Finding Gold in Intranet Data: A Comparison of Mined and Surveyed Social Networks

Chen Zhao[1] **Baozhong Liu**[2]

[1]Microsoft Research Asia Zhichun Road No.49, Beijing, 100190, P.R. China
[2]Peking University Yiheyuan Road No. 5, Beijing, 100871, P.R. China
[1]chzha@microsoft.com, [2]baozhong123@yahoo.com.cn

Abstract

The goal of this research is to identify to what extent a social network extracted from public intranet data sources represents the actual interactions among people as reflected by a self reported survey. We describe a case study we conducted within a large multinational software company comparing social networking information gathered from a traditional socio centric network analysis (SNA) survey with various public intranet sources. Our results suggest that the value which automatically harvested social networks represent good percentage of actual ties with less privacy concern and low cost.

Introduction

Traditionally, social network information has been collected through social science methods, such as interviews and surveys. With the evolution of the internet, the automatic collection of social network information has became more feasible, with email being the most popular source for this automated collection. For instance, analyzing email logs provides a rich source of relationships (Grippa et al., 2006; Guy et al., 2008; Lada, & Eyten, 2005). The use of such information raises, however, concerns of privacy and security that are hard to allay. There are more and more valuable public sources for social network information, both on the internet and within intranets, which present opportunities to collect social network information in more privacy-preserving ways than with mining personal email. Many studies mine social network data by searching the web for various data sources, such as co-authors of academic papers, links on homepages, organizational charts, and net-news archives (Kautz et al., 1997).

Given the rapid growth of online social networks, researchers have studied the network that a traditional social network analysis (SNA) survey provides and those that can be created from electronic or other sources. Guy et al. compared public social network information, extracted from eight different data sources across the IBM organizational intranet, with private social network information extracted from email and the results shed light on the richness of public social network information (Guy, et al. 2008). Grippa et al. compared the complete networks implied by four different media (e-mail, face-to-face, chat and phone) and found some biases in the e-mail mining method, such as the overestimation of communication between peers with technological expertise, and a lack in representing ties between peripheral co-located team members (Grippa, et al. 2006). Although Grippa et al. compared traditional survey-based social networks to those derived from email, there exists little research aimed at understanding to what extent public social network information can represent the actual social interactions among people. In this study, we fill this gap by assessing the level of approximation of socio-centric network information that is achieved from a traditional self-report survey with the public intranet information within a large multinational software company. No technology can be perfect in capturing the actual social networking among people. To increase the reliability of data collected, we avoided using binary options in the survey and Name Generator techniques. We use socio-centric network analysis which is generally considered to give a complete picture of relations in the population.

Grippa et al. found that email covered 72% of the density of the combined networks obtained from email, face-to-face, chat and phone. Guy et al. showed the public social network covered on average over 30% overlap of the private email network. Our own results show that the network captured by public intranet data on average covered 51% of the density of a traditional SNA survey. Thus, suggests there is value in public intranet data which represent good percentage of actual ties without infringing on privacy by mining email. This can be especially valuable for obtaining the social network of an entire organization which is not possible with traditional survey techniques as they do not scale. Applications of this type of data abound, but include social search, expertise finding, social path suggestion, and collaborative filtering. For our study, we analyzed data from the Acing system (Enterprise Social Searching), which was developed by Microsoft Research Asia. Acing allows the sharing and aggregation of social network information across the organization from representative public intranet data sources, which are available in many other organizations, such as ours includes (1) Microsoft Sharepoint collaboration sites, 3.6 million documents, composed in our case of ASP, HTML, Word, and PowerPoint documents; (2) Email discussion lists (0.65 million email threads); (3) organizational charts; (4) homepages of groups, projects and special topics; (5) employees' public personal information. The Acing system has proven to be an effective information extraction technology when evaluated with users making actual queries (Hang, et al. 2005).

