INVITED PAPER

Social Network Analysis in Enterprise

This paper focuses on the challenges and solutions in mining and analyzing social networks in enterprises; the authors base their study on a social network analysis tool called SmallBlue.

By Ching-Yung Lin, *Fellow IEEE*, Lynn Wu, Zhen Wen, *Senior Member IEEE*, Hanghang Tong, Vicky Griffiths-Fisher, Lei Shi, *Member IEEE*, and David Lubensky

ABSTRACT | Social network analysis (SNA) has been a research focus in multiple disciplines for decades, including sociology, healthcare, business management, etc. Traditional SNA researches concern more human and social science aspects—trying to undermine the real relationship of people and the impacts of these relationships. While online social networks have become popular in recent years, social media analysis, especially from the viewpoint of computer scientists, is usually limited to the aspects of people's behavior on specific websites and thus are considered not necessarily related to the day-to-day people's behavior and relationships. We conduct research to bridge the gap between social scientists and computer scientists by exploring the multifacet existing social networks in organizations that provide better insights on how people interact with each other in their professional life. We describe a comprehensive study on the challenges and solutions of mining and analyzing existing social networks in enterprise. Several aspects are considered, including system issues; privacy laws; the economic value of social networks; people's behavior modeling including channel, culture, and social inference; social network visualization in large-scale organization; and graph query and mining. The study is based on an SNA tool (SmallBlue) that was designed to overcome practical challenges and is based on the data collected in a global organization of more than 400 000 employees in more than 100 countries.

KEYWORDS | Atlas; behavior analysis; computational social science; enterprise; graph analysis; large-scale network; organization; SmallBlue; social capital; social network analysis (SNA); social network visualization

I. INTRODUCTION

In recent years, we have witnessed a drastic uptick in the growth of information. With the recent advance of social media and the growing use of social networking tools, organizations are increasingly interested in understanding how individuals, teams, and organizations harvest value from their social networks. As estimated in 2006, the amount of digital information created, captured, and replicated is 161 billion GB, about three million times the information in all the books ever written [12]. Thus, the simultaneous explosion of social media, knowledge management, and networking tools is not a mere coincidence, as these technologies have played an important role in sharing and disseminating the vast amount of information recently created. However, before formulating network strategies on how one leverages social networks to achieve superior outcomes, it is crucial to understand how and why networks create advantages. It should be also noted that a major difference of social network analysis (SNA) in enterprise and in online social media is its stronger interest in finding the "actual" social networks and productivity and security impacts rather than the friending networks.

Drawing from the field of economic sociology, social network researchers have long predicted that certain network positions are more advantageous than others. One particular network that has perceived a tremendous amount of attention is structural holes. Actors spanning multiple structural holes are theorized to have more information and control advantage than their peers. For example, bankers with structurally diverse networks are

Manuscript received June 30, 2011; revised February 6, 2012; accepted March 18, 2012. Date of publication July 26, 2012; date of current version August 16, 2012.
C.-Y. Lin, **Z. Wen**, **H. Tong**, and **D. Lubensky** are with IBM T. J. Watson Research Center, Hawthorne, NY 10532 USA (e-mail: chingyung@us.ibm.com).
L. Wu is with the Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA.
V. Griffiths-Fisher is with IBM U.K., London UB6 0AD, U.K.
L. Shi is with the Chinese Academy of Science, Beijing 100864, China.
Digital Object Identifier: 10.1109/JPROC.2012.2203090

more likely to be recognized as top performers [7]. Similarly, employees in research and development positions maintaining diverse contacts outside the team are more productive than their peers [32]. Interestingly, findings in structural holes transcend beyond individual levels. Projects, teams, and firms that span structural holes are also correlated with higher work performance. McEvily and Zaheer find greater access to competitive ideas when firms have access to nonredundant sources of advice beyond the firm [25]. Stuart and Podolny show that firms are more likely to create innovative products when they establish alliances with organizations outside their own technical area [37]. Though these studies are largely correlations, the results collectively show that structural holes seem to affect performance regardless of the setting, the industry, or the level of analysis.

SNA has been an important scientific research focus in management, sociology, and healthcare for decades. However, traditional SNA relied heavily on manual methods, such as questionnaires and interviews, to construct social networks. The results are usually static and the scope has been limited. Today, workers frequently interact digitally. Because of the limitation of meaningful data acquisition, especially from academics, more systematic ongoing large-scale researches are still waiting to be done to leverage the ample data that are created by people's interactions, such as e-mail, call logs, text messaging, document repositories, and web 2.0 tools in organizations. It is very difficult to conduct large-scale cross-modality or multimodality analysis, e.g., examining how personal network structures affect revenue. This gap is problematic, because the literature on organizational networks suffers from the same deficits that much of the social network literature does. It has to focus on small, static networks, because electronic traces reside in heterogeneous places.

In most countries, employee data generated through company assets belong to the company. Company, as a legal identity, is obligated to the data generated by its employees and thus has strong legitimate needs to collect and store all work-related data. Employees are supposedly not allowed to use the company assets for personal use. However, it is common that employees browse the Internet and receive/send personal e-mails using company computers and networks. Privacy law, telecommunication law, and labor law in many countries prohibit the collection, aggregation, and use of such data that reside in scattered servers.

SmallBlue went live in 2006 for enterprise collaborations [10], [23] and is the first major system that overcame the challenges and paves way to scientific insight for large-scale dynamic SNA through continuous multimodality data acquisition. *SmallBlue* has been deployed in more than 70 countries to quantitatively infer the social networks of 400 000 employees within IBM organization. We have deployed 15 000 *social sensors* in volunteers' machines to gather, crawl, and mine more than 25 million messages, including content and properties of individual e-mails and instant message (IM) communications. Here, an important solution is to gather data from users, not servers, in order to be compliant to privacy laws, and it is important to get explicit consents. Furthermore, we also gathered information such as the organizational hierarchical structure, project and role assignment, employee performance measurement, personal and project revenue, etc. Except the small-scale studies based on surveys, there was no precedent in literature being able to link these data involving all three aspects of capital: *financial capital*, *human capital*, and *social capital*. Mining "existing" social networks in organizations can be used for various applications such as expertise and knowledge search, social proximity and collaboration, social recommendations, marketing, and cybersecurity. *SmallBlue* was originally used for global collaborations of enterprise employees [23], which requires solving the issues of data gathering, privacy laws, structure and economic analysis, culture analysis, and visualization. Since 2010, it has been extended to accommodate other applications such as information browsing, cybersecurity and data leakage detection, anomaly detection, and content recommendation. These applications require analysis and infrastructure for large graph storage, mining, and visualization.

This paper describes and provides overviews of the various aspects that an enterprise SNA system needs to consider in practice. It is organized as follows. First, in Section II, we introduce the data we collected in organization and the data privacy laws that guide us to the system design. We show the guidelines that are usually needed for enterprise to collect data about employees and the required balance between company's goals and employee's privacy. In Section III, we describe a few studies on economic studies of social network impacts toward employee performance. Specially, we will report on a new study of the financial impact of social media tool, such as *SmallBlue*, in enterprise. In [48], we reported that adding a person in one's practical social network,[1] on average, contributes to additional $948 annual revenue to enterprise. People with strong e-mail ties with a manager, or a more diverse circle of correspondents, enjoyed greater financial success than those who were more aloof. Teams with an even mix of genders also performed well financially. Individuals have more diverse networks and thus have more people who are reachable within two social steps (i.e., your friends' friends' friends), which is valuable. Too intensive communications with the same people have a negative impact, perhaps because of the repetitive redundant information exchange. We also discovered that the common expression of "too many cooks spoil the broth" really is

[1]Note that the theoretical cognitive limit of the number of people with whom an individual can maintain stable social relationships is bounded by a commonly used value of 150, which is usually called Dunbar's number [9]. This upper bound is clearly observed in our data set, while only one out of more than 15 000 people we studied exceeds this bound, and about 87% of people are not maintaining stable social relationships of more than 100.

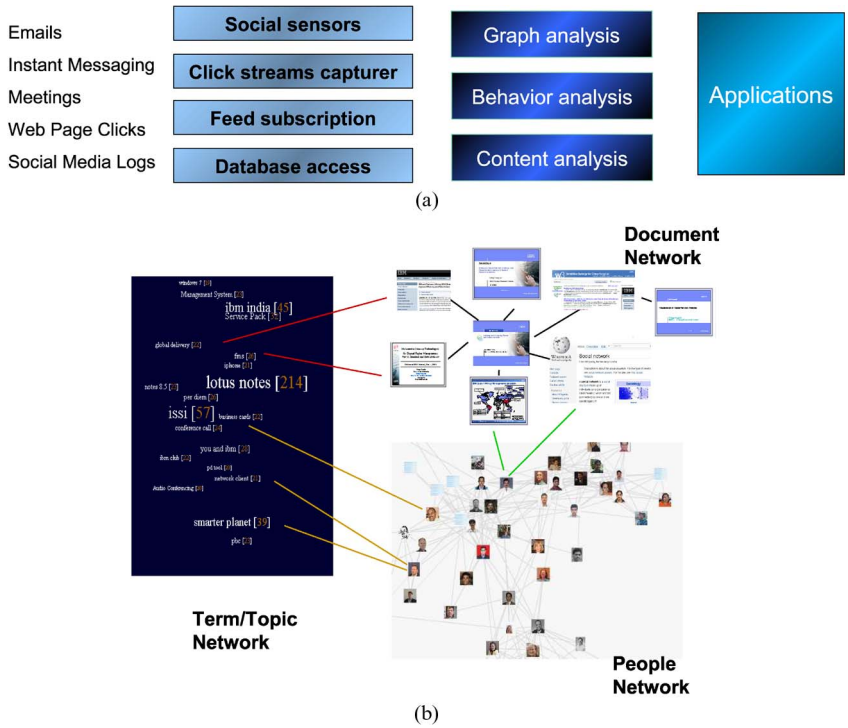


Fig. 1. ***System for SNA in enterprise: (a) flowchart, (b) generating tripartite relationship networks via data mining.***

true—with less success attributed to projects with too many managers.

Channel, culture, and influence issues in people relationships are also of strong interest in global organizations. We model the dynamic and evolutionary people's relationships as multilayer networks. Section IV describes how the layers of people's behavior can be considered as graphical models, including a person's networks, characteristics of network edges between a pair of people, and the dynamic graph representation of the intrinsic network of a person. We will also show some new analyses on culture aspects of social networks. Section V describes several of our network visualization tools, including visualization of large-scale networks based on hierarchical clustering. In Section VI, we will address the next step applications of SNA in graph mining. We will discuss future directions and conclusions in Section VII. Note that, although a commercial version of *SmallBlue*—*IBM Atlas*—has been deployed in several global enterprises, the empirical data analyses reported in this paper are limited to our internal deployment of *SmallBlue*, because data in other companies are not accessible.

