Social Network Analysis and Mining for Business Applications

FRANCESCO BONCHI, CARLOS CASTILLO, ARISTIDES GIONIS,
and ALEJANDRO JAIMES, Yahoo! Research Barcelona

Social network analysis has gained significant attention in recent years, largely due to the success of online social networking and media-sharing sites, and the consequent availability of a wealth of social network data. In spite of the growing interest, however, there is little understanding of the potential business applications of mining social networks. While there is a large body of research on different problems and methods for social network mining, there is a gap between the techniques developed by the research community and their deployment in real-world applications. Therefore the potential business impact of these techniques is still largely unexplored.

In this article we use a business process classification framework to put the research topics in a business context and provide an overview of what we consider key problems and techniques in social network analysis and mining from the perspective of business applications. In particular, we discuss data acquisition and preparation, trust, expertise, community structure, network dynamics, and information propagation. In each case we present a brief overview of the problem, describe state-of-the art approaches, discuss business application examples, and map each of the topics to a business process classification framework. In addition, we provide insights on prospective business applications, challenges, and future research directions. The main contribution of this article is to provide a state-of-the-art overview of current techniques while providing a critical perspective on business applications of social network analysis and mining.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Human Factors, Algorithms, Economics

Additional Key Words and Phrases: Social networks, community structure, networks dynamics and evolution, influence propagation, viral marketing, expert finding

ACM Reference Format:
Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. 2011. Social network analysis and mining for business applications. ACM Trans. Intell. Syst. Technol. 2, 3, Article 22 (April 2011), 37 pages.
DOI = 10.1145/1961189.1961194 http://doi.acm.org/10.1145/1961189.1961194

1. INTRODUCTION

Social network analysis emerged as an important research topic in sociology decades ago [Degene and Forse 1999; Scott 2000; Wasserman and Faust 1994; Freeman 2004], with the first studies focused on the adoption of medical and agricultural innovations [Coleman et al. 1966; Valente 1955]. It is an interdisciplinary topic that has attracted researchers from psychology, anthropology, economics, geography, biology, and epidemiology, just to mention a few.

Most of the early works were conducted on data collected from individuals in particular social settings, in order to study specific social phenomena. The analysis was

This research is partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" (www.cenitsocialmedia.es).
Authors' addresses: F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes (corresponding author), Yahoo! Research Barcelona, Avinguda Diagonal 177, 8th floor, Barcelona, Catalunya, Spain; email: ajimes@yahoo-inc.com.

DOI 10.1145/1961189.1961194 http://doi.acm.org/10.1145/1961189.1961194

usually carried out as a "field study" on small communities, gathering data through questionnaires, interviews, and other labor-intensive methods. A prominent example is the famous Travers and Milgram experiment [Travers and Milgram 1969].

The analysis of networks and networked systems, however, has a long tradition in economics, and an even longer history of graph theory in discrete mathematics [Ahuja et al. 1993; Bollobas 1998; West 1996]. From the late 1990s onwards, research on social networks has branched onto a number of fields, and has been generally carried out under the umbrella term of *complex networks*, a new emerging area in which networks are studied in several domains, using data from a wide variety of sources. The classes of networks studied include computer, biological, financial, medical, physical, and transportation networks, among many others. The goal of this research has mainly been to understand the general properties of such networks, often by analyzing large datasets collected with the aid of technology. The data is often abstracted at the level at which the networks are treated as large graphs, often with little or no concern on whether the nodes represent people, computers, or other entities. Such an abstraction is possible because in many ways the problems addressed in complex network research are similar across different domains. Relevant problems include understanding of the structure of the networks (i.e., by identifying underlying properties of the link and edge structures), the evolution of such structures (i.e., how the networks change over time), and how information propagates within the networks.

In recent years, social network research has been carried out using data collected from online interactions and from explicit relationship links in online social network platforms (e.g., Facebook, LinkedIn, Flickr, Instant Messenger, etc.). The ability to collect this kind of data by technological means has implied a significant shift in social network research, leading to the emergence of a "new," "computational social science" [Lazer et al. 2009; Watts 2004]. On one hand, it has brought a huge increase in the availability and in the size of social network data, and on the other hand it has completely redefined the types of data that can be collected and analyzed. This shift in the ability to collect data has also broadened the variety of disciplines contributing to the advance of social network research.

While traditionally social network analysis has had a strong synergy with business models in certain industries (e.g., in the telecommunications industry where rates are carefully engineered to take into account who is called and the operator of the person being called), there is still a clear gap between the social network mining techniques recently developed and their applicability in several business processes. Indeed, most research on social network analysis has focused on the general problems stated before without specific business applications in mind. As a consequence, there is little understanding of the potential application to business of social network analysis and mining methods.

In this article we give an overview of what we consider the most relevant problems in social network analysis from a business perspective.[1] In particular, we discuss data acquisition and preparation, community structure and network dynamics, propagation, and expert finding. In each case we give a brief overview of the problem, describe state-of-the art approaches, and give business application examples. In addition, we provide insights on future research directions with a particular focus on business impact. The main contribution of the article is thus to give the reader a state-of-the-art overview of key techniques while providing a critical perspective on business applications of mining social networks. More specifically, the main contributions of this article can be summarized as follows.

[1]We do not claim to cover *every* important technical research area nor all relevant industries.

—We present a state-of-the-art overview of the main social network analysis and mining problems and techniques of interest.
—We provide insights into business applications of social network analysis and mining methods.
—We detail future research directions in social network analysis and mining from the perspective of business applications.

As stated before, our goal is not to present a full survey. Instead, we aim at providing the interested reader with sufficient references to follow up on any of the subareas. For example, although recommender systems have gained significant attention in recent years, we limit our coverage to mentioning application areas (see Amatriain et al. [2010] for an overview of data mining methods for recommender systems).

Finally, it is worth mentioning that there are many ways of organizing topics in social network research. In particular, many of the techniques discussed in this article could be placed under the umbrella of predictive modeling, which may be considered the single most important business application of social network analysis and mining. Predictive modeling can be used for targeted marketing and advertising (see Provost et al. [2009]), churn prediction, and several others. Given the wide scope of predictive modeling, we have chosen not to create a separate section for it. However, the reader should keep this in mind through the article.

The rest of the work is organized as follows. In Section 2 we introduce a business process framework and outline the topics covered in this article in the context of business processes. In Section 3 we discuss data preparation, which includes acquisition and anonymization. Section 4 focuses on reputation, trust, and methods of finding experts and assembling teams. In Section 5 we discuss the detection of communities in social networks, models of graph evolution, and link formation. Section 6 focuses on information propagation in social networks, considering influence, information propagation, and churn. Finally, in Section 7 we summarize potential business applications and future research directions. Conclusions and future work are presented in Section 9.

2. BUSINESS PROCESSES

The tools and techniques developed for analyzing and mining social networks can be used in a wide range of processes across the enterprise. In this section we examine different business processes in which the techniques discussed in this article could have an impact, and we highlight some of the main challenges.

The APQC Process Classification Framework. There is a large body of research on business process management, and several business process classifications exist. Here we opt for APQC's Process Classification Framework (PCF),[2] which serves as a high-level, industry-neutral enterprise process model that allows organizations to see their business processes from a cross-industry viewpoint. The PCF has been in continuous development since 1992, when it involved over 80 organizations. In 2008 the APQC worked with IBM to enhance the cross-industry PCF and create a number of industry-specific process frameworks. The PCF was developed as an open standard to facilitate improvement through process management and benchmarking, regardless of industry, size, or geography. The PCF organizes operating and management processes into 12 enterprise-level categories, including process groups and over 1,000 processes and associated activities. The 12 enterprise-level process *categories* (first column in Tables I and II) include *process groups*, followed by *processes*, and finally by *activities*. The

[2]The PCF and associated measures and benchmarking surveys are available for download and completion at no charge from the Open Standards Benchmarking Collaborative Web site at www.apqc.org/OSBCdatabase.

Table I. Operating Processes

Process Category	*Process group or Activity*	*Technical area*
1. Vision & stategy	Social Networking	SN Support tools
2. Products & services	Product Recommendations	Recommenders
	Social product search	Social search
3. Market & selling	Social CRM	SN support tools
	Trend spotting	SN monitoring
	Product quality	SN monitoring
	Social marketing	SN support tools
	Loyalty programs	Influence
	Direct marketing	Influence, communities
	Advertising	Influence
	Business intelligence	Churn, propagation, etc.
	Churn prediction	Influence, propagation
	Reputation monitoring	Monitoring
4. Delivery	Production scheduling	Mining of customer data
5. Customer Service	Customer Support	Expert routing

Table II. Management and Support Processes

Process Category	*Process group or Activity*	*Technical area*
6. Human Capital	Internal social networking	SN Support tools
	Professional development	Expert routing
	Recruiting	Social search
7. Information Technology	Resource allocation	Measurement
	Information sources	Data preparation
	Content Management	Privacy
8. Financial Resources	Customer & product strategies	SN Mining
	Customer-product mix	Community
	Manage internal controls	Community
9. Property Management	N.A.	N.A.
10. Environmental issues	N.A.	N.A.
11. External Relationships	Public relations program	Monitoring
	Legal and ethical issues	Privacy
	Social networking	SN support tools
12. Knowledge Management	Knowledge sharing	Internal social networks
	Strategic KM	SN Mining

process categories are organized in two groups: *operating processes* and *management and support processes*.

Social Network Analysis and Mining in Business Processes. Tables I and II highlight the categories of the framework in which the social network analysis techniques described in this article could potentially be used.

A company's *vision and strategy* can be highly influenced by social networks, thus we dedicate a separate section (7) to this category, focusing on *social networking*, which encompasses several activities around social networks.

In our opinion, *products and services* is clearly a category in which there will be significant opportunities, in particular, in offering products and services that make use of a users' social network to improve their experience. In online products strong impact may be obtained from the use of tools and techniques for recommendations and for social search, among others.

A second category that we believe presents significant opportunities is *marketing and selling of products*. This category includes many activities for which social network mining is crucial. For instance, Social CRM, a new emerging area, consists of leveraging the power of social media for customer relationship management. The addition of "social" to CRM includes trend-spotting to anticipate customer needs and

future business opportunities, as well as reputation monitoring. In the same realm of monitoring we include keeping track of products delivered (e.g., by detecting customer complaints or negative comments in online networks), as well as all aspects of business intelligence. This includes churn prediction, and community detection and evolution, among others. Activities here also include marketing and advertising: the main difference with traditional methods is the ability to do direct and social marketing and advertising, which takes advantage of many of the results provided by social network analysis and mining (detection of influential nodes, propagation, etc.).

Several activities in the category of *delivery of products and services* can benefit from the techniques discussed in this article. Relevant activities in this category include collaboration with customers, forecasting, and creation and management of production schedules (e.g., by using insights obtained from mining customer social network data).

Social network analysis tools for expert finding and reputation can be of great importance in the *customer service* category. In particular, with an internal social network in place, customer calls and emails can be routed more effectively. Reputation and trust scores can be assigned to customers (e.g., customer x usually posts legitimate questions, whereas "customer" y appears to be an automated agent), and such scores can even be assigned internally to customer service representatives.

Next we describe process categories in the *management and support processes* group (Table II). In the *human capital* category, the techniques described in this article can be used for internal social networking, for professional development, and recruiting. Techniques for *information technology management* could be developed that view equipment resources as nodes in a graph (e.g., in the telecommunications domain, to measure resources). This category also includes activities related to all aspects of data preparation, such as the definition of information strategies and policies (Section 3).

In the *management of financial resources* category, we find a couple of activities of potential impact. In *tracking and performance of new costumer and product strategies*, for instance, mining information from the social network could be beneficial, as it is in *optimizing customer and product mix* (e.g., is this the right strategy for a customer given how others in his community are reacting to an offer?). In addition, this category encompasses the management of internal controls which includes *defining and communicating the code of ethics* which is so important in dealing with social network data.

We include *management of property* and *environment, health, and service* only for completeness as there appear to be no direct applications of social networks in these categories (except perhaps in the real estate and similar industries in which properties or other items can be represented by graphs). *Management of external relationships*, on the other hand, includes several high-impact groups. *Management of the public relations program*, for example, includes activities such as managing community relationships and media relationships. These activities could clearly be supported by techniques to perform reputation monitoring in online social networks and some of the other techniques used in the marketing and selling of products and services category. Business processes within this category are responsible for creating ethics policies and for ensuring compliance; legal and ethical issues play an important role in considering external relationships because, as described earlier, privacy preservation in social networks can be more challenging.

Finally, social network analysis and mining have an important role in the category of processes to *manage knowledge, improvement, and change*, particularly in *designing processes for knowledge sharing, capture, and use* which could be supported by business process mining (e.g., considering the social networks that exist in organizational structures).

3. SOCIAL NETWORK DATA

Social network data can often be derived from multiple data sources, thus the preparation of social network data deserves special attention as it continues to be a major hurdle in industrial applications. In this section we very briefly mention some of the most important issues.

3.1. Explicit and Implicit Connections

In the most basic framework the social network is represented as a graph $G = (V, E)$. Each node in the set V represents a user or customer in the network, and an edge (u, v) in the set E models a certain type of interaction between the users or customers represented by nodes u and v. Depending on the type of relationship modeled the edges may be *directed* or *undirected*.

In many domains, the social network structure includes links that are explicitly declared by users and links that are implicit and have to be inferred. For instance, in online social networking platforms, individuals can declare explicitly their "friends" or connections, "join" a group, "follow" a user, accept a "friendship" request, etc. However, these explicitly declared links may be incomplete and not describe entirely all of the relationships in the network.

Implicit connections can be discovered from user's activities by analyzing extensive and repeated interactions between users. In social media sites, this may include voting, sharing, bookmarking, tagging, and/or commenting items from a specific user or set of users, or other type of repeated interactions between individuals. In telecommunications networks, repeated calls or SMS between individuals can be extracted from call-detail records and interpreted as relationships. Similar issues arise in email and financial networks [Duan et al. 2009]. In physical spaces, proximity can be interpreted as interaction; and this data can be obtained from GPS location logs or from RFID tags used experimentally in conventions and other social events.

