Netw Spat Econ (2016) 16:545–578
DOI 10.1007/s11067-015-9288-4

# Organization Mining Using Online Social Networks

**Michael Fire · Rami Puzis**

Published online: 25 April 2015

**Abstract** Complementing the formal organizational structure of a business are the informal connections among employees. These relationships help identify knowledge hubs, working groups, and shortcuts through the organizational structure. They carry valuable information on how a company functions de facto. In the past, eliciting the informal social networks within an organization was challenging; today they are reflected by friendship relationships in online social networks. In this paper we analyze several commercial organizations by mining data which their employees have exposed on Facebook, LinkedIn, and other publicly available sources. Using a web crawler designed for this purpose, we extract a network of informal social relationships among employees of targeted organizations. Our results show that it is possible to identify leadership roles within the organization solely by using centrality analysis and machine learning techniques applied to the informal relationship network structure. Valuable non-trivial insights can also be gained by clustering an organization's social network and gathering publicly available information on the employees within each cluster. Knowledge of the network of informal relationships may be a major asset or might be a significant threat to the underlying organization.

**Keywords** Organizational data mining · Social network data mining · Social network privacy · Organizational social network privacy · Facebook · LinkedIn · Machine learning · Leadership roles

M. Fire (✉) · R. Puzis
Telekom Innovation Laboratories and Department of Information Systems Engineering,
Ben Gurion University of the Negev, Beer Sheva, PO Box 653, Israel
e-mail: mickyfi@bgu.ac.il

R. Puzis
e-mail: puzis@bgu.ac.il

## 1 Introduction

In recent years, online social networks have grown in scale and variety and today offer individuals the opportunity to publicly present themselves, exchange ideas with friends or colleagues, and network more widely. For example, the Facebook[1] social network has more than 1.32 billion monthly active users, with new users signing up each month (Facebook 2014). According to recent statistics published by Facebook (Constine 2013), on average 655 million Facebook users log onto this site on a daily basis, and more than 4.75 billion pieces of content are shared each day (web links, news stories, blog posts, notes, photo albums, etc.). On the personal level, social networks create new opportunities to develop friendships, share ideas, and conduct business. Many social network users expose personal details about themselves and their social connections via their profile pages (Acquisti and Gross R 2006; Boshmaf et al. 2011), as well as location data, sensitive business information, and details about their place of employment. On the global level, the abundance of information provides oportunities for mining data about almost any entity in our lives. For example, social network data was analyzed recently by Zhan et al. (2014) to infer urban land use.

In this study, we analyze publicly available social network data in order to infer the internal organizational structure of six high-tech companies of different scales. A similar analysis has been performed by Tyler et al. (2005) on the Hewlett-Packard organization. However, their analysis was based on protected organizational data, i.e., email logs. We show that it is possible to use only publicly available data, such as from Facebook and LinkedIn,[2] in order to achieve similar results for multiple organizations.

### 1.1 Our Approach in a Nutshell

In this work we employ the power of complex network analysis methods (Ducruet and Beauguitte 2014) for mining information about commercial companies. The mining methods proposed in this paper were applied to six well-known high-tech companies of various sizes, ranging from small companies with several hundred employees to large-scale companies with hundreds of thousands of employees. For each company, the mining process included three major steps. First, we acquired the organization's informal social network topology from publicly available information, as detailed in Section 3. As part of this process, we collected information about the company's structure as exposed by the company's employees on Facebook. The presented method for organizational data mining can yield a wide range of organization social network topologies which were not available to the research community in the past.

Next, we used different centrality measures to detect the hidden leadership roles inside each organization. In Section 4, we highlight the centrality measures with the

[1] http://www.facebook.com

[2] http://www.linkedin.com

highest accuracy in pinpointing the leaders. We additionally used machine learning algorithms to classify management roles in each organization.

In the third step, we used a state-of-the-art algorithm to cluster the organization's social network into disjoint communities, and we cross-referenced the disclosed leaders and communities with information obtained from LinkedIn (see Section 5). This enabled us to derive the roles of many communities within an organization, providing important insights about the organization's structure and communication patterns. Such insights included, for example, the relationships between divisions, and the assimilation patterns of employees from previously acquired companies. These details also highlight the need for organizations to be aware of their social networking vulnerability and to establish policies to control this exposure as necessary.

### 1.2 Contributions

The contributions of this paper are fourfold: First, we present a method for uncovering an organization's informal social network topology based solely on publicly available data. The crawled social network topology can later be utilized to investigate a wide range of organizational phenomena, such as to study the diffusion of information inside the organization (Chesney and Fire 2014); to uncover organizational structural problems, such as structural holes (Burt 1995) and fragile structures (Krackhardt and Hanson 1993); and to study how vulnerable organizations are to socialbot attacks (Elyashar et al. 2013; Paradise et al. 2014). Second, we use the organization's structure to discover hidden leadership roles within the organization. Pinpointing employees with leadership roles and analyzing these employees' inner organizational links can assist in constructing better working groups and improving the formal organizational structure. Third, we utilize the organization's structure to identify communities inside the organization. This could identify which communities are dominant or weak, and which ones are functioning well or poorly. Lastly, we perform a qualitative analysis of these leadership roles and communities and demonstrate that it is possible to obtain significant insights into the organization and the role of each community without having any access whatsoever to the organization's internal data.

### 1.3 Organization

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of previous relevant studies on social networks with a special focus on organizational social networks analysis. In Section 3, we describe the methods used to obtain the organizational social network structure, and we show the different organizational datasets obtained. In Section 4, we present methods for identifying an organization's leadership roles. Next, our methods used to discover the communities' roles inside each organization are described in Section 5. In Section 6, we discuss our obtained results. Then, in Section 7, we offer future research directions. Lastly, in Section 8, we present our conclusions.

## 2 Related Work

In this section, we describe previous work in the fields of online social networks and organizational social networks. We also provide an overview of studies that have used different types of data to reveal informal connections among an organization's employees in order to discover the organization's social network.

### 2.1 Online Social Networks

In recent years, the use of online social networks has grown exponentially. Online social networks such as Facebook, Twitter,[3] LinkedIn, Flickr,[4] and YouTube[5] serve millions of users on a daily basis. With this increased usage, new privacy concerns have been raised. These concerns result from the fact that online social network users frequently publish information about themselves and their workplaces. In 2007, a study carried out by Dwyer et al. (2007) determined that 100 % of people who participated in the study had used real names on their Facebook accounts and 98.6 % had added photographs of themselves to their Facebook accounts. Moreover, in 2011, Boshmaf et al. (2011) collected and analyzed more than 250GB of Facebook users' data and evaluated the amount of personal information exposed by users. They concluded that many Facebook users disclose detailed personal information, including date of birth, place of work, email address, relationship status, and phone number. By using publicly available data from Facebook and cross-referencing it with other public data sources on the web, such as Google[6] and LinkedIn, one can infer further details about a Facebook user, such as specific work experience and areas of expertise. For example, Pipl[7] and PeekYou[8] are able to search for information about a person across different social networks. These people search engines aggregate the obtained results and present a fully detailed personal profile.

In this study, we used publicly available data from Facebook in order to identify which Facebook users worked for a specific organization. We then cross-referenced the users' details with LinkedIn, Google search results, and the company's own web page in order to reveal the users' positions in the organization.

### 2.2 Organizational Social Networks

In the past six decades, a considerable amount of research has gone into analyzing and understanding communication patterns between individuals inside organizations. Jacobson and Seashore (1951) were among the first researchers to study communication patterns among federal agency employees throughout the levels of the

[3] http://www.twitter.com

[4] http://www.flickr.com

[5] http://www.youtube.com

[6] http://www.google.com

[7] http://pipl.com

[8] http://www.peekyou.com

management hierarchy, across and within work groups, and between people holding similar and complementary positions. Based on a series of questionnaires, Jacobson and Seashore proposed an approach to identify sub-groups within an organization, the communication patterns of these groups, the position of individuals in the communication structure, and the power structure of the organization. Today, after more than half a century, such information has not lost its importance, but now it can be obtained semi-automatically based on publicly available data as described in this paper.

In 1968, Pugh et al. (1968) studied five primary dimensions of the organizational structure of 52 different organizations in England. By utilizing these five dimensions, they succeeded in constructing a profile characteristic of the structure of an organization, in comparing it to other organizations, and in measuring the organizations' structural differences. In 1969, Allen and Cohen (1969) studied technical communication patterns and their influences within two research and development laboratories. By using questionnaires and personal interviews, they observed that the informal organization of the laboratory can occupy an important position in the transfer of information. Additionally, they noted that there were "sociometric 'stars' in the technical communication network who provided other members of the organization with information, either making greater use of individuals outside the organization or reading the literature more than other members of the laboratory." In 1979, Tichy et al. (1979) presented a method for analyzing organizations using a network framework which included many network structural properties, such as centrality, clustering, and density. Tichy et al. used this framework to perform a comparative analysis of two organizations with several hundred employees.

In 1993, Krackhardt and Hanson (1993) demonstrated the high importance of an organizational informal social network in better understanding the organization's inner workings. They presented several real-life scenarios in which analyzing an organization's informal social networks helped managers tackle unexpected problems, such as communication problems among employees. Moreover, Krackhardt and Hanson asserted that by utilizing the organizational informal social network, managers can identify and solve problems such as *fragile structures* in which members of a group communicate only among themselves or perhaps with employees in one other division, and *holes in the network* in which expected links between employees are missing.

In 2003, Campbell et al. (2003) presented algorithms for expertise identification using email communication patterns. In their study, they used relatively small sample of emails collected from two organizations, and they showed that "a graph-based algorithm that takes account of communication patterns does a better job of identifying who knows most about specific topics than a content-based algorithm." Since 2004, after the release of about 500,000 Enron employees' emails (Shetty and Adibi 2004), many researchers have utilized this internal email dataset to better understand the Enron Corporation's social network and to discover various insights about the organization (Diehl et al. 2007; Diesner et al. 2005; McCallum et al. 2005; Shetty and Adibi 2005; Wilson and Banzhaf 2009). For example, (Shetty and Adibi 2005) successfully tested their entropy model which can identify "the most interesting and important nodes" in a graph on the Enron dataset.

Previous studies have also demonstrated how organizational social network analysis can assist law enforcement units in fighting organized crime, conspiracies, and terror. In 1991, Sparrow (1991) presented a method for using social network structural analysis to better understand criminal organizations. In his study, Sparrow suggested using six centrality measures, such as degree, betweenness (Freeman 1977), and closeness to identify central, vital, key, or pivotal individuals in the criminal organization and to target them for removal or surveillance. In 1993, Baker and Faulkner (1993) analyzed the social organization of three well-known price-fixing conspiracies in the heavy electrical equipment industry. They discovered that "the structure of illegal networks is driven primarily by the need to maximize concealment, rather than maximize efficiency." Additionally, Baker and Faulkner utilized various individuals' features, such as individuals' management levels and centrality measures in the conspiracy network, to construct a model for predicting an individual's verdict (guilt or innocence), sentence, and received fine. In 2002, after the tragic events of September 11, 2001, Krebs (2002) primarily used publicly released information reported in major newspapers to study Al-Qaeda's organizational network structural properties and succeeded in identifying the conspiracy leader by using the degree and closeness structural properties of vertices.