II. DATA ACQUISITION AND PRIVACY ISSUES

Fig. 1(a) describes the fundamental structure of our system. We implemented several methods to collect various aspects of people's activities in enterprise, including: 1) social sensors; 2) clickstream capture; 3) feed subscriptions, and 4) database access. Then, we conducted three types of analysis: graph, behavior, and content semantics. Various applications such as expertise search, people and content recommendation, social search, social path access, etc., are some of the sample applications.

A. Data Acquisition

Social sensors [23] are based on a distributed front-end analysis mechanism that is installed in individual volunteer's machines. Its usage is twofold. First, this mechanism can distribute the computational workload by placing first level of data gathering and feature extractions. Second, this is an important mechanism for privacy compliance. In several countries, it is illegal to conduct data analysis within the communication channel. Communication providers cannot process data for the purpose other than providing communication services. Social sensors solve this legal issue by processing the copy of the data that are stored in an individual's computer, instead of gathering data from communication servers. This mechanism can resolve several legal challenges. Furthermore, via distributed sensors, our system can distribute the first level of feature extraction functions of content analysis such as stop word removal, stemming, one-gram and bi-gram statistics, etc., in an individual's machine to avoid the liability of storing the original communication content in a

centralized server. Many features were designed to protect the human rights on privacy and free speech.

Clickstream capturers were implemented and embedded in several enterprise webpages to capture users' web browsing and clickthroughs inside enterprise. With this mechanism, the server captures information directly through users' browser, which sends a small packet to the server for each user click. Feed subscription is used for getting user data that are provided from the service provider via JSON or ATOM feeds. It includes the server logs of user behaviors as well as the server data on the content. Database access is also included because some of the user activities are conducted through traditional databases without going through the Web.

Table 1 shows the data we have collected in our enterprise. We generate tripartie social information networks in organization, as shown in Fig. 1(b). Several different types of networks are generated: the dynamic and evolutionary relationship network captured from communications, the relationship of documents that are captured via the content similarity as well as linkage generated via common authors, readers, etc. (such as collaborative filtering), and the term/topic networks that are generated by people's search terms in session, terms used in a communication, etc. Afterwards, the system conducts graph analysis, behavior analysis, and semantic analysis.

Table 1 Data Description

Data Collection	Source/Type	Size
Email	Enterprise emails, from 15,000 employees in 76 countries. The data include sender, receiver, subject, timestamp, and content represented by term frequencies.	More than 25,000,000 of emails over 3 years.
Instant Messaging	Enterprise instant messages from 10,000 employees. The data include sender, receiver, timestamp, and content.	7 million messages over 2.5 years.
Calendar Meetings	Enterprise meeting information from 10,000 employees, including meeting host, attendees, timestamp.	over 2.5 years.
Blogcentral	Internal blog hosting system.	50,900 blogs.
Social Bookmarking	Internal social bookmarking [26] data (who bookmarked which url by what kinds of tags at when)	350,000 social bookmarking.
Knowledge Database Access	The access information and document content of IBM internal knowledge base (who read which document at when, and the document might be tagged and rated)	3 million of records.
File Sharing	Internal file sharing information, who posted/shared/accessed which document at when	60,000 shared files.
ClickStream and Search Terms	ClickStream and SearchTerms on several webpages	about 100,000 daily records.
Financial Contribution	Information about the financial consulting revenue a person or a project team earns for the company	200,000 employees with monthly data.
Employee Directory	Employee names, demographics and reporting structure, dynamics in the group	Over 400,000 employees

A key issue of protecting privacy is to detach the personal identifiable information from the collected data, if any analysis is done without the explicit personal consent. Sensitive data need to be hashed. Furthermore, content collection of e-mails and IMs needs to avoid capturing the original sentences which can be deanonymized, in comparison to the statistics of one-gram or bi-gram frequencies. Users also need to be able to set up controls on what/when contents should not be captured and have the right to modify any incorrect inference. Section II-B will describe the key issues about the privacy law. Note that for some specific types of sensitive information, such as health histories used in healthcare industry, a stronger protection based on cryptography is required. A new research thread is merging, based on the end-to-end encrypted domain data mining mechanism to protect sensitive information using full-homomorphism cryptography while allowing data mining applications such as recommendations without decryption [34]. That method can prevent system owners from accessing the content and thus provide strongest protection to privacy and is better immune to system cyberattacks.

B. Privacy Laws

Privacy is a fundamental human right, as described in the United Nations Universal Declaration of Human Rights in 1948. Article 12 specifies:

> "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such inference or attacks."A fundamental element of privacy is data privacy, the ability to control one's personal information (PI), where PI is defined as any information that relates to a living individual who can be identified from that data, or from that data plus other information which is in possession of, or is likely to come in possession of the data controller. Data-privacy-related legislation varies widely across the world, a critical legislative element being the European Union (EU) Personal Data Protection Directive 95/46/EC, where there is the added complication of interpretation and enforcement of the legislation varying in each of the 27 member states. Fig. 2 shows the current status of privacy laws worldwide.

In an organizational setting, other factors related to employment legislation also had to be considered. The employer/employee relationship can compromise the

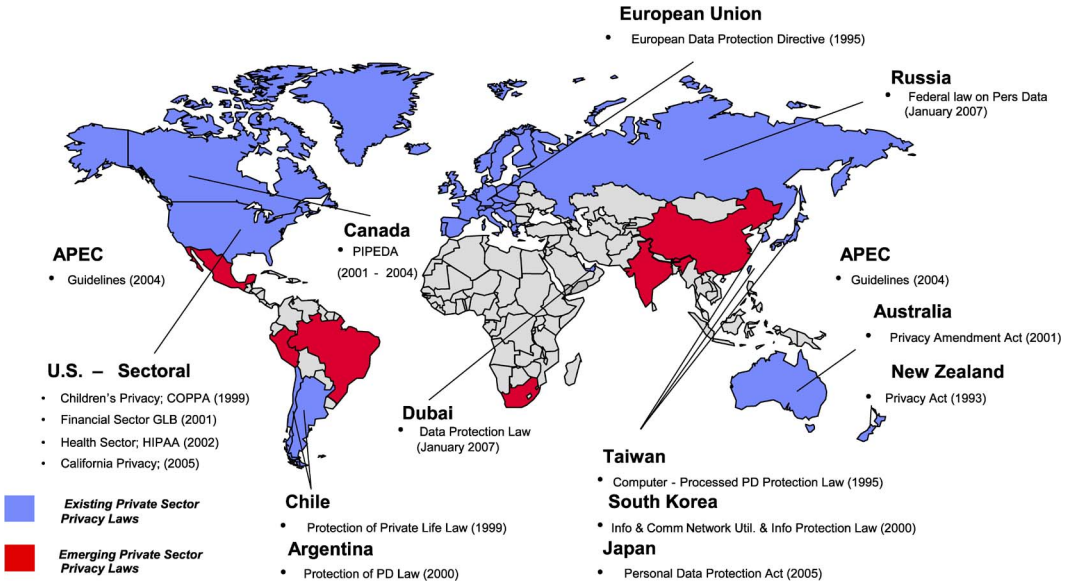


Fig. 2. ***Worldwide privacy laws. In Europe, most countries have also derived their own privacy law based on the EU data protection directive.***

ability to gain free and informed consent of participants in some countries. There are strict limitations around, and in some countries (e.g., Germany and Austria), prohibitions on employee monitoring at work. Together these mean that social software features that present few or no issues in an Internet setting can present significant issues in an enterprise setting [36].

In order to make the social network mining system a practical and valid application in a global organization, several aspects of privacy had to be considered. The first was the maturity of the organization with respect to privacy, in terms of privacy polices and the ability of the organization to accommodate the processing and global transfer of data in line with applicable legislation. The second was to design the platform so that the different legislative and privacy related requirements of applicable geographies can be accommodated. The system was designed to have a flexible set of user types with differing characteristics for data capture, sources, processing, and application use. These were made configurable at a number of different levels (e.g., country, division) so a privacy policy and organization-segment-driven approach to implementation is possible. The third was to adopt the EU data protection principles [16] of *notification*, *purpose*, *consent*, *access*, *information standards*, and *security*, as shown in Table 2, into the design of the platform, the result being the appropriate balance of maximum utilization of data with the ability for users to fully control their participation level, visibility in the system, as well as the data used to represent them in the system.

From a practical aspect, the system had to be approved by the data privacy officers responsible for each country with applicable legislation in addition to labor union (works council) approval in some EU countries. Engaging with the privacy and legal departments early on was critical, with some of the configurable features required to protect the privacy being the product of a collaborative design process with privacy practitioners making *SmallBlue*, to our knowledge, the first system in literature to be legally deployed globally for enterprise SNA and a unique privacy preserving system.

III. VALUE OF SOCIAL NETWORK

SmallBlue allows us to track how individuals' networks evolve over time. To evaluate the performance implications of social networks, we also obtained the performance

Table 2 EU Data Protection Principle, Adopted by the SmallBlue System Design

Type	Description
Notification	Data subjects should be given notice before their data is collected and processed.
Purpose	Data subjects should be informed of the data collected and the specific purpose(s) it will be used for. It should not be used any other (incompatible) purpose.
Consent	Data should not be collected, processed or disclosed without the data subject's free and informed consent.
Information standards	Data should be: accurate and up-to-date; adequate, relevant and not excessive; not kept longer than necessary.
Access	Data subjects should have to access their data and be able to have inaccurate data corrected or destroyed.
Security	Data should be kept secure from accidental or deliberate compromise or misuse.

metrics of these individuals. The longitudinal nature of the analyses enables us to explore the potential causal linkage between social networks and performance and observe the micromechanisms of how networks drive productivity. The detailed recording of electronic communication archives also helps reducing the potential biases derived from using surveys and self-reports. Often, networks constructed using self-reports are subject to memory errors and related biases. For example, recent interactions are more memorable than distant interactions. *SmallBlue* alleviates this type of error because each electronic communication exchange is recorded with a timestamp and the content of these messages is also encoded and stored in archives. The system has a perfect memory of all the electronic communication records. Social networks derived from such data are thus rarely subject to memory errors or recall biases that often mire the validity of survey instruments in earlier network studies [24]. However, social networks instantiated using electronic communications are also not always a perfect representation of a person's overall network. After all, face-to-face interactions, especially impromptu encounters around water coolers, cannot be recorded easily, and accordingly, networks generated from electronic communications do not include impromptu face-to-face interactions, thus potentially biasing the real social network of individuals. Furthermore, what constitutes a tie also differs in the online world as opposed to the offline world. When two people e-mail each other once, it does not necessarily mean that a real network tie exists between the two of them. They may not ever communicate again. Thus, we have to be extremely careful in determining what constitutes a tie in electronic communications.