Implicit connections can also be discovered from user's similarity. For instance, in social media sites, users that use the same tags often can be described as similar and connected through links, and such implicit connections can be used for business applications. For example, Provost et al. [2009] construct "quasisocial networks" from online visitations to social network pages and use a predictive modeling framework for advertising.

3.2. Data Acquisition and Preparation

In the early days of social network analysis research the biggest hurdle was collection of relevant data. There were no "automatic" methods to collect data and, as in most of social science research, data collection was done by performing interviews and often small-scale group studies with volunteers. Nowadays, the collection of raw data collection from online sources (e.g., Web) and offline sources (e.g., call data) is much easier, and while data quality has always been an important issue and approaches to address it have been studied since the early 1950's [Winkler 2003], there are new challenges specific to social networks that include the computational complexity in analyzing networks of millions or billions of nodes and the integration of multiple data sources in treating implicit connections. In addition, due to the sensitivity of information on social relationships, additional privacy issues arise (e.g., when you reveal who your friends are, you are revealing information that may not be sensitive to you, but that may be sensitive to your friends).

From a practical perspective, particularly in the context of various of the business processes outlined in the previous section, identification of data sources is often difficult in industry settings. This often pertains to the organizational structure and the

environment in which the data is collected and used. In a typical company, for instance, there may be organizational silos (e.g., marketing, business intelligence, database administration, etc.) and each one may have different levels of access to the data, interests, and requirements. Thus it is often difficult to have a clear picture of what data is available, where is it, who has rights over the data, and what is the format of the data. In many cases, it is possible for relevant data to be missing, or to be poorly documented. Even in cases where data useful for social network mining may seem very valuable, the structure of the data may be rather complex. Data records (for instance call detail records in the telecommunications industry [Phithakkitnukoon and Dantu 2008]) have a complex syntax and structure and contain many fields that are irrelevant for SNA.

Social network data from online networks may suffer additional problems including the following:

—duplicate nodes, for example, a single person having two email addresses;
—inactive nodes: individuals who do not explicitly remove their profile, but no longer access it (one case occurs when people pass away and their profiles remain active[3]);
—Artificial nodes for example, automated agents, possibly malicious ones.

Data cleaning includes the elimination of duplicates, verification of values in the proper range, and others. Rahm and Do [2000] classify the problem into two categories: *single-source* and *multiple-source* problems. These are further divided into *schema* and *instance* levels. Data cleaning is then characterized as having several phases: data analysis, definition of transformation workflow and mapping rules, verification, transformation, and backflow of cleaned data. Although their framework is not specific to social networks, the issues in data cleaning for social network analysis can be clearly identified from their perspective.

Finally, in some countries, storing call data over a period of a few months is required by law in case the data is needed in future legal inquiries. At the same time, there are laws that prevent storage and use of such data for periods exceeding a few months (even if it is backed up for future legal use).

3.3. Anonymization

As data mining algorithms are becoming ubiquitous and as data are continuously collected and shared within organizations, *privacy-preserving data mining* [Agrawal and Srikant 2001a; Vaidya et al. 2006] has been proposed as a paradigm of performing data mining tasks while protecting the personal information of individuals.

The graph of social connections of users can be a rich source of information and may be used to discover personal information about users. Even if personally identifiable information like names or social security numbers are removed from the data, this is far from being sufficient. As shown by Backstrom et al. [2007], the mere structure of the released graph may reveal the identity of the individuals behind some of the nodes. Hence, one needs to apply a more substantial procedure of sanitization on the graph before its release.

The objective of protecting the privacy of individuals represented in databases was formulated by Dalenius [1977] in 1977. Since then, many approaches have been suggested for finding the right path between data hiding and data disclosure. Such approaches include query auditing [Kleinberg et al. 2003], output perturbation [Blum et al. 2005], secure multiparty computation [Aggarwal et al. 2004], and data sanitization [Agrawal and Srikant 20001; Evfimievski et al. 2003].

A basic operation in data anonymization is to perturb the data so that individual values are hidden, while still being able to recover useful information, such as the

[3]http://www.nytimes.com/2010/07/18/technology/18death.html.

distribution of the data values or rules and patterns in the data. In one of the first papers that introduced the concept of privacy-preserving data mining, Agrawal and Srikant [2000b] propose the idea of perturbing the data by adding random values from an a priori known distribution and they show that it is possible to reconstruct the original distribution of the data. Privacy is preserved because one cannot make an inference about any individual value in the data. In their paper, Agrawal and Srikant also show how to use the perturbed data for the problem of *classification*. In particular, they show how to use the reconstructed distributions in order to build classification trees that achieve accuracy close to the one achieved with the original (unperturbed) data. Following up in the work of Agrawal and Srikant, researchers provided more examples of how to perform data mining tasks while preserving the privacy of individual data records. For instance, Evfimievski et al. [2002] and Rizvi and Haritsa [2002] employ the use of randomization in order to discover frequent itemsets and association rules in transactional data.

3.4. Anonymization of Graphs and Social Networks

The methods of identity obfuscation in graphs fall into three main categories. The methods of the first category [Liu and Terzi 2008; Wu et al. 2010; Zhou and Pei 2008] provide k-anonymity in the graph via a deterministic procedure of edge additions or deletions. The methods of the second category [Hanhijarvi et al. 2009; Hay et al. 2007; Ying and Wu 2008, 2009a, 2009b] add noise to the data, in the form of random additions, deletions or switching of edges, in order to prevent adversaries from identifying their target in the network, or from inferring the existence of links between nodes. The methods of the third category [Campan and Truta 2008; Hay et al. 2008; Zheleva and Getoor 2007] do not alter the graph data like the methods of the two previous categories; instead, they group together nodes into supernodes of size at least k, where k is the required threshold of anonymity, and then publish the graph data in that coarse resolution.

Hay et al. [2007] investigate methods of random perturbations in order to achieve identity obfuscation in graphs. They concentrated on reidentification of nodes by their degree. By performing experimentation on the Enron dataset, they found out that in order to achieve a meaningful level of anonymity for the nodes in the graph, the random perturbation methods need to add and remove too many edges in the graph. Those methods were revisited by Ying et al. [2009], in which they compare the random-perturbation method to the method of k-degree anonymity due to Liu and Terzi [2008]. Based on experimentation on two modestly sized datasets (`Enron` and `Polblogs`) they arrived at the conclusion that the deterministic approach of k-degree anonymity preserves the graph features better for given levels of anonymity. On the other hand, Bonchi et al. [2011] provided an information-theoretic look on the strategy of random additions and deletions of edges, and they showed that randomization techniques may achieve meaningful levels of obfuscation while still preserving characteristics of the original graph. They also showed that due to small-world phenomena, only deleting edges maintains better the characteristics of the graph than adding and deleting edges. Overall, the problem of anonymization of social networks is open and still under investigation.

3.5. Business Applications

The main reason for anonymization of social network data is to protect the privacy of the individuals whose data is being collected. Collecting and aggregating personal information from many people creates data that has to be handled with extreme care. Writer and activist Cory Doctorow has compared collections of private electronic data

held by governments and businesses with weapons-grade plutonium[4] in their tenacity and longevity, and in the fact that once data are refined and stockpiled they become much more dangerous and difficult to contain.

In recent years, laws in many countries have started to demand more stringent requirements on how this information is handled. Besides the protection of users' privacy rights, anonymization may also be needed to share data across different business units, or to provide particular services to users.

Telecommunication and Computer Networks. Monitoring and measuring are crucial for determining where and how large companies should invest to enhance their infrastructure. This holds for all industries in general and for the telecommunications industry in particular. In this case, it is necessary to have reliable and timely information about network traffic and other operational parameters.

This data is routinely shared among organizations for research, regulatory, or business reasons. For instance, sharing of network logs is necessary to improve network security, because network attacks cross organizational boundaries. Also, companies may share data with government agencies for national security purposes, a booming industry [Soghoian 2008], increasing the need to properly anonymize data when needed.

Additionally, in the telecommunications industry, social network analysis is used for fraud detection (e.g., an offensive node can be identified based on its outgoing links or on behavioral patterns [Fawcett and Provost 1997; Cortes et al. 2001]), as well as for marketing purposes. The telecommunication operators often outsource these or other operations, sharing data with third parties that provide the relevant services, for example, to estimate churn, identify influential nodes, communities, fraud, etc.

There are legal requirements to protect privacy as there are substantial risks (and financial impact, both legal and otherwise) from customer information leakage. Without proper anomymization, for instance, of host identities, user behaviors, network topologies, etc., and without appropriate security practices, enterprise networks are vulnerable to attack [Coull et al. 2009].

Online Communities. As we describe in subsequent sections, one of the driving business applications of social network analysis is marketing. As a consequence, many current online social network platforms share data with third parties for advertising purposes. Furthermore, as part of their business model, many social network platforms provide open APIs that allow third parties to create applications that often access user profiles (or profiles of "friends"), possibly breaching user's privacy [Narayanan and Shmatikov 2009]. It is undoubtedly in the interest of these companies to properly anonymize the data shared or made accessible through the API, but, as discussed in Narayanan and Shmatikov [2009], there are still many challenges in accurately anonymizing social network data. One possible alternative is to use "privacy friendly" techniques at the time of collecting information from social network sites. Provost et al. [2009], for instance, build high brand-affinity audiences by selecting the social network neighbors of existing brand actors identified via covisitation of social networking pages, without saving any information about the identities of the browsers or content of the pages.

Anonymization is of significant importance in general business data management, but it is even more crucial when it comes to social network data. As pointed out by Narayanan and Shmatikov [2009], the increase in user overlap between different online social networks (e.g., Flickr and Twitter in their study) and the growth in

[4]http://www.guardian.co.uk/technology/2008/jan/15/data.security.

number of third parties with access to potentially sensitive anonymized social network data may result in major privacy breaches, and any potential solution would appear to require a fundamental shift in business models and practices and clearer privacy laws on the subject of personally identifiable information [Narayanan and Shmatikov 2009].

Search Engines and Other Online Platforms. Search engines and other online platforms posses large logs that record the interaction of users with the system. Increasingly, such interactions go beyond search and include sharing with friends and tagging actions. These logs contain very valuable information about the behavior of the users and their interests. Mining these logs has immediate impact in a wide variety of applications: improving search results, building user models, making recommendations to users, understanding trends and the market, and many more. Thus it is of interest to the search engines to be able to share their user log data so that they can benefit from data mining results and research methodologies developed for the data. However, data sharing can not be a reality until secure anonymization techniques that protect the privacy of users are developed.

4. REPUTATION, TRUST, AND EXPERTISE

4.1. Definitions

According to the taxonomy presented by Ziegler and Lausen [2005a], there are two basic types of trust computation: *global* and *local*. In global trust computation, the trustworthiness of each agent is computed from the perspective of the whole network, and thus each agent is associated to a single trust value. We use the term "reputation" to refer to "global trust."

In local trust computation, trust inferences are done from the perspective of another agent, and thus each agent in the network can have multiple trust values. Depending on the context, it may be important to compute local trust, global trust, or both. For instance, in large-scale social networks, from the point of view of the entire system, establishing the reputation or global trust of users is very important when aggregating information, to lessen the impact of malicious activities. On the other hand, from the point of view of specific users, establishing local trust efficiently is important when exchanging information or collaborating, particularly in decentralized environments.

Expertise can be understood as reputation with respect to a given topic. Finding an expert may help in cases where users need to access nondocumented information, or need some contextual information that is not provided by documents alone. An expert-finding system may help users whenever they cannot specify their information need, or want to be efficient in terms of minimizing group effort, as it may be easier for the expert than for the nonexpert to locate a particular piece of information. Others may simply prefer asking an expert instead of interacting with documents and systems [seid and Kobsa 2002].

4.2. Computing Trust from Social Ties

We first describe a set of metrics that estimate trust based only on links. The next section incorporates other factors.

Trust relationships can be naturally represented as a graph. The concept of a "web of trust" was first introduced in large-scale systems during the design of key management protocols for PGP (*Pretty Good Privacy*). A web of trust is a directed graph where nodes are entities, and arcs indicate a trust (or distrust) relationship between two entities. The web of trust in a large community tends to be very sparse. Any given agent interacts only with a small fraction of the members of the community, and thus can only assess the trustworthiness of a handful of other agents. A natural way of alleviating this

sparsity problem is to *aggregate* the ratings given by several people, usually through the use of some sort of propagation mechanism.

There are many link-based methods for finding authoritative, influential, central, reputable nodes on a network. These methods are very general in the sense that they can be applied to any type of network, including networks of connected inviduals or linked documents. The output is a ranking that can be interpreteded in the case of social networks as social prestige or reputation.

One of the best-known methods of this family is Katz's index [Katz 1953], which was proposed for ranking individuals in a social network since the 1950's. In its original formulation, each person in the network chooses another participant and "votes" for him/her; votes are pased transitively with a certain attenuation factor. Katz's index and similar kinds of link-based reputation metrics are known nowadays as *spectral ranking* methods, as they rely on computing an eigenvector of a matrix that represents votes or endorsements.

The most popular variant of this family of method is PageRank [Page et al. 1998], which is used to rank Web pages. PageRank also has a probabilistic interpretation as a "*random surfer*" who wonders the network following links at random: the score of a node is the fraction of time spent at that node by the random surfer in the limit. PageRank has been studied extensively during the last decade; for a survey, see Langville and Meyer [2003].

A nice property of PageRank is that it can be easily adapted to boost the scores of certain nodes and those connected to them. For instance, Haveliwala [2002] proposes to compute a series of *topic-sensitive* PageRank scores by executing independent random walks on the graph in which the restarting probabilities of each walk are biased towards pages on a given topic. Applying the same general principle, Gyöngyi et al. [2004] propose *TrustRank*, in which they use a small seed set of nonspam (trusthworthy) pages that are carefully selected by human editors, and then compute a global trust score by performing short random walks with restart to the seed set.