In recent years with the increasing prevalence of online social network usage, many studies have addressed the use and benefits of both public and internal social networking services to organizations. In 2009, Steinfield et al. (2009) studied the connection between social capital and the use of social networking services deployed inside organizations. In the same year, (Rooksby et al. 2009) published a detailed report on how online social networks are used in the context of the workplace. Comprehensive reviews on organizational social networks can provide further insights (Kilduff and Tsai 2003; Provan et al. 2007; Kilduff and Brass 2010).

### 2.3 Discovering an Organization's Social Network

The work reported in this paper is closely related to a 2004 internal study on the Hewlett-Packard organization carried out by Tyler et al. (2005). By analyzing the organization's email corpus, which contained more than one million messages, they discovered the organizational social network topology and identified communities inside the organization. The authors used the betweenness-centrality measure to detect leadership roles within the organization. They also applied a version of the (Wilkinson and Huberman 2004) algorithm which partitions the organization's social network into communities. The results were evaluated by interviewing several employees about the community in which they were automatically placed by the community detection algorithm. Naddafa and Mutyalab (2010) presented a similar study. They demonstrated a method for extracting informal social networks based on employees' email records. They tested their method on a large public sector client and identified the authority of the employees by using the PageRank measure (Page et al. 1999). Moreover, Naddaf and Mutyala used the Fast Modularity algorithm (Clauset et al. 2004) to identify communities in their client's organizational social network.

As can be observed from the studies reviewed in this section, although researchers have been studying organizational structures for over sixty years, until now there have been no easy-to-implement methods of assessing the informal social network of an organization. So far, elicitation methods have included questionnaires (Allen and Cohen 1969), interviews (Allen and Cohen 1969; Pugh et al. 1968), analysis of newspaper articles (Krebs 2002), and analysis of internal email logs (Tyler et al. 2005; Shetty and Adibi 2005), which in most cases are troublesome and are unavailable to parties outside the investigated organization. In our current study, we present a data collection method for obtaining the social network of an organization using publicly available information only. Furthermore, we show that expertise, leadership, and the roles of communities can be identified using data sources of which an organization has no control.

## 3 Organization Social Network Crawler

Many different types of web crawlers have been developed to collect data from large scale online social networks (Mislove et al. 2007; Boshmaf et al. 2011; Gjoka et al. 2011; Fire et al. 2013). Social networks crawlers usually start from several seed profiles and gradually expand the set of acquired profiles using, for example, Breadth-First-Search (BFS) crawling or other methods, such as Random-Walks (Gjoka et al. 2011).

Unfortunately, standard social network crawling techniques are insufficient for performing data collection which focuses on a specific organization. During a preliminary study performed using BFS crawling, many irrelevant profiles were collected and Facebook users who worked in our targeted organization were often skipped.[9] To tackle the problem of targeted acquisition of profiles from online social networks, we developed an organization crawler which optimizes data collection from users associated with a specific group or organization. Our organization crawler utilizes the homophily principle (McPherson et al. 2001). According to the homophily principle, it is more likely that a person has been employed by a certain organization if many of his or her friends have been employed by the same organization as well.

In order to mine the social network for the profiles of employees from a selected target organization, our crawler worked according to the algorithm depicted in Algorithm 1. We will refer to this crawler as Version 1. The crawl starts from a set of seed profile pages initially identified as belonging to employees of the targeted organization. The initial set of seeds can be obtained using a search engine. The crawling proceeds by iteratively processing the profile pages having the highest number of

[9] Although, in theory, using a BFS crawler can return the target organization's complete social network, in practice, a BFS crawler is not usable for crawling organizational social networks due to a BFS crawler's low precision rates in identifying relevant profiles and due to many social networks providers' limitations on the number of page requests (Twitter 2013). Moreover, due to BFS algorithm properties, there can be cases in which the BFS crawler collects employee profiles with several thousands of Facebook friends, and even though most of these are not employees of the targeted organization, the crawler will still need to collect the profile pages of each one of these friends before continuing its crawling process and moving to more relevant profiles. Nevertheless, in cases where the organization is relatively small with a few dozen employees, a BFS crawler may be sufficient to crawl the entire organization.

social network friends employed by the target organization. The relevance of a profile to the organization was determined by matching a set of keywords against the semi-structured data that appears in the user's publicly available Facebook profile. For example, in order to identify users from Ben-Gurion University's Information System Engineering Department, the crawler searched for strings such as "Ben-Gurion Information System Engineering," "BGU ISE," or "ISE BGU" in the collected profile page.

We also evaluated an optimized version of the organization crawler which tracked the number of friends within the targeted organization for each user profile in the priority queue and also the number of organization employees discovered during the last iterations. We stopped the crawling process if all users in the priority queue had at most one friend in the targeted organization and if the last thousand profiles acquired from Facebook did not belong to those of the organization's employees. We will refer to the crawler with this stricter stopping condition as Version 2.

**Algorithm 1:** Organization Social Network Crawler (Version 1)

```
Input: A set of seed URLs (S) to Facebook profile pages of organization's employees.
A set of crawling organization target names, N.
Output: A set of Facebook profiles and their connections.
$Q \leftarrow$ Priority-Queue()
$\forall_{URL \in S}, Q.Enqueue(URL : 1)$
$Crawled \leftarrow \emptyset$
while ($Q \neq \emptyset$) do
    $URL \leftarrow Q.Dequeue()$
    $Crawled \leftarrow Crawled \cup \{URL\}$
    $Page \leftarrow DownloadProfilePage(URL)$
    if $Page\ contains\ N$ then
        F_URLs ← Extract list of friends from Page
        F_URLs ← F_URLs − Crawled
        for ( $F\_URL \in F\_URLs \cap Q$ ) do
            Increase priority (Q, F_URL)
        end
        for ( $F\_URL \in (F\_URLs - Q)$ ) do
            Q.Enqueue(F_URL:1)
        end
    end
end
return Collected pages
```

## 3.1 Ethical Considerations

During this study, we used our organization crawlers to collect a considerable amount of data from public sources regarding the studied organizations and their employees. To the best of our knowledge, Ben-Gurion University of the Negev regulations do not require explicit approval by an ethics committee for studies that involve publicly collected data. Nevertheless, in order to protect the privacy of the organizations' employees and the discovered confidential details of the organizations, we anonymized the organizations' names throughout this paper. Additionally, in the attached published datasets, we anonymized the employees' Facebook identities by randomly replacing the users' Facebook IDs with a series of contiguous integers.

**Table 1** Organization crawling results

| Org. | Crawler Version | #Total Crawled Profiles | #Org. Crawled Profiles | Precision |
|---|---|---|---|---|
| S1 | Version 1 | 22, 992 | 165 | 0.7 % |
| S2 | Version 2 | 3, 312 | 320 | 9.6 % |
| M1 | Version 1 | 11, 247 | 1, 429 | 12.7 % |
| M2 | Version 2 | 7, 422 | 3, 862 | 52.0 % |
| L1 | Version 1 | 13, 505 | 5, 793 | 42.9 % |
| L2 | Version 2 | 18, 810 | 5, 524 | 29.3 % |
| Total | – | **77,288** | **17,096** | **22.1** % |

## 3.2 Collected Organization Datasets

In order to test the methods of organization data collection reported in Section 3, we used our organization social network crawler to collect publicly available data from six commonly known high-tech companies.

The organization crawling results are depicted in Table 1, where all the organizations' data were obtained during 2012.

We grouped the companies based on their size: Small (S), currently employing 500 to 2,000; Medium (M), employing 4,000 to 20,000; and Large (L) having more than 50,000 employees. Data on one company of each scale was acquired using each version of the organization crawler. We refer to the three companies targeted by Version 1 and Version 2 of the crawler as S1, M1, L1; and S2, M2, and L2, respectively.

In the following subsections, we describe in detail the properties of each collected organization dataset (see Table 2). We used Cytoscape (Shannon et al. 2003) software to visualize the social networks formed by the employees of each organization.[10] The vertex colors in Figs. 1, 2, 3, 4, 5 and 6 represent various cluster roles, as will be explained in Section 5. The analysis results of these networks are reported in Sections 4 and 5.

### *3.2.1 Small Hardware Company (S1)*

The S1 company is a publicly held company that specializes in network hardware development. According to the company's web page, they employ 500 to 1,000 individuals and have one head office in North America and another in Asia. We used the organization crawler to identify 726 informal links among 165 Facebook users who, according to their Facebook page, worked for the company (see Fig. 1). We also collected information on 93 employee positions inside the company. Out of these 93 employees, we identified 21 in management positions. Most of the discovered

[10] All the organizations' graphs presented throughout this paper are embedded as Scalable Vector Graphics (SVG) images, which enable the reader to zoom in and view each node in each graph.

**Table 2** Collected organization datasets

| Org. | Size | Discovered Employees | Links | Employees Disclosing Positions on Facebook |
|---|---|---|---|---|
| S1 | 500–1K | 165 | 726 | 53(32.1 %) |
| S2 | 1K–2K | 320 | 2,369 | 104(32.5 %) |
| M1 | 2K–10K | 1,429 | 19,357 | 383(26.8 %) |
| M2 | 10K–20K | 3,862 | 87,324 | 1,529(39.6 %) |
| L1 | 50K+ | 5,793 | 30,753 | 1,599(27.6 %) |
| L2 | 50K+ | 5,524 | 94,219 | 1,131(20.5 %) |
| Total | – | **17,093** | **234,748** | **4,799**(**28.1** %) |

company employees held R&D positions, and most of the identified managers were R&D team leaders.