To achieve this goal, we tested various criteria to best represent a tie between two people and matched the results against a survey we conducted about people's relationships and interactions. We find that a network tie exists between two individuals only when they have communicated enough to pass a certain communication threshold. This threshold may differ across individuals because it incorporates the propensity to use electronic communication in the calculation. If a person who e-mails frequently requires a higher threshold to register a tie than someone who rarely uses e-mail

$$X'_{i,j} = \begin{cases} 0, & X_{i,j} \leq 3 + \log(X_{i,j}) \\ X_{i,j}, & \text{otherwise.} \end{cases}$$

The above formulation indicates that a tie exists between people only when they have communicated on at least three occasions. The tie strength is approximated by the log of total electronic communications between persons i and j, i.e., $\log X'_{i,j}$. We calculated a normalized tie strength $p_{i,j}$, which presents the faction of the network strength i has devoted to j, $p_{i,j}$. It is then used to calculate the structural holes [6]

$$p_{i,j} = \frac{\log\left(X'_{i,j}\right)}{\sum_k \log\left(X'_{i,k}\right)}$$

$$\text{Structural Holes}_j = 1 - \sum_j \left(p_{i,j} + \sum_q p_{i,q} p_{q,j} \right)^2, \quad q \neq i, j.$$

Structural holes measure the degree to which a person's network is redundant. If a person's social connections are all connected with each other, then this person has a maximally constrained network and all her contacts are redundant in the sense that all her friends can access the same resource she has. The structural holes measure for this person is very low. However, if a person's connections are not connected, her structural hole measure would be high, indicating that her network is not redundant.

A. Network Effects on Personal Revenues

To leverage the longitudinal nature of our network data, we created a panel of networks using both three- and six-month intervals with a sliding window of one month. We matched these time-varying network data with consultants' performance as measured by billable revenue. We also gathered information about these consultants such as their *gender*, *division*, *hierarchy within the firm*, *seniority*, *job role* as well as *the type of work and the industry* these consultants typically work for. These factors serve as the control for our econometric analysis to eliminate confounding factors such as more senior consultants are more likely to generate more billable revenue.

We leveraged both random-effect and fixed-effect econometric models to eliminate many confounding factors that are unobservable in our data, such as *personality traits* or *inherent abilities*. For example, if certain individuals are very social and they also happen to be the star performers, the positive relationship between diverse networks and performance may be spurious because both are the results of an underlying personality trait, instead of having a real causal nexus. Similarly, a person could have a diverse network because her positions and hierarchical order require her to reach out to many people. Again, the positive relationship between performance and network positions is a result of the person's inherent job role, as opposed to network positions actually enabling performance. By eliminating these factors using panel data, we greatly reduce this type of bias in estimating the effect of social networks on performance. Recording the network change of individuals over a long period of time (over three years), as we have done in *SmallBlue*, allows us to explore how the change in networks relates to performance.

We found that certain network characteristics are highly correlated with performance. Using both random-effect and fixed-effect models, we found that structural holes are highly correlated with performance in a statistically significant way at 5% p-value, after controlling for seasonal shocks and demographics. Specifically, we found that one standard deviation of structural holes is associated with billing $882.4 of additional monthly revenue for the company. We controlled for seasonal shocks because it is possible that a person is able to bill more simply because it is a good market during holiday seasons and her work is in high demand. Similarly, we controlled for demographics because economic conditions in a certain region can be better than in others and consultants residing in those well-off regions can naturally bill more than others.

We explored how network size affects performance. We found that each communication exchange in the form of e-mail, IM, and calendar event has a negligent effect on performance, and one extra person that communicated has a modest return on performance. Overall, results indicated that the network structure rather than the network size or communication volume dominates the return on performance, even after eliminating confounding factors such as individual abilities, personality traits, positions within the organization, and seasonality shocks that may bias the estimates.

B. Network Effects on Projects

We explored the implication of structural holes at the project level where each node in the network represents a project and each link in the network represents the communication instances exchanged between the two projects forming the link. Similar to the findings at the individual level, project networks that span structural holes are associated with positive increases in a project revenue, after controlling for the total number and the type of people in each project, temporal and regional shocks such as business cycle at various regions, and the line of business the project is in. We also employed random and fixed-effect specifications to eliminate other time invariant factors. Specifically, a one standard deviation of structural holes at the project level is associated with additional billing of $776 in revenue.

Interestingly, we found that the number of managers in projects is positively correlated with the overall project revenue, probably because more managers may send positive signals to the client that the firm is staffed with its best employees for the project. However, the relationship exhibits an inverse-U shape that having too many managers involved in a project can actually hurt the project's revenue. We studied 1029 consultants (including 66 managers) and 2952 projects in 39 countries from June 2007 to July 2008. The coefficient on quadratic of managers is negative, implying a concave relationship, such that more managers in a project team are associated with greater revenue to a point, after which there are diminishing marginal returns, and then negative returns to increased number of managers

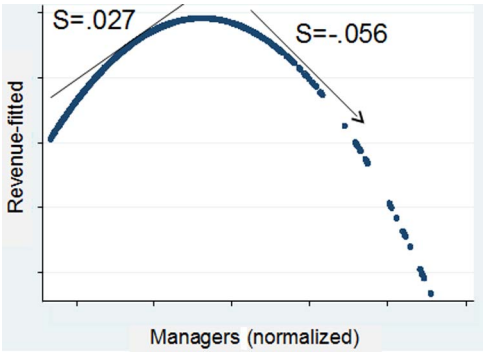


Fig. 3. ***The fitted curve of the revenue versus the number of managers in a project.***

$$\text{rev} = \alpha + \beta_1 \text{mgr} + \beta_2 \text{mgr}^2 + \gamma_1 \text{factor}_1 + \cdots + \gamma_k \text{factor}_k + \epsilon.$$

Using linear regression, we got coefficient of β_1 being $2733.9, with the heteroscedasticity-consistent standard error being $537.5, and β_2 being $-$$682.02 with the standard error of $215.3. The best fitting curve is shown in Fig. 3. The result is statistically significant with p-value < 0.001. Perhaps, lacking a clear leadership role intensifies internal debates among managers and derails the consultant from making progress. Our interview with consultants further confirms our hypotheses.

Overall, we show that certain configuration of network ties can have a positive effect on work performance both at the individual level and at the project level. These results inspire us to explore whether social media can play a role in helping individuals achieve superior work performance.

C. Impact of Social Networking Tool

If social networking tools can facilitate the process of finding the right resources that are critical to the task at hand, they could have tremendous implications for organizations, especially for creating strategies on how to invest and use these technologies to improve firms' bottom line.

We studied 2038 anonymized global business consultants for two years. In Fig. 4, we plot the relationship between individual work performance as measured by billable revenue and the number of months since the adoption of *SmallBlue*. We controlled for factors including temporal shocks, individual characteristics such as job roles and hierarchies within the organizations, and the characteristics of each project such as the line of business and the region when the project was initiated. After eliminating these confounding factors, we graph the relationship between the time since the adoption of *SmallBlue* and

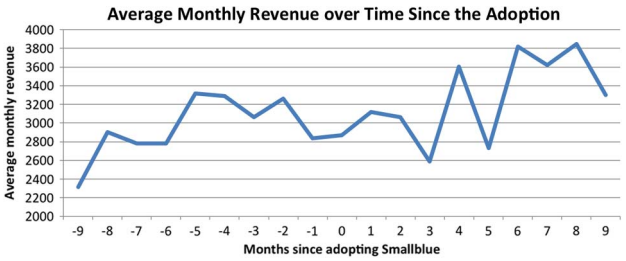


Fig. 4. ***Normalized average monthly revenue increases since the adoption of SmallBlue.***

the work performance of these consultants. The *X*-axis labels the number of months since a person has adopted *SmallBlue*; the zero value indicates a person has just adopted *SmallBlue*; negative values indicate the number of months before a person has adopted *SmallBlue*; and positive values indicate the number of months after a person has adopted *SmallBlue*. The *Y*-axis indicates the extra revenue generated in each month since the adoption. As indicated in the graph, the billable revenue of a person gradually increases after the person has adopted *SmallBlue*. Prior to the adoption, the coefficient estimates range between $2300 and $3300 in monthly billable revenue. After the adoption, the coefficient estimates for each month grow gradually from $2500 to $3600 in the first five months, and then increase to more than $3600 in the eighth month. It is important to note that this graph is a comparison within an individual over time, measuring on average how a person's billable revenue has increased over time since the adoption of *SmallBlue*.

These results show that networking tools can play a critical role in facilitating individuals to locate resources and expertise within the firm. On average, we saw an increase of $584.15 revenue per month (i.e., $5257 in nine months) for an adopter. As a result, their overall performance improved significantly over time, especially after a few months of adopting the tool. The lagged effect may come from the fact that it takes time for the newly found resources to translate into tangible performance metrics. One caveat is that we use the time since adoption as a way to proximate use of the tool. Obviously, it is possible that a person can sign up to use *SmallBlue* but never actually use it. In that case, we would expect *SmallBlue* to have no effect or a negligent effect on the overall performance for these people and the overall effect from using *SmallBlue* would be downwardly biased. We have been designing and conducting causality studies that will pinpoint the exact effect of each change and can test the reference groups. We acknowledge that no strong conclusion can be confirmed until such rigid social scientific experiments are completed, which can take up to several years. The fact that we found a positive relationship between performance and time since adoption underscores the importance of social media on work performance. Overall, these results indicate the importance of leveraging social media as an important corporate strategy to facilitate information-intensive work.

IV. CHANNEL, CULTURE, AND SOCIAL INFERENCE

Another aspect of the study is to understand how employees behave in terms of the channel, culture, and influence perspectives. In [50], Yang *et al.* showed the culture being the most significant factor in shaping perception and behavior via a survey of near 1000 people from four countries in an organization. For instance, Chinese and Indian users are more likely to use online social network tools for Q&A, in comparison to the U.S. and U.K. users. Culture is particularly important for a large organization that operates globally. Employees from different job roles, geography, nation, culture, gender, and educational backgrounds need to work together to achieve common goals. Understanding these issues is definitely a key to a successful global collaboration [51].