Another link-based reputation metric is HITS, introduced by Kleinberg [1999] also in the context of Web pages. In this method, two scores are computed for each node: a hub score and an authority score. Intuitively, a node has a high authority score if it is endorsed by many good nodes with a high hub score, and a node has a high hub score if it endorses many authoritative nodes. Despite the apparent circularity of the definition, the HITS scores can be computed by an eigenvector computation.

Methods such as PageRank, TrustRank, and HITS have been used extensively to find relevant documents in linked document collections, as well as to find relevant people in a social network [Pujol et al. 2002] or good askers/answerers of questions in a question-answering portal [Jurczyk and Agichtein 2007].

4.3. Computing Trust from Social Ties and Other Factors

In this section we describe trust metrics that use a graph of social connections and some extra information, such as feedback provided about other peers (either as scores or positive/negative judgments) or other properties of the agents.

Incorporating Negative Feedback. In many communities the base assessments from which trust is computed include both positive (trust) and negative (distrust) assessments. However, negative assessments are not used as often as positive assessments. First, the semantics of trust propagation, for example, "the friend of my friend is my friend," are clear and effective in practice, while the semantics of distrust propagation, for example, "the enemy of my enemy is my friend," have been shown less effective in practice. According to the results of Guha et al. [2004a], a good method for global trust computation uses an iterative (multistep) direct propagation of trust, but only

a single-step direct propagation of distrust. Second, in many communities positive assessments are dominant, as people are much more cautious when providing negative judgments for fear of retaliatory negative feedback, or simply to avoid further unpleasant interactions [Resnick et al. 2000].

Local Trust. There are many examples of trust computations that are local: computed from the perspective of a user, and not from the perspective of the entire network. Among others, they include the reputation system implemented in the *Advogato* community, based on maximum flows [Levien and Aiken 1998], a model based on weighted paths due to Mui et al. [2002], and *Appleseed*, a system based on spreading activation [Ziegler and Lausen 2005a].

The email exchanges of a person with his/her peers (a *personal email network*) can also be used to generate a local trust score, to be used for email spam filtering or other tasks. This topic is studied, among other authors, by Boykin and Roychowdhury [2004] where the subgraphs induced by legitimate and spam email messages are shown to be clearly different. A related study is due to Gomes et al. [2005].

Decentralized Computations. Another setting in which local trust computations take place are peer-to-peer (P2P) networks. Trust propagation in P2P networks require decentralized trust computations to establish the quality of the files offered by each peer for download. A taxonomy of P2P reputation systems is introduced by Marti and Garcia-Molina [2006]. This taxonomy considers factors such as how the information is gathered and aggregated and what the actions taken by the system are with respect to inauthentic peers.

Incorporating the Effect of Time. The SocialTrust model by Caverlee et al. [2008] is an example of a model that incorporates the notion of time. The design principle is to mitigate the effect of users who accumulate a good reputation over time, and then take advantage of that reputation to behave maliciously.

Exchanging Trust Information. Finally, social network trust can also be shared across different services. The FaceTrust protocol by Sirivianos et al. [2009] provides a general mechanism for verifying the credentials of a user. The objective of the system is to create an environment in which users can assure new online services that they are "good netizens" by providing credentials from their previous activities in other social networks.

4.4. Expert-Finding Methods

Early expert-finding methods can be classified into two complementary approaches having either a strong information retrieval component or a strong social networks component.

The information retrieval approach is exemplified by the *P@NOPTIC Expert* system described by Craswell et al. [2001]. In this system, first a collection of all the documents authored by an individual is collected; then, documents are concatenated to create a "person-document." Finally, a standard document search system is run over these person-documents, and the people corresponding to the highest-ranked person-documents are returned as "experts" for the input query.

A refinement of this approach is shown by Balog et al. [2006] who consider the relevance of documents for a query, so that not all documents authored by a person are considered equally. The authors compare two approaches: one in which they attempt to model user expertise on a topic directly, and one in which they first collect relevant documents and then use them to locate experts. In their experimental evaluation the second method shows better results.

The social network approach is exemplified by an early presentation about *Verity* by Abrol et al. [2002]. A rich representation of users and documents is used in which documents are linked to their authors, and people and documents are connected through interaction histories, search queries, keywords, and explicit feedback.

Refinements of this approach can be found in Zhang et al. [2007], where the authors build a network from threads in an online forum in which nodes are users and arcs connect users starting a thread with users replying to them; variants of PageRank and HITS are tested in this graph. Campbell et al. [2003] run HITS on a graph created from email exchanges.

Currently, many effective expert-finding systems use a combination of the approaches described before. For instance, the Expert Finding Demo[5] described by Deng et al. [2008] identifies scientists' expertise on a topic using their published articles. It considers the ranking of documents retrieved for a query, as well as the citation information in order to prefer highly cited documents.

The topic of expert finding gained considerable attention in the research community since the TREC 2005 competition included an expert-finding task. To get more insights about how different techniques compare to each other, the interested reader can read the overview of TREC 2005 and related competitions [Craswell et al. 2005].

4.5. Assembling Teams of Experts

A natural generalization of the expert-finding problem is to find not one but many experts that can form a team. To solve this problem, we first take into account the *topical profile* of individuals, describing their expertise in terms of topics. Next, we consider their *social profile* that includes their social connections [Balog and De Rijke 2007] and describes their compatibility with others.

There are many possible ways of formalizing the team formation problem. Lappas et al. [2009] consider that a good team for a particular problem must cover all the required skills for the problem (must contain at least one expert in each of the topics in which the problem requires expertise). Also, the members of a good team must span a subgraph of the social network that has good connectivity properties, for instance, a subgraph whose diameter is small or that has a low-cost spanning tree.

4.6. Business Applications

Techniques and theories related to reputation, trust, and expertise have been developed and applied in a number of offline and online business settings. After all, in many ways these topics form the foundations of organized efforts in corporate and noncorporate settings alike. Trust, for instance, has been studied extensively in organizational theory [Kramer 1999], while concepts like expertise capitalization/leveraging, skill mining, competence management, intellectual capital management, expertise networks, and knowledge sharing systems have also been studied extensively in the knowledge management discipline [Yimam-seid and Kobsa 2002]. The business impact of techniques to address these topics is therefore understood (although not always easily quantifiable).

We can argue that all Web-scale systems incorporate a reputation layer. For instance, search engines cannot function without measures to reduce spam: "without such measures, the quality of the rankings suffers severely" [Henzinger et al. 2002].

Online marketplaces, on the other hand, such as e-Bay, incorporate explicit community feedback mechanisms that can be used effectively to compute reputation scores. Making such scores public has been effective in "filtering" as in many ways the community regulates itself. For example, it has been observed that buyers pay a small but measurable premium for buying items from high-reputation sellers [Melnik and Alm

[5]http://expertfinding.net/.

2002], increasing these sellers' revenue, visibility, and motivation to keep high reputation scores by effectively delivering what they promise. It has also been observed that some users try to game the reputation system by creating fake identities to inflate the reputation of particular nodes in order to engage in auction fraud; however, some mechanisms have been designed to prevent this type of attack [Pandit et al. 2007].

In marketplaces like e-Bay or even the stockmarket in general there are several communities (e.g., sellers of particular types of goods, buyers, etc.), so although it may not be immediately obvious, the social component plays an extremely important role. Therefore, in terms of business impact, techniques that leverage social networks are likely to be more successful. For instance, applications that detect fraud in financial statements [Virdhagriswaran and Dakin 2006] may be able to find outliers in the statements themselves. However, more advanced applications have demonstrated that it is useful to exploit multiple sources of information, for instance, in the the case of securities fraud by considering the relationships between firms, branches, and brokers [Neville et al. 2005].

Aside from regulators seeking to prevent fraud, companies themselves can use social network mining to detect customers likely to purchase services that they do not intend to pay. One example application has been developed by Detica to prevent telecom subscription fraud [Detica 2006]. Fawcett and Provost [1997], for example, detect fraud by uncovering suspicious changes in user behavior using a rule-learning program. The system has been applied to the problem of detecting cellular cloning fraud based on a database of call records.

Reputation systems can also be used for trend spotting, public relations, for monitoring the reputation of the enterprise, and in general for Customer Relationship Management (CRM) tasks. For instance, reputation systems can be used for filtering unsolicited commercial email. They can also be used to prevent spam in blogs and other publicly writable spaces. Akamai[6] and Mollom[7] offer commercial services of this kind. In trend spotting, or in managing public relations, it is important to consider the reputation of the individuals or organizations generating information. A public complaint by a highly reputable source merits a very different corporate response from a response to a malicious action by an nonreputable source.

Expert finding is crucial in large corporate environments because, when faced with problems that require collaboration, it is extremely difficult for any one single person or department to have a complete and accurate view of skills and availability of everyone else in the organization. Therefore, expert-finding methods have been proposed for enterprise search systems. This is the case of the K2 product developed by Verity [Abrol et al. 2002] acquired by Autonomy Corporation in 2005; or the Colleague Search system demonstrated by Milette et al. [Davitz et al. 2007] that allows to exploit the social ties to find experts in an organization.

There are also several commercial services for finding experts, examples include Community of Science[8] to find scientists, Profnet[9] to find professional journalists, and Expert Witnesses[10] to find expert witnesses for trials. Hettich and Pazzani proposed such a system to match proposals with reviewers in the U.S. National Science Foundation [Hettich and Pazzani 2006].

The dynamics of expertise in an organization is a relevant and current research topic. Expertise in particular, and knowledge and information in general, are not static

[6]http://akamai.com/.
[7]http://mollom.com/.
[8]http://www.cos.com/.
[9]https://profnet.prnewswire.com/.
[10]http://www.expertwitness.com.

aspects of an organization but change and disseminate by the interaction of coworkers. In this process, aspects such as the bandwidth available as well as the diversity of people's connections are important, among other factors [Aral and Van Alstyne 2010].

For finding experts in the "open" Web, Kaiser et al. [2007] presented the EXPOSE system to find people or companies that are experts on a topic. The question-answering community Aadvark[11] [Horowitz and Kamvar 2010] offers a system for locating experts who can answer questions posed by the community.

Aside from finding experts, social networks can also be exploited in the context of knowledge management in a large organization. For instance, the POLESTAR system described by Pioch and Everett [2006] allows analysts to have access to other people's assertions about the document they read, for example, a document or an entire information source can be flagged as "discredited" and this flag will be visible for other people inside an organization. Social networks can also be used to organize the collaborative production of content, for example, a Frequently Asked Questions document [Davitz et al. 2007].

In Section 7 we will discuss how, in enterprises, social networks are placing an even stronger emphasis on unified collaboration and communication. Techniques for expert finding can therefore be used to enhance professional development (e.g., by helping employees find the right mentors for particular tasks), for human capital tasks (e.g., recruiting and other HR functions), and to mine and improve organizational structure.

5. COMMUNITY STRUCTURE AND NETWORK DYNAMICS

Grouping related elements is a basic operation in many domains such as Web analysis, bioinformatics, ecology, and telecommunications, among others. Substantial effort in social network analysis has been devoted to discovering communities in large social graphs and the problem has attracted attention not only among computer scientists, but also among statisticians and applied physicists.

The objective of community detection methods is to find groups of users for which, intuitively, the set of edges is dense within the group and sparse outside the group. For example, a community may consist of a team within a company, whose members exchange a large number of emails with each other, or of a set of users of a blogging site who are interested in a certain topic and contribute blog posts about it and comment on each others' posts. One of the main difficulties is that how a community is defined depends a lot on the task and the types of links between nodes that are considered.

Naturally, community detection algorithms take advantage of graph-theoretic concepts. Indeed, the community detection problem is closely related to the problem of *graph clustering*, but there are important differences that require novel approaches not traditionally considered in the graph clustering domain.

—*Definition of interaction among users*. As noted in Section 3.1, the interaction among users in a social network can be defined in various ways; users often have profiles consisting of heterogeneous information, and there are complex ways of interaction among users and between users and the system. Furthermore, such interactions are dynamic (i.e., implicit links may be ephemeral);
—*Scalability*. Dealing with real social networks that have millions of users limits the applicability of many traditional graph clustering algorithms in practical scenarios.

A topic related to the detection of communities is how such communities change over time. Traditionally, however, the analysis of social networks has focused only on a

[11]http://vark.com/.—acquired in February 2010 by Google.

single snapshot of a network. The fact that social networks follow power-law degree distributions [Faloutsos et al. 1999], have small diameter (i.e., the maximum possible distance between two nodes measured as length of the *shortest path*), exhibit *small-world* structure [Watts and Strogatz 1998], and community structure [Girvan and Newman 2002], are only few of the ubiquitous properties that many researchers have verified. Attempts to explain some of the properties of social networks have lead to dynamic models inspired by the *preferential attachment* model [Barbási and Albert 1999], which predicts that new nodes arriving to the network will connect to existing nodes with a probability proportional to the number of connections already present in the graph. This is regarded as an instance of a multiplicative process, also known as Yule process, or simply the *rich-get-richer* process.

In the following sections we cover these two closely related topics: community structure and network dynamics. We then briefly describe business applications, but it is worth noting that most of the published work on business applications in these two areas is found in the context of influence propagation (for marketing), and churn, which are discussed in more detail throughout Section 6, and in particular in Section 6.5. The reader might also want to keep in mind some of the business applications outlined in Section 4, which also relate to the topics addressed in this section.

5.1. Community Structure

In this section we present some of the methods for discovering communities. Our survey is by no means complete, and the reader interested in more details is referred to the thorough survey of Fortunato [2010].