The main goal of this study is to identify to what extent a network mined from intranet public data represents the social network implied by traditional self-report social network analysis (SNA). The question of how similar self-report social network data and automatically harvested social behavioral data are similar or different is important to informing how we understand social behavior and build new tools that use social relationships in tasks such as social search, social path suggestion and collaborative filtering. This is an important and yet largely unexplored topic. The rest of this paper is organized as follows. The next section describes our research methods and data extraction. This is followed by description of our results. The last section presents our conclusion and the implications of our findings.

Methods

We conducted a case study in which we compared traditionally derived socio-centric networks with the automatically collected intranet social networks of employees in an R&D organization. The socio-centric network information was obtained from a survey. The Acing system was used to extract online social network from public intranet sources.

Data Collection Methods

Our research sample was two research groups under different departments at Microsoft Research Asia. The selection criteria were that (1) the group had to have between 20 and 40 people; (2) the group had to have formal sub-groups; (3) contacts and communication were supported by a variety of electronic media. The team members included engineers, junior researchers, senior researchers and managers. In the survey respondents were provided with a roster of all the people working in the group. The survey, which was adapted from a previous study (Ehrlich, et al. 2006), consisted of 11 items, including questions on emotional support, information sharing, information acquisition and learning and innovation. Respondents were asked to report: 1 *"How likely do you seek general information/help for your routine job from this person?"* (based on a 4 point scale: most unlikely, unlikely, likely, most likely); 2 *"How often do you communicate with this person for new ideas or better ways to perform your task? "* (based on a 5 point scale: not at all, less than once a month, at least once a month, at least once a week, daily) and so on. We were interested in instrumental or work-related relationships as well as expressive relationships that address emotional matters, which have been shown to play an important role in team performance, work efficiency and employee satisfaction (Ibarra, 1993). So we examined the above four networks which are important for an R&D organization and wondered if certain networks are better represented by the online networks. We then extracted the same group's data from Acing. The Acing mined three relationships within the Microsoft intranet from different sources: Acing-coauthor, Acing-co-occurrence and Acing-colleague. Acing-coauthor and co-occurrence use the co-author and the co-occurrence of names in close proximity in documents publicly available on the Microsoft intranet as evidence of a direct relationship. Acing-colleague mines the social network from an organizational chart and employees' public personal information. This data is all binary: relations being absent (coded zero) and ties being present (coded one).

For each of the two research group networks we studied, we compared the SNA-derived network to the automatically mined network by examining network correlations. The differences in the results for each group were comparable, so here we only report the results for one group. The sample group we report here had 34 members and contained three sub-groups. The group has regular interactions, such as weekly lunch meeting and quarterly group all hands meeting that enable cross sub-group communication. The group members are located on the same floor of an office building, which supports ad-hoc interactions. Group communication is also supported by a variety of electronic media, such as email, newsletters, discussion lists, team websites, etc. We handed out 34 questionnaires and 34 were returned.

Data Analysis

We dichotomized the survey data for each network according to the below coding: "unlikely/likely"; "unaware/aware"; "hard to reach/easy to reach", "less than once a month/at least once a month". We have four adjusted social networks after the integration of items for: emotional support, information sharing, information acquisition and learning and innovation. Network data was represented as adjacency matrices, with both columns and rows representing group members, and the cells representing the absent/present relations between group members from self-report survey responses and Acing, respectively. Ucinet 6 for Windows (version 6.212) was used to calculate network property density and the correlation between networks. As social network data is

Type of network		Emotional Support (survey)	Information Sharing (survey)	Information Acquisition (survey)	Learning & Innovation (survey)	Complete network (survey)
	Density	0.52	0.89	0.52	0.32	0.92
		Jaccard Coefficient between the each paired network by QAP correlation				
Acing-coauthor (mined)	0.03	0.04**	0.03	0.05***	0.06***	0.03
Acing-co-occurrence (mined)	0.03	0.05***	0.04 *	0.06***	0.08 ***	0.04
Acing-colleague (mined)	0.24	0.31***	0.26***	0.37***	0.45***	0.25***
Acing combined (mined)	0.25	0.32***	0.27 ***	0.40***	0.45***	0.26***

Table 1. Densities of networks and their correlation * $p \le .05$, **$p \le .01$, ***$p \le .001$

not independent and does not satisfy the assumptions of statistical inference, we used the QAP Jaccard to run the correlations. QAP Correlation computes correlation and other similarity measures between entries of two square matrices and is principally used to compare correlation between a pair of networks. The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the sample sets (Hubert & Schultz, 1976).