We first focus on computationally modeling people's interaction behavior in an organization as multilayer networks, shown in Fig. 5. These models allow us to investigate how people spread information to achieve their objectives, such as stronger ties with colleagues and higher productivity. Based on such understanding, *SmallBlue* can then build applications to recommend people/activities to

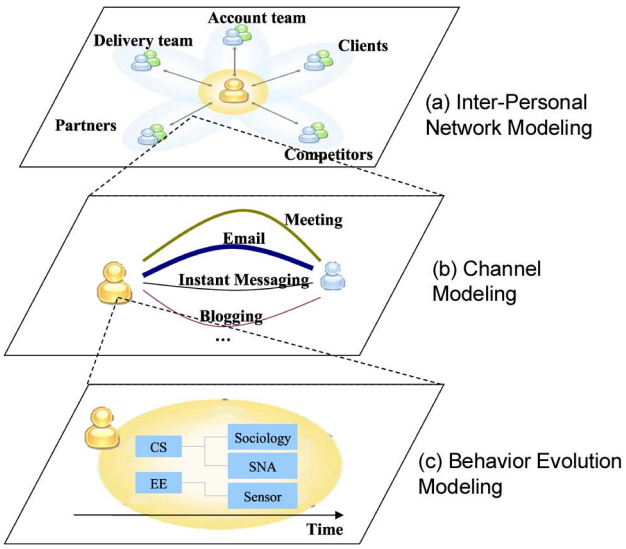


Fig. 5. ***Three layers of human behavior in networks. The granularity increases from the top to the bottom. (a) Complex people networks can be filtered to multiple personalized relationships (composed of individuals or communities) for different needs. (b) A relationship is built through multiple means of interaction between people, tailed to an individual capabilities. (c) An individual's capabilities can be modeled as evolving expertise networks.***

improve collaboration, or detect abnormal behavior to ensure security.

In complex people networks, an individual usually develops different types of relationships with different sets of people, and for various objectives. Moreover, to build such a relationship, the person may leverage multiple means of interaction, which we call *channels*, such as e-mail, meetings, friending activities, etc. Therefore, *SmallBlue* considers three layers of granularity of human behavior in networks. The coarsest layer considers multiple relationships involving multiple information sources from multiple friends as well as multiple types of social networks. The next level considers multichannel communication methods between two people. In the finest level, we consider a person's inner interest or expertise represented as network graphs (e.g., the ExpertiseNet model [35]).

Toward this goal, *SmallBlue* provides large-scale multifaceted data in which people's behavior can be observed in multiple different sources in enterprises. In contrast, on the Internet, it is very difficult to collect comparable multifaceted activity data of individuals. In particular, it is extremely challenging to match user ID across different sites, and most user activity logs have to be anonymized before they are processed. Therefore, *SmallBlue* offers a unique opportunity to study how a people network forms and changes in multifaceted activities. In this section, we focus on community and channel level modeling.

A. Channel Level Modeling

People usually use multiple channels to interact with each other. Their behavior can be different in different channels. *SmallBlue* models channel-specific behavior, as well as cross-channel behavior.

1) Channel-Specific Modeling: Here we describe one such model: the information spreading model in e-mail. More specifically, we focus on modeling structural properties of information spreading trees in which e-mails are forwarded by users to their contacts. We found, in contrast to the spreading trees observed in the Internet chain letter and virus diffusion, the trees in e-mail are brushy and ultra shallow. The information fans out, but quickly dies out. Previous models, such as Galton–Watson branching model [44], which can be used to model Internet chain letter with selection bias [14], cannot explain such ultra-shallow depth, or the stage dependency of the branching factor (i.e., the number of children each node has depends on the distance from the root). *SmallBlue* models the distribution of κ, a random variable of the number of each node's children. A good model can predict the observed tree structure (e.g., width and size), if the distribution, denoted as $P(\kappa)$, can generate a random tree structure with similar properties. Let us denote the stage-independent distribution of the number of recipients for all the e-mails being sent as $P'(\kappa)$. We model the stage-dependent branching factor distribution as

$$P(\kappa|d=0) = \mathfrak{A}(1-(1-p)^{\kappa})P'(\kappa)$$
$$P(\kappa|d>0) = P'\kappa)$$

where $\mathfrak{A}$ is the normalization factor, and p is the probability that a piece of information is forwarded. $P'(\kappa)$ and p can be independently measured in the experiments [43]. The correction term for $d=0$ means that the original e-mail with more recipients is more likely to get forwarded, as there will be more people to make a decision on whether to pass on the information. This model explains more than 2000 e-mail spreading trees in *SmallBlue* data very well, as illustrated in Fig. 6. This indicates that, despite the complexity in real life, the macroscopic structures of information spreading processes can be well captured and explained via simple mechanisms.

2) Cross-Channel Modeling: The way people choose different communication channels reveals their preferences, either intentionally by context or unintentionally by

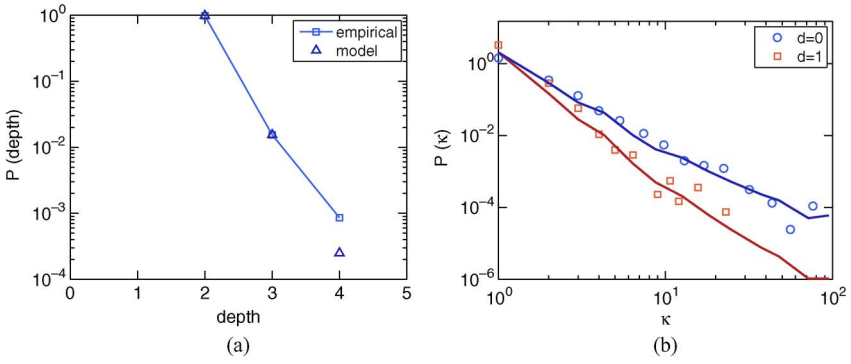


Fig. 6. *(a) Distributions of depth of the trees. (b) $P(\kappa)$ predicted by our model (solid lines) versus the experiment measure (scattered squares and circles).*

culture, education, etc. In a global enterprise, culture plays important roles. We investigate how culture factor affects people's preferences over the multiple channels [51]. To the best of our knowledge, this is the first large-scale empirical study of cultural differences in terms of social interaction channels in organizational context. We consider three channels: e-mail, IM, and calendar meetings (coordinated through the format of e-mail). For this analysis, we select eight countries with a sufficient number of users (e.g., > 200), including Japan (JP), United Kingdom (UK), the United States (US), Canada (CA), Germany (DE), Brazil (BR), India (IN), and China (CN). We also control confounding variables (e.g., job role and gender) to ensure that the results of each culture are statistically significant.

We measured the ratios of a user's contacts that she communicates using different combinations of the three channels. A contact is a colleague who she has communicated with through one of the channels. Fig. 7 shows the comparisons of the ratios for users in different countries. Countries such as China and India are more likely to use IM to communicate with their contacts, whereas Western countries have very few contacts through IM. Japan is an exception for Asian countries, as it is the least likely to use IM. In contrast, JP shows a preference for communicating through e-mail for most contacts. Comparing calendar meeting and IM, we see an opposite usage pattern: Western countries tend to adopt calendar meeting for more contacts. Since the preference might be affected by a person's job role, we examined this difference on a controlled group: business consultants, and we found that the patterns are consistent. The observed significant cultural differences echo the inherent cultural characteristics as suggested by cultural theories [42], and warrant the incorporation of the culture variable into people's behavior models.

In Fig. 8, we show the analysis on the sentiment of the words a person uses in her communications. In this study, we focus on English communications, since English is the primary language in which business is conducted by the company. We use an open source English dictionary of the positive and negative words and calculate the percentage of these words that are used in one's composed e-mails and IMs [51]. We can see people in most countries fall in a similar positive/negative ratio. English speaking countries are mostly more willing to express their feelings, while people in the United States are more often showing positive sentiments. We also observed a clear outliner that, in Germany, people are more willing to express negative sentiments.

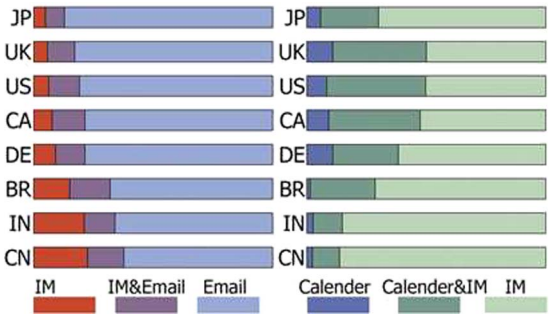


Fig. 7. *Relative preferences—ratio of contacts being allocated, between: IM versus e-mail, calendar meeting versus IM. The chart contrasts the proportion of contacts reached by two separate individual communication channels versus proportion reached by both channels together.*

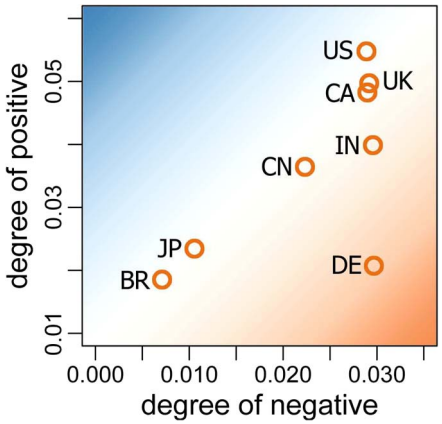


Fig. 8. *Sentiment analysis of the type of content by people in different countries.*

B. Community Level Modeling

Users typically interact with different sets of people for different purposes. In other words, they can participate in a diverse set of communities where interests/attributes in each community may be quite different. Users within a community usually share certain attributes, which makes it possible to infer one's interests from her social neighbors [27]. Such inference can help to improve personalized services and privacy control. It is, however, challenging to obtain consistently high-quality results in inferring user interests from social neighbors [45]. Even in the same community, especially a weak community, users may still have diverse attributes [27]. To address this challenge, we proposed an optimization-based method to exploit the correlation among multiple attributes of a user, in addition to the social correlation of an attribute among a group of users [46]. It first uses social correlation to obtain initial inference from neighbors. Then, it refines the inference using attribute correlation to remove less likely attribute combinations (e.g., "social software" and "VLSI").