### *3.2.2 Small Software Company (S2)*

The S2 company is an international publicly held company that specializes in software development. According to public sources, the company has between 1,000 and

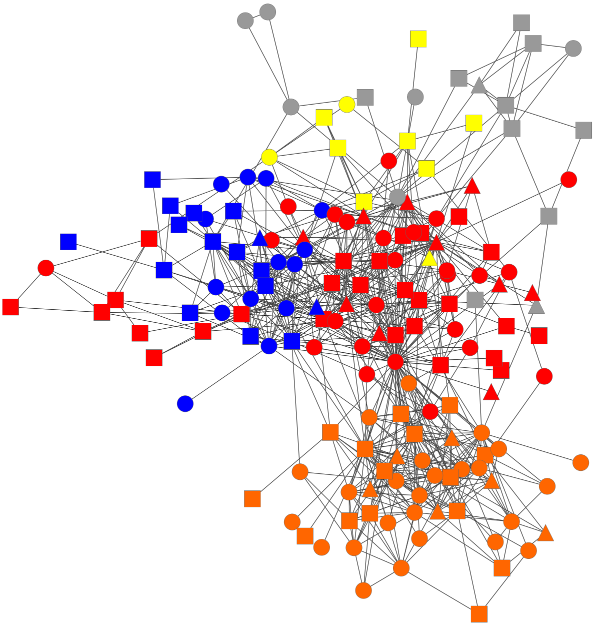

**Fig. 1** S1 Company: *Blue nodes* - R&D and administration groups in Asia. *Red nodes* - primarily hardware verification engineers and chip designers in Asia. *Yellow nodes* - Hardware R&D. *Orange nodes* - acquired startup company. *Gray nodes* - R&D in Asia

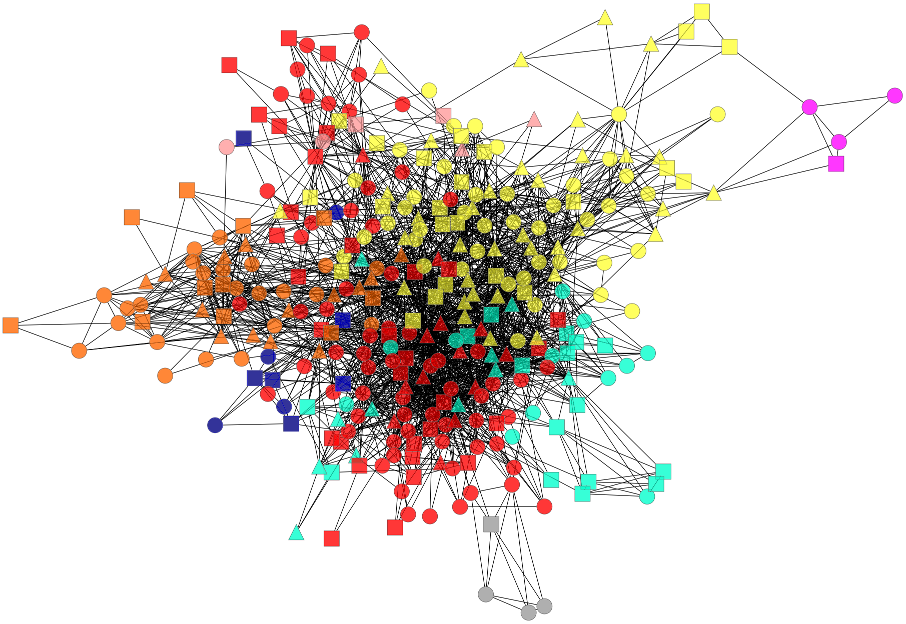

**Fig. 2** S2 Company: *Blue nodes* - IT group in the Middle East. *Red and Orange nodes* - R&D groups in the Middle East. *Purple nodes* - North American group. *Yellow nodes* - managers and international project managers. *Cyan nodes* - R&D teams in Australia and the Middle East. *Gray nodes* - European group

2,000 employees and maintains offices in North America, Europe, Asia, Australia, and the Middle East. We used our organization crawler to identify 2,369 informal links among 320 Facebook users who stated that they worked for the company in their Facebook profiles (see Fig. 2). We also collected information on the positions of 164 company employees. Out of these 164 individuals, 69 were in management positions. While many of the company employees held project manager (PM) positions, we also identified a number of developers, quality assurance (QA) positions, and support employees.

### *3.2.3 Medium Telecommunication Service Company (M1)*

M1 is an international technology company located in North America that specializes in telecommunication services. According to the company's web page, M1 currently has between 2,000 and 10,000 employees. We used the organization crawler and identified 19,357 informal links among 1,429 Facebook users who, according to their Facebook profile page, worked for the company (see Fig. 3). When we also collected information on the positions of 456 employees, we learned 225 held management positions. A wide range of positions inside the company were identified during the crawl: senior management positions, sales and marketing employees, PMs, developers, IT engineers, support engineers, technical writers, etc.

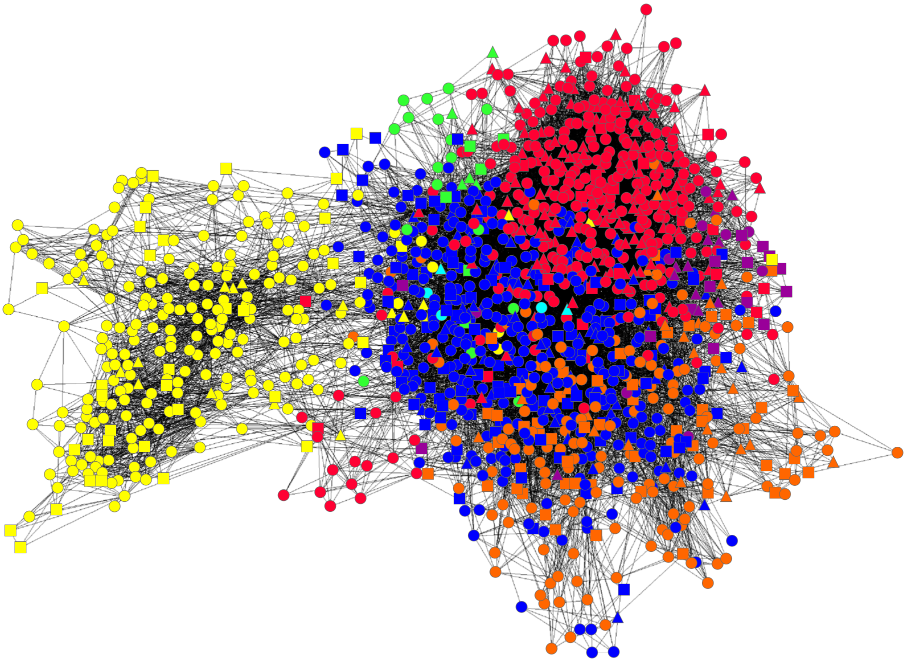

**Fig. 3** M1 Company: *Blue and Orange nodes* - R&D divisions. *Red nodes* - senior management. *Yellow nodes* - international consultants and support engineers. *Green nodes* - North American headquarter employees. *Purple nodes* - Middle East R&D and QA employees

### *3.2.4 Medium Software Provider and Outsourcing Company (M2)*

M2 is an international software and outsourcing provider that specializes in telecommunication services and serves a global customer base. The company's web page indicates its size as 10,000 to 20,000 employees. We used the organization crawler to focus on the company headquarters, located in South Asia. We stopped the crawling process after identifying 87,324 informal links among 3,862 Facebook users who state that they work for M2 in their Facebook profiles (see Fig. 4). We also succeeded in collecting information on the positions within the company for 1,510 employees. During the crawl, a variety of positions were identified: senior managers, developers, sales and marketing positions, IT, PMs, support engineers, technical writers, etc. Out of the 1,510 employees, 233 held management positions.

### *3.2.5 Large Information Technology Corporation (L1)*

L1 is an information technology corporation that provides products and services to customers around the world. As indicated on the company's web page, L1 currently employs more than 50,000 people. Our organization crawler collected data on corporation employees in North and South America, Asia, and Eastern Europe. We identified 30,753 informal links among 5,793 Facebook users who, according to their

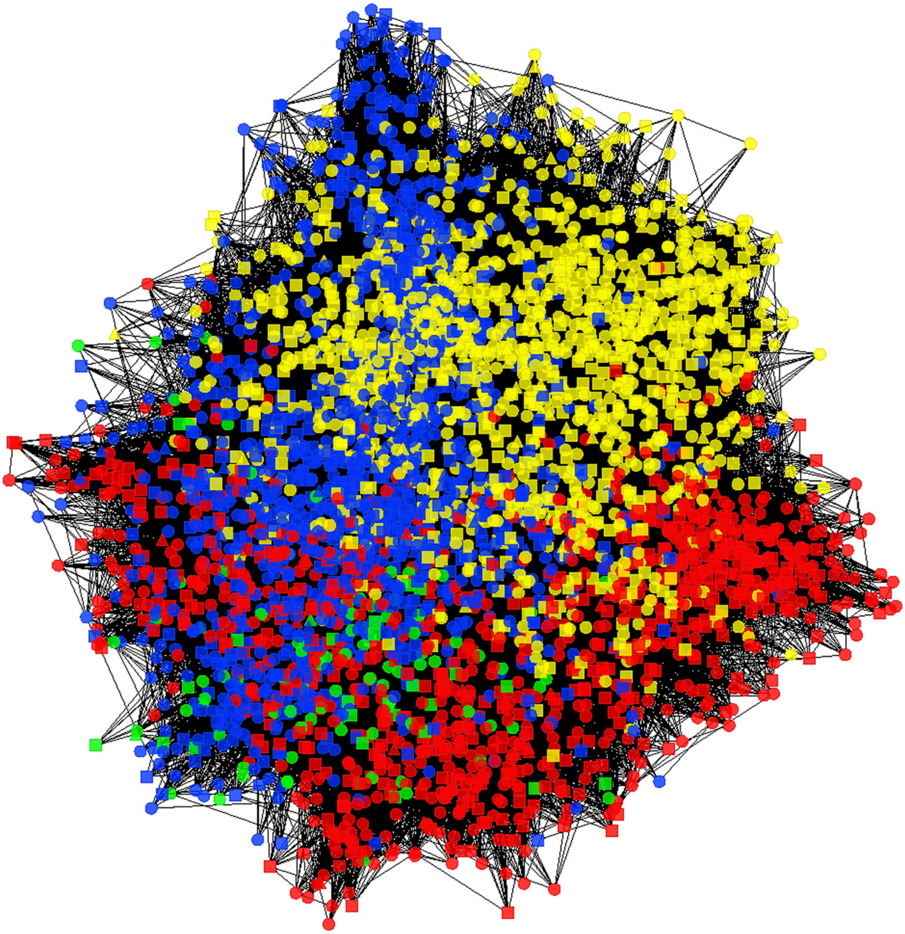

**Fig. 4** M2 Company: *Blue and Green nodes* - R&D and Specific Domain Experts (SDE) connected to North American and Asia employees. *Red nodes* - R&D and SDE connected to Australia, Europe and North America. *Yellow nodes* - R&D and SDE connected to Africa, North America, and Asia

Facebook profile page, worked for the corporation (see Fig. 5). We also were able to gather information on the positions of 1,619 employees. Out of these 1,619 employees, we succeeded in identifying 463 holding management positions. A broad range of positions were identified, spread throughout the world: senior managers, sales and pricing positions, marketing positions, technical writers, developers, IT, PMs, support engineers, etc.