Hierarchical Algorithms. A basic family of algorithms for finding communities is based on building a *hierarchical decomposition* of the nodes of the social network. Such hierarchical methods have been used traditionally in sociology. A property of these methods is that they return not just a flat partitioning of the network into communities, but a hierarchy of communities and subcommunities. Such a hierarchy can be represented by a *dendrogram*. The general approach requires definition of a similarity function between two sets of nodes in the network. A special case is when the sets are singletons, where the similarity function is defined among two nodes. Typical methods to define similarity functions among sets of nodes include notions such as *shortest-path distance*, and similarity measures involving sets of neighboring nodes such as *cosine similarity* and *Jaccard coefficient*. One starts by first computing the similarity value between every pair of nodes in the network. The general algorithm proceeds recursively in an agglomerative fashion: initially each node is alone in its own set, then the sets with the largest similarity value are merged into one new set, and the similarity of the new set with all existing sets is computed. The algorithm terminates when only one community remains. Instances of this generic framework are the *single-linkage* algorithm and the *complete-linkage* algorithm.

A different approach to hierarchical community detection was presented by Girvan and Newman [2002]. Instead of merging nodes in a bottom-up fashion, the method proceeds top down. It starts with the whole network as a single group, and at each iteration it removes one edge from the network. Some of these edge removals may partition a connected component into smaller connected components, thus defining a hierarchy of communities. To completely specify the algorithm one needs to define how to remove edges. Girvan and Newman suggest to rank the edges of the network with respect to a measure called *edge betweenness*, and remove edges with decreasing order of the value of this measure. The edge betweenness of an edge is defined as the number of pairs of nodes in the network for which the edge lies on a shortest path. The intuition

is that edges with large edge betweenness value lie between communities and thus they should be removed first in order to reveal the communities.

Modularity Maximization. Girvan and Newman [2002] proposed a measure of evaluating the quality of a partitioning of a network into communities, and selecting the best community partitioning from a hierarchal decomposition. The measure is called *modularity*, and is defined as the fraction of edges that fall within communities minus the same fraction if edges were assigned at random. A nice property of the modularity measure is that it is not optimized for an extreme value ($k = |V|$ or $k = 1$, as most clustering measures do), thus optimizing modularity gives a natural way of selecting the number of communities in the network. Girvan and Newman [2002] proposed to optimize modularity directly, instead of evaluating modularity at the end of the community discovering algorithm. For this modularity maximization problem, they presented an algorithm with running time $O(|V|(|V| + |E|))$. The algorithm of Girvan and Newman was further improved by Clauset et al. [2004] to $O(|V| \log^2 |V|)$. Many researchers have studied and developed algorithms for the modularity measure. Brandes et al. [2008] showed that it is NP-hard to optimize modularity, Fortunato and Barthelemy [2007] identified the resolution-limit problem, according to which the optimization point of modularity depends on the size of the network. White and Smyth [2005] follow a spectral approach to optimize modularity and Agarwal and Kempe [2008] develop a mathematical programing algorithm, among many other algorithms.

Graph-Partitioning Algorithms and Spectral Partitioning. As we mentioned earlier, many community detection methods employ techniques based on graph theory. Flake et al. [2000, 2002] define a community to be a set of nodes that have more edges to nodes of the community than to nodes outside the community, and they develop algorithms based on the notions of *minimum cut* and *maximum flow*.

Another approach to clustering graphs is based on *spectral* partitioning. The main idea is to project the nodes of the network onto a low-dimensional Euclidean space and then cluster the projected Euclidean points using standard clustering algorithms, such as the *k-means* algorithm [Lloyd 1982]. Details on the properties of spectral embeddings of graphs and spectral clustering algorithms can be found in Chung [1997], Korean [2003], and Ng et al. [2001]. A popular suite of graph-partitioning algorithms, which is accompanied by high-quality software, is the METIS algorithm [Karypis and Kumar 1998]. METIS tries to find the good separator while minimizing the number of edges cut in order to form two disconnected components of relatively similar sizes.

5.2. Network Dynamics

Models of Graph Evolution. Recently several researchers have turned their attention to the dynamics and evolution of social networks.

The copy-model [Kumar et al. 2000] states that a new node that connects to a network selects some nodes to which to connect by the preferential attachment rule, but also picks an existing node at random and "copies" some of its out-links.

Leskovec et al. [2005] empirically observed that networks become denser over time, in the sense that the number of edges grows superlinearly with the number of nodes. Moreover, the densification follows a power-law pattern. In the same paper they also report another surprising observation: the network diameter often shrinks over time, in contrast to the conventional wisdom that such distance measures should increase slowly as a function of the number of nodes.

The triangle-closing model [Leskovec et al. 2005, 2008] states that new nodes have a tendency to complete triangles on a network, in other words that they may connect to an existing node and to some of that node's neighbors. The forest-fire model [Leskovec

et al. 2007b] is in some sense a generalization of the triangle-closing model: when a new node connects to an existing node, it picks a subgraph containing the existing node (by running a process that starts at the existing node and resembles a fire spreading from it through the network) and connects to all the nodes in that subgraph.

While some effort has been devoted to analyze global properties of the evolution of social networks, not much work has been done to study graph evolution at a microscopic level. A first step in this direction is the work of Leskovec et al. [2008], investigating a wide variety of network formation strategies, and showing that edge locality plays a critical role in the evolution of networks.

Other recent papers present algorithmic tools for the analysis of evolving networks. Tantipathananandh et al. [2007] focus on assessing the community affiliation of users and how this changes over time. The algorithms proposed to solve this problem are based on dynamic programming, exhaustive search, maximum matching, and greedy heuristics. Sun et al. [2007] apply the MDL principle to the discovery of communities in dynamic networks, developing a parameter-free framework. This is the main difference with previous work such as Aggarwal and Yu [2005] and Sun et al. [2006]. However, as in Tantipathananandh et al. [2007], the focus lies on identifying approximate clusters of users and their temporal change. No exact patterns are found, nor is time part of the results obtained with these approaches. Ferlez et al. [2008] use the MDL principle for monitoring the evolution of a network.

Mining Evolving Graphs. A different approach to the analysis of network evolution, which follows the paradigm of *association-rule mining* and *frequent-pattern mining* is presented by Berlingerio et al. [2009]. By introducing *graph evolution rules*, a novel type of frequency-based patterns, Berlingerio et al. consider the problem of searching for typical patterns of structural changes in dynamic networks. They first compute a set of frequent graph patterns that describe "typical" evolution mechanisms and then they find graph evolution rules that satisfy a given minimum confidence constraint.

Desikan and Srivastava [2004] study the problem of mining temporally evolving Web graphs. Three levels of interest are defined: single node, subgraphs, and whole graph analysis, each of them requiring different techniques. They study changes of properties on each of the three levels under investigation. Inokuchi and Washio [2008] propose a fast method to mine frequent subsequences from graph sequence data defining a formalism to represent changes of subgraphs over time. However, the time in which the changes take place is not specified in the patterns. Liu et al. [2008] identify subgraphs changing over time by means of vertex importance scores and vertex-closeness changes in subsequent snapshots of the graphs. The most relevant subgraphs are hence not the most frequent, but the most significant based on the two defined measures.

Borgwardt et al. [2006] represent the history of an edge as a sequence of 0's and 1's representing the absence and presence of the edge, respectively. Then conventional graph mining techniques are applied to mine frequent patterns. The employed mining algorithm GREW does not mine all the frequent patterns, but it employs heuristics.

Link Formation Prediction. Models of graph evolution are typically developed with the aim of estimating the overall statistical properties of existing graphs. One can also consider whether two particular nodes are likely to become connected in the future. This basic computational problem underlying social network evolution in time is known as the *link prediction problem*, introduced by Liben-Nowell and Kleinberg [2003].

Given a snapshot of a social network at time t and a future time t_0, the problem is to predict the new links that are likely to appear in the network in the time interval $[t, t_0]$. As Liben-Nowell and Kleinberg state, the link prediction problem is about modeling the evolution of a social network using network-intrinsic features. In fact, Liben-Nowell

and Kleinberg consider only the features that are based on the link structure of the network, including statistics such as number of common neighbors, geodesic distance, personalized PageRank and hitting time in the social network, and in general methods that compute some notion of similarity or closeness in the social network.

Taskar et al. [2003] apply link prediction to a social network of universities. They rely on machine learning techniques and use personal information of users (music, books, etc.) to increase the accuracy of predictions. Following a similar approach, O'Madadhain et al. [2005] focus on predicting events between entities and use the geographic location as a feature. Clauset et al. [2008] apply link prediction to biology and physics using hierarchical models in order to detect links that have not been observed during experimentation.

Several probabilistic models such as Markov logic [Domingos and Richardson 2004], relational Markov networks [Taskar et al. 2003], Markov random fields [Chellappa and Jain 1993], and probabilistic relational models [Getoor et al. 2003] have been used to capture the relations existing in data.

Other approaches focus instead on properties of the users themselves. According to Kumar et al. [2004], many connections in a large social networks (the blogosphere, in this case) can be explained by matching demographic groups, topical interests in common, or geographical proximity.

5.3. Business Applications

The traditional approach in business intelligence and marketing has been to treat customers as individuals, or to group them into sets (segments) with certain characteristics. One of the most important shifts brought by the advent of social network research is to start thinking of customers as forming communities, or to put it another way, as individuals that *belong* to communities. A single individual may belong to multiple communities and those communities may even partially overlap.

While traditional customer segmentation methods to partition a customer base are still valid and widely used, considering communities arising from social graphs, has shown its potential for creating new marketing strategies as well as in new product offerings in online social networks.

We can summarize some of the main business applications of community structure detection as follows.

—Social recommendations in online social networks. The business models of many companies (e.g., Amazon, Pandora, Last.fm, iLike, and many others) are strongly linked to generating useful recommendations. In businesses such as these, implicit links between users are the norm and thus communities are not explicitly defined and must be discovered. Schifanella et al. [2010], for example, analyze Flickr and Last.fm tags and find that friend suggestions constructed from implicit semantic similarity of user generated tags on Last.fm capture friendship more accurately than Last.fm's suggestions based on listening patterns.

—Social search. Modern search engines try to exploit as much context as possible from the query to provide relevant results. Context may include, for example, the identities of the people executing the search as well as their connections. Ronen et al. [2009] introduced a system for enterprise search that allows finding people and documents that are somehow connected to the user who executes a search, for example, documents that are authored by contacts-of-contacts (see Marlow [2003] for community discovery from blogs). Google recently added "results from your social circle" to the search results.[12] These features may have a positive impact on knowledge-intensive

[12]http://googleblog.blogspot.com/2010/01/search-is-getting-more-social.html.

industries, as there are measurable effects of social information seeking behavior on the productivity of knowledge workers [Aral et al. 2006, 2007]. Watts et al. [2002], for instance, present a model to explain social network searchability along a set of social dimensions, with possible applications in many network search problems.

—Marketing in offline settings. In the telecommunications industry and in other industries that have rewards programs for customer loyalty, network structure plays a significant role in helping identify target groups and allocation policies for such rewards (see work of Richardson and Domingos [2002] and Hill et al. [2006] described in Section 6).

—Security. For companies that provide security consulting, or for governments fighting criminal or terrorist organizations, identifying communities and network structure is of extreme importance, whether it is in online or in offline social networks.

All of these applications must take into account that users may declare only some of their connections to groups or other users, so the data provided is incomplete. In general, there may be a general perception that the association between users and groups is often explicitly declared, but from a practical business perspective, it is often the case that these communities must be discovered from the data.

An extreme example can be found in an application in the fields of journalism and intelligence: Krebs [2002] describes how to mine known relationships between Al-Qaeda operatives, discovering communities in this network that matched actual roles taken during the September 11 attack. There are several other examples of social network analysis for journalism at the IRE[13](Investigative Reporters and Editors) Web site, and as described in Section 4 identifying communities and monitoring network dynamics can also be used to identify fraud (e.g., malicious nodes tend to show certain behavioral properties [Fawcett and Provost 1997]) and to fight organized crime and terrorism activities. Cortes et al. [2001], for example, propose data structures that are useful for detecting telecommunications fraud that are based on communities of interest, that use the fact that fraudulent account nodes tend to be closer to other fraudulent nodes than random accounts are to fraudulent account nodes. In other words, relatively few legitimate accounts are directly adjacent to fraudulent accounts.

In addition, understanding network dynamics is a task of extreme importance from a business perspective when the network itself is highly integrated in the business model. For companies that do marketing on social networks, it is clear that the network's structure and evolution are critical factors for success because they determine, for instance, how to execute the campaign (i.e., deciding how many and which nodes to target and where in the network). For companies that produce third-party applications that run on these platforms, a basic understanding of the network's structure can make a difference between the successful adoption of an application and a failure. In addition, models of graph evolution can be applied to provision services because knowing how the network is going to change allows businesses to make the right infrastructure investments. The techniques described in this section can also be used for knowledge discovery. Helander et al. [2007], for example, analyzed the social network and dynamics of interaction in the IBM Innovation Jam, a moderated online discussion between IBM worldwide employees and external contributors.

Finally, for the operators of social networking platforms it is crucial to have a clear understanding of how the network may be growing (or shrinking) and why, and detecting communities is crucial not just for offering advertising and new services, but also for growing the networks via friend suggestions: link prediction in online social networks is useful by itself as a service to the users of the network, to generate link

[13]http://www.ire.org/sna/.

recommendations (e.g., "people you may know"), in making product or service recommendations, and in marketing. Link prediction models can also be used to predict customer behavior in spreading information and adopting new services.

6. PROPAGATION AND VIRALITY

The study of the spread of influence through a social network has a long history in the social sciences. The first studies focused on the adoption of medical and agricultural innovations [Coleman et al. 1066; Valente 1955]. Later, marketing researchers investigated the "word-of-mouth" diffusion process for *viral marketing* applications [Bass 1969; Goldenberg et al. 2001; Maharajan et al. 1990; Jurvetson 2000]. The idea behind viral marketing is that by targeting the most influential users in the network we can activate a chain-reaction of influence driven by word-of-mouth, in such a way that with a very small marketing cost we can actually reach a very large portion of the network. Selecting these key users in a wide graph is an interesting learning task that has received a great deal of attention in recent years (more extensive surveys can be found in the paper of Wortman [2008] and in the Chapter 19 of the recent book by Easley and Kleinberg [2010]).