Given an initial set of attributes $\mathbf{A}_0$ inferred based on social correlation for user i, the method tries to find the optimal subset $\mathbf{A} = \langle a_1, \ldots, a_n \rangle$ by balancing three constraints. First, the total degree of interests on attributes in $\mathbf{A}$ should be maximized. Next, the overall pairwise attribute correlation in $\mathbf{A}$ should be maximized. Finally, smaller size of $\mathbf{A}$ should be favored, if everything else is

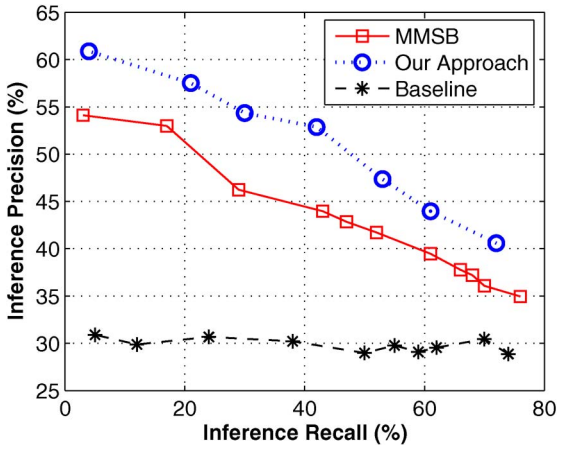


Fig. 9. *The performance comparison of our approach with MMSB [3] and the baseline method [45] in inferring explicit user interests.*

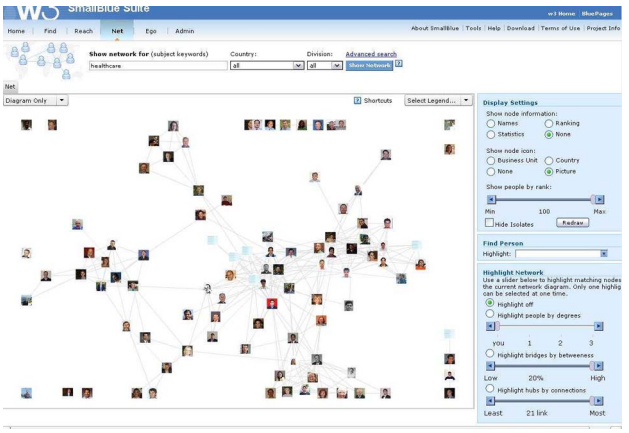


Fig. 10. *SmallBlue Net visualization showing the subject network and supporting various filtering functions.*

equal (i.e., Occam's razor principle). This optimization can be formulated as a generalized assignment problem, which is NP-hard [11]. Therefore, we use a greedy approach to approximating the optimization. Fig. 9 shows that our method significantly outperforms a previous community clustering method called mixed membership stochastic blockmodels (MMSBs) [3] and an inference method only using social correlation [45]. Currently, our method does not take into account the temporal evolution of attributes. We plan to incorporate this factor into our method so that it can address the changes of people's attributes over time.

V. VISUALIZATION

Network visualization is the most intuitive way to bring network analytic results to general users. There is no one-fit-all network visualization design, as each user may have a rather specified information need. A famous site VisualComplexity.com [2] has accumulated hundreds of network visualizations over diverse network data and scenario. For quite many cases, the standard node-link representations are employed; for others, the matrix visualization is also applied, mainly for networks with dense connections so as to increase the network readability [13]. We follow this general guideline and design network visualizations to meet users' two key requirements: 1) understanding the entire social network within the enterprise; and 2) understanding the personal network subject to the user, *a.k.a.* the user's ego network. As the *SmallBlue* networks are essentially large and dynamic, we also develop techniques to transform and interactively represent such huge and evolving networks.

A. Social Network Visualization

In *SmallBlue Net*, we depict the fundamental social network of people in enterprise in the traditional node-link form, as shown in Fig. 10. The major challenge here is how to deal with the potential network size up to half a million nodes. We take a straightforward approach: to filter the network into a smaller size, so that the network layouts can be computed in a reasonable time and the graph complexity can be controlled on a readable scale.

We introduce a two-stage filtering strategy for the visualization, including a first-stage in-disk query and a second-stage in-memory filtering. In the query stage, a search interface is provided, as shown in the upper side of Fig. 10, which includes the input for subject keywords and drop box selections for the country site and division of the company. More search terms such as the category of connections in the network can also be specified in the advanced search mode. Upon the search, only the network of people working on these topics and in this country/division is fetched from the graph database to the *SmallBlue Net* visualization engine. In the filtering stage, the nodes in the queried network are further ranked in memory, typically by a kind of centrality. The user could manipulate the rank slider in the right panel of Fig. 10 to show more or less network information. Highlights of part of the network are also supported with the interaction toolbox given in the right panel of Fig. 10. A group of highlighted people can be selected according to their betweenness centrality, degree centrality, the degrees to the user querying the network, and also directly through a name search.

B. Ego Network Visualization

In *SmallBlue Ego*, the personal social capital management tool, we present a dynamic ego network view to help the user access his connectivity to the collaborators, their profiles, and his ego network evolvements, hence to increase the user's social situational awareness in the enterprise. In its static mode, as shown in Fig. 11, the ego network visualization is designed as a disk in the background divided into multiple slices indicating different

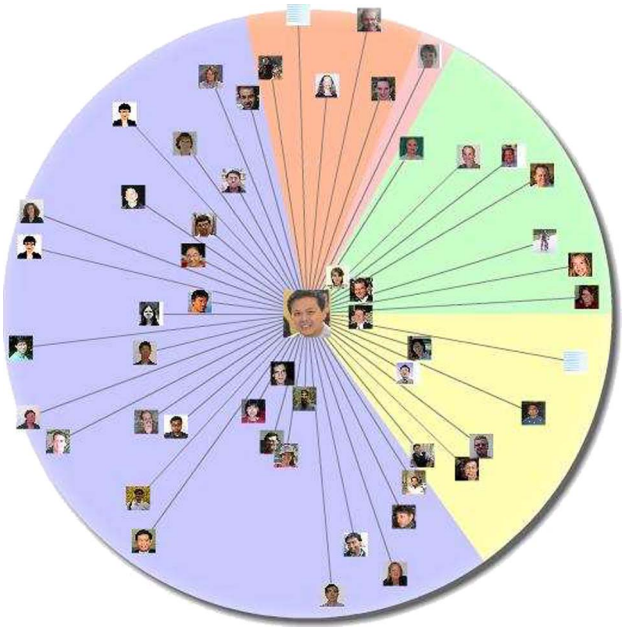


Fig. 11. ***SmallBlue Ego visualization for social dynamics of the SmallBlue user.***

countries or divisions. The current user is placed to the center of the disk with his direct collaborators scattered out in slices according to their country/division attributes. The closer the collaborators are placed to the center, the stronger they connect to the current user.

The *SmallBlue Ego* visualization supports a movie mode [1]. Upon the user's click on the "play" button, the user's ego network in each time period is shown frame by frame. Staged animations between each of the two consecutive frames are added. In the first stage, the disappearing collaborators in the current time frame fade out with animations, then in the second stage, the remaining collaborators move to the updated locations from the previous time frame, and finally in the last stage, the new collaborators emerging in the current time will appear and move to their locations from the center. These animations work to smooth out the display between frames and help to keep the user's visual momentum.

C. Huge Graph Visualization

Although *SmallBlue Net* is shown to be useful in presenting topic-centric network information, there is still a need to understand the whole picture of the enterprise social network, the communities within it, and the interconnections and structural holes of the network. The simplex methods to filter the huge graph into readable size bring the side effect of losing the overall topology of the network, and more importantly prohibit the access to network details, which could be critical in the user's navigation tasks. In *SmallBlue*, we introduce a novel technique called *HiMap* [33] to more effectively visualize huge graphs up to millions of nodes. The design goals of *HiMap* are fourfold. 1) Each graph view of the network should be visualized adaptively in a readable manner, easy to be comprehended, independent with its scale, topology, and the screen size to display. 2) A suite of navigation methods should be provided so that it is capable to visualize and diagnose every detail of the network. 3) Smooth animations should be presented between any view changes, so as to keep the user's momentum [28]. 4) The visualization system should run fast enough and keep lightweight to catch up with the animation speed.

The basic idea is to combine the well-studied hierarchical graph clustering algorithm with the interactive visualization to present the user with a top view of the network and also allow the user to freely traverse the graph hierarchy to access the details, quite coherent with the information visualization mantra: "*Overview first, zoom and filter, then details-on-demand.*" It is divided into two major stages: the offline data manipulation and the online adaptive visualization. The core to the offline data manipulation part is the hierarchical graph clustering module. Initially, the graph is clustered with the method in [29] into a binary tree. To obtain balanced hierarchical clustering structure, we invoke the algorithm recursively until the predefined maximal tree depth is reached. It is well known that the online social network possesses a highly clustered and self-similar community structure. The clustered graph, which could reveal their built-in hierarchy information, is one of the best ways to visualize it, let alone providing semantic abstraction that makes the readable visualization possible.

A snapshot visualization is given in Fig. 12. The clusters are drawn as circles without an explicit frame, and background color for each cluster is painted from the center in a descending lightness along the radius to indicate its boundary. The clusters not capable of showing their internal structure due to the screen constraints are drawn as a much smaller circle without any subclusters (nodes) in it. When the subcluster only represents one leaf node, it is shown as a people icon instead. The background color of the cluster is set according to its depth in the entire tree structure, rendering darker for the deeper depth. By default, the edges between any two leaf nodes are drawn in the view by the straight line. To reduce the visual clutter commonly found in densely connected graphs, we also introduce two edge bundling methods: the geometric edge bundling and the hierarchical edge bundling. The geometric bundling implements the solution similar to [8]. It works by carefully selecting some control points in the graph and forcing all the edges to traverse them. The other method bundles all the edges between any two upper hierarchy cluster together and only shows the intraedges inside each cluster for the lower hierarchy subclusters. We have implemented three kinds of zooming operations: the hierarchical zooming that navigates through different hierarchies, the semantic zooming that focuses the view on a

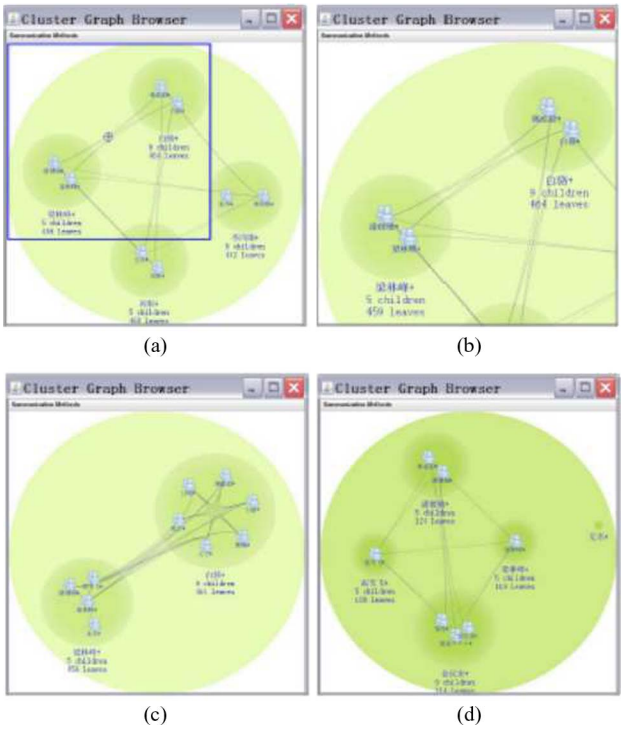


Fig. 12. ***SmallBlue huge graph visualization. Zooming operation: (a) initial view, the blue frame shows the virtual zoom-in window; (b) the view after geometric zoom-in; the items are signified; (c) the view after semantic zoom-in, more items are visualized adaptively; and (d) the view after drilling in the cluster in the left of (a).***

smaller/larger portion of the previous graph and adaptively reload graph data, and the traditional geometric zooming operation. To maintain visual momentum, we have also designed customized animations for each interaction upon view changes.