### *3.2.6 Large Technology Corporation (L2)*

The L2 corporation provides hardware and software products, infrastructure, and other technology services to global customers. According to the company's web page, there are currently more than 50,000 employees. We used our organization

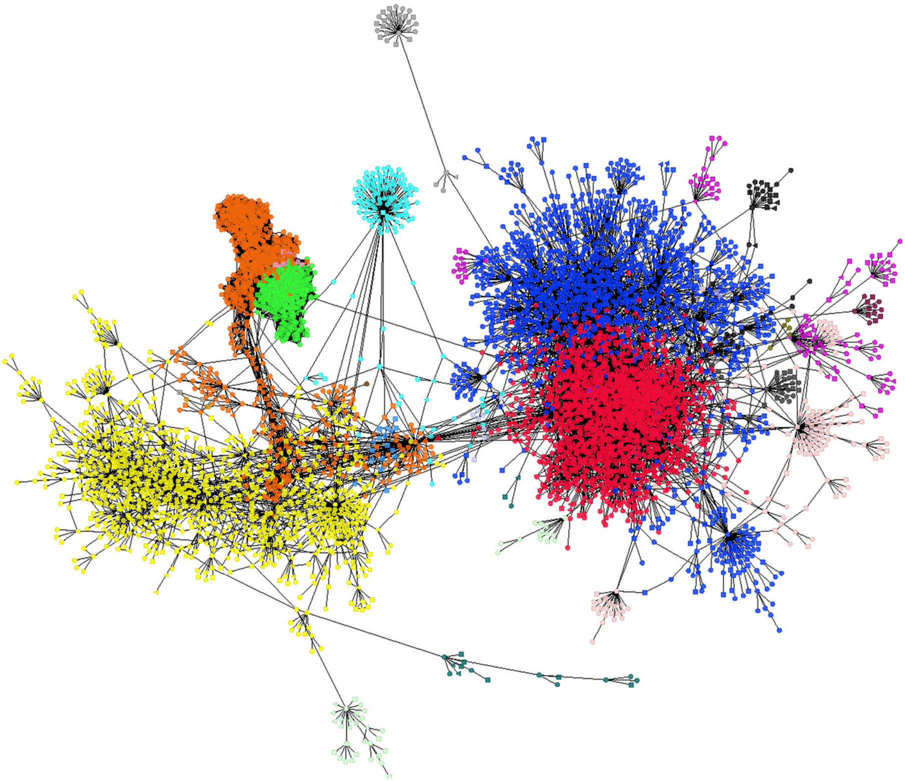

**Fig. 5** L1 Corporation: *Blue nodes* - South American support engineers. *Red nodes* - South American Branch (IT, support engineers, analysts, and PMs). *Orange nodes* - South Asia Analysts, Analysts and Eastern Europe Sales. *Yellow nodes* - Middle East R&D, Europe QA, Pricing. *Green nodes* - Eastern European Sales and Pricing. *Purple nodes* - East Asian - R&D. *Cyan nodes* - Management Positions. *Black nodes* - South American Analysts

crawler to accumulate data on corporation employees in North and South America, Asia, and Eastern Europe. We stopped the crawling process after identifying 94,219 informal links among 5,524 Facebook users who indicated on their Facebook profiles that they worked for the corporation (see Fig. 6). We also succeeded in collecting information on the company positions of 808 employees, out of which 398s held management positions. During the crawling, we found a wide range of positions inside the company: senior management positions, PMs, sales and marketing positions, developers, IT, support engineers, technical writers, etc.

## 4 Identifying Organizational Leadership Roles

After the organization crawler completes collecting data from the Facebook profiles of employees of a targeted organization, we can analyze the organizational social network created by the informal Facebook connections. Previous studies have illustrated

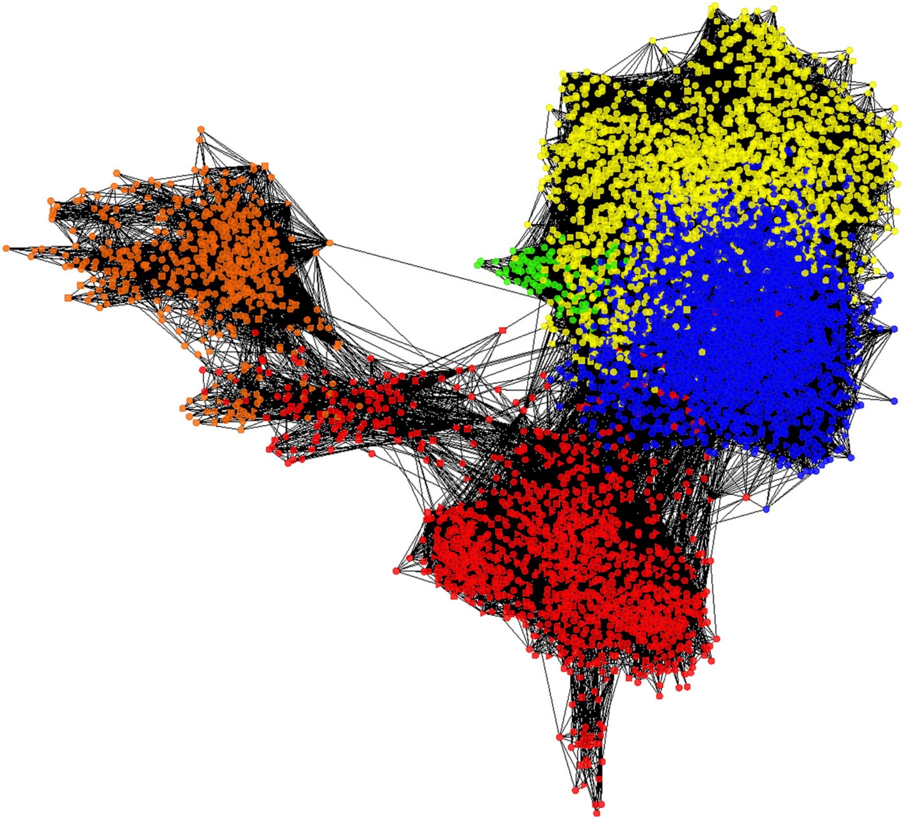

**Fig. 6** L2 Corporation: *Blue nodes* - East Asia Headquarter (management and consultants). *Red nodes* - international senior management and researchers. *Yellow nodes* - East Asian headquarters (R&Ds and consultants). *Green nodes* - the company's amateur sports team

that this type of organizational social network can be used to infer many interesting and important insights regarding how the organization functions (Krackhardt and Hanson 1993) and regarding specific employees' significance to the organization (Sparrow 1991; Tyler et al. 2005). In this section, we demonstrate that it is possible to pinpoint leadership roles solely by analyzing the structure of the informal social network of an organization's employees.

Let $G =< V, E >$ represent the informal social network, where node $v \in V$ is a Facebook user who worked in the target organization and $(u, v) \in E$ represents a Facebook friendship link between two users. To pinpoint leadership roles we performed the following steps: First, with the suggestion by Sparrow (1991) in mind – that central and vital key players in the organization can be detected by utilizing various centrality measures – we calculated eight centrality measures for each user $v \in V$ in the informal social network. Next, for each centrality measure, we examined the top 10, the top 20, and the top 50 users who received the maximal score. Then, to measure the precision of our centrality results, we reviewed the selected employees'

Facebook and LinkedIn profile pages and manually classified whether or not the user held a management position.[11] In many cases, however, the user's profile information was not sufficient to reveal the user's specific position inside the organization. To obtain more information, we cross-referenced the user's personal details with other publicly available online sources, such as the Google search engine. By using these methods, in many cases we succeeded in revealing the user's position within the organization and could then calculate precision values for our centrality measures.

After labeling a subset of organizational employees as managers and non-managers, we selected the centrality measure that resulted in the largest number of managers among the 50 most central users to investigate more deeply. We then extracted for each employee an additional 8 features, using the employee's organizational informal social network and analyzing the employee's Facebook profile. Lastly, we used several machine learning algorithms to build classifiers that can automatically identify management roles inside an organization. By using these classifiers, we can recall a wider range of management roles that answer complex centrality measures criteria. It is important to note that these types of classification methods can be used to compromise users' privacy by exposing non-public positions inside the organization. Furthermore, similar methods can assist in revealing various statistics about the organization, thereby disclosing and potentially compromising the organization's privacy. For example, using the above methods, we estimated the percent of management positions and the number of employees inside each organization (see Tables 2 and 4). In many privately held companies, this type of data may be confidential organizational information.

### 4.1 Manually Identifying Managers

To calculate the various precision values of the centrality measures and to construct machine learning classifiers which can predict the likelihood of an organization's employee holding a management role, we needed to manually classify a sufficient number of employees and determine if each employee held a management position or not.[12] We succeeded in manually classifying a sufficient number of employees by quickly reviewing the users' data extracted from their Facebook profiles. By analyzing the Facebook profiles of the crawled organizations' users, we discovered that an average of 28.1 % of the collected users had inserted at least partial information[13] about their previous and current employment positions into their Facebook

[11] In contrast to the (Baker and Faulkner 1993) study in which, based on the employees' titles and the company's organization charts, the company's employees were divided into three management categories (top executives, middle managers, or junior managers), we decided to divide each organization's employees only into two dichotomous groups – managers and non-managers – where to the best of our judgment employees in the manager's group held some type of management role in the organization, ranging from sales manager to the organization's CEO.

[12] In all six organizations, our methods discovered more than 17,000 employees. Therefore, to manually determine if each employee in each organization held a management role or not was impractical; consequently, we evaluated our algorithms on 4,650 manually classified employees' profiles.

[13] In this study, we considered position information to be partial if the description field in the employee's Facebook profile was not empty.

profiles (see Table 2). For each user who had included his or her previous or current work experience, we attempted to determine if the user held a management role inside the organization. In some cases we also did a deeper inspection of the user by cross-referencing the user's work experience with data obtained from other sources, such as LinkedIn. Using this method, we reviewed and classified 4,650 users' profiles (referred as *Manually Classified Employees* (MCE)). Out of the 4,650 manually classified positions, we identified 1,409 users who held management positions (see Table 4). These manually classified employees provided us with a ground truth group of managers in an organization, with which we could check the results of our centrality measures (Section 4.2) and machine learning algorithms (Section 4.3) as we proceeded in detecting hidden leadership roles within a company.

## 4.2 Centrality Measures

Using the organization datasets described Section 3.2, we proceeded to identify leadership roles within the organization using several centrality measures. For each node in the informal organization social network, we calculated eight centrality measures:[14] Degree centrality (DG), Closeness centrality (CL), Betweenness centrality (BC) (Freeman 1977), HITS (H) (Kleinberg 1999), PageRank (PR) (Page et al. 1999), Eigenvector centrality (EC) (Newman 2008), Communicability centrality (CC) (Estrada and Rodriguez-Velazquez 2005), and Load centrality (LC) (Newman and et al. 2001). We then sorted the crawled organization's users' list according to the different centrality measures. We inspected the top 50 user profiles according to each centrality measure in order to infer employees' positions within the target organization. Since a large fraction of Facebook users do not disclose their positions on their profile page, for many of profiles we used other online sources, such as LinkedIn or results returned by Google's search engine, in order to manually classify whether or not a particular employee held a management position. These classified employees provided us with a confirmed group of managers and non-managers in each organization, with which we could measure the various centrality measures' precision. We will refer to managers who do not report their position on Facebook as concealing their management position. Afterwards, for each centrality measure $C$, and for each organization $O =< V_O, E_O >$, we calculated the centrality measure's precision at $k(precision@k)$ in $O$ using the following equation:

$$precision_O@k(C) := \frac{|\{u_i \in Top^k_{C,O}|\text{Is-Manager}(u_i)\}|}{|\{Top^k_{C,O} \cap MCE_O\}|},$$

where $Top^k_{C,O}$ is defined as the top $k$ profiles which have the maximal $C$ measure values in $O$, $MCE_O$ are the manually classified employees in the organization $O$, and Is-Manager$(u)$ returns a true value if we identified $u \in MCE$ as an employee who held a management position in $O$.