Results

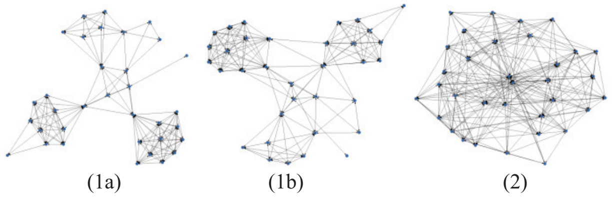
Table 1 provides the density of each mined network and QAP correlations between each survey network and the online mined networks. Among the four survey social networks, the information sharing network had the highest value of group density (0.89), followed by the information acquisition network and the emotional support network (0.52), and the learning and innovation (0.32) network. Comparing the three mined networks, the Acing-colleague network captured more ties (density = 0.24), relatively speaking. Acing-coauthor and co-occurrence networks' densities are very low, at 0.03 only. The Acing combined network (density = 0.25) is based on the combination of the three Acing networks and covers 51% of density of the four traditional SNA survey networks on average (range from 28%, Acing combined / Information Sharing network to 78%, Acing combined / Learning & Innovation network). We also combined the four survey networks to get a complete survey network (density = 0.92). The total edges of the information sharing network explained 97% of the overall network density. The Acing general network participated with 27% of the connections in the complete network implied by traditional SNA survey.

For the network correlation and similarity measures, the correlation between the Acing-colleague network and the learning and innovation network is the highest among all the networks (Jaccard Coefficient =.45, p<.001). The Acing-colleague network also has higher correlations with each survey network. The Acing-co-author network and the Acing-co-concurrence network have very low correlations with the four networks we obtained from the survey. Despite the low similarities, most correlations between the social networks obtained via survey and the online networks we obtained by data mining are significant, except for the Acing co-author network with the information sharing network.

Discussion

As we discussed above, the main objective of this study is to identify to what extent a network mined from internet public data provides the same view of a network that traditional social network analysis (SNA) surveys do, and how accurately digital traces interpret these relationships. Our results show that the network captured by public intranet data on average covered 51% of the density of the traditional SNA survey. This result suggests there is value in public intranet data which represent good percentage of actual ties with less privacy concern and low cost. The public intranet data we examine in this study are available to many organizations without infringing on privacy by mining email. The value of social networks has been widely recognized and more and more tools provide social network analysis using internet or intranet public data, such as papers, documents, bibliographic databases, organizational charts, and net-news archives mined by web search (Kautz, 1997). Our results suggest that those social networks built based on public data can represent a valuable part of the actual relationships; the maximal overlap identified in our study is over 45% for the Acing-colleague network and the learning and innovation network.

Density is perhaps the most common way to index network structure as a whole; it reflects the level of interrelatedness, or reticulation, among all possible social ties. The four networks we obtained from the survey have a high density (over 0.52), except the learning & innovation network. The three automatically mined networks have lower densities. The combined Acing network has the highest degree of correlation (45%) with the learning and innovation network. The density difference (0.07) between those two networks is the smallest among all paired networks in this study. Among the four survey-based networks, the density of the information sharing network is significantly higher than the other three networks. Compared to the other survey-based networks, the information sharing network is relatively easy to capture by survey data. Our research sample is an R&D organization, where learning and innovation is considered very important. The 45% correlation between the learning and innovation network and the Acing combined network indicates that an online social network might be relatively more effective at identifying an organizational learning and innovation network. Figure 1 shows graphs of different networks implied by the intranet data and SNA survey.