We also used several network visualization functions, such as the information flow modeling in complex networks [21], in which we proposed to model user node transitions as susceptible–active–informed (SAI) states and edge transitions as a Markov model with susceptible–dormant–active–removed (SDAR) stages. We can then predict information flows in a social network. Demos of network visualizations can be found in [1].

VI. NETWORK GRAPH MINING

In enterprise, relationships of entities can be mined from various data sources and form networks of millions or billions of nodes and edges. Networks of people can reach millions, if all internal and external contacts are includes. When information content represents the node of a graph, then it is very easy to achieve graphs of billions of nodes and edges. In a newer version of *SmallBlue*, we addressed the scalability issue mainly through two complementary efforts: 1) *system support*—we built a general and scalable graph mining foundation that can support a variety of core operations on large graphs, based on Hadoop, stream processing system, and/or data warehouses; and 2) *algorithms design*—we designed scalable algorithms specific applications, such as anomaly detection and diversity enhancement.

A. GBase: Graph Database for Hadoop Framework

Numerous applications (e.g., neighborhood search, PageRank, subgraphs, proximity, etc.) are common to network graph analysts. Our goal is to develop a general and scalable graph mining framework for *SmallBlue* to support a variety of common core operations on large graphs. The design objective is threefold: 1) efficiently store and manage huge graphs in parallel, distributed settings to answer graph queries efficiently; 2) define common, core algorithms to satisfy various graph applications; and 3) exploit the efficient storage and general algorithms to execute queries efficiently.

Large graphs cannot be fit in main memory or at least the disk of a single workstation, on which most of existing graph algorithms have been built. Thus, we need to rethink those algorithms, and to develop scalable, parallel ones, to manage graphs that span terabytes and beyond. Moreover, these methods have to be scalable with respect to indexing time, storage cost, as well as online query time.

The second challenge lies in the *application heterogeneity*. Different graph applications require different kinds of inputs as well as (seemingly) different types of operations on the graphs. As a result, most, if not all, of the existing graph indexing techniques have to restrict themselves to a particular type of applications. Here, the goal is to find a set of popular, "core" graph operations, that most applications require.

We designed GBase [17] to address the above challenges. As shown in Fig. 13, first, we propose to index large graphs on homogeneous block levels. By exploring this community-like property that exists in many real graphs, we can largely reduce the storage cost. In addition, by organizing the graph in such a blockwise structure, it also helps with online query response. Second, to handle the challenge of application heterogeneity, we proposed a unified query execution engine which unifies the different types of queries on graphs by generalized matrix–vector multiplications. There are two key advantages of the proposed graph management system: 1) it is *scalable* in the sense that it is linear in indexing and querying time; and 2) it is *general* in the sense that it can support a broad range of queries, spanning from node-level queries to the graph-level query, with the community-level query in the middle.

Table 3 summarizes the queries (the first column) that are supported by our graph mining system. These queries construct the main building blocks for a variety of important graph applications (Table 3). For example, the PageRank [30] query provides a natural ranking function

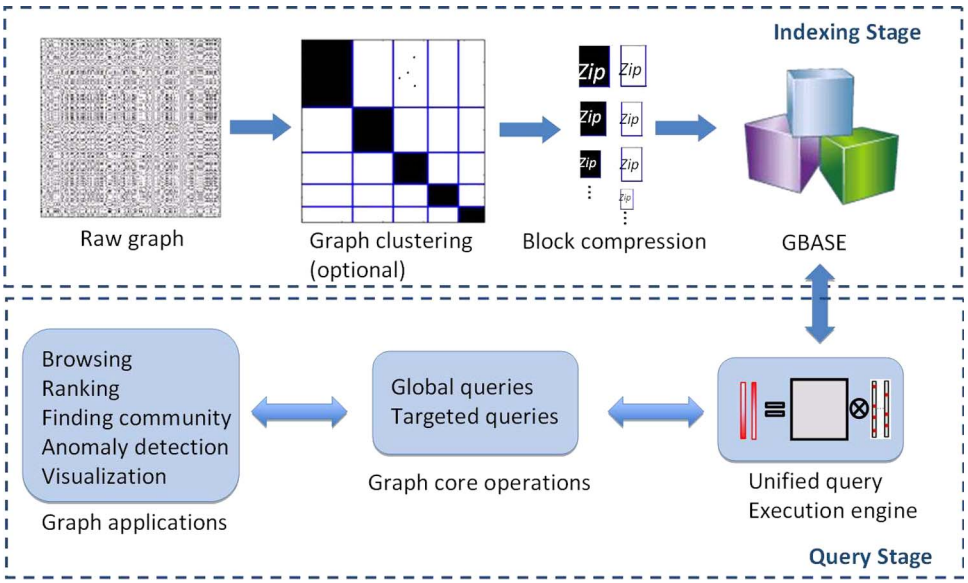


Fig. 13. *Overall framework of GBase. 1) Indexing Stage: Raw graph is clustered and divided into compressed blocks. 2) Query Stage: Global and targeted queries from various graph applications are handled by a unified query execution engine.*

to find important nodes on graphs. The diversity of random walk with restart (RWR) [41] scores among the neighborhood of a given edge/node is a strong indicator of abnormality of that node/edge [38]. The ratio between the number of edges (or the summation of edge weights) and the number of nodes within the egonet can help find abnormal nodes on weighted graphs [4]. The K-cores and cross edges can be used for visualization and finding communities in large graphs [5].

B. Scalable Algorithms Design: Case Studies

Orthogonal to the general graph mining system, we also designed application-specific scalable algorithms. Here, we present two case studies: 1) nonnegative residual matrix factorization for interpretable graph anomaly detection [40]; and 2) diversified ranking on large graphs [39].

Table 3 Applications of Our Graph Mining System. Notice That it Can Support a Wide Range of Both Global (Top Three Rows) and Targeted (Bottom Five Rows With Bold Fonts) Queries With Applications, for Example, in Browsing [22], [30], [41], Ranking [30], [41], Finding Communities [19], [22], Anomaly Detection [4], [18], [19], [38], and Visualization [5], [22]

Query \ Applications	Browsing	Ranking	Finding Community	Anomaly Detection	Visualization
Connected Comp.			✓	✓	
Radius				✓	✓
PageRank, RWR	✓	✓		✓	
Induced Subgraph	✓		✓		✓
(K)-Neighborhood	✓		✓		✓
(K)-Egonet	✓		✓	✓	✓
K-core			✓		✓
Cross-edges				✓	✓

1) Nonnegative Residual Matrix Factorization: Matrix factorization (i.e., to decompose the adjacency matrix of the graph by the multiplication between two low-rank matrices plus a residual matrix) is powerful to find graph patterns. For instance, the two low-rank matrices often capture the community structure of the graph; and the residual matrix is often a good indicator for anomalies on graphs. A new application of *SmallBlue* aims at finding anomalies in enterprise. In this scenario, matrix factorization is important.

Among others, it is now widely recognized that nonnegativity is a highly desirable property for interpretation since negative values are usually hard to interpret. For example, for the task of community detection, the so-called nonnegative matrix factorization often leads to part-based, or subcommunity-based decomposition [15], [20]. However, most, if not all, of these methods impose the nonnegativity constrain on the two *low-rank matrices*. Consequently, these existing methods are tailored for the task of community detection. It is not clear how to improve the interpretation for the task of anomaly detection from the algorithmic aspect. Can we impose similar constraints (e.g., nonnegativity) on the *residual matrix* to improve the interpretation for graph anomaly detection?

In response to such challenges, we proposed a new matrix factorization in [40]. While the exact/global optimal solution is hard to obtain due to the nonconvexity of the problem, we proposed an efficient approximate

algorithm. The core idea is to recursively find a rank-1 approximation for the current residual matrix; in the meanwhile, it requires that the new residual matrix is always nonnegative. One major advantage of this method is the scalability—it enjoys linear complexity with respect to the size of the graph in both time and space cost. More specifically, our method solves the following optimization problem:

$$\begin{aligned} \arg\min_{\mathbf{F},\mathbf{G}} \quad & \sum_{i,j,\mathbf{A}(i,j)>0} (\mathbf{A}(i,j) - \mathbf{F}(i,:)\mathbf{G}(:,j))^2 \\ \text{s.t.} \quad & \text{for all } \mathbf{A}(i,j) > 0 : \\ & \mathbf{F}(i,:)\mathbf{G}(:,j) \leq \mathbf{A}(i,j) \end{aligned}$$

where $\mathbf{F}$ and $\mathbf{G}$ are the two low-rank matrices, and $\mathbf{A}$ is the adjacency matrix for the graph.

2) Diversified Ranking on Networks: Diversified ranking on graphs is a key factor to address the uncertainty and ambiguity in an information need, and to cover the different aspects of the information need [31]. A recent study shows that diversity is also positively associated with personnel performance and job retention rate in a large organization [47].

Two important questions remain open. The first challenge is the *measure*: For a given top-k ranking list, how can we quantify its goodness? Intuitively, a good top-k ranking list should capture both the relevance and the diversity. In an enterprise setting, given a task which typically requires a set of different skills, if we want to form a team of experts, not only should the people in the team have relevant skills, but also they should somehow be "different" from each other so that the whole team can benefit from the diversified, complementary knowledge and social capital. However, there does not exist such a goodness measure for the graph data in the literature. The second challenge lies in the *algorithmic* aspect: How can we find an optimal, or near-optimal, top-k ranking list that maximizes the goodness measure?