[14] The centrality measures were calculated by using the Networkx Python package (Hagberg et al. 2008).

Table 3 presents the leadership identification *precision at k for* the top 10 (T-10), top 20 (T-20), and top 50 (T-50) user profiles for the various centrality measures.[15] The results indicate that each of the calculated centrality measures can assist in identifying managers inside the organizations. The HITS, Eigenvector centrality, and Communicability centrality measures demonstrated the highest average precision at 50 (0.72), while Load centrality received the lowest score (0.608).

The role of Betweenness centrality and its derivatives, such as Load, is commonly exaggerated in the analysis of social networks. Even in transport networks, where shortest paths are particularly important, Betweenness is found to be a poor importance indicator (Cats and Jenelius 2014). Thus, it is not surprising that it was outperformed by HITS as an indicator for management positions.

### *4.2.1 HITS Measure*

The HITS measure resulted in the highest $precision@k$ across the companies. Next, we investigated the relation between HITS values and managerial positions in the companies. Table 4 shows the number of managers out of the MCE for the top 50 most central users in the organizations. In large organizations, there are many fewer MCEs among the top 50 most central users. Nevertheless, we can see that organizations S2 and L1 as well as M1 and M2 have roughly similar fractions of managers compared to non-managers in the set of Top-50 users. In M1 and M2 we have found the highest density of managers among the most central users. For instance, 20 and 17 of the most central MCEs are managers in M1 and M2 compared to very few in other companies. S1 has the lowest fraction of managers among the most central MCEs, and finally, in L2 the number of MCEs among the 50 most central users is too low to be conclusive.

In order to avoid bias caused by the size of an organization, we checked the fraction of managers, along various percentiles of the most central manually classified employees. Figure 7 presents $precision_O@k(HITS)$ of various organizations along the Y-axis as a function of $\frac{k}{|MCE_O|}$ (in percents) along the X-axis. Organizations L2 and M1 have the largest overall fraction of managers and all organizations position their managers at the top percentile of the MCE. Nevertheless, we found that in M2, managers consistently hold more central positions in the informal social network than in other organizations.

We conclude the analysis of HITS by plotting its ROC (Receiver Operating Characteristics) curve in Fig. 8. ROC is an objective tool for comparing the predictive power of binary classifiers. For arbitrary $k$ it depicts the trade-off between the true positive rate (TPR), the fraction of managers holding central positions:

$$TPR = \frac{|\{u_i \in Top^k_{HITS,O} \cap MCE_O | \text{Is-Manager}(u_i)\}|}{|\{u_i \in MCE_O | \text{Is-Manager}(u_i)\}|}$$

[15] During the precision at T-10, T-20, and T-50 calculations, we only took into account the employees that we succeeded to manually classify. For example, if among the top 10 employees in S2 that received the highest HITS measure, we could only manually classify 6 employees out of which 4 employees held management positions. Then S2's precision at T-10 would be equal to $\frac{4}{6} = 0.66$.

**Table 3** Management positions percentage based on centrality measures (Precision at 10/20/50)

| Org. | Cat. | DG | CL | BC | H | PR | EC | CC | LC |
|---|---|---|---|---|---|---|---|---|---|
| S1 | T-10 | 0.25 | 0.29 | **0.50** | 0.22 | 0.29 | 0.22 | 0.22 | 0.43 |
| | T-20 | 0.31 | 0.31 | **0.33** | 0.29 | 0.31 | 0.29 | 0.29 | 0.31 |
| | T-50 | 0.28 | 0.21 | 0.22 | **0.32** | 0.22 | **0.32** | **0.32** | 0.19 |
| S2 | T-10 | 0.63 | **0.67** | 0.63 | 0.63 | 0.50 | 0.63 | 0.63 | 0.57 |
| | T-20 | 0.62 | **0.64** | 0.62 | 0.62 | 0.62 | 0.62 | 0.62 | 0.58 |
| | T-50 | 0.58 | 0.50 | 0.55 | **0.59** | 0.54 | **0.59** | **0.59** | 0.55 |
| M1 | T-10 | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| | T-20 | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| | T-50 | **0.95** | **0.95** | 0.90 | 0.91 | **0.95** | 0.91 | 0.91 | 0.90 |
| M2 | T-10 | **1** | **1** | **1** | **1** | **1** | **1** | **1** | 0.80 |
| | T-20 | 0.89 | 0.91 | 0.80 | **1** | 0.88 | **1** | **1** | 0.80 |
| | T-50 | 0.81 | 0.85 | 0.67 | **0.90** | 0.81 | **0.90** | **0.90** | 0.65 |
| L1 | T-10 | – | **1** | **1** | – | 0.67 | – | – | **1** |
| | T-20 | 0.33 | **1** | 0.80 | **1** | 0.86 | **1** | **1** | 0.8 |
| | T-50 | 0.60 | 0.75 | **0.79** | 0.60 | 0.57 | 0.60 | 0.60 | 0.78 |
| L2 | T-10 | **1** | 0.67 | 0.33 | – | 0.50 | – | – | 0.33 |
| | T-20 | 0.67 | 0.80 | 0.4 | **1** | 0.33 | **1** | **1** | 0.40 |
| | T-50 | 0.67 | 0.67 | 0.64 | **1** | 0.67 | **1** | **1** | 0.58 |
| Avg. | T-10 | **0.776** | 0.772 | 0.743 | 0.713 | 0.660 | 0.74 | 0.713 | 0.688 |
| | T-20 | 0.637 | 0.777 | 0.658 | **0.818** | 0.667 | **0.818** | **0.818** | 0.648 |
| | T-50 | 0.648 | 0.655 | 0.628 | **0.720** | 0.627 | **0.720** | **0.720** | 0.608 |

Highest values in bold

and the false positive rate (FPR), the fraction of non-managers holding central positions:

$$FPR = \frac{|\{u_i \in Top^k_{HITS,O} \cap MCE_O | \neg\text{Is-Manager}(u_i)\}|}{|\{u_i \in MCE_O | \neg\text{Is-Manager}(u_i)\}|}$$

Once again we can see that HITS works well for M1 and M2 fails but has little correlation with the formal organizational structure in L2.

### *4.2.2 Hidden Management Positions*

We determined that the location of an employee in the informal social network of the organization reveals his or her management role within the organization with high precision, even though it was not reported on Facebook. Table 4 reports the number of concealed management roles that can be detected by using the closeness measure. Out of 76 managers detected by focusing on the top 50 Facebook users with the highest HITS within the informal social network of their organization, 32.89 % did

**Table 4** Organizations' hidden management positions

| Org. | Classified Employee Positions | Classified as Management Positions | HITS T-50 Management Positions | Hidden T-50 Management Positions |
|---|---|---|---|---|
| S1 | 93 | 21 (22.58 %) | 9 (out of 28 MCE) | 5 (55.56 %) |
| S2 | 164 | 69 (42.07 %) | 17 (out of 29 MCE) | 8 (47.06 %) |
| M1 | 456 | 225 (49.34 %) | 21 (out of 23 MCE) | 9 (42.86 %) |
| M2 | 1,510 | 233 (15.43 %) | 19 (out of 21 MCE) | 0 (0 %) |
| L1 | 1,619 | 463 (28.60 %) | 6 (out of 10 MCE) | 2 (33.33 %) |
| L2 | 808 | 398 (49.26 %) | 4 (out of 4 MCE) | 1 (25.00 %) |
| Total | **4,650** | **1,409** (**30.30** %) | **76** (**out of 115 MCE**) | **25** (**32.89** %) |

not report their positions on Facebook. Looking at Table 4, we can also observe that for many of the T-50 profiles it was unclear if a specific employee held a management role inside the organization. This could be due to the employee's privacy settings or due to language barriers, making it difficult to determine management roles.

According to these results, high centrality within the informal social network of an organization is a good indication of a leadership role within the organization. However, this straightforward, general method can only identify management roles of employees with relatively high centrality measures; other management roles with more complex centrality criteria will not be identified using this technique. To overcome the problem of complex centrality criteria, we used machine learning algorithms to classify management roles in each organization (see Section 4.3).

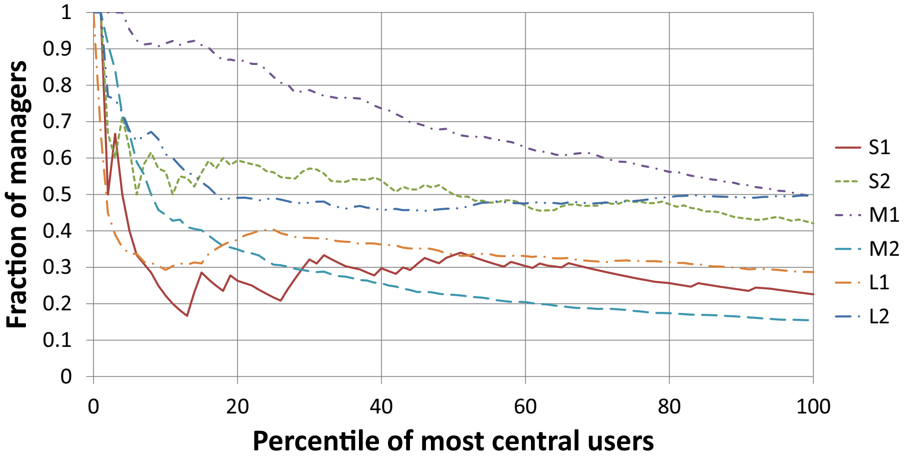

**Fig. 7** HITS-Precision (i.e. fraction of management position) as a function of the most central users percentile

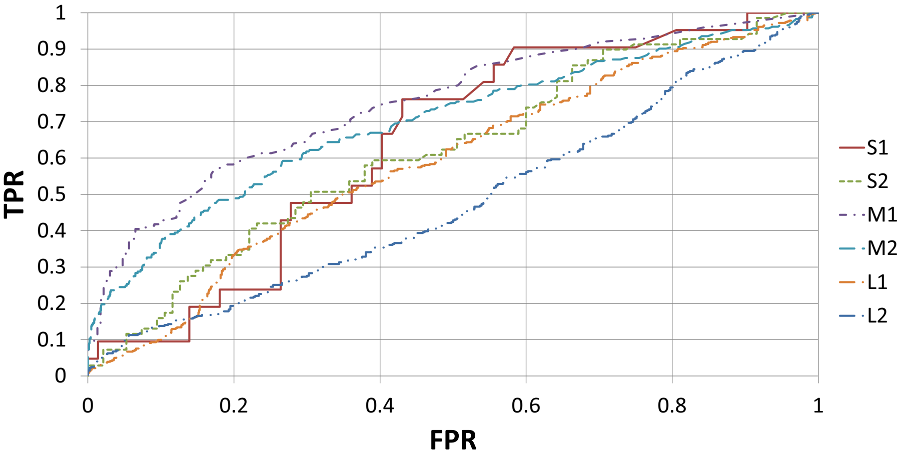

**Fig. 8** ROC curve of the HITS measure

### 4.3 Machine Learning

Using machine learning techniques, we constructed classifiers that can identify management positions inside each organization. This allowed us to identify employees with management roles who satisfied more complex centrality criteria. Moreover, using similar methods and techniques assists in distinguishing different types of positions, such as senior management positions versus R&D engineers.