1a) network built on Acing colleague, (1b) network built on Acing, (2) learning and innovation network

Figure 1. Graphs of different networks implied by intranet data and SNA survey

There are several limitations to our study. We only studied two R&D groups and report one here which may raise the generalization concern. We argue that the densities of self-reported survey network in our study are higher compared to many other literatures. In this way, our findings can be supported in stronger way in term of the coverage of ties. In the future we would like to extend the samples with other business units to make it more representative. In this study, we have only analyzed the network density that is the most common way to index network structure as whole. We may also want to consider other network properties, such as betweeness, small world and structure hole. In Guy et al.'s study comparing intranet public data with private email, they include data sources from Beehive, a social networking site, and Fringe, a friending and people tagging system within the enterprise (Guy, et al. 2008). We believe the automatically mined public intranet social network could be even more comprehensive if information sources like those found in Beehive and Fringe can be aggregated. We look forward to including this type of information when it's available.

Finally, we have confirmed that powerful algorithms can mine actual social networks. Meanwhile we must notice there is richness in the off-line interactions that is not captured by data mining and might be crucial to understanding what the interactions on-line are about. We should be aware of the representation when we take advantage of automatically mined social network information. In future work, we are interested in understanding what is missing so as to help us design new tools to capture, extract and mine holistic social interactions. Since there are fewer ties in the Acing combined than in the self-reported survey networks, we can ask who were left out? Were ties included more central? Do the passive networks pick up on the tip of iceberg (in which cases they are useful) or do they simply pick up on any assortment of this network (which is not as useful)? Some new technologies such as mobile phone tools for tracking people's daily activities make it promising to analyze what is lost in the mined networks.

In this work, we compared the degree of correlations among different pairs of social networks obtained from both a traditional survey and by the mining of intranet public data, such as web documents, discussion lists and web pages. We used the Acing system to obtain the public online social network information. Results show that the network captured automatically on average covered 51% of the density of the traditional SNA survey-based network, and a 26% correlation with the complete socio-centric network and a 45% correlation with the learning and innovation network implied by the self-reported survey. We have proposed a specific question: To what degree do online social networks represent the actual social networks? Our analysis shows the value of automatically harvested public intranet social networks present. More and more social networks mined from automatic collection of digital data are being used for various purposes, such as visualization of social maps, identifying central individuals or suggestions of paths for finding experts. It is important to know how self-report social network data and automatically harvested social behavioral data are similar or different because this will help us understand social behavior and facilitate the design of tools which will mine and use socio-centric social networks.

References

Ehrlich, K. & Chang, K. Leveraging expertise in global software teams: Going beyond boundaries. *In Proc. IEEE International Conference on Global Softwar* (2006).

Grippa F., Zilli A., Laubacher R., & Gloor P. E-mail may not reflect the social network. *In Proc. International Sunbelt Social Network Conference,* (2006).

Guy, I., Jacovi, I., Meshulam, N., Ronen, I., & Shahar, I. Public vs. Private – Comparing Public Social Network Information with Email. *In Proc. CSCW*. ACM (2008), 393-402.

Hang, L., Yunbo C., Jun X., Hunhua H., Shenjie L. & Dmitriy, M. A New Approach to Intranet Search Based on Information Extraction. *In Proc. International conference on Information and knowledge management,* ACM (2005).

Hubert, L. J. & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology, 29,* 190-241.

Ibarra, H. (1993). Network centrality, power and innovation involvement: determinants of technical and administrative roles. *Academy of Management Journal 38,* 471–501.

Kautz, H., Selman, B., & Shah., M. (1997). ReferralWeb: Combining social networks and collaborative filtering. *Comm. of the ACM*, 40, 3, 63-65.

Lada, A. & Eyten, A. (2005). How to search a social network. *Social Networks*, 27(3):187—203.