Bringing diversity into the design objective implies that we need to optimize on the set level. In other words, the objective function for a subset of nodes is usually not equal to the sum of objective functions of each individual node. It is usually very hard to perform such set-level optimization. For instance, a straightforward method would need exponential enumerations to find the exact optimal solution, which is infeasible even for medium size graphs. This, together with the fact that real graphs are often of large size, reaching billions of nodes and edges, poses the challenge for the optimization algorithm: How can we find a near-optimal solution in a scalable way?

We addressed these challenges from an optimization point of view in [39] by a new goodness measure

$$\mathrm{f}(\mathcal{S}) = 2\sum_{i\in\mathcal{S}} \mathbf{r}(i) - \sum_{i,j\in\mathcal{S}} \mathbf{B}(i,j)\mathbf{r}(j)$$

where $\mathcal{S}$ is the subset of the nodes in the ranking list; $\mathbf{r}(i)$ is the ranking score for the node i; and $\mathbf{B}$ is the *personalized* adjacency matrix for the query. It intuitively captures both 1) the relevance between each individual node in the ranking list and the query node; and 2) the diversity among different nodes in the ranking list. The core idea of our algorithm is to explore the so-called submodular (i.e., diminishing return) property of the goodness measure.

In this section, we have introduced the use of graph mining algorithms for enterprise uses such as browsing, ranking, community finding, security, and visualization. Specifically, we introduced two use cases that show more details in anomaly detection and diversified ranking that are especially useful for cybersecurity and personalized search and recommendation.

VII. CONCLUSION

We have discussed various challenges and solutions for conducting SNA in enterprise. We considered multimodality aspects of people relationships, including social aspect, financial aspect, and human property aspect. We also discussed various system challenges such as large-scale graph mining and large-scale network visualization. This paper focused on the fundamental research and system issues. It did not discuss the various applications of enterprise SNA such as collaboration (e.g., enterprise location, social proximity access, social recommendation, social search, etc.), cybsersecurity (e.g., anomaly detection, fraud detection, etc.), and commerce (e.g., social marketing and selling).

On the scientific aspect, there are many unsolved issues. For instance, despite having collected the largest enterprise data set in literature for employee interactions, we still have not obtained the teleconference data (although it can be approximated by the calendar info) and the face-to-face interaction data [49] in a large setting. The accuracy of inferred social networks obtained in this study is much better than that from internal social media which is still in the early stage of wide adoption in enterprise. Our preliminary studies show that the use of e-mail and IM is so intense that even people who meet face-to-face or on teleconferences spend comparable time in communicating with each other by e-mail and IM.

One other challenge is to understand the economic impact of social networks on people other than consultants. Due to the sensitivity of data, we could only obtain

the revenue generated by about 200 000 employees with external consultant revenue. We did not get the detailed performance impact of sales, hardware, and software developers. Using a relative small set of 532 salespersons, we observed different impact of social networks. For sales, long-term strong ties between networks demonstrate to be more important than the diversity of networks. This is probably due to the nature of "trust" needed in sales. We will continue pursuing this thread if we received more comprehensive data.

Causality study on the impact of the "control of network" is one of our current research threads. It involves the strategies in influencing people to change their networks and then observing their real-world revenue impacts. This type of researches requires several years of studies. We will report on that in the future.

Based on the knowledge we gained from studying multimodality data for people multinetworks, we will extend our research to the study of the difference on these networks and the online social media networks. We will also report on their differences in the future. Network visualization, especially visualizing it in large-scale and evolutionary manner, is a challenge. Studies on how to associate network visualization with browsing and machine learning tasks are emerging. With the rapid growth of network data, how to read, write, store, and query dynamic and large-scale data in large-scale and/or stream environments are also very challenging issues. ■

REFERENCES

[1] Smallblue Network Visualization Demo Videos. [Online]. Available: http://smallblue.research.ibm.com/demos/

[2] Visual Exploration on Mapping Complex Networks. [Online]. Available: http://www.visualcomplexity.com/vc/

[3] E. Airoldi, D. Blei, S. Fienberg, and E. Xing, "Mixed membership stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, 2008.

[4] L. Akoglu, M. McGlohon, and C. Faloutsos, "Oddball: Spotting anomalies in weighted graphs," in *Proc. Pacific-Asia Conf. Knowl. Disc. Data Mining*, 2010, pp. 410–421.

[5] I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, *k-core decompositions: A tool for the visualization of large scale networks*. [Online]. Available: http://arxiv.org/abs/cs.NI/0504107

[6] R. Burt, *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard Univ. Press, 1992.

[7] R. Burt, "Structural holes and good ideas," *Amer. J. Sociol.*, vol. 110, no. 2, pp. 349–399, 2004.

[8] W. Cui, H. Zhou, H. Qu, P. C. Wong, and X. Li, "Geometry-based edge clustering for graph visualization," *IEEE Trans. Inf. Visual. Comput. Graph.*, vol. 14, no. 6, pp. 1277–1284, Nov. 2008.

[9] R. I. M. Dunbar, "Neocortex size as a constraint on group size in primates," *J. Human Evol.*, vol. 22, no. 6, pp. 469–493, 1992.

[10] K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher, "Searching for experts in the enterprise: Combining text and social network analysis," in *Proc. ACM Conf. Supporting Group Work*, 2007, pp. 117–126.

[11] L. Fleischer, M. Goemans, V. Mirrokni, and M. Sviridenko, "Tight approximation algorithms for maximum general assignment problems," in *Proc. ACM/SIAM Symp. Discrete Algorithm*, 2006, pp. 611–620.

[12] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, D. Schlichting, and A. Toncheva, "The diverse and exploding digital universe," White Paper. International Data Corporation, Framingham, MA, Mar. 2008. [Online]. Available: www.emc.com/collateral/analyst-reports/diverse-exploding-digitaluniverse.pdf

[13] M. Ghoniem, J.-D. Fekete, and P. Castagliola, "A comparison of the readability of graphs using node-link and matrix-based representations," in *Proc. InfoVis*, 2004, pp. 17–24.

[14] B. Golub and M. O. Jackson, "Using selection bias to explain the observed structure of Internet diffusions," *Proc. Nat. Acad. Sci.*, vol. 107, no. 24, p. 10833, 2010.

[15] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[16] Information Commissioner's Office (ICO), Improving user interest inference from social neighbors, U.K., 2012.

[17] U. Kang, H. Tong, J. Sun, C.-Y. Lin, and C. Faloutsos, "GBASE: A scalable and general graph management system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2011, pp. 1091–1099.

[18] U. Kang, C. E. Tsourakakis, A. P. Appel, C. Faloutsos, and J. Leskovec, "Radius plots for mining tera-byte scale graphs: Algorithms, patterns, and observations," in *Proc. SIAM Int. Conf. Data Mining*, 2010, pp. 548–558.

[19] U. Kang, C. E. Tsourakakis, and C. Faloutsos, "PEGASUS: A peta-scale graph mining system implementation and observations," in *Proc. Int. Conf. Data Mining*, 2009, pp. 229–238.

[20] D. D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Ing. Process. Syst.*, 2000, pp. 556–562.

[21] C.-Y. Lin, "Information flow prediction by modeling dynamic probabilistic social network," in *Proc. Int. Conf. Netw. Sci.*, May 2007.

[22] C.-Y. Lin, N. Cao, S. Liu, S. Papadimitriou, J. Sun, and X. Yan, "Smallblue: Social network analysis for expertise search and collective intelligence," in *Proc. Int. Conf. Data Eng.*, 2009, pp. 1483–1486.

[23] C.-Y. Lin, K. Ehrlich, V. Griffiths-Fisher, and C. Desforges, "Smallblue: People mining for expertise search," *IEEE Multimedia Mag.*, vol. 15, no. 1, pp. 78–84, Jan.–Mar. 2008.

[24] P. Marsden, "Network data and measurement," *Annu. Rev. Sociol.*, vol. 16, pp. 435–463, 2009.

[25] B. McEvily and A. Zaheer, "Bridging ties: A source of firm heterogeneity in competitive capabilities," *Strategic Manage. J.*, vol. 20, no. 4, pp. 1133–1156, 1999.

[26] D. Millen, J. Feinberg, and B. Kerr, "Dogear: Social bookmarking in the enterprise," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2006, pp. 111–120.

[27] A. Mislove, B. Viswanath, P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in *Proc. Web Search Data Mining*, 2010, pp. 251–260.

[28] K. Misue, P. Eades, W. Lai, and K. Sugiyama, "Layout adjustment and the mental map," *J. Vis. Lang. Comput.*, vol. 6, no. 2, pp. 183–210, Jun. 1995.

[29] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, 066133, 2004.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford Univ., Stanford, CA, Stanford Digital Library Technologies Project, 1998.

[31] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims, "Redundancy, diversity and interdependent document relevance," *SIGIR Forum*, vol. 43, no. 2, pp. 46–52, 2009.

[32] R. Reagans and E. Zuckerman, "Networks, diversity, and productivity: The social capital of corporate r and d teams," *Organizat. Sci.*, vol. 12, no. 4, pp. 502–262, 2001.

[33] L. Shi, N. Cao, S. Liu, W. Qian, L. Tan, G. Wang, J. Sun, and C.-Y. Lin, "HiMap: Adaptive visualization of large-scale online social networks," in *Proc. Pacific Vis. Symp.*, 2009, pp. 41–48.

[34] J.-R. Shieh, C.-Y. Lin, and J.-L. Wu, "Recommendation in the end-to-end encrypted domain," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 915–924.

[35] X. Song, B. Tseng, C. Lin, and M.-T. Sun, "Expertisenet: Relational and evolutionary expert modeling," in *Proc. Int. Conf. User Model.*, 2005, pp. 99–108.

[36] J. Stanton and K. Stam, *The Visible Employee*. New York: CyberAge Books, 2006.

[37] T. Stuart and J. Podolny, "Positional causes and correlates of strategic alliances in the semiconductor industry," *Res. Sociol. Organizat.*, vol. 16, pp. 161–182, 1999.

[38] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Neighborhood formation and anomaly detection in bipartite graphs," in *Proc. Int. Conf. Data Manage.*, 2005, pp. 418–425.

[39] H. Tong, J. He, Z. Wen, and C.-Y. Lin, "Diversified ranking on large graphs: An optimization viewpoint," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2011, pp. 1028–1036.