In order to use the machine learning algorithm, we first needed to create a training set consisting of sufficient training instances. Every training instance represents a collected employee in the organization. The target attribute is a binary attribute which indicates whether or not the employee held a management role inside the organization, while the instance features are divided into the following three categories:

- *Centrality measure features* - for each employee, we calculated all eight centrality measures which were defined in the previous section.
- *Clustering coefficients* - for each employee, in each organizational social network, we also calculated the employee's *clustering coefficient* (Saramäki et al. 2007) and the employee's *squares clustering coefficient* (Lind et al. 2005).
- *Facebook profiles features* - for each employee, we extracted the following six features from his or her Facebook profile:

  1. *Gender* - the employee's gender.
  2. *Birth year* - the employee's birth year.
  3. *Current city* - the city in which the employee currently lives according to the "Current City" field in his or her Facebook profile.
  4. *Hometown* - the employee's hometown according to the "Hometown" field in his or her Facebook profile.

5. *Relationship Status* - the employee's relationship status (single, married, etc.) according to the employee's "Relationship Status" field in employee's Facebook profile.
6. *Number of friends* - the employee's number of Facebook friends according to his or her Facebook profile.

We constructed the classifiers' training datasets for each organization by using the 4,650 manually classified employees and extracting for each employee the 16 features described above. All these profiles were fed into WEKA (Hall et al. 2009), a popular suite of machine learning software, as training instances, where all the features missing values were replaced with question marks, prompting WEKA to treat these features as missing values. Using WEKA, we tested many different machine learning algorithms, such as *ZeroR*(ZR), *OneR* (OR), *K-Nearest-Neighbors* (IBk) with $K \in \{1, 3, 10\}$, *Naive-Bayes* (NB), *Decision tree* (J48), *RandomForest* (RF), and *RotationForest* (ROF). For each of these algorithms, most of the configurable parameters were set to their default values. Whereas, for the J48 classifier, we set the minimum number of instances per leaf to be equal 8. For the RandomForest algorithm, we selected the number of trees to be equal to 50, and for the RotationForest algorithm, we set the iteration number to be equal to 100, with J48, with minimum number of instances per leaf equal to 16, as a base classifier. Lastly, we evaluated each classifier by using the 10-folds cross validation method and calculating the classifier's *True-Positive Rate* (TP), *False-Positive Rate* (FP), *F-measure*, and *AUC* (Area Under the ROC curve) (see Table 5). It can be observed that the RandomForest classifier returned higher AUC values than 0.5, i.e., the AUC of a random classifier. Moreover, for S2 and L1 the relatively simple IBk classifiers obtained the maximal AUC results.

## 5 Communities Formed by Employees

### 5.1 Community Detection Algorithm

In order to better understand the structure of each organization, we used Cytoscape's Girvan-Newman fast greedy algorithm implementation (Clauset et al. 2004) to separate each informal social network into disjointed communities. Each community is marked with a different color in Figs. 1–6. The node shapes in these figures indicate whether or not the particular employee held a management position in the organization. Triangle nodes represent those who, to the best of our knowledge, held management positions, while square nodes represent users who did not hold any management position. Circles represent employees holding an unknown position within the organization.

### 5.2 Community Role Analysis

After separating the informal social network of each organization into disjoint communities, we calculated the partition's modularity (Newman 2006), the number

**Table 5** Machine learning classifiers results

| Org. | Measure | ZR | OR | J48 | NB | IBK | IBK | IBK | RF | ROF |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | K=1 | K=3 | K=10 | | |
| S1 | TP | 0 | 0.21 | 0.1 | 0.21 | **0.47** | 0.32 | 0.05 | 0.07 | 0 |
| | FP | 0 | 0.16 | 0.08 | 0.28 | 0.36 | 0.28 | 0.06 | 0.06 | **0** |
| | F-measure | 0 | 0.1 | 0.1 | 0.16 | **0.33** | 0.26 | 0.06 | 0.09 | 0 |
| | AUC | 0.50 | 0.41 | 0.51 | 0.45 | 0.46 | 0.43 | 0.42 | **0.62** | 0.51 |
| S2 | TP | 0 | 0.43 | 0.06 | 0.34 | 0.49 | 0.31 | **0.61** | 0.36 | 0.32 |
| | FP | 0 | 0.38 | **0.07** | 0.21 | 0.44 | 0.31 | 0.45 | 0.24 | 0.25 |
| | F-measure | 0 | 0.43 | 0.06 | 0.39 | 0.45 | 0.33 | **0.54** | 0.41 | 0.37 |
| | AUC | 0.50 | 0.37 | 0.49 | 0.47 | **0.61** | 0.49 | 0.36 | 0.57 | 0.54 |
| M1 | TP | 0 | 0.60 | 0.42 | 0.36 | **0.64** | 0.62 | 0.44 | 0.58 | 0.58 |
| | FP | 0 | 0.29 | 0.11 | **0.09** | 0.50 | 0.48 | 0.25 | 0.22 | 0.22 |
| | F-measure | 0 | 0.63 | 0.54 | 0.49 | 0.59 | 0.58 | 0.52 | **0.64** | **0.64** |
| | AUC | 0.50 | 0.37 | 0.47 | 0.44 | 0.37 | 0.22 | 0.20 | **0.62** | 0.59 |
| M2 | TP | 0 | 0.20 | 0.12 | **0.30** | 0.16 | 0.13 | 0.13 | 0.12 | 0.19 |
| | FP | 0 | 0.20 | 0.01 | 0.08 | 0.05 | 0.02 | **0.01** | **0.01** | **0.01** |
| | F-measure | 0 | 0.18 | 0.19 | **0.34** | 0.22 | 0.21 | 0.22 | 0.20 | 0.30 |
| | AUC | 0.50 | 0.50 | 0.57 | 0.73 | 0.56 | 0.60 | 0.72 | 0.72 | **0.75** |
| L1 | TP | 0 | 0.24 | 0.02 | 0.12 | **0.29** | 0.27 | 0.22 | 0.07 | 0.06 |
| | FP | 0 | 0.21 | **0.01** | 0.06 | 0.22 | 0.17 | 0.13 | 0.04 | 0.03 |
| | F-measure | 0 | 0.27 | 0.03 | 0.19 | **0.32** | 0.31 | 0.27 | 0.11 | 0.11 |
| | AUC | 0.50 | 0.60 | 0.43 | 0.62 | 0.78 | **0.83** | 0.63 | 0.63 | 0.59 |
| L2 | TP | 0 | 0.69 | 0.05 | 0.27 | 0.49 | 0.51 | **0.68** | 0.51 | 0.62 |
| | FP | 0 | 0.76 | **0.04** | 0.19 | 0.42 | 0.41 | 0.55 | 0.37 | 0.42 |
| | F-measure | 0 | 0.55 | 0.07 | 0.35 | 0.51 | 0.53 | **0.60** | 0.54 | **0.60** |
| | AUC ‘ | 0.50 | 0.47 | 0.50 | 0.62 | 0.54 | 0.57 | 0.60 | 0.61 | **0.63** |

of nodes and links in each community, and the community diameter.[16] Then, we analyzed the role of all the major communities within the organization (see Table 6). We cross-referenced the community members with position descriptions and residence locations from their Facebook profile pages. We also randomly chose several dozen users from each community. For these selected users, we manually inspected their positions within the organization by using publicly available sources, such as LinkedIn. The role of each community in the organization was determined by the majority of the community members' positions, geographic locations, and employment histories. For example, if most of the sampled community users lived in New

[16] The diameter of a graph $G$ is defined as the maximum eccentricity, where the eccentricity of a node $v$ is the maximum distance from $v$ to all other nodes in $G$.

York City and worked as software developers within the organization, then we determined that the community was part of the organization's R&D division in New York City. By understanding the role of each community, we inferred details about the organization and the people it employed. The roles of the different communities within the targeted organizations are presented in Sections 5.2.1–5.2.6.

### 5.2.1 S1 Communities

The community detection algorithm separated the S1 organization's social network into five main communities with a modularity value of 0.475, an average size of 33 nodes, minimum size of 10 nodes, and maximum size of 62 nodes (see Fig. 1). Community role analysis revealed that S1 has several branches in Asia, most of them consisting of R&D employees. There were four R&D communities consisting of employees with different sets of skills. While three communities included mainly software developers (blue, red, and gray communities in Fig. 1), one community consisted mainly of hardware developers (yellow community). Moreover, by reviewing the users' publicly available employment history, we identified a previously acquired start-up company (orange community) and the social connections between the acquired company's employees and S1 employees.

### 5.2.2 S2 Communities

The S2 organizational social network was separated into seven communities by the clustering algorithm, with a modularity value of 0.40, an average size of 45.71 nodes, minimum size of 4 nodes, and maximum size of 109 nodes (see Fig. 2). By reviewing the S2 employees' positions within the organization and the user residence locations, we discovered that S2 has one headquarter office in the Middle East (blue, red, and orange communities in Fig. 2) and another in North America (purple community). We also discovered that the company has worldwide activities occurring on four continents. Project managers (yellow community) are living in more than seven different major cities in the world. The S2 communities' structures indicate that S2 has two headquarters that focus on R&D and worldwide operations, which are managed by the different projects managers in each country.

### 5.2.3 M1 Communities

Our clustering algorithm separated the M1 organizational social network graph into seven connected communities, with a modularity value of 0.453, an average size of 204.14 nodes, minimum size of 5 nodes, and maximum size of 467 nodes (Fig. 3). We discovered two of the company's headquarters, both located in North America (green community in Fig. 3), and also two large R&D divisions (blue and orange communities). Moreover, we succeeded in detecting the company's senior management community (red community) and found the informal connections among the company's senior managers. Identifying the senior management community may assist in inferring key positions inside the M1 organization that in many cases were not available through publicly available resources.

### *5.2.4 M2 Communities*

The M2 organizational social network graph was separated into four closely connected communities, with a modularity value of 0.422, an average size of 965.5 nodes, minimum size of 114 nodes, and maximum size of 1,348 nodes (Fig. 4). Each community represents a group of R&D and Specific Domain Expert (SDE) employees who work in the company's South Asia branch. Each one of the four employee groups was well connected to other employee groups within the same company that were located in different parts of the globe. For example, the South Asian yellow employees group had close ties with another employee group that was located in Africa, while the red employees group was well connected to employees in Australia, Europe, and North America.

