iSchools and Social Identity – A Social Network Analysis

Arvind Karunakaran
The Pennsylvania State University, College of Information Sciences & Technology, University Park, Pennsylvania-16802, USA
1-814-206-4189
axk969@psu.edu

Hyun-Woo Kim
The Pennsylvania State University, College of Information Sciences & Technology, University Park, Pennsylvania-16802, USA
1-814-360-0992
Hxk263@psu.edu

Madian Khabsa
The Pennsylvania State University, Department of Computer Science, University Park, Pennsylvania-16802, USA
1-814-321-7666
mxk479@psu.edu

ABSTRACT

We analyze the publication co-authorship network of an iSchool faculty community using 'Social Identity Theory' as the theoretical lens. Initially, we discuss the need for a theoretical framework to analyze and interpret social network data. Then, we find out the patterns in the levels of interaction happening within the faculty community at an inter-group level. We grouped faculty members into different clusters according to several parameters such as their educational backgrounds, affiliations with research centers/labs, and h-indices. We based our analysis on this classification and we try to understand the relationship among social identity, group affiliation and academic collaborations. We conclude with the remarks that one could avoid idiosyncratic ways of interpreting social network data by using a proven theoretical lens like 'Social Identity Theory',

Categories and Subject Descriptors

J.3 [Social and Behavioral Sciences]: Psychology, Sociology

General Terms

Theory, Measurement

Keywords

Social Identity, iSchool, Social Network Analysis, Academic Collaboration

1. INTRODUCTION

'Social identity theory' [7] was originally developed to understand inter-group behavior and discrimination. According to this theory, an individual derives part of his self-concept from the groups that the individual is affiliated with. Thus, the social identity perspective is in part a critique of individualistic notions of self, and argues that individual's social identity is a self-concept derived from the membership with social groups [4], [5]. The behavior of the individual is triggered by the social context and the individual's perceived affinity with a specific group that has a significant influence on the behavior. Also, other sub-theories within the social identity perspective such as *the self-categorization theory* [8] and *the social categorization theory* [9] argue that individuals tend to continuously bracket themselves and others into categories and in this process, they identify themselves with certain groups and not identify with certain other groups. This identification is critical, and helps them to develop a coherent self-image and self-esteem [7]. However, this does create a bias, since such categorizations would lead to the formulation of in-groups and out-groups, the former they increasingly identify with, while the latter, they do not identify with.[7],[8].

On an interdisciplinary research setting like iSchools [2], the concept of *social identity* plays a very important role. One of the distinguishable characteristics of an iSchool would be its faculty body consisting of those with diverse educational and professional backgrounds; their disciplines include computer science, information science, management science, organizational behavior, physics, cognitive science, sociology, cultural anthropology, social psychology, etc. Thereby, the individual faculty member's exclusive identification with a specific group—depending on their perceived *self-categorization* and *social categorization*—might influence their research collaborations, which could in-turn impact on the climate of inter-disciplinary research within iSchools.

However, past studies of publication collaboration in terms of a social network analysis have a tendency to overlook the influence of *social identity*. A few studies did use the term *social identity* [10], but they did not use the theoretical framework put forth by Tajfel and Turner [7].

In this poster, we adopt *social identity theory* as our lens to analyze a publication collaboration network. We take an instance of a particular iSchool and analyze the influence of *social identity* and *self categorization* over the publication collaboration among its faculty members by using social network analysis techniques.

2. RESEARCH METHOD

2.1 Social Network Analysis

The term "*Social Network"* initially started out as a *metaphor* for complex relationships or the sets of complex relationships among different entities at different scales. For example, the nature of a relationship between an individual and a community can be illustrated and explained as a social network. Similarly, the nature of a relationship among different organizations forming a cartel to

Conference'04, Month 1–2, 2004, City, State, Country.

indulge in a trade agreement, and the complex web of relationships among countries involved in a nuclear deal could all be as well illustrated as social networks.