[40] H. Tong and C.-Y. Lin, "Non-negative residual matrix factorization with application to graph anomaly detection," in *Proc. SIAM Int. Conf. Data Mining*, 2011, pp. 143–153.

[41] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Proc. Int. Conf. Data Manage.*, 2006, pp. 613–622.

[42] M. Varnum, I. Grossmann, S. Kitayama, and R. Nisbett, "The origin of cultural differences in cognition: The social orientation hypothesis," *Current Directions Psychol. Sci.*, vol. 10, no. 1, pp. 9–13, 2010.

[43] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabási, "Information spreading in context," in *Proc. Int. Conf. World Wide Web*, 2011, pp. 735–744.

[44] H. W. Watson and F. Galton, "On the probability of the extinction of families," *J. Anthropol. Inst. Great Britain Ireland*, pp. 138–144, 1875.

[45] Z. Wen and C. Lin, "On the quality of inferring interests from social neighbors," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2010, pp. 373–382.

[46] Z. Wen and C.-Y. Lin, "Improving user interest inference from social neighbors," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 1001–1006.

[47] L. Wu, "Social network effects on performance and layoffs: Evidence from the adoption of a social networking tool," Job Market Paper, 2011.

[48] L. Wu, C.-Y. Lin, S. Aral, and E. Brynjolfsson, "Value of social network—A large-scale analysis on network structure impact to financial revenues of information technology consultants," presented at the Winter Inf. Syst. Conf., Salt Lake City, UT, 2009.

[49] L. Wu, B. Waber, S. Aral, E. Brynjolfsson, and A. Pentland, "Mining face-to-face interaction networks using sociometric badges: Evidence predicting productivity in it configuration," presented at the Int. Conf. Inf. Syst., 2008.

[50] J. Yang, M. R. Morris, J. Teevan, L. Adamic, and M. S. Ackerman, "Culture matters: A survey study of social q&a behavior," presented at the Int. Conf. Weblogs Social Media, 2011.

[51] J. Yang, Z. Wen, L. A. Adamic, M. S. Ackerman, and C.-Y. Lin, "Collaborating globally: Culture and organizational computer-mediated communications," presented at the Int. Conf. Inf. Syst., 2011.

ABOUT THE AUTHORS

Ching-Yung Lin (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1991 and 1993, respectively, and the Ph.D. degree in electrical engineering from Columbia University, New York, NY, in 2000.

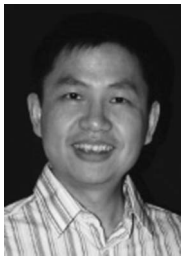
He joined IBM T. J. Watson Research Center, Hawthorne, NY, in 2000, where he is the Lead and Principal Investigator of Social Network Analytics Research. He was an Affiliate Assistant/Associate Professor at the University of Washington, Seattle (2003–2009), and has been an Adjunct Associate/Full Professor at Columbia University, New York, NY, since 2005. His research interest mainly focuses on multimodality signal analysis, complex network analysis, and computational social and cognitive sciences. He leads several large research projects, including more than 35 Ph.D. researchers in IBM Research and ten U.S. universities, to advance fundamental research of network science and people analytics, as well as applied researches on collaboration, security, and commerce. The research goals are to 1) explore and investigate scientific challenges on large-scale network graph processing; 2) quantify value of networks; and 3) understand multichannel behaviors of people, from cognitive level to societal level. He is an author or coauthor of more than 150 research papers. He initiated the first large-scale video semantic annotation task including 23 global institutes and 111 researchers in 2003. His invention SmallBlue (IBM Atlas) has been featured in more than 120 press articles, including appearing four times in *BusinessWeek* magazine and being the Top Story of the Week in April 2009.

Dr. Lin is a recipient of the 2011 Association of Information System ICIS Best Theme Paper Award, the 2003 IEEE CAS Society Young Author Award, the 2011 IBM Corporation Outstanding Innovation Award, the 2005 IBM Research Division Award, and the IBM Invention Achievement Award in 2001, 2003, 2007, 2010, and 2011. In 2010, IBM Exploratory Research Career Review selected him as one of the researchers "most likely to have greatest scientific impact for IBM and the world." He was the Editor of the Interactive Magazines (EIM) of the IEEE Communications Society (2004–2006), and a Guest Editor of the PROCEEDINGS OF THE IEEE Special Issue on Digital Rights Management (2004), the *EURASIP Journal on Applied Digital Signal Processing* Special Issue on Visual Sensor Network (2006), the IEEE TRANSACTIONS ON MULTIMEDIA Special Issue on Communities and Media Computing (2009), the IEEE JOURNAL ON SELECTED AREA IN COMMUNICATIONS Special Issue on Network Science (2013), and the *Journal of Multimedia* Special Issue on Social Multimedia Computing (2013). He was the Chair of the IEEE International Conference on Multimedia and Expo 2009 and the Chair of Circuits and Systems Society Multimedia Technical Committee (2010–2011), and is the Chair of the Steering Committee of ACM SIG Health Informatics (IHI) (2009–2012). He is a member of the Academy of Management.

Lynn Wu received the B.S. and M.S. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002 and 2003, respectively, and the Ph.D. degree in management science from MIT Sloan School of Management in 2011. She also received the B.S. degree in finance, minor in economics, from MIT, in 2002.

She has been an Assistant Professor at the Wharton School, University of Pennsylvania, Philadelphia, since 2011. She spent four years working in the MIT Computer Science and Artificial Intelligence Laboratory. She worked in IBM Silicon Valleny Lab as a Research Engineer (2003–2005) and has been with IBM T. J. Watson Research Center, Hawthorne, NY, as a Research Affiliate since 2008. She is interested in studying how information and information technology impact the productivity of information workers, organizations, and broad sectors of economy. Specifically, her work follows two streams. In the first stream, she studies how social networks and information derived from social networks affect individual's performance and long-term career trajectories. In her second stream of research, she examines the role of investment in IT and complementary organizational practices to explain how firms can achieve greater business value from IT. She has published articles in economics, management, and computer science. Her work has been featured by the *Wall Street Journal*, *BusinessWeek*, and *The Economist*.

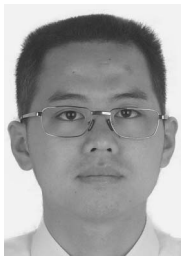
Dr. Wu was the winner of the 2008 and 2010 Google Research Awards, the 2008 HP Labs Innovation Research Award, and the 2008 Best Paper Award of the Association of Information System ICIS.



Zhen Wen (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign, Urbana, in 2004.

He is currently a Research Staff Member at IBM T. J. Watson Research Center, Hawthorne, NY. He is a Co-Principal Investigator of the Social Network Analytics Research in IBM T. J. Watson Research Center. He leads both basic and applied research efforts on modeling human behavior in social networks, which are sponsored by various funding agencies. His past research at IBM includes context-sensitive visualization for visual analytics. Specifically, he has worked on generating visualization that is appropriate for user analytic tasks using contextual cues. His work has been used in a U.S. Department of Homeland Security project on monitoring and analyzing shipment through U.S. Customs, as well as in a project on business intelligence (e.g., IBM Cognos). He has broad interests in data mining, signal processing, and human–computer interaction with applications on social network analysis and multimedia analysis.

Dr. Wen has served as an organizing committee member and a technical committee member at various IEEE/ACM conferences. He is an Area Chair of Social Media at the 2012 ACM Conference on Multimedia. He received the 2011 Association of Information System ICIS Best Theme Paper Award, the 2005 Best Paper Award at the ACM Conference on Intelligent User Interfaces (IUI), the 2005 IBM Research division award, and the 2007, 2010, and 2011 IBM invention achievement awards.



Hanghang Tong received the B.E. degree in automation technology and the M.E. degree in pattern recognition and intelligent systems from Tsinghua University, Beijing, China, in 2002 and 2005, respectively, and the M.Sc. and Ph.D. degrees in machine learning from Carnegie Mellon University, Pittsburgh, PA, in 2008 and 2010, respectively.

He has been a Research Staff Member at IBM T. J. Watson Research Center, Hawthorne, NY, since 2010. Before that, he was a Postdoctoral Fellow at Carnegie Mellon University. His research interest is in large-scale data mining for graphs and multimedia.

Dr. Tong has received several awards, including the best research paper at the 2006 IEEE International Conference on Data Mining (ICDM) and the best paper award at the 2008 SIAM International Conference on Data Mining (SDM). He has published over 40 refereed articles and served as a program committee member of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), and the International Conference on World Wide Web (WWW). He was a Section Editor in Social Network Applications in Homeland Security in *Encyclopedia of Social Network Analysis and Mining* (2012), and a Guest Editor in *Data Mining and Knolwedge Discovery*'s Special issue on Data Mining Technologies for Computational Social Science in 2011.

Vicky Griffiths-Fisher received the B.A. degree with honors in earth sciences from Oxford University, Oxford, U.K., in 1993.

She is the Privacy Officer of IBM United Kingdom and Ireland, London, U.K., responsible for compliance issues and advisory work around privacy and associated impact on new technologies/business processes. Her advisory work also includes data privacy and analytics, especially on enterprise social computing technologies, for the worldwide IBM corporation. Her areas of expertise include: privacy and data protection compliance, privacy impact assessments, European Union Data Protection and related Legislation, privacy by design, requirements analysis, application design, information architecture, and user usability design. She was a Project Manager in IBM Global Business Services (GBS) Learning and Knowledge, and led the internal strategic impact effort of SmallBlue. She joined PwC as an IT Consultant in 1995, and became the Head of e-learning Design and Production team in PwC Consulting in 2002.

Lei Shi (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2003, 2006, and 2008, respectively.

He is currently an Associate Research Professor at the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science, Beijing, China. Previously, he was a Research Staff Member and Research Manager at IBM Research–China from 2008 to 2012, working on information visualization and visual analytics. His research interests span information visualization, visual analytics, network science, networked systems, and human–computer interaction. He has published over 40 papers in refereed conferences and journals.

Prof. Shi is the recipient of the IBM Research Accomplishment Award on "Visual Analytics" and the 2010 VAST Challenge Award.

David Lubensky received the B.S. degree in computer science and the M.S. degree in electrical engineering from Drexel University, Philadelphia, PA, in 1984 and 1987, respectively.

He is a Senior Manager of Collaborative Technologies and Analytics at IBM T. J. Watson Research Center, Hawthorne, NY. He currently leads the development of innovative technologies and solutions in the area of mobile Internet, big data analytics, and e-commerce. Prior to IBM, he worked at Verizon Science and Technology and Siemens Research.