In the rest of this section we provide a brief overview of influence propagation and discuss related business applications. In particular, in Section 6.1 we discuss some work that provides evidence of influence propagation and viral phenomena in social networks. In Section 6.2 we present influence-propagation models and algorithms for maximizing the spread of influence, which is the basic computational problem behind viral marketing. In Section 6.3 we discuss the same problem but for the case of multiple competitive products. Finally, in Section 6.5 we discuss open research problems and we provide an overview of viral marketing applications in the real world.

6.1. Influence and Information Propagation Analysis

The idea of influence in social networks is rather straightforward: when users see their social contacts performing an action they may decide to perform the action themselves (e.g., people buy items their friends buy). Influence for performing an action, may come (i) from outside the social network, (ii) because the action is popular, or (iii) by the social contacts in the network [Friedkin 1998]. Influence from inside the social network can be leveraged for a number of applications, the most famous among which is viral marketing. Other applications include personalized recommendations [Song et al. 2006, 2007] and feed ranking in social networks [Samper et al. 2006]. Besides, patterns of influence can be taken as a sign of user trust and exploited for computing trust propagation in large networks and in P2P systems [Guha et al. 2004b; Ziegler and Lausen 2005b; Golbeck and Hendler 2006; Taherian et al. 2008].

While many of the applications mentioned earlier essentially assume that influence exists as a real phenomenon, questions have been raised on whether there is evidence of genuine influence in real social network data. Watts and Dodds [2007], Watts [2007], and Watts and Peretti [2007] challenge the very notion of influential users but argue that viral campaigns still can be effective if a large-enough seed set is targeted. The question of similarity versus social influence is also addressed by Hill et al. [2006], who use a matched-sampling approach to attempt to deal with it. In particular Hill et al. show that the social network can be used to target a particularly effective set, and that the neighbors of that set can be targeted explicitly, thus "guided" viral propagation can be created without needing social influence if there is data on the social network and data on adoption of the product or service in question.

Anagnostopoulos et al. [2008] have developed techniques for showing that influence may *not* be genuine: while there is substantial social correlation in tagging behavior it cannot be attributed to influence. Another work highlighting the importance of

separating influence-based contagion from homophily-driven diffusion is Aral et al. [2009] where it is observed that the former can be overestimated if not measured correctly. Moreover, the strength of the different factors affecting the propagation of a piece of information may vary depending on what type of information (e.g., news, or discussion topic) is being propagated [Aral et al. 2007].

On the other hand, many researchers have analyzed social network data to find patterns of influence in various domains.

One domain in which a lot of analysis has been done is the blogging and microblogging domain [Gruhl et al. 2004; Adar and Adamic 2005]. Gruhl et al. characterize four categories of individuals based on their typical posting behavior within the lifecycle of a topic, then they develop a model for information diffusion based on the theory of the spread of infectious diseases capturing how a new topic spreads from blog to blog [Gruhl et al. 2004]. They also devise an algorithm to learn the parameters of the model based on real data, and apply the algorithm to blog data, thus being able to identify particular individuals who are highly effective at contributing to the spread of infectious topics. Backstrom et al. [2006] show that bloggers are more likely to join a group that many of their friends joined, especially if those friends belong to the same clique. Similar studies have been performed for the blogosphere: Song et al. [2007] show that blogs are likely to link to content that other blogs have linked to, while Agarwal et al. [2008] study the problem of identifying influential bloggers. Cha et al. [2010] analyzed Twitter data and concluded that the number of followers is not a metric of influence, when influence is defined on the basis of number of retweets that one user's posts receive.

In another domain, Leskovec et al. discover patterns of influence by studying person-to-person recommendations for books and videos, finding conditions under which such recommendations are successful [Leskovec et al. 2006, 2007a], and Cha et al. [2009] analyze how photo popularity is distributed across the Flickr social network, characterizing the role played by social links in information propagation. Their analysis provides empirical evidence that the social links are the dominant method of information propagation, accounting for more than 50% of the spread of favorite-marked photos. Moreover, they show that information spreading is limited to individuals who are within close proximity of the uploaders, and that spreading takes a long time at each hop, contrary to the common expectations about the quick and wide spread in the word-of-mouth effect. Lerman and Jones [2006] also show that the photos users view in Flickr are often the ones they can observe their friends consuming.

An additional piece of support on the hypothesis that network linkage can directly affect product/service adoption is presented by Hill et al. [2006], who analyze the adoption of a new telecommunications service and show that it is possible to predict with a certain confidence whether a customer will sign up for a new calling plan once one of their phone contacts does the same.

Aral and Walker [2010] is a study that measures the effect of adding "viral" features to a product in the diffusion of such product. Viral product features are basically of two types: (a) personalized referrals, including easy ways of inviting your friends to use the product (b) automatic broadcasting, meaning whenever you use the product you automatically post an update or send an email so that other people that are your friends know about this. Aral [2010] is a list of open research questions related to product diffusion using "viral" features.

Bakshy et al. [2009] present an empirical study of user-to-user content transfer occurring in the context of a time-evolving social network in Second Life, a massively multiplayer virtual world. They identify and model social influence based on the change in adoption rate following the actions of friends and find that the social network plays a significant role in the adoption of content. Their study also highlights that sharing

among friends occurs more rapidly than sharing among strangers. Moreover, some users play a more active role in distributing content than others, but these influencers are distinct from the early adopters.

Crandall et al. [2008] analyze the interactions between social influence and user similarity over the social networks of Wikipedia and LiveJournal editors. Their work confirms a feedback effect between users' similarity and social influence, and that combining features based on social ties and similarity is more predictive of future behavior than either social influence or similarity features alone. In other words, their work suggests that both social influence and one's own interests are drivers of future behavior and that they operate in relatively independent ways.

Finally, Lahiri et al. [2008] find that influential users and influence itself are both very sensitive to structural changes in the network.

6.2. Influence Maximization

Consider a social network in which we have accurate estimates of reciprocal influence among users. Suppose now that we want to launch a new product in the market, and consider that in a campaign we can target an initial set of users in order to advertise the product. The data mining problem of *influence maximization* is to select the initial set of users so that they eventually influence the largest number of users in the social network.

The first to consider the propagation of influence and the problem of identification of influential users from a data mining perspective were Domingos and Richardson [2001; Richardson and Domingos 2002]. In that work the problem is modeled by means of *Markov random fields* and heuristics are given for choosing the users to target. The function to maximize is the global expected lift in profit, that is, intuitively, the difference between the expected profit obtained by employing a marketing strategy and the expected profit obtained using no marketing at all [Chickering and Heckerman 2000]. A Markov random field is an undirected graphical model representing the joint distribution over a set of random variables, where nodes are variables and edges represent dependencies between variables. It is adopted in the context of influence propagation by modeling only the final state of the network at convergence as one large global set of interdependent random variables.

Kempe et al. [2003] approach the problem using discrete optimization methodology and they obtain approximation algorithms for two preexisting models coming from mathematical sociology, namely, the *linear threshold model* and the *independent cascade model*. Kempe et al. showed that for the two aforementioned propagation models the influence maximization problem is NP-hard. On the other hand, they argued that the objective function of influence spread is *monotone* and *submodular*, and thus a greedy algorithm gives a constant-factor approximation for the problem.

Leskovec et al. study the propagation problem from a different perspective, namely *outbreak detection*: how to select nodes in a network in order to detect the spread of a virus as fast as possible? They present a general methodology for near-optimal sensor placement in these and related problems [Leskovec et al. 2007]. They also prove that the influence maximization problem of Kempe et al. [2003] is a special case of their more general problem definition. By exploiting submodularity they develop an efficient algorithm based on a "lazy forward" optimization in selecting new seeds, achieving near-optimal placements while being 700 times faster than the simple greedy algorithm. Regardless this big improvement over the basic greedy algorithm, their method still faces serious scalability problems as shown in Chen et al. [2009]. In that paper, Chen et al. improve the efficiency of the greedy algorithm and propose new degree discount heuristics that produce influence spread close to that of the greedy algorithm but much more efficiently.

Tang et al. introduce the novel problem of topic-based social influence analysis Tang et al. [2009]. They propose a *topical-affinity propagation* approach to describe the problem using a graphical probabilistic model. They also deal with the efficiency problem by devising a distributed learning algorithm under the map-reduce programming model.

Ever-Dal and Shapira [2007] study the influence maximization problem under the so-called *voter model*, which is one of the most basic and natural probabilistic models to represent the diffusion of opinions in a social network [Clifford and Sudbury 1973; Holley and Liggett 1975]. In the voter model, the social network is an undirected graph with self-loops. At each time step, each node chooses one of its neighbors uniformly at random and adopts its opinion. The voter model is similar to the threshold model as it has the same property that a person is more likely to adopt the opinion which is held by most of his neighbors, but it is very different as it allows nodes to change opinion. This makes the voter model more suitable in scenarios in which progressiveness is undesirable (e.g., studying phenomena such as infection processes) and has the nice property that it is guaranteed to converge to a consensus (either everyone chooses the new action A or everyone chooses the incumbent action B) with probability 1. Even-Dar and Shapira [2007] show that when the cost of marketing to each individual in the network is the same, the obvious heuristic solution of marketing to those individuals with the highest degree is in fact optimal in this setting, and give a fully polynomial-time approximation scheme that works when this is not the case.

The voter model can also capture the case of different target times while previous models [Kempe et al. 2003] considered only the status of the network in the limit case of convergence to the steady state. Another advantage of the voter model is that it naturally captures viral marketing in a competing environment scenario, which is the topic of the next subsection.

Ienco et al. [2011] introduce the meme ranking problem, where meme refers to brief text updates or micromedia such as photos, video, or audio clips. The problem requires to select which k memes (among the ones posted their contacts) to show to users when they log into the system. The objective is to maximize the overall activity of the network, that is, the total number of reposts that occur. This problem is in a sense the converse of the influence maximization problem. In the latter, it is given a single piece of information and the problem is that of identifying k users from which to start the propagations so to maximize the expected spread. Oppositely in the meme ranking problem it is given a single user and we want to select k memes to show him in order to maximize the virality of the system.

6.3. Competitive Viral Marketing

The model by Kempe et al. assumes that there is only one player introducing only one product in the market [Kempe et al. 2003]. However, in the real world, it is more likely for multiple players to be competing with comparable products in the same market. For instance, in videogame consoles (X-Box versus Playstation), or reflex digital cameras (Canon versus Nikon) it is very unlikely for the average consumer to adopt more than one of the competing products. Thus it makes sense to formulate the influence maximization problem in terms of *mutually exclusive and competitive products*. Historically, competition between two products has largely been addressed from an economic modeling perspective and focused on areas such as market equilibrium. For example, in Arthur [1989] and David [1975], primarily network-independent properties are employed to model the propagation of two technologies through a market. Tomochi et al. [2005] offer a more game-theoretic approach which relies on the network for spatial coordination games. However, they do not address the problem of taking advantage of the social network and viral marketing when introducing a new technology into a market.

In the computer science literature, independently and concurrently two papers have approached this problem in 2007 [Bharathi et al. 2007; Carnes et al. 2007]. Bharathi et al. [2007] propose a natural extension of the independent cascade model for the competitive case. The model is related to competitive facility location and Voronoi games [Ahn et al. 2001; Cheong et al. 2004]. Bharathi et al. [2007] show that the last player to select the set of nodes to activate can apply the usual greedy algorithm to obtain a constant-factor approximation to the optimal strategy.

Carnes et al. [2007] also study the algorithmic problem of influence maximization in a competitive social network by what they call the "follower's perspective," that is, when the follower is the player trying to introduce a new product into an environment where a competing product is also being introduced, keeping itself hidden from a competitor until the moment of introduction. They assume that the company has a fixed budget for targeting consumers and knows who its competitor's early adopters are, and propose two alternative models for the diffusion of competitive products: the *distance-based model* and the *wave propagation model*. Both of these models reduce to the independent cascade model if there is no competition in the network. For both models Carnes et al. show that the decision version of the influence maximization problem under these models is NP-hard, but also that the corresponding influence function is nonnegative, monotone, and submodular. Thus they can apply the usual greedy algorithm to obtain a constant-factor approximation to the optimal strategy for the follower. Additionally, they generalize the allowed subsets to be limited based upon cost rather than simply size, hence allowing different costs to be associated with targeting different subsets of customers. They show that a company can obtain a larger market share than its unsuspecting competitor even if the competitor has a much larger marketing budget.

6.4. Churn

Churn is a business term that refers to the loss of customers. As such, it is of interest in many industries (financial, telecommunications, subscription services, etc.) and is probably the most important business application of social network analysis, particularly in industries in which the service being offered to consumers is strongly linked to their social network (e.g., telecommunications).

In general, churn is measured in terms of a rate that refers to the number of individuals leaving a customer base (e.g., measuring the number of individuals that leave their contracts, either to sign up with other companies or who simply rescind their contracts for other reasons). More recently, the term has been applied in a more general sense, to measure the number of customers that stop using any service.

From a business perspective, the goal of churn analysis is twofold. On one hand, it is to understand why customers churn so that appropriate customer relationship management measures can be taken, and on the other hand to predict individual churn so that appropriate measures can be taken. The measures can be financial (e.g., determining where to invest) or involve marketing (e.g., offers can be made to customers or customer segments predicted to churn).

In industries such as telecommunications, social networks play a major role because customers pay different fees depending on who they call. The implication of this is that customers often make service decisions based on the operators used by people in their network. From a business perspective, then, churn analysis encompasses many of the techniques discussed in this article: network structure has an influence on information propagation, which is related to influence, which in turn has a big impact on customer decisions to leave a service or to acquire it. With the advent of online social networks we expect churn prediction based on social network analysis and mining to gain importance.

Most of the work on churn analysis based on social networks to date has been done in the context of the telecommunications industry. For example, in Dasgupta et al. [2008] an activation algorithm is used to predict churn using social network analysis.