With the advent of some visualization tools for social networks such as *Netdraw* and *Pajek*, the field of social network analysis is picking up its momentum. Also, a lot of researchers in many disciplines are finding it very useful so that they make use of the concepts of social network analysis and adopt it into their research nowadays.

By leveraging the visualization capabilities offered by these tools, we try to understand the effects of *social identity* on *clique behavior* [6] at an "inter-group" level.

2.2 Data Processing

We followed these techniques to extract and process data from different sources.

2.2.1 Data from CiteSeer

CiteSeer is a public search engine and digital library for scientific and academic papers. We chose CiteSeer because it contains publications information in a highly *structured* format. Besides, the system accepts logical operators in a query string. To get information from CiteSeer systematically, we developed a '*bot*' that queries CiteSeer a publication list of each faculty member and fetches the results. A returned result contains title, publication venue, year of publication, and co-author names. These are stored in our database for post-processing. Although CiteSeer does not cover all up-to-date publications, it has more than 1,100,000 articles in a well formatted structure. Moreover, it becomes helpful for us to deal with the data that CiteSeer has made much effort to disambiguate the author names.

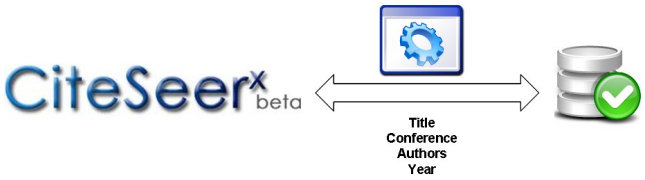

Figure 1. Data processing from CiteSeer

CiteSeer was used as an *augmenter* for our dataset, since some faculty members do not have the list of their publications available on their web pages or they have an incomplete list. CiteSeer helped in collecting more publication data for those faculty members. For example, one of the faculty members had only 4 papers listed on his/her webpage, whereas we got 24 papers from CiteSeer. Sometimes, for the worse, it is difficult to parse the list of published papers when the list is not constantly well structured while it is relatively easy to parse results from CiteSeer as it contains all their publications in the same format. For this reason, we decided to get publication information from CiteSeer as the only data source for those faculty members. The number of publications collected from CiteSeer is given in Table 1.

2.2.2 Data from Faculty Members Homepages

Collecting publications data from CiteSeer was not enough, because not every faculty member's publication was indexed in CiteSeer's database. To solve this problem, we *crawled* the websites of the faculty members and tried to extract publications information from it. Most of the faculty members follow some pattern in listing their publications. We tracked these patterns and wrote regular expressions that capture the publication data. For example, a popular pattern is listing authors first followed by a single dot, then year of publication, after that the title and the conference name.

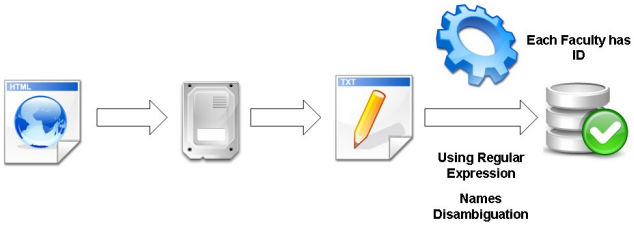

Figure 2. Data processing from Faculty Homepage

We started by downloading the publications page of each faculty member. Then, we used a tool to convert the HTML pages into text files. This would only keep the text and eliminate HTML tags. Then, we developed a program that takes these text files as an input and extracts publications information in accordance with the regular expressions given before. Finally, this program stores the data in the same database.

One major challenge was *name disambiguation*. We developed a simple heuristics to match the name. i.e., if the last name of a faculty member appeared in another faculty member's *publication* list, then it is more likely that this last name *belongs* to a faculty member in the same iSchool. Consequently, we considered an author as the one whose last name appears on some colleague's publication list. To make it easy, when dealing with authors, we assigned an ID for each faculty member. Since each faculty member's name is listed with some variation in spelling (sometimes full name and sometimes initials), we defined regular expressions for recognizing their different spellings. These regular expressions would map different spellings to the same ID.

