suggested using a concept map in conjunction with SNA to be able to answer questions such as "Who do I know that has experience with robotics?" and to insure that employees who are working on similar subjects are aware of one another. Some coarse trust-related decisions can also be made by observing who a person has collaborated with and what subjects they worked on.

REFERENCES

[1] T. Casciaro and M. Lobo. Competent jerks, lovable fools, and the formation of social networks. *Harvard Business Review*, page 92, June 2005.

[2] H. Kautz, B. Selman, and M. Shah. Referralweb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[3] F. Khan, T. Fisher, L. Shuler, T. Wu, and W. Pottenger. Mining chat-room conversations for social and semantic interactions, 2002.

[4] M.E.J. Newman. The structure of scientific collaboration networks. In *Proceedings of the National Academy of Science*, volume 98, pages 404–409, January 2001.

[5] M.E.J. Newman, D. Watts, and S. Strogatz. Random graph models of social networks. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 99, pages 2566–2572, February 2002.

[6] M. Schwartz and D. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78–89, Aug 1993.

[7] John Scott. *Social Network Analysis: A Handbook*. SAGE Publications, 1991.

[8] F. Wu, B. Huberman, L. Adamic, and J. Tyler. Information flow in social groups. *Physica A*, 337:327–335, 2004.