Customer churn affects the bottom line of all businesses, thus many of the business applications of social network analysis, from a business perspective, can be seen to converge on this particular problem. In preventing churn, for example, it is desirable to identify customer communities, identify influential nodes, and understand how information propagates.

6.5. Towards Viral Marketing for the Real World

The simple idea behind viral marketing is very attractive; as Watts and Peretti [2007] state, "it seems like the ultimate free lunch." However, influence propagation research has mainly focused on graph-theoretic approaches, assuming a propagation model, a graph with edges labeled by the probability with which a user's action will be influenced by neighbor's actions, and the optimization of an objective function. Unfortunately, many additional factors determine the outcomes of a campaign in the real world.

Finding the optimal marketing strategy, moreover, is known to be NP-hard. Hartline et al. [2008] propose a very simple marketing strategy, dubbed *influence-and-exploit* that is shown to be a good approximation of the optimal strategy. They argue that in the real world *revenue maximization* is a more natural objective than influence maximization and propose considering the sequence in which buyers are made offers, as well as the prices, so ideally influential buyers buy first, even if at lower prices. In their approach the item is given for free to a selected set of influential users, then randomly offered to the remaining buyers in a random sequence. The goal is to maximize the revenue that can be extracted from each buyer by offering the optimal price.

Along similar lines, Arthur et al. [2009] propose a model assuming a cascading propagation of sales through the network where the seller can use product price and "referral bonuses" to influence propagation. The idea is that recommendations from friends (who have incentives) are more effective than direct marketing by advertisers. The cascade model assumed is a natural extension of both linear threshold and independent cascade models.

In order to develop effective viral marketing solutions in the real world, it is important to take advantage of the information recorded in past action logs to detect the real extent of influence and propagation mechanisms.

Goyal et al. [2008, 2009] mine logs to discover frequent patterns of influence to identify the leaders and their tribes of followers in a social network. Log mining has also been used to determine the parameters of the influence maximization problem a la Kempe et al. [2003]. Saito et al. [2008] formally define the likelihood maximization problem and then apply a EM algorithm to solve it, but at each iteration the influence probability associated to each edge is updated so the approach is not scalable. Goyal et al. [2011] propose a variety of probabilistic models of influence showing that all of them satisfy submodularity while all, with the exception of one, satisfy *incrementality*, which is a desirable property for efficient computation. They also introduce the temporal dimension in the models, and show that the proposed time-dependent models can predict the time at which a user will perform an action with a very good error margin.

Kim and Srivastava [2007] study how social influence data can be used by e-commerce Web sites to aid the user decision-making process. They also provide a summary of technologies for social network analysis and identify the research challenges of measuring and leveraging the impact of social influence on e-commerce decision making.

Buzz-based recommender systems analyze query logs in e-commerce platforms in order to detect bursts in query trends. These bursts are linked to external entities like news and inventory information to find the queries currently in demand. A simple

system for buzz-based recommendation in the context of eBay is presented in [Nguyen et al. 2008]. The system follows the paradigm of limited quantity merchandising, in the sense that on a per-day basis the system shows recommendations around a single buzz query with the intent of increasing user curiosity and improving activity and stickiness on the site.

7. SOCIAL NETWORKING

In several of the business process categories discussed in this article, the use of social networking platforms is an important strategy. In fact, currently the deployment of social networking platforms is perhaps the most widespread contribution of social networks to businesses. Although the activities that fall under the umbrella of social networking could be viewed as separate from analysis and mining, we foresee many research opportunities because ultimately, tools built using the techniques described in this article could support many of the social networking activities and have significant business impact.

The white paper published by AT&T [Demailly and Silman 2008] identifies 10 opportunities and challenges of social networking. Based on that white paper, we highlight the following examples of ways in which social networking tools can be leveraged for business purposes:

—promotion of products and services in online social networks;
—trend monitoring;
—mechanisms for interaction with customers;
—research of new product ideas;
—creation and follow-up of customer user groups;
—advertising;
—sponsoring of interactive content;
—creation and monitoring of online focus groups.

In the current social networking paradigm people in the organization, for the most part manually, undertake the tasks described (e.g., marketing agents may promote products in online social network sites by manually posting information, monitoring, or responding to customer complaints). As pointed out in Demailly and Silman [2008], however, the potential business impact of social networking is wide and covers different dimensions.

In particular, given the volumes of data and quick spread of information in online social networks, it is clear that the creation of tools for some of the tasks mentioned before would significantly simplify the social networking process, lowering costs and, if effective, contributing to more streamlined and effective networking.

The authors of Demailly and Silman [2008] predict what they consider the most important opportunities for social networking business impact in the corporate world. We highlight how the techniques discussed in our article can contribute to success in seizing the opportunities outlined by Demailly and Silman [2008].

—*"Corporations will change the way they communicate; being visible and personalizing communication are the silver bullets."* Techniques for expert finding and mining can be used to make communication between companies and customers much more effective.
—*"Corporations will change their vision, defining a strategy of unified collaboration and communication: employees will rely more on the enterprise culture, and search for it."* This implies providing tools for social search and analysis within the corporation will be crucial.

—*"Corporations will change their organization, managers will need to adapt and become social networking evangelists: the IT group will need to work much more closely with knowledge managers and users to enable new applications."* This means that effective mining tools to gain knowledge insights internally and externally will be of great importance.
—*"Collective intelligence and customer experience will lead innovation, the process of collective innovation needs to be formalized and customer needs should be anticipated."* This means that social mining techniques to identify novel collective ideas (from employees, customers, and partners) and facilitate collective thinking are likely to have a strong impact.
—*"Networking will be key to employee excellence: as social networks open access to multiple advisors and mentors, the networks can be used as important employee development tools, and mobility will be increased."* For this opportunity techniques for social search, expert finding, and reputation will play a significant role.
—*"Corporations will adapt their motivation and career path systems, which will develop through collaboration and social influence."* Again, mining tools used internally will facilitate this shift.
—*"Intranets will become richer, personalizable, presence features and user rating will invade almost every application."* This implies that expert finding and reputation techniques will be important, as will be tools for generating recommendations.
—*"Social networking may allow increased revenue."* The enterprise will be more visible and accessible to its market, so a new strategy may allow the following:
  —*"expanded reach"*
  —*"conversion of direct marketing from static to dynamic to better targeting prospects"*
  —*"transformation of CRM in personalizing the contact with customers and reconnecting Web, call center, and online service centers for better customer experience and retention"*
  —*"facilitation of external channel management."*

We close this section by emphasizing that the last few items predicted in Demailly and Silman [2008] are poised to have perhaps the strongest business impact. Interestingly, upon close examination it is fairly straightforward to map the techniques described in this article to the foreseen changes.

8. CHALLENGES

We have painted a very positive outlook for social network analysis and mining from a business perspective, and given an overview of the technical areas we consider most relevant to future business impact. But the field is really still in its infancy, and there are many challenges, on one hand technical, and on the other hand human and social. We highlight a few in each area.

Technical Challenges. Each of the topics covered in this article contains a number of research issues. We highlight those that we think are most relevant in terms of business impact.

—Data preparation. In spite of many advances in interoperable standards, open-source, and Web-friendly formats, large-scale data management in most organizations remains inefficient at best and often nonexistent at worst. Technical challenges include development of methods to facilitate streamlining of data management and reuse (cleaning, documentation, annonymization, etc.)
—Network dynamics. The majority of early work on social networks assumed static networks. But networks are constantly evolving, and from a business perspective, being able to react to changes quickly is crucial. However, research in this domain is

still young so a lot more work needs to be done in creating network evolution models and in understanding how such evolution impacts particular business goals.
—Reputation, trust, and expertise. From a technical standpoint the challenges here include accurate user modeling (to be able to properly match experts to tasks), and accurate rating methods (to properly assign reputation scores), among others.
—Propagation and virality. Developing accurate propagation models is crucial in effectively taking business actions in the social networking space (e.g., marketing, etc.). Although there has been some interesting work in this direction, this is by far the area of which we know the least: it is largely unclear why certain information propagates while other information does not, measuring influence remains a difficult task (in large part because all social network data is partial), and successful application of models depends on a number of external factors that are difficult to quantify.
—Evaluation. In many cases it is difficult to choose an evaluation metric on a principled way, as often data cannot be shared, and even if it is publicly available, collecting ground truth is difficult. Similarly, the business impact of applying social network analysis techniques can be measured (e.g., in financial terms), but given that there are so many actors and external factors involved, it is unclear how results from one experiment can be generalized or how benchmarking can be accurately performed.

Human and Social Challenges. Social network analysis and mining inherently require an interdisciplinary approach at every level. While, as stated in the Introduction, many of the approaches consider problems in this domain by abstracting them to graphs, when it comes to business the application of these tools has to be informed by a clear understanding of the role that human issues play. These include the following.

—Cultural factors. While it is recognized that culture (corporate and otherwise) is likely to be a factor, it is unclear how to quantify it and how to incorporate it in the design of algorithms and systems.
—Privacy expectations. Privacy is by no means a static, objective concept, and expectations vary depending on the situation, the individual, or organization, etc. It is unclear how to quantify these differences and how to make the right balance.
—Legal and ethical issues. Laws regarding data vary from country to country and even across industries, in some cases placing severe limitations on what can be done and in others insufficiently protecting individuals. Ethical policies within the enterprise have to be designed and communicated in a way that transcends and impacts all of the technical work.
—Community structure. Although in many applications links are explicitly defined, there are many open issues starting with a better understanding of what *really* constitutes a link, how such links are to be interpreted (e.g., what frequency or type of email contacts imply "friendship"), and at what level or levels communities and subcommunities need to be considered.

In summary, while there are many opportunities for social network analysis and mining, both in terms of technical research and for business impact, the field is still very young. Its development in terms of having practical impact will require the careful integration of techniques and views from multiple disciplines.

9. CONCLUSIONS AND FUTURE WORK

In this article we provided an overview of what we consider key problems and techniques in social network analysis from the perspective of business applications. We started by outlining each area of research in the context of a specific business processes classification framework (The APQC process classification framework), and then focused on several areas, giving an overview of the main problems and describing

state-of-the-art approaches. We discussed data acquisition and preparation, trust, expertise, community structure, network dynamics, and information propagation. In each case we highlighted the main business application areas. Finally, we highlighted business impact opportunities as well as future research directions.

Social network analysis and mining constitute a very large, interdisciplinary area of study that is evolving fast. Therefore, our analysis is by no means complete. However, our goal in this article is to provide an overview of the main technical research areas in relation to business impact.

Future work will focus on going deeper into examining the relationship between the techniques described and existing processes. This may include a mapping for one or two particular industries and specific applications.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their feedback. In addition, we would like to thank Sinan Aral, Francoise Soulie Fogelman, Foster Provost, and Duncan Watts for their additional comments.

REFERENCES

ABROL, M., MAHADEVAN, U., MCCRACKEN, K., MUKHERJEE, R., AND RAGHAVAN, P. 2002. Social networks for enterprise webs. In *Proceedings of the International World Wide Web Conference (WWW'02).*

ADAR, E. AND ADAMIC, L. A. 2005. Tracking information epidemics in blogspace. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'05).*

AGARWAL, G. AND KEMPE, D. 2008. Modularity-Maximizing network communities via mathematical programming. *Euro. Phys. J. B 66*, 3.

AGARWAL, N., LIU, H., TANG, L., AND YU, P. S. 2008. Identifying the influential bloggers in a community. In *Proceedings of the 1st International Conference on Web Search and Web Data Mining, (WSDM'08).*

AGGARWAL, C. C. AND YU, P. S. 2005. Online analysis of community evolution in data streams. *In Proceedings of the SIAM International Data Mining Conference (SDM'05).*

AGGARWAL, G., MISHRA, N., AND PINKAS, B. 2004. Secure computation of the kth-ranked element. In *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT).*

AGRAWAL, R. AND SRIKANT, R. 2000a. Privacy-Preserving data mining. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data (SIGMOD).* 439–450.

AGRAWAL, R. AND SRIKANT, R. 2000b. Privacy-Preserving data mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data.*

AHN, H.-K., CHENG, S.-W., CHEONG, O., GOLIN, M. J., AND VAN OOSTRUM, R. 2001. Competitive facility location along a highway. In *Proceedings of the 7th Annual International Conference on Computing and Combinatorics, (COCOON'01).*

AHUJA, R. K., MAGNANTI, THOMAS L., AND ORLIN, J. B. 1993. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, Englewood Cliffs, NJ.

AMATRIAIN, X., JAIMES, A., OLIVER, N., AND PUJOL, J. M. 2010. *Data Mining Methods for Recommender Systems.* Springer, Chapter 2, 39–100.

ANAGNOSTOPOULOS, A., KUMAR, R., AND MAHDIAN, M. 2008. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08).*

ARAL, S. 2010. Identifying social influence: A comment on opinion leadership and social contagion in new product diffusion. SSRN eLibrary .

ARAL, S., BRYNJOLFSSON, E., AND VAN ALSTYNE, M. W. 2006. Information, technology and information worker productivity. SSRN eLibrary .

ARAL, S., BRYNJOLFSSON, E., AND VAN ALSTYNE, M. W. 2007. Productivity effects of information diffusion in networks. SSRN eLibrary.

ARAL, S., MUCHNIK, L., AND SUNDARARAJAN, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Nat. Acad. Sci. 106*, 51, 21544–21549.

ARAL, S. AND VAN ALSTYNE, M. W. 2010. Networks, information and brokerage: The diversity-bandwidth tradeoff. SSRN eLibrary.

ARAL, S. AND WALKER, D. 2010. Creating social contagion through viral product design: A Randomized trial of peer influence in networks. SSRN eLibrary.

ARTHUR, D., MOTWANI, R., SHARMA, A., AND XU, Y. 2009. Pricing strategies for viral marketing on social networks. In *Proceedings of the 5th International Workshop on Internet and Network Economics (WINE'09).* 101–112.

ARTHUR, W. B. 1989. Competing technologies, increasing returns, and lock-in by historical events. *Econ. J. 99,* 394.