In total, our final dataset had 1357 publications of 41 faculty members.

Table 1. Publications Summary

From CiteSeer	***From Faculty Members' website***	***Total no. of Publications***
742	615	1357

2.3 The Need for a Theoretical Lens

Past studies which used Social Network Analysis methods to analyze academic publication collaboration tend to overlook the influence of social identity on clique behavior [7][8][9]. A few studies [10] did use the term *social identity*, but they did not use it in the context of the theoretical framework put forth by Tajfel and Turner [7].

2.3.1 Analysis of Student Project Teams

To illustrate the drawbacks of those approaches which did not use any theoretical lens to interpret their social network data, we did an exploratory study of team project reports on social network analysis. Eighteen team project reports, written by iSchool graduate students, were obtained from a course instructor with the team names blinded. The goal of the project was to conduct a

social network analysis on the collaboration in publishing research work among their faculty members.

The rationale for selecting these teams is as follows:

- All teams had almost the same dataset of publication collaborations.
- All teams deployed computational methods with the help of social network analysis software (like *Pajek or UCINet*) and visualization software (e.g. *Netdraw*).
- Except one team, all other student teams did not use any 'theoretical lens'

Each team was asked to start out with various hypotheses from different research questions. However, there were few common research questions across multiple teams. We chose one such research question concerning 'inter-disciplinarity', and analyzed how a team's 'grouping of faculty members into arbitrary clusters' affected their findings, and how their idiosyncratic interpretation of those findings without a theoretical lens led to disparate conclusions.

2.4 Categorization

We begin by categorizing faculty members according to their:

1. Educational backgrounds
2. *h-indexes*
3. *Affiliations* with Research Centers and Labs

Based on the categorization above, we did a sub-network and a cross-network analysis, and tried to find out if there is any clique behavior [6], and the influence of *social identity* on such clique behavior. For example, do faculty members sharing the same educational background tend to collaborate more with each other than those with different backgrounds? What is the influence of *h-index* [3] on collaboration patterns? Do faculty members with higher *h-indexes* tend not to collaborate much with those with low *h-indexes*?

We use different centrality measures such as degree centrality, betweeness centrality and closeness centrality to look into the influence of Social Identity on publication collaborations. We would share the results of our findings and analysis in our poster.

3. CONCLUSION

From the analysis above, a strong correlation among *group affiliations, clique behavior* and *closed collaboration networks* have been observed. When we tried to interpret the results by using the 'Social Identity Theory' lens, we found that faculty members' 'self' and 'social' categorizations - of themselves and of the groups around them - did influence collaboration patterns.

Membership within a specific group is a factor of the "*zone of intersection"* between an individual's "personal" and "social" identities. However, an individual would feel part of the large community if there is a considerable overlap between one's "personal and social identities" with that of the overall community's identity. This identification with the 'larger community', than an exclusive identification with their respective research centers and labs, would foster collaboration between discrete disciplines, leading to a vibrant inter-disciplinary research climate within iSchools.

We conclude that one could thereby avoid idiosyncratic ways of interpreting social network data by using a theoretical lens like 'Social Identity Theory'.

4. TOOLS USED

"Harzing's Publish or Perish" was used for computing the h-indexes of the faculty members. **Pajek, UCINet** and **Netdraw** were used for the social network analysis and the visualization of our dataset.

5. REFERENCES

[1] Cronin, B., & Meho, L. 2006. Using the h-index to rank influential information scientists. Journal of the American Society for Information Science and Technology, 57(9), 1275–1278.

[2] Green, M. and M. E. Bridget Warbington. 2004. The Information School Diversity Appraisal January 2004 Report, Information School,University of Washington.

[3] Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. PNAS 102 (46): 16569–16572. doi:10.1073/pnas.0507655102

[4] Hogg, M. A., & Vaughan, G. M. 1998. Social psychology (2nd ed.). Hemel Hempstead: Prentice-Hall

[5] Hogg, M. A. 2003. Social identity. In M. R. Leary & J. P. Tangney (Eds.), Handbook of self and identity (pp. 462–479). New York: Guilford.