### *5.2.5 L1 Communities*

Using the community detection algorithm, we separated the L1 social network into 25 communities, with a modularity value of 0.683, an average size of 231.72 nodes, minimum size of 2 nodes, and maximum size of 1,617 nodes (Fig. 5). Out of these 25 communities, we identified 4 large communities which consist of over 80 % of the crawled L1 employees (see Table 6). By examining the residence and position information of the employees in these seven communities, it is possible to pinpoint the group of support engineers in South America (blue community in Fig. 5). We also succeeded in detecting the company's marketing and sales division in Eastern Europe (orange and green communities) and the company's R&D divisions in East Asia (purple community). Moreover, we discovered that part of the company's R&D group is in the Middle East, and it is connected to a QA engineering group in Europe and to a group of American managers (yellow community).

### *5.2.6 L2 Communities*

Our community detection algorithm separated the L2 social network into five communities, with a modularity value of 0.509, an average size of 1,104.8 nodes, minimum size of 78 nodes, and maximum size of 2,286 nodes (Fig. 6). Two well-connected communities contain many of the company's R&D employees, consultants, and managers in the East Asia headquarters (blue and yellow communities in Fig. 6). We also revealed one of the company's amateur sports teams (green community). Moreover, we were successful in detecting the corporation's international senior management and their informal connections across four continents and more than 20 major cities (red community). By analyzing the company's international senior management community, we could discover the cross-Atlantic connections between the different corporate branches.

**Table 6** Organizations' communities

| Org. | Modularity | Color | #Users | #Links | Diameter | #Facebook Profiles with Positions | Number of MCE Users | Description |
|---|---|---|---|---|---|---|---|---|
| S1 | 0.475 | Blue | 30 | 96 | 4 | 8 | 16 | R&D and administration groups in Asia |
| | | Red | 62 | 234 | 5 | 24 | 37 | Mainly hardware verification engineers and chip designers in Asia |
| | | Yellow | 10 | 13 | 4 | 3 | 8 | Hardware R&D |
| | | Orange | 46 | 197 | 4 | 13 | 21 | Acquired startup company |
| | | Gray | 17 | 29 | 5 | 5 | 11 | R&D in Asia |
| S2 | 0.400 | Blue | 10 | 16 | 6 | 5 | 6 | IT group in the Middle East |
| | | Red | 109 | 645 | 5 | 25 | 45 | R&D groups in the Middle East |
| | | Orange | 48 | 230 | 4 | 16 | 26 | R&D groups in the Middle East |
| | | Yellow | 100 | 576 | 5 | 39 | 58 | Managers and international PM |
| | | Purple | 4 | 5 | 2 | 1 | 1 | Group in North America |
| | | Gray | 4 | 6 | 1 | 1 | 1 | European group |
| | | Cyan | 39 | 155 | 5 | 15 | 27 | R&D teams in Australia and the Middle East |
| M1 | 0.453 | Blue | 464 | 4,410 | 5 | 99 | 162 | R&D division |
| | | Red | 425 | 7,022 | 5 | 86 | 130 | Senior management |
| | | Orange | 217 | 1,496 | 7 | 46 | 75 | R&D divisions |
| | | Yellow | 254 | 1,797 | 7 | 47 | 53 | International consultants and support engineers |
| | | Green | 28 | 61 | 7 | 4 | 11 | North American Headquarters |
| | | Purple | 36 | 170 | 10 | 27 | 11 | Middle East R &D and QA |

**Table 6** (continued)

| Org. | Modularity | Color | #Users | #Links | Diameter | #Facebook Profiles with Positions | Number of MCE Users | Description |
|---|---|---|---|---|---|---|---|---|
| M2 | 0.422 | Blue | 1,329 | 23,549 | 5 | 504 | 498 | R&D and SDE connected to North American and Asian employees |
| | | Red | 1,071 | 16,637 | 5 | 437 | 430 | R&D and SDE connected to Australia, Europe, and North America |
| | | Yellow | 1,348 | 24,080 | 4 | 556 | 551 | R&D and SDE connected to Africa, North America, and Asia |
| | | Green | 114 | 1,058 | 6 | 33 | 32 | R&D and SDE connected to North America and Asia employees |
| L1 | 0.683 | Blue | 1,396 | 1,846 | 11 | 434 | 451 | South America support engineers |
| | | Red | 1,617 | 6,604 | 8 | 406 | 425 | South American Branch (IT, PM, Support engineers, and Analysts) |
| | | Orange | 774 | 10,243 | 12 | 164 | 175 | South Asia Analysts, and Eastern Europe Sales |
| | | Yellow | 921 | 1,442 | 11 | 406 | 247 | Middle East R&D, Europe QA, and American Managers |
| | | Green | 443 | 8,127 | 4 | 85 | 110 | Eastern European (Sales and Pricing) |
| | | Purple | 108 | 107 | 6 | 29 | 42 | East Asian - R&D |
| | | Cyan | 143 | 145 | 8 | 9 | 9 | Management Positions |
| | | Black | 47 | 47 | 11 | 14 | 18 | South American Analysts |

**Table 6** (continued)

| Org. | Modularity | Color | #Users | #Links | Diameter | #Facebook Profiles with Positions | Number of MCE Users | Description |
|---|---|---|---|---|---|---|---|---|
| L2 | 0.509 | Blue | 2,286 | 42,011 | 5 | 228 | 139 | East Asian Headquarters (management and consultants) |
| | | Red | 1,135 | 14,221 | 7 | 525 | 382 | International Senior management (Senior management, Senior researchers) |
| | | Orange | 448 | 5,602 | 6 | 99 | 72 | Middle East software and IT engineers, and management |
| | | Yellow | 1,577 | 18,762 | 6 | 275 | 218 | East Asian Headquarters (R&Ds and consultants) |
| | | Green | 78 | 1,478 | 3 | 4 | 1 | The company's amateur sports team |

## 6 Discussion

To our knowledge, this study is the first study to date which utilizes publicly available data from online social networks to study various organizations' properties. The algorithms and methods presented throughout this study, which were evaluated on six organizations, reveal several interesting results.

First, in contrast to the BFS social network crawler which inefficiently collected organization data, the organizational social network crawler presented in this paper succeeded in efficiently collecting data from 17,096 social networks users from six organizations with an average precision rate of 22.1 % (Table 1). However, the organizational crawler has several limitations. Namely, the crawler can only collect publicly available data, and the crawler can only identify employees who post their place of employment on their Facebook profiles. Moreover, although over a billion people have active Facebook profiles, there are still people without such profiles. Therefore, the organizational crawler can only collect a partial projection of the targeted organization's informal social network. Nevertheless, as we have demonstrated throughout this study in many cases the collected information is sufficient to obtain various perceptive and non-trivial insights regarding a target organization and its employees.

Second, although many employees disclose their job titles on Facebook (see Table 2). The percentage of employees who expose their organizational positions varies significantly from organization to organization. For example, 39.6 % of M2 employees disclosed their positions on Facebook, while only 20.5 % of L2 employees disclosed their positions (see Table 2). From manually examining several thousand social network profiles during this study, we assume that this variation may be influenced by culture differences. For example, we noticed that many employees from South Asia included their job titles on Facebook, while employees from East Asia tended not to publish their positions. We hope to verify this assumption in a future study. Furthermore, one portion of this study included manually classifying 4,650 employee positions as management positions and non-management positions. Using this classification, we can observe that the percentage of management positions in each organization varies considerably from organization to organization, even if two companies are in related industries. For example, we identified 28.60 % of managers among M2 employee profiles, while we identified 49.34 % managers among M1 employee profiles (see Table 4). This difference can indicate different management perspectives which exist in each organization. Alternately, this difference could also be a result of our crawling process limitation; i.e., it is possible that in some organizations fewer managers publicly disclose their personal details.

Third, we discovered that those individuals who received relatively high values in one of the centrality measures were more likely to hold management positions inside the organization (see Table 3). Additionally, the HITS, Eigenvector centrality, and Communicability centrality measures presented the best precision at 20 and 50 results, with an average precision at 20 and 50 of 81.8 % and 72.0 %, respectively (see Table 3). Furthermore, we can observe that each organization includes employees with relatively high centrality measures. However, these employees may not hold management positions in the organization (see Table 3 and Fig. 7). This result agrees with the observation by Krackhardt and Hanson (1993) regarding the considerable existing differences between the informal organizational structure and the formal organizational structure. Also, using the HITS measure, we identified 76 management positions where 32.89 % of these positions were hidden management roles and did not appear in the individuals' Facebook profiles (see Table 4).

Fourth, using machine learning algorithms, we were able to construct models which could predict if an employee held a management position or not. Our models presented better than random performances, with AUCs up to 0.62, 0.61, 0.62, 0.77, 0.83, and 0.67 for the S1, S2, M1, M2, L1, and L2 organizations, respectively (see Section 4.3 and Table 5). In most cases, the maximal or near maximal AUC and F-measure results were obtained by the RandomForest and RotationForest classifiers. Interestingly, for S2 and L1 the relatively simple K-Nearest-Neighbors classifiers obtained the maximal AUC results. This may indicate that in the S2 and L1 organizations, employees with similar positions have similar features. We hope investigate this issue more deeply in a future study.

Fifth, we demonstrated that by using a community detection algorithm it is possible to reveal interesting insights on an organization as a whole and also on the disjoint communities within it (see Section 5). By identifying the positions of more than four thousand employees in the organizations studied, we discovered specific community's roles and geographic locations, according to the positions and residences of the majority of community users. Using this method, we succeeded in inferring many observations about each organization, such as the geographic locations of its branches and the common employees' qualifications in each branch. We also discovered further non-trivial insights about each company. For example, although sample company S1 acquired a start-up R&D company, the acquired company still performed as a separate company with almost no social connections to S1 as a whole. This type of discovery can be used by an organization's management to identify problems within the social structure of the company, such as structural holes (Burt 1995).

We notice that the majority of communities include members that reside in the same geographic location. This result is consistent with the observations of Illenberger et al. (2013), that the probability of a social tie is inversely proportional to the geographical distance between people. Nevertheless, we also see that all companies except S1 and M2 have a cross border community formed by managerial staff. In companies M1 and L2, we uncovered the senior management community and their notable informal friendship connections. Detecting an organization's senior management community can assist in identifying undisclosed management roles and key positions inside the organization. Furthermore, by understanding the relationships between a company's senior managers, we can reveal the connections among the organization's different branches. In the Asian branch of M2, we could infer methods of work where each discovered group inside this branch consisted primarily of R&D and Specific Domain Expert employees who interacted with company employees in other continents.

Lastly, by investigating the organizational structures of the various communities, we can observe that different communities have different structural properties with varying diameters and link densities (see Table 6). This observation is also true for communities within the same organization. For example, in M2 the communities' diameters varies from 4 to 6, and in L1 the communities' diameters varies from 4 to 12, suggesting that M2 has a more closely connected social network. We believe that the varying structural properties of each community can be utilized to better understand how an organization functions. We hope to verify this assumption in future study.