BACKSTROM, L., DWORK, C., AND KLEINBERG, J. M. 2007. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the International World Wide Web Conference (WWW'07).* 181–190.

BACKSTROM, L., HUTTENLOCHER, D. P., KLEINBERG, J. M., AND LAN, X. 2006. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06).*

BAKSHY, E., KARRER, B., AND ADAMIC, L. A. 2009. Social influence and the diffusion of user-created content. In *Proceedings 10th ACM Conference on Electronic Commerce (EC'09).*

BALOG, K., AZZOPARDI, L., AND DE RIJKE, M. 2006. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM Press, New York, 43–50.

BALOG, K. AND DE RIJKE, M. 2007. Determining expert profiles (with an application to expert finding). In *Proceedings of the 20th International Joint Conference on Artifical Intelligence.* Morgan Kaufmann, San Francisco, CA, 2657–2662.

BARABASI, A. L. AND ALBERT, R. 1999. Emergence of scaling in random networks. *Sci. 286,* 5439, 509–512.

BASS, F. 1969. A new product growth model for consumer durables. *Manag. Sci. 15,* 215–227.

BERLINGERIO, M., BONCHI, F., BRINGMANN, B., AND GIONIS, A. 2009. Mining graph evolution rules. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD'09).* Lecture Notes in Computer Science, vol. 5781. Springer, 115–130.

BHARATHI, S., KEMPE, D., AND SALEK, M. 2007. Competitive influence maximization in social networks. In *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE'07).*

BLUM, A., DWORK, C., MCSHERRY, F., AND NISSIM, K. 2005. Practical privacy: The SuLQ framework. In *Proceedings of the ACM-SIGMOD Symposium on Principles of Database Systems (PODS).* 128–138.

BOLLOBAS, B. 1998. *Modern Graph Theory*. Springer.

BONCHI, F., GIONIS, A., AND TASSA, T. 2011. Identity obfuscation in graphs through the information theoretic lens. In *Proceedings of the International Conference on Data Engineering (ICDE'11).*

BORGWARDT, K. M., KRIEGEL, H.-P., AND WACKERSREUTHER, P. 2006. Pattern mining in frequent dynamic subgraphs. In *Proceedings of the IEEE International Conference on Data Mining.* 818–822.

BOYKIN, O. P. AND ROYCHOWDHURY, V. 2004. Personal Email networks: An effective anti-spam tool. Condensed Matter cond-mat/0402143.

BRANDES, U., DELLING, D., GAERTLER, M., GORKE, R., HOEFER, M., NIKOLOSKI, Z., AND WAGNER, D. 2008. On modularity clustering. *IEEE Trans. Knowl. Data Engin. 20*, 2, 172–188.

CAMPAN, A. AND TRUTA, T. 2008. A clustering approach for data and structural anonymity in social networks. In *Proceedings of the International Workshop on Privacy, Security and Trust in KDD (PinKDD'08).*

CAMPBELL, C. S., MAGLIO, P. P., COZZI, A., AND DOM, B. 2003. Expertise identification using email communications. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03).* ACM, New York, 528–531.

CARNES, T., NAGARAJAN, C., WILD, S. M., AND VAN ZUYLEN, A. 2007. Maximizing influence in a competitive social network: A follower's perspective. In *Proceedings of the 9th International Conference on Electronic Commerce (ICEC'07).*

CAVERLEE, J., LIU, L., AND WEBB, S. 2008. Socialtrust: Tamper-Resilient trust establishment in online communities. In *Proceedings of the Joint Conference on Digital Libraries (JCDL).* 104–114.

CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

CHA, M., MISLOVE, A., AND GUMMADI, P. K. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09).*

CHELLAPPA, R. AND JAIN, A. 1993. *Markov Random Fields: Theory and Application.* Academic Press, Boston, MA.

CHEN, W., WANG, Y., AND YANG, S. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09).*

CHEONG, O., HAR-PELED, S., LINIAL, N., AND MATOUSEK, J. 2004. The one-round voronoi game. *Discr. Comput. Geom. 31,* 1, 125–138.

CHICKERING, D. M. AND HECKERMAN, D. 2000. A decision theoretic approach to targeted advertising. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence (UAI'00).*

CHUNG, F. R. K. 1997. *Spectral Graph Theory.* American Mathematical Society.

CLAUSET, A., MOORE, C., AND NEWMAN, M. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature 453*, 98–101.

CLAUSET, A., NEWMAN, M. E. J., AND MOORE, C. 2004. Finding community structure in very large networks. *Phys. Rev. E ,* 1–6.

CLIFFORD, P. AND SUDBURY, A. 1973. A model for spatial conflict. *Biometrika 60*, 3, 581–588.

COLEMAN, J., MENZEL, H., AND KATZ, E. 1966. *Medical Innovations: A Diffusion Study.* Bobbs Merrill.

CORTES, C., PREGIBON, D., AND VOLINSKY, C. 2001. Communities of interest. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA'01).* Springer, 105–114.

COULL, S. E., MONROSE, F., REITER, M. K., AND BAILEY, M. D. 2009. The challenges of effectively anonymizing network data. In *Proceedings of the Cybersecurity Applications and Technology Conference For Homeland Security (CATCH '09).* 230–236.

CRANDALL, D. J., COSLEY, D., HUTTENLOCHER, D. P., KLEINBERG, J. M., AND SURI, S. 2008. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08).*

CRASWELL, N., DE VRIES, A. P., AND SOBOROFF, I., EDS. 2005. *The 14th Text Retrieval Conference, TREC 2005*. Information Technology Laboratory's NIST Special Publications, vol. 500-266. NIST.

CRASWELL, N., HAWKING, D., MARIE VERCOUSTRE, A., AND WILKINS, P. 2001. P@noptic expert: Searching for experts not just for documents. In *Proceedings of the AusWeb Conference.*

DALENIUS, T. 1977. Towards a methodology for statistical disclosure control. *Statistik Tidskrift 15,* 429–444.

DASGUPTA, K., SINGH, R., VISWANATHAN, B., CHAKRABORTY, D., MUKHERJEA, S., NANAVATI, A. A., AND JOSHI, A. 2008. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th International Conference on Extending Database Technology.* ACM, New York, 668–677.

DAVID, P. A. 1975. *Technical Choice, Innovation and Economic Growth*. Cambridge University Press.

DAVITZ, J., YU, J., BASU, S., GUTELIUS, D., AND HARRIS, A. 2007. ilink: Search and routing in social networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, New York, 931–940.

DEGENNE, A. AND FORSE, M. 1999. *Introducing Social Networks.* Sage Publications.

DEMAILLY, C. AND SILMAN, M. 2008. At&t white paper: The business impacts of social networking. White paper. http://www.business.att.com/content/whitepaper/WP-soc_17172_v3_11-10-08.pdf.

DENG, H., KING, I., AND LYU, M. R. 2008. Formal models for expert finding on dblp bibliography data. In *Proceedings of the IEEE International Conference on Data Mining*. 163–172.

DESIKAN, P. AND SRIVASTAVA, J. 2004. Mining temporally changing web usage graphs. In *Proceedings of the International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles (WebKDD'04).* 1–17.

DETICA. 2006. Detecting telecoms subscription fraud. Tech. rep., Detica Information Intelligence.

DOMINGOS, P. AND RICHARDSON, M. 2001. Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01).*

DOMINGOS, P. AND RICHARDSON, M. 2004. Markov logic: A unifying framework for statistical relational learning. In *Proceedings of the ICML'04 Workshop on Statistical Relational Learning and its Connections to Other Fields.* 49–54.

DUAN, D., LI, Y., JIN, Y., AND LU, Z. 2009. Community mining on dynamic weighted directed graphs. In *Proceedings of the 1st ACM International Workshop on Complex Networks meet Information and Knowledge Management.* ACM, New York, 11–18.

EASLEY, D. AND KLEINBERG, J. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

EVEN-DAR, E. AND SHAPIRA, A. 2007. A note on maximizing the spread of influence in social networks. In *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE'07).*

EVFIMIEVSKI, A., GEHRKE, J., AND SRIKANT, R. 2003. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the ACM-SIGMOD Symposium on Principles of Database Systems (PODS'03).* 211–222.

Evfimievski, A. V., Srikant, R., Agrawal, R., and Gehrke, J. 2002. Privacy preserving mining of association rules. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Faloutsos, M., Faloutsos, P., and Faloutsos, C. 1999. On power-law relationships of the internet topology. In *Proceedings of the ACM SIGCOMM Data Communications Festival.* 251–262.

Fawcett, T. and Provost, F. J. 1997. Adaptive fraud detection. *Data Min. Knowl. Discov. 1,* 3, 291–316.

Ferlez, J., Faloutsos, C., Leskovec, J., Mladenic, D., and Grobelnik, M. 2008. Monitoring network evolution using mdl. In *Proceedings of the International Conference on Data Engineering (ICDE'08).*

Flake, G. W., Lawrence, S., and Giles, C. L. 2000. Efficient identification of web communities. In *Proceedings of the International SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'00).*

Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. M. 2002. Self-Organization and identification of web communities. *Comput. 35,* 3, 66–71.

Fortunato, S. 2010. Community detection in graphs. *Phys. Rep. 486*, 3–5, 75–174.

Fortunato, S. and Barthelemy, M. 2007. Resolution limit in community detection. *Proc. Nat. Acad. Sci. 104,* 1.

Freeman, L. 2004. *A History of Social Network Analysis*. Empiric Press.

Friedkin, N. E. 1998. *A Structural Theory of Social Influence*. Cambridge University Press.

Getoor, L., Friedman, N., Koller, D., and Taskar, B. 2003. Learning probabilistic models of link structure. *Mach. Learn. 3*, 679–707.

Girvan, M. and Newman, M. E. J. 2002. Community structure in social and biological networks. *Proc. Nat. Acad. Sci. USA 99*, 12, 7821–7826.

Golbeck, J. and Hendler, J. 2006. Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Technol. 6*, 4, 497–529.

Goldenberg, J., Libai, B., and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Market. Lett. 12*, 3, 211–223.

Gomes, L. H., Almeida, R. B., Bettencourt, L. M. A., Almeida, V., and Almeida, J. M. 2005. Comparative graph theoretical characterization of networks of spam and legitimate email. http://www.arxiv.org/abs/cs.CR/0504012.

Goyal, A., Bonchi, F., and Lakshmanan, L. V. S. 2010. Learning influence probabilities in social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10).*

Goyal, A., Bonchi, F., and Lakshmanan, L. V. S. 2008. Discovering leaders from community actions. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'08).*

Goyal, A., On, B.-W., Bonchi, F., and Lakshmanan, L. V. S. 2009. Gurumine: A pattern mining system for discovering leaders and tribes. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE'09).*

Gruhl, D., Guha, R. V., Liben-Nowell, D., and Tomkins, A. 2004. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04).*

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. 2004a. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04).* ACM Press, New York, 403–412.

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. 2004b. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04).*

Gyongyi, Z., Garcia-Molina, H., and Pedersen, J. 2004. Combating Web spam with Trust-Rank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB).* Morgan Kaufmann, 576–587.

Hanhijarvi, S., Garriga, G., and Puolamaki, K. 2009. Randomization techniques for graphs. In *Proceedings of the SIAM Conference on Data Mining (SDM).*

Hartline, J. D., Mirrokni, V. S., and Sundararajan, M. 2008. Optimal marketing strategies over social networks. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08).*

Haveliwala, T. H. 2002. Topic-Sensitive pagerank. In *Proceedings of the 11th World Wide Web Conference*. ACM Press, 517–526.

Hay, M., Miklau, G., Jensen, D., Towsley, D. F., and Weis, P. 2008. Resisting structural re-identification in anonymized social networks. *Proc. VLDB*, 102–114.

Hay, M., Miklau, G., Jensen, D., Weis, P., and Srivastava, S. 2007. Anonymizing social networks. Tech. rep. 07, 19, University of Massachusetts.

Hel, M., Lawrence, R., Liu, Y., Perlich, C., Reddy, A., and Rosset, S. 2007. Looking for great ideas: Analyzing the innovation jam abstract. In *Proceedings of the International SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'07).*

HENZINGER, M. R., MOTWANI, R., AND SILVERSTEIN, C. 2002. Challenges in Web search engines. *SIGIR Forum 37*, 2.

HETTICH, S. AND PAZZANI, M. J. 2006. Mining for proposal reviewers: Lessons learned at the national science foundation. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM, New York, 862–871.

HILL, S., PROVOST, F., AND VOLINSKY, C. 2006. Network-Based marketing: Identifying likely adopters via consumer networks. *Statist. Sci. 21*, 2, 256–276.

HOLLEY, R. AND LIGGETT, T. 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann. Probab. 3*, 643–663.

HOROWITZ, D. AND KAMVAR, S. D. 2010. The anatomy of a large-scale social search engine. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, 431–440.

IENCO, D., BONCHI, F., AND CASTILLO, C. 2010. The meme ranking problem: Maximizing microblogging virality. In *Proceedings of the SIASP Workshop at IEEE International Conference on Data Mining (ICDM'10)*.

INOKUCHI, A. AND WASHIO, T. 2008. A fast method to mine frequent subsequences from graph sequence data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'08)*.

JURCZYK, P. AND AGICHTEIN, E. 2007. Hits on question answer portals: Exploration of link analysis for author ranking. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 845–846.

JURVETSON, S. 2000. What exactly is viral marketing? *Red Herr. 78*, 110–112.

KAISER, F., SCHWARZ, H., AND JAKOB, M. 2007. Expose: Searching the web for expertise. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 906.

KARYPIS, G. AND KUMAR, V. 1998. Multilevel algorithms for multi-constraint graph partitioning. In *Proceedings of the ACM/IEEE Conference on Supercomputing* (CDROM).

KATZ, L. 1953. A new status index derived from sociometric analysis. *Psychometrika 18,* 39–43.

KEMPE, D., KLEINBERG, J. M., AND TARDOS, E. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*.