[6] McPherson, J. M., Smith-Lovin, L., & Cook, J. M. 2001. Birds of a feather: Homophily in social networks. In J. Hagan & K. S. Cook (Eds.), *Annual review of sociology*, vol. 27: 415–444. Palo Alto, CA: Annual Reviews.

[7] Tajfel, H. and Turner, J. C. 1986. The social identity theory of inter-group behavior. In S. Worchel and L. W. Austin (eds.), Psychology of Intergroup Relations. Chigago: Nelson-Hall.

[8] Tajfel, H. 1972. Social categorization. English manuscript of ''La cate´gorisation sociale.''In S. Moscovici (Ed.), Introduction a` la psychologie sociale (Vol. 1, pp. 272–302). Paris:Larousse.

[9] Tajfel, H., Billig, M., Bundy, R. P., & Flament, C. 1971. Social categorization and intergroup behaviour. European Journal of Social Psychology, 1, 149–177.

[10] Xu, J., Chau, M. 2006. The social identity of IS: analyzing the collaboration network of the ICIS conferences (1980–2005). Twenty-Seventh International Conference on Information Systems, Milwaukee

APPENDIX

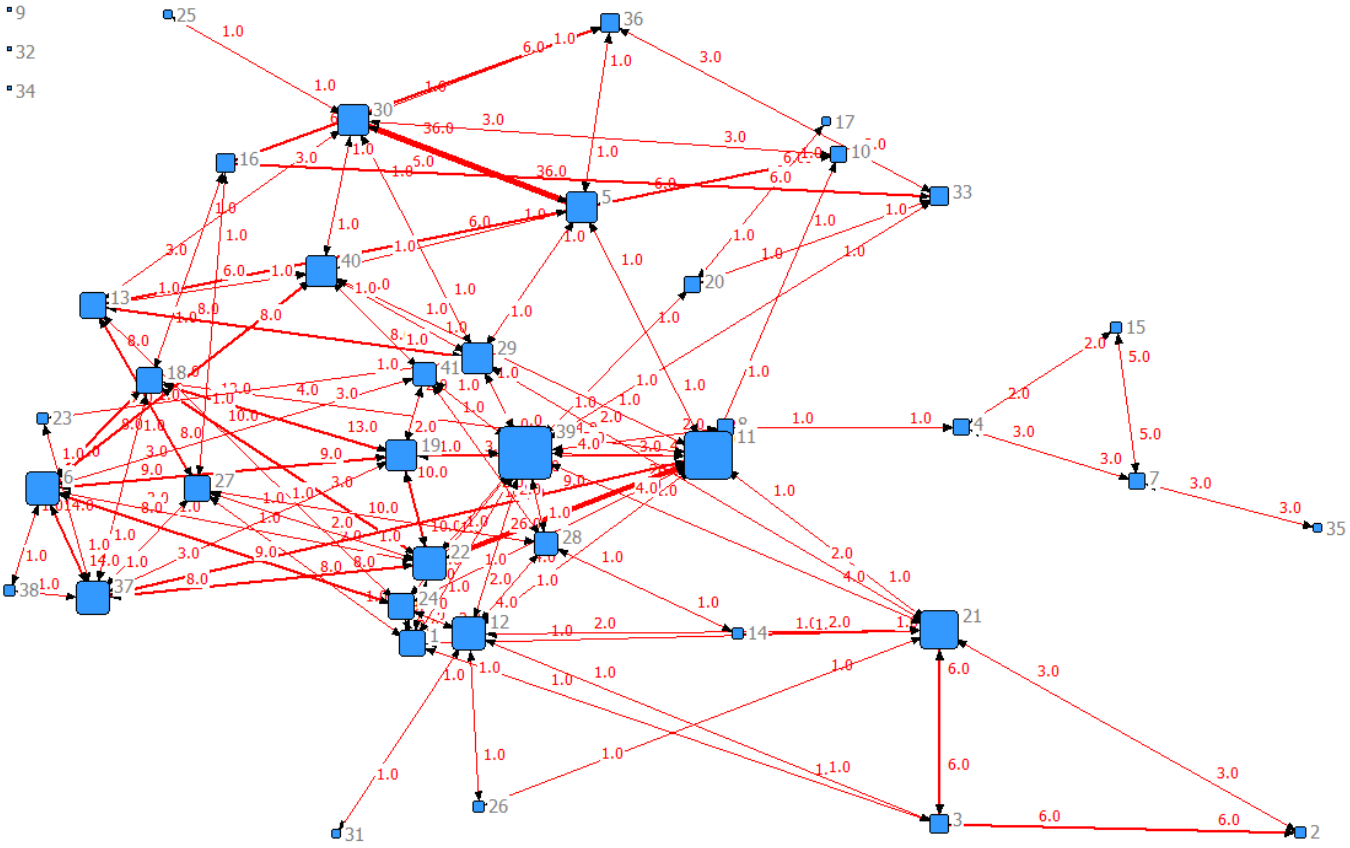

Figure 3. Overall Publication Collaboration Network

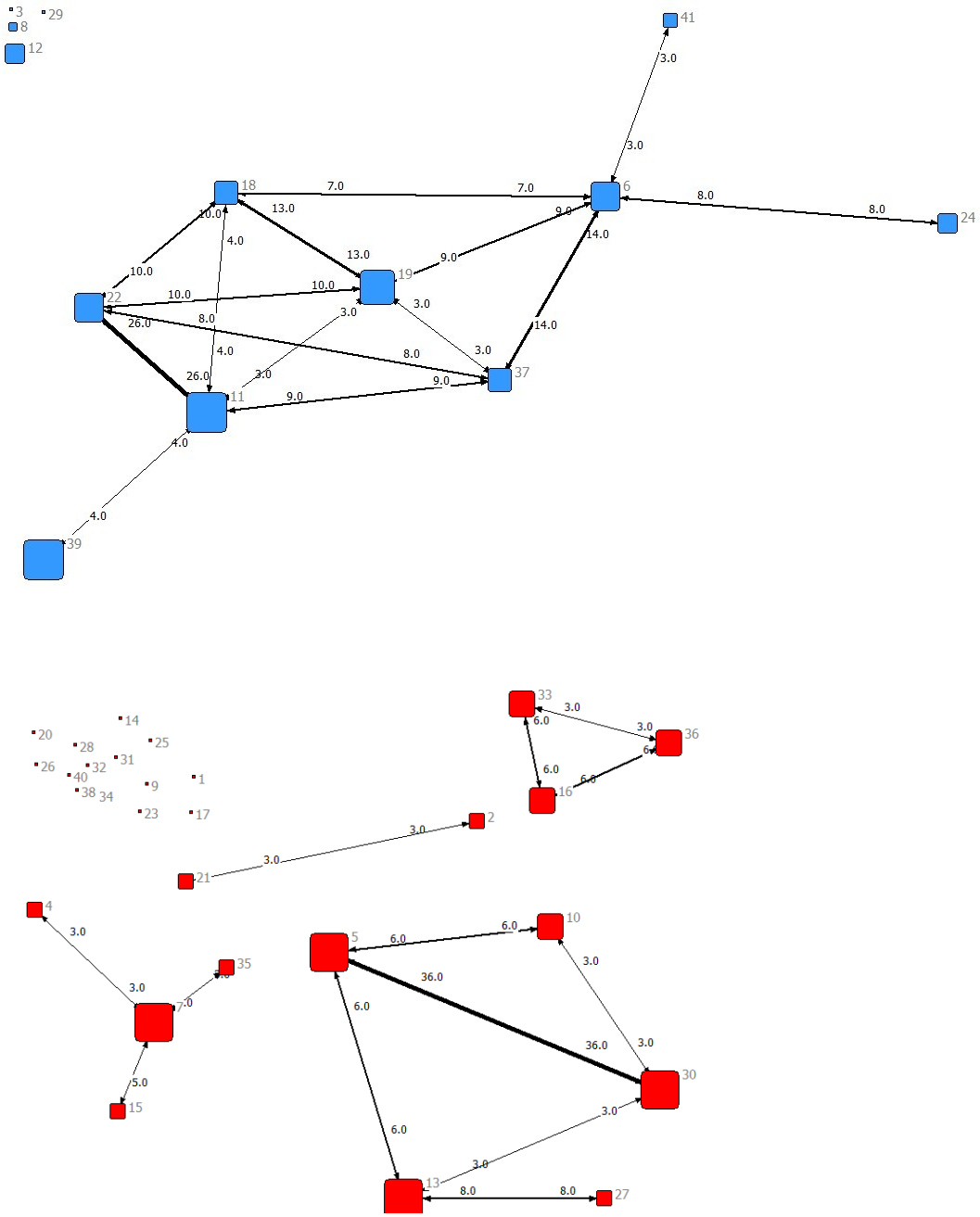

Figure 4. Sub-Network Analysis (Degree Centrality > 2)

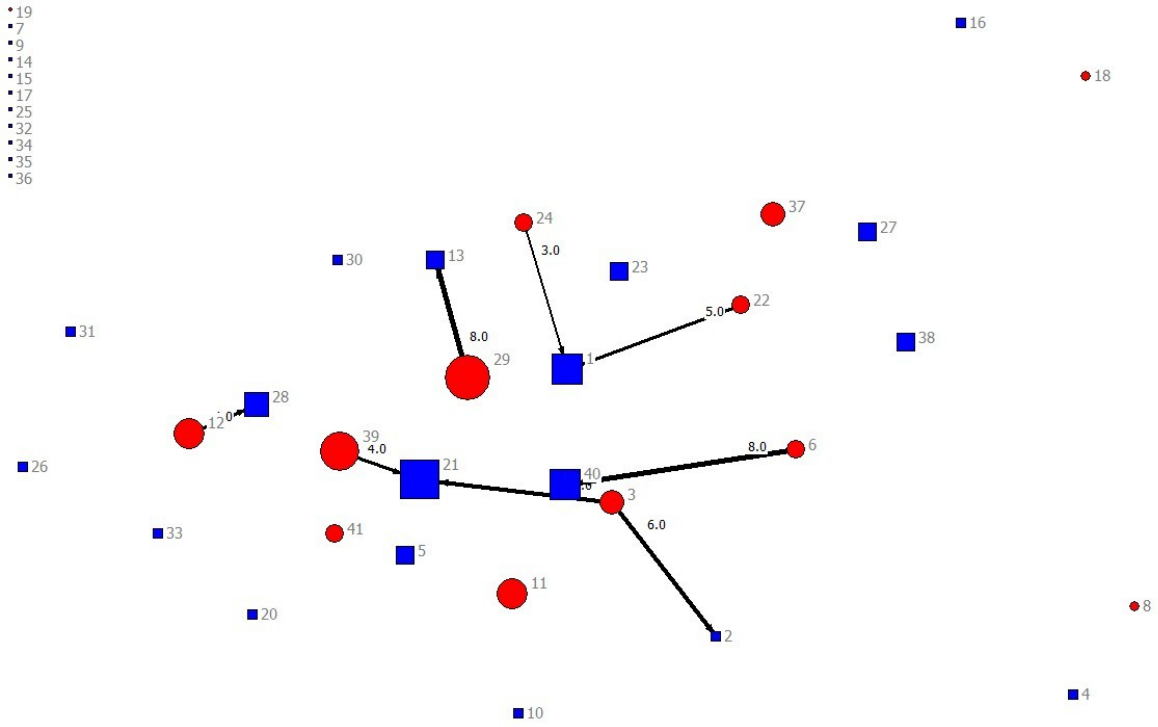

Figure 5. Cross-Network Analysis (Degree Centrality > 2)