## 7 Future Research

The study presented here is innovative and offers many future research directions to pursue. One possible direction is to create multi-label organizational social networks by cross-referencing an organization's online social network with other social networks associated with that organization, such as the network created by the organization's emails (Tyler et al. 2005). These multi-label social networks can provide valuable insights and assist in better understanding the organization as a whole. An additional possible future research direction is to improve the organizational crawler precision. By taking a similar approach to those of Lesser et al. (2013) and developing machine learning classifiers for identifying an organization's employees, we believe that it is possible to identify employees that work in the target organization, even if they did not specifically state their employment organization name on their Facebook profiles.

Another possible direction is to combine different community detection algorithms in order to improve an organization's community detection results and reveal more communities inside each organization. Yet another possible direction is to enrich an organization's user-collected data by automatically adding user data from different publicly available data sources, such as LinkedIn and people search engines. Automatically adding more details to the collected organization's users can improve the results when identifying community roles within an organization.

A further future direction for this study, which was purposed by Lindsay (2013), is to use the collected organizational social network to identify isolated teams inside an organization. We also believe that certain positions within commercial organizations exhibit specific connection patterns in the informal social networks. Machine learning techniques can be applied to identify the patterns associated with specific positions inside the organization, such as developers, sales representatives, administrative officers, or senior managers.

One additional possible future research direction includes performing a similar study to that of Pugh et al. (1968), using the organizational crawler for collecting a considerable number of organizational structures. Then, it is possible to utilize the collected data to compare the structure properties of several organizations.

A research direction we have already started to pursue is to examine the implications of malicious users utilizing the collected organizational social network data. Such users might perform a series of friend requests to company employees

(Elishar et al. 2012) or attack a specific employee inside a targeted organization (Elyashar et al. 2013).

## 8 Conclusions

This paper presents methods and algorithms that can be used to collect data from publicly available sources and analyze organizations' informal social networks. By collecting data posted on Facebook, we developed a web crawler to extract profiles of employees from six targeted organizations. We created a social network topology for each organization, and we utilized centrality measures and machine learning algorithms to detect hidden leadership positions within each company. Additionally, our algorithms disclosed the social community clusters inside these organization, providing insight into the organizational structure and communication network of each company.

Organizations should be aware that outsiders can employ similar online crawling techniques to gather employee information. On the one hand, these methods can offer valuable insights to those leading an organization by revealing the strengths and weaknesses in their organizational structure. Potential problem areas can be determined and corrected. Unrecognized leaders can be acknowledged and placed in more effective roles. Information can be distributed more efficiently and resources used more effectively.

On the other hand, publicly accessing the informal online social network of an organization can be used to expose private and sometimes sensitive data regarding the target organization and its employees. For example, in this study, we demonstrated that is possible to infer various organizational properties of a targeted organization, such as the employees' specific roles and their geographic distribution (see Sections 4 and 5). Organizations wanting to conceal such things as their internal structure, the identity of their most influential leaders, and the effect of social communities within their offices should enforce policies to control the use of social media by their employees.

Regardless, these methods offer researchers many research opportunities, not existing before the recent prevalence of online social networks, to collect, study, and analyze the structures of various types of organizations. As Krackhardt and Hanson observed, "If the formal organization is the skeleton of a company, the informal is the central nervous system driving the collective thought processes, actions, and reactions of its business units." It is vital to understand and analyze both the formal and the informal organizational structures of a company in order to optimize its operation.

## 9 Data Availability

Anonymous versions of the six organizations' social network topologies used in our study were created by randomly replacing the employees' Facebook IDs with a series of contiguous integers. This is available for other researchers to use and can be found on our research group website http://proj.ise.bgu.ac.il/sns/.

## References

Acquisti A, Gross R (2006) Imagined communities: Awareness, information sharing, and privacy on the facebook. In: Privacy enhancing technologies. Springer, pp 36–58

Allen T, Cohen S (1969) Information flow in research and development laboratories, Administrative Science Quarterly

Baker WE, Faulkner RR (1993) The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry, American sociological review, pp 837–860

Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2011) The socialbot network: when bots socialize for fame and money. In: Proceedings of the 27th Annual Computer Security Applications Conference. ACM, pp 93–102

Burt R (1995) Structural holes: rhe social structure of competition. Harvard University Press

Campbell C, Maglio P, Cozzi A, Dom B (2003) Expertise identification using email communications. In: Proceedings of the twelfth international conference on Information and knowledge management. ACM, pp 528–531

Cats O, Jenelius E (2014) Dynamic vulnerability analysis of public transport networks: mitigation effects of real-time information. Netw Spatial Economics 14(3):435–463

Chesney T, Fire M (2014) Diffusion through networks of heterogeneous nodes in a population characterized by homophily. Nottingham University Business School Research Paper, pp 2014–05

Clauset A, Newman M, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066,111

Constine J (2013) Facebooks growth since ipo in 12 big numbers. TechCrunch

Diehl CP, Namata G, Getoor L (2007) Relationship identification for social network discovery. AAAI 22:546–552

Diesner J, Frantz TL, Carley KM (2005) Communication networks from the enron email corpus it's always about the people. enron is no different. Comput Math Org Theory 11(3):201–228

Ducruet C, Beauguitte L (2014) Spatial Science and Network Science: Review and Outcomes of a Complex Relationship. Netw Spatial Economics 14(3):297316

Dwyer C, Hiltz S, Passerini K (2007) Trust and privacy concern within social networking sites: A comparison of facebook and myspace. In: Proceedings of AMCIS. Citeseer, pp 1–12

Elishar A, Fire M, Kagan D, Elovici Y (2012) Organizational intrusion, ASE Cyber Security Conference (CyberSecurity)

Elyashar A, Fire M, Kagan D, Elovici Y (2013) Homing socialbots: intrusion on a specific organization's employee using socialbots. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, pp 1358–1365

Estrada E, Rodriguez-Velazquez J (2005) Subgraph centrality in complex networks. Phys Rev E 71(5):056,103

Facebook (2014) Company info. [last accessed on July 27th, 2014]. http://newsroom.fb.com/company-info/

Fire M, Tenenboim-Chekina L, Puzis R, Lesser O, Rokach L, Elovici Y (2013) Computationally efficient link prediction in a variety of social networks. ACM Trans Intell Syst and Technol (TIST) 5(1):10

Freeman L (1977) A set of measures of centrality based on betweenness, Sociometry :35–41

Gjoka M, Butts C, Kurant M, Markopoulou A (2011) Multigraph sampling of online social networks. Selected Areas in Communications. IEEE J 29(9):1893–1905

Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkx. In: Proceedings of the 7th Python in Science Conference (SciPy2008)

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. SIGKDD Explor Newsl 11:10–18. doi:10.1145/1656274.1656278

Illenberger J, Nagel K, Flötteröd G (2013) The Role of Spatial Interaction in Social Networks. Netw Spatial Economics 13(3):255–282

Jacobson E, Seashore S (1951) Communication practices in complex organizations. J Soc Issues 7(3): 28–40

Kilduff M, Brass D (2010) Organizational social network research: Core ideas and key debates. Acad Manag Ann 4(1):317–357

Kilduff M, Tsai W (2003) Social networks and organizations. Sage Publications Ltd

Kleinberg J (1999) Authoritative sources in a hyperlinked environment. J ACM (JACM) 46(5):604–632

Krackhardt D, Hanson JR (1993) Informal networks: the company behind the chart. Harv Bus Rev 71(4):104–11

Krebs V (2002) Mapping networks of terrorist cells. Connections 24(3):43–52

Lesser O, Tenenboim-Chekina L, Rokach L, Elovici Y (2013) Intruder or welcome friend: inferring group membership in online social networks. In: Social Computing, Behavioral-Cultural Modeling and Prediction. Springer, pp 368–376

Lind PG, González MC, Herrmann HJ (2005) Cycles and clustering in bipartite networks. Phys Rev E 72(5):056,127

Lindsay G (2013) Engineering serendipity. New York Times

McCallum A, Corrada-Emmanuel A, Wang X (2005) Topic and role discovery in social networks. Computer Science Department Faculty Publication Series, p 3

McPherson M, Smith-Lovin L, Cook J (2001) Birds of a feather: Homophily in social networks, Annual review of sociology

Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and Analysis of Online Social Networks. In: Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07), San Diego

Naddafa Y, Mutyalab S (2010) Social network analysis and community mining in organizations based on email records

Newman M (2008) The mathematics of networks, The New Palgrave Encyclopedia of Economics

Newman M, et al. (2001) Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Phys Rev Ser E-64(1; PART 2):16,132–16,132

Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23): 8577–8582

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking, Bringing order to the web

Paradise A, Puzis R, Shabtai A (2014) Anti-reconnaissance tools: Detecting targeted socialbots. Internet Computing IEEE PP(99):1–1. doi:10.1109/MIC.2014.81

Provan K, Fish A, Sydow J (2007) Interorganizational networks at the network level: A review of the empirical literature on whole networks. J Manag 33(3):479–516

Pugh D, Hickson D, Hinings C, Turner C (1968) Dimensions of organization structure, Administrative science quarterly, pp 65–105

Rooksby J, Kahn A, Keen J, Sommerville I, Rooksby J (2009) Social networking and the workplace, The UK Large Scale Complex IT Systems Initiative, pp 1–39

Saramäki J, Kivelä M, Onnela JP, Kaski K, Kertesz J (2007) Generalizations of the clustering coefficient to weighted complex networks. Phys Rev E 75(2):027,105

Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504

Shetty J, Adibi J (2004) The enron email dataset database schema and brief statistical report. Information sciences institute technical report. University of Southern California, p 4

Shetty J, Adibi J (2005) Discovering important nodes through graph entropy the case of enron email database. In: Proceedings of the 3rd international workshop on Link discovery. ACM, pp 74–81

Sparrow M (1991) The application of network analysis to criminal intelligence: An assessment of the prospects. Soc Networks 13(3):251–274

Steinfield C, DiMicco J, Ellison N, Lampe C (2009) Bowling online: Social networking and social capital within the organization. In: Proceedings of the fourth international conference on Communities and technologies. ACM, pp 245–254

Tichy N, Tushman M, Fombrun C (1979) Social network analysis for organizations. Academy of Management Review, pp 507–519

Twitter (2013) Rest api rate limiting in v1.1. [last accessed on August 3th, 2014]. https://dev.twitter.com/docs/rate-limiting/1.1

Tyler J, Wilkinson D, Huberman B (2005) E-mail as spectroscopy: Automated discovery of community structure within organizations. Inf Soc 21(2):143–153

Wilkinson D, Huberman B (2004) A method for finding communities of related genes. Proc Natl Acad Sci USA 101(Suppl 1):5241

Wilson G, Banzhaf W (2009) Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis. In: IEEE Congress on Evolutionary Computation, 2009. CEC'09. IEEE, pp 3256–3263

Zhan X, Ukkusuri SV, Zhu F (2014) Inferring urban land use using large-scale social media check-in data. Netw and Spatial Economics 14(3):647–667