KIM, Y. A. AND SRIVASTAVA, J. 2007. Impact of social influence in e-commerce decision making. In *Proceedings of the 9th International Conference on Electronic Commerce: The Wireless World of Electronic Commerce (ICEC'07)*.

KLEINBERG, J., PAPADIMITRIOU, C., AND RAGHAVAN, P. 2003. Auditing boolean attributes. *J. Comput. Syst. Sci. 6,* 244–253.

KLEINBERG, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM 46*, 5, 604–632.

KOREN, Y. 2003. On spectral graph drawing. In *Proceedings of the 9th International Computing and Combinatorics Conference*. Springer, 496–508.

KRAMER, R. M. 1999. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Ann. Rev. Psychol. 50*, 569–598.

KREBS, V. 2002. Uncloaking terrorist networks. *First Monday 7*, 4.

KUMAR, R., NOVAK, J., RAGHAVAN, P., AND TOMKINS, A. 2004. Structure and evolution of blogspace. *Comm. ACM 47,* 12, 35–39.

KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., SIVAKUMAR, D., TOMKINS, A., AND UPFAL, E. 2000. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society Press, 57–65.

LAHIRI, M., MAIYA, A. S., SULO, R., HABIBA, AND BERGER-WOLF, T. Y. 2008. The impact of structural changes on predictions of diffusion in networks. In *Workshops Proceedings of the 8th IEEE International Conference on Data Mining (Workshops of ICDM'08)*.

LANGVILLE, A. N. AND MEYER, C. D. 2003. Deeper inside pagerank. *Internet Math. 1*, 3, 335–380.

LAPPAS, T., LIU, K., AND TERZI, E. 2009. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, 467–476.

LAZER, D., PENTLAND, A., ADAMIC, L., ARAL, S., BARABASI, A., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., GUTTMANN, M., JEBARA, T., KING, G., MACY, M., ROY, D., AND ALSTYNE, M. V. 2009. Computational social science. *Sci, 323*, 5915, 721–723.

LERMAN, K. AND JONES, L. 2006. Social browsing on flickr. CoRR abs/cs/0612047.

LESKOVEC, J., ADAMIC, L. A., AND HUBERMAN, B. A. 2007a. The dynamics of viral marketing. *ACM Trans. Web 1*, 1.

LESKOVEC, J., BACKSTROM, L., KUMAR, R., AND TOMKINS, A. 2008. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference* textiton Knowledge Discovery and Data Mining. ACM, New York, 462–470.

LESKOVEC, J., KLEINBERG, J., AND FALOUTSOS, C. 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD)*. ACM, New York, 177–187.

LESKOVEC, J., KLEINBERG, J., AND FALOUTSOS, C. 2007b. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data 1*, 1, 2.

LESKOVEC, J., KRAUSE, A., GUESTRIN, C., FALOUTSOS, C., VANBRIESEN, J., AND GLANCE, N. S. 2007c. Cost-Effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*.

LESKOVEC, J., SINGH, A., AND KLEINBERG, J. M. 2006. Patterns of influence in a recommendation network. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'06)*.

LEVIEN, R. AND AIKEN, A. 1998. Attack-Resistant trust metrics for public key certification. In *Proceedings of the 7th USENIX Security Symposium*. 229–242.

LIBEN-NOWELL, D. AND KLEINBERG, J. 2003. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*. ACM Press, New York, 556–559.

LIU, K. AND TERZI, E. 2008. Towards identity anonymization on graphs. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 93–106.

LIU, Z., YU, J. X., KE, Y., LIN, X., AND CHEN, L. 2008. Spotting significant changing subgraphs in evolving graphs. In *Proceedings of the 8th International Conference on Data Mining (ICDM'08)*.

LLOYD, S. 1982. Least squares quantization in pcm. *IEEE Trans. Inf. Theory 28*, 2.

MAHAJAN, V., MULLER, E., AND BASS, F. 1990. New product diffusion models in marketing: A review and directions for research. *J. Market. 54*, 1, 1–26.

MARLOW, C. 2003. Classifying emergent communitites through diffusion. In *Proceedings of the Sunbelt International Social Networks Conference XXIII*.

MARTI, S. AND GARCIA-MOLINA, H. 2006. Taxonomy of trust: Categorizing P2P reputation systems. *Comput. Netw. 50*, 4, 472–484.

MELNIK, M. I. AND ALM, J. 2002. Does a seller's ecommerce reputation matter? Evidence from ebay auctions. *J. Industr. Econ. 50*, 3, 337–349.

MUI, L., MOHTASHEMI, M., AND HALBERSTADT, A. 2002. A computational model of trust and reputation. In *Proceedings of the 35th Hawaii International Conference on System Science (HICSS)*.

NARAYANAN, A. AND SHMATIKOV, V. 2009. De-Anonymizing social networks. In *Proceedings of the 30th IEEE Symposium on Security and Privacy*.

NEVILLE, J., SIMSEK, O., JENSEN, D., KOMOROSKE, J., PALMER, K., AND GOLDBERG, H. 2005. Using relational knowledge discovery to prevent securities fraud. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, 449–458.

NG, A. Y., JORDAN, M. I., AND WEISS, Y. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*. MIT Press, 849–856.

NGUYEN, H., PARIKH, N., AND SUNDARESAN, N. 2008. A software system for buzz-based recommendations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, 1093–1096.

O'MADADHAIN, J., HUTCHINS, J., AND SMYTH, P. 2005. Prediction and ranking algorithms for event-based network data. *ACM SIGKDD Explor. Newslett. 7*, 2, 23–30.

PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. 1998. The PageRank citation ranking: Bringing order to the Web. Tech. rep., Stanford Digital Library Technologies Project.

PANDIT, S., CHAU, D. H., WANG, S., AND FALOUTSOS, C. 2007. Netprobe: A fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, New York, 201–210.

PHITHAKKITNUKOON, S. AND DANTU, R. 2008. Adequacy of data for characterizing caller behavior. In *Proceedings of the 2nd Workshop on Social Network Mining and Analysis (SNA-KDD'08) in Conjunction with the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

PIOCH, N. J. AND EVERETT, J. O. 2006. Polestar: Collaborative knowledge management and sensemaking tools for intelligence analysts. In *Proceedings of the 15th ACM Inter-National Conference on Information and Knowledge Management*. ACM, New York, 513–521.

PROVOST, F. J., DALESSANDRO, B., HOOK, R., ZHANG, X., AND MURRAY, A. 2009. Audience selection for on-line brand advertising: Privacy-Friendly social network targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

PUJOL, J. M., SANGUESA, R., AND DELGADO, J. 2002. Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems*. ACM, New York, 467–474.

RAHM, E. AND DO, H. H. 2000. Data cleaning: Problems and current approaches. *IEEE Data Engin. Bull. 23*, 4, 3–13.

RESNICK, P., KUWABARA, K., ZECKHAUSER, R., AND FRIEDMAN, E. 2000. Reputation systems. *Comm. ACM 43*, 12, 45–48.

RICHARDSON, M. AND DOMINGOS, P. 2002. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02).*

RIZVI, S. AND HARITSA, J. R. 2002. Maintaining data privacy in association rule mining. In *Proceedings of the International Conference in Very Large Databases (VLDB).*

RONEN, I., SHAHAR, E., UR, S., UZIEL, E., YOGEV, S., ZWERDLING, N., Carmel, D., GUY, I., HAR'EL, N., AND KOIFMAN, S. O. 2009. Social networks and discovery in the enterprise (sand). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 836.

SAITO, K., NAKANO, R., AND KIMURA, M. 2008. Prediction of information diffusion probabili- ties for independent cascade model. In *Proceedings of the 12th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'08).*

SAMPER, J. J., CASTILLO, P. A., ARAUJO, L., AND GUERVOS, J. J. M. 2006. Nectarss, An rss feed ranking system that implicitly learns user preferences. CoRR abs/cs/0610019.

SCHIFANELLA, R., BARRAT, A., CATTUTO, C., MARKINES, B., AND MENCZER, F. 2010. Folks in folksonomies: Social link prediction from shared metadata. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM).*

SCOTT, J. 2000. *Social Network Analysis: A Handbook*. Sage Publications.

SEID, D. Y. AND KOBSA, A. 2002. *Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach*. MIT Press, Cambridge, MA, 327–358.

SIRIVIANOS, M., YANG, X., AND KIM, K. 2009. FaceTrust: Assessing the credibility of online personas via social networks. Tech. rep., Duke University. http://www.cs.duke.edu/~msirivia/publications/facetrust-tech-report.pdf.

SOGHOIAN, C. 2008. Widespread cell phone location snooping by nsa? http://voices.allthingsd.com/20080909/exclusive-widespread-cell-phone-location-snooping-by-nsa/.

SONG, X., CHI, Y., HINO, K., AND TSENG, B. L. 2007. Information flow modeling based on diffusion rate for prediction and ranking. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07).*

SONG, X., TSENG, B. L., LIN, C.-Y., AND SUN, M.-T. 2006. Personalized recommendation driven by information flow. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06).*

SUN, J., FALOUTSOS, C., PAPADIMITRIOU, S., AND YU, P. S. 2007. Graphscope: Parameter-Free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, 687–696.

SUN, J., TAO, D., AND FALOUTSOS, C. 2006. Beyond streams and graphs: Dynamic tensor analysis. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, 374–383.

TAHERIAN, M., AMINI, M., AND JALILI, R. 2008. Trust inference in web-based social networks using resistive networks. In *Proceedings of the 3rd International Conference on Internet and Web Applications and Services (ICIW'08).*

TANG, J., SUN, J., WANG, C., AND YANG, Z. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09).*

TANTIPATHANANANDH, C., BERGER-WOLF, T., AND KEMPE, D. 2007. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, 717–726.

TASKAR, B., WONG, M., ABBEEL, P., AND KOLLER, D. 2003. Link prediction in relational data. *Neural Inf. Process. Syst. 15.*

TOMOCHI, M., MURATA, H., AND KONO, M. 2005. A consumer-based model of competitive diffusion: The multiplicative effects of global and local network externalities. *J. Evolut. Econ. 15,* 273–295.

TRAVERS, J. AND MILGRAM, S. 1969. An experimental study of the small world problem. *Sociometry 32*, 4, p425–443.

VAIDYA, J., ZHU, Y. M., AND CLIFTON, C. 2006. *Privacy Preserving Data Mining*. Springer-Verlag.

VALENTE, T. 1955. *Network Models of the Diffusion of Innovations*. Hampton Press.

VIRDHAGRISWARAN, S. AND DAKIN, G. 2006. Camouflaged fraud detection in domains with complex relationships. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, 941–947.

WASSERMAN AND FAUST, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

WATTS, D. 2007. Challenging the influentials hypothesis. In *WOMMA Measuring Word of Mouth*, Vol. 3. 201–211.

WATTS, D. AND DODDS, P. 2007. Influential, networks, and public opinion formation. *J. Consum. Res. 34,* 4, 441–458.

WATTS, D. AND PERETTI, J. May 2007. Viral marketing for the real world. *Harvard Bus. Rev.,* 22–23.

WATTS, D. J. 2004. The "new" science of networks. *Ann. Rev. Sociol. 30*, 243–270.

WATTS, D. J., DODDS, P. S., AND NEWMAN, M. E. J. 2002. Identity and search in social networks. *Sci. 296*, 1302–1305.

WATTS, D. J. AND STROGATZ, S. H. 1998. Collective dynamics of 'small world' netwroks. *Nature 393,* 440–442.

WEST, D. 1996. *Introduction to Graph Theory*. Prentice Hall.

WHITE, S. AND SMYTH, P. 2005. A spectral clustering approach to finding communities in graph. In *Proceedings of the SIAM International Conference on Data Mining (SDM'05).*

WINKLER, W. E. 2003. Methods for evaluating and creating data quality. *Inf. Syst. 29,* 531–550.

WORTMAN, J. 2008. Viral marketing and the diffusion of trends on social networks. Tech. rep. MS-CIS-08-19, University of Pennsylvania. May.

WU, W., XIAO, Y., WANG, W., HE, Z., AND WANG, Z. 2010. k-Symmetry model for identity anonymization in social networks. In *Proceedings of the 13th International Conference on Extending Database Technology.*

YIMAM-SEID, D. AND KOBSA, A. 2002. Expert finding systems for organizations: Problem and domain analysis and the demoir approach. *J. Orgiz. Comput. Electron. Commerce 13*, 2003.

YING, X., PAN, K., WU, X., AND GUO, L. 2009. Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing. In *Proceedings of the 3rd SNA-KDD Workshop.*

YING, X. AND WU, X. 2008. Randomizing social networks: A spectrum preserving approach. In *Proceedings of the SIAM Conference on Data Mining (SDM'08).* 739–750.

YING, X. AND WU, X. 2009a. Graph generation with prescribed feature constraints. In *Proceedings of the SIAM Conference on Data Mining (SDM'09).*

YING, X. AND WU, X. 2009b. On link privacy in randomizing social networks. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'09).*

ZHANG, J., ACKERMAN, M. S., AND ADAMIC, L. 2007. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web.* ACM, New York, 221–230.

ZHELEVA, E. AND GETOOR, L. 2007. Preserving the privacy of sensitive relationship in graph data. In *Proceedings of the International Workshop on Privacy, Security and Trust in KDD (PinKDD'07).* 153–171.

ZHOU, B. AND PEI, J. 2008. Preserving privacy in social networks against neighborhood attacks. In *Proceedings of the International Conference on Data Engineering (ICDE'08).* 506–515.

ZIEGLER, C.-N. AND LAUSEN, G. 2005a. Propagation models for trust and distrust in social networks. *Inf. Syst. Front. 7*, 4-5, 337–358.

ZIEGLER, C.-N. AND LAUSEN, G. 2005b. Propagation models for trust and distrust in social networks. *Inf. Syst. Front. 7,* 4-5, 337–358.

Received May 2010; revised July 2010; accepted October 2010