

Organization Mining Using Online Social Networks

Michael Fire, Rami Puzis, and Yuval Elovici*

Telekom Innovation Laboratories at Ben-Gurion University of the Negev
 Department of Information Systems Engineering, Ben-Gurion University

Abstract

Mature social networking services are one of the greatest assets of today's organizations. This valuable asset, however, can also be a threat to an organization's confidentiality. Members of social networking websites expose not only their personal information, but also details about the organizations for which they work. 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 a given target organization. Our results, obtained using centrality analysis and Machine Learning techniques applied to the structure of the informal relationships network, show that it is possible to identify leadership roles within the organization solely by this means. It is also possible to gain valuable non-trivial insights on an organization's structure by clustering its social network and gathering publicly available information on the employees within each cluster. Organizations wanting to conceal their internal structure, identity of leaders, location and specialization of branches offices, etc., must enforce strict policies to control the use of social media by their employees.

Keywords. Organizational data mining, Social network data mining, Social networks privacy, Organizational social network privacy, Facebook, LinkedIn, Machine learning, Leadership roles detection

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 Facebook¹ social network has more than 1.11 billion monthly active users, with new users signing up each month. According to recent statistics published by Facebook, 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.) [7]. On the one hand, social networks create new opportunities to develop friendships, share ideas, and conduct business. On the other hand, many social network users expose personal third-party details about themselves and their social connections via their profile pages [1, 3], as well as sensitive business information and details about their place of employment.

*Email:{mickyfi,puzis,elovici}@bgu.ac.il

¹<http://www.facebook.com>

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. on the Hewlett-Packard organization [42]. 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.

The contributions of this paper are threefold: First, we present a method for uncovering an organization’s informal social network topology based solely on publicly available data. Second, we use the organization’s structure to discover hidden leadership roles within the organization and to identify communities inside the organization. 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.

1.1 Our Approach in a Nutshell

The organizational 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 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. Such insights included, for example, the geographic deployment of the organization, the structure of the organization’s different divisions, the relationships between divisions, companies that were previously acquired, and the research focus of the organization. These details can help us better understand both the structure and the communication patterns within organization. They also highlight the need for organizations to be aware of their social networking vulnerability and to establish policies to control this exposure as necessary.

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of previous relevant studies on social network analysis with a special focus on organizational social network 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. Lastly, in Section 6, we present our conclusions and offer future research directions.

²<http://www.linkedin.com>

2 Background

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 YouTube⁵ 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 work-places. In 2007, a study carried out by Dwyer et al. [10] 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. [3] 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 Google⁶ and LinkedIn, one can infer further details about a Facebook user, such as specific work experience and areas of expertise. For example, Pipl⁷ and PeekYou⁸ 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. In 1951, Jacobson and Seashore [19] were among the first researchers to study communication patterns among federal agency employees. In 1968, Pugh et al. [34] studied five primary dimensions of organizational structure applied to 52 different organizations in England. In 1969, Allen and Choens [2] studied technical communication patterns and their influences within two research and development laboratories at MIT. In 1979, Tichy et al. [41] 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 1991, Sparrow [39] presented a method for using social network structural analysis to better understand criminal organizations. In 2002, after 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>

tragic events of September 11, 2001, Krebs [23] studied 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 2003, Campbell et al. [5] presented algorithms for expertise identification using email communication patterns. Their algorithms were evaluated on two different organizations. In our own study, we show that expertise, leadership, and the roles of communities can be identified using publicly available data sources even without having access to internal organization data, such as email logs. Since 2004, after the release of about 500,000 Enron employees’ emails [37], 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 [8, 9, 25, 38, 44].

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. [40] studied the connection between social capital and the use of social networking services deployed inside organizations. In the same year, Rooksby et al. published a detailed report on how online social networks are used in the context of the workplace [35]. Comprehensive reviews on organizational social networks can provide further insights [21, 33, 20].

2.3 Discovering an Organization’s Social Network from Informal Connections

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. [42]. 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 [15] to detect leadership roles within the organization. They also applied a version of the Wilkinson and Huberman algorithm [43] which partitions the organization’s social network into communities. The results were evaluated by interviewing several employees about the community they were automatically placed in by the community detection algorithm. Naddaf and Mutyala [28] presented a similar study in 2010. They demonstrated a method for extracting informal social networks formed by employees of an organization based on the 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 [32]. Moreover, Naddaf and Mutyala used the Fast Modularity algorithm [6] to identify communities in their client’s organizational social network.

3 Organization Social Network Crawler

Many different types of web crawlers have been developed to collect data from large scale online social networks [27, 3, 16, 14]. 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 [16].

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, we collected many irrelevant profiles and often skipped Facebook users who worked in our targeted organization. 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 or-

ganization. Our organization crawler utilizes the homophily principle [26]. 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. 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. These seeds are used to initialize a priority queue (line 2). All seeds have an initial priority of zero. Later, the priority of profile pages in the queue is increased with every friend that is employed by the target organization (lines 11-12).

We proceeded by iteratively processing the next profile page with the highest priority (i.e., likelihood of being an employee of the target organization) until no potentially valuable profiles were left in the priority queue (lines 4-5). Every processed profile page was downloaded (line 7) and automatically analyzed. We employed a heuristic in an attempt to discover whether or not the currently processed profile page belonged to an employee of the target organization. This heuristic matched various keywords associated with the organization to 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. In case we did not find a match to any of the keywords, we continued on to the next profile in the priority queue. If, however, we did find that the dequeued profile page belonged to an individual who worked in the targeted organization, we collected the list of his or her Facebook friends (lines 8-9).

Profile pages of Facebook friends that were already processed were ignored (line 10). We increased the priorities of all friends waiting in the priority queue (lines 11-13). Afterwards, we inserted all newly encountered Facebook friends of the currently processed profile into the priority queue, with a priority of one (lines 14-16). This process repeated with the next profile page extracted according to the updated priorities.

The crawler whose pseudo code is described by Algorithm 1 stops when the queue is empty. We will refer to this crawler as Version 1. We also evaluated an optimized version of the organization crawler. This version 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.

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

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.

```
 $Q \leftarrow \text{Priority-Queue}()$ 
 $\forall URL \in S, Q.\text{Enqueue}(URL : 1)$ 
 $Crawled \leftarrow \emptyset$ 
while ( $Q \neq \emptyset$ ) do
     $URL \leftarrow Q.\text{Dequeue}()$ 
     $Crawled \leftarrow Crawled \cup \{URL\}$ 
     $Page \leftarrow \text{DownloadProfilePage}(URL)$ 
    if  $Page$  contains  $N$  then
         $F\_URLs \leftarrow \text{Extract list of friends from Page}$ 
         $F\_URLs \leftarrow F\_URLs - Crawled$ 
        for ( $F\_URL \in F\_URLs \cap Q$ ) do
            Increase priority ( $Q$ ,  $F\_URL$ )
        end
        for ( $F\_URL \in (F\_URLs - Q)$ ) do
             $Q.\text{Enqueue}(F\_URL:1)$ 
        end
    end
end
return Collected pages
```

the employees' Facebook identities by randomly replacing the users' Facebook IDs with a series of contiguous integers.

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 [36] software to visualize the social networks formed by the employees of each organization.⁹ The vertex colors in Figures 1-6 represent various cluster roles, as will be explained in Section 5. The analysis results of these networks are reported in Sections 4 and 5.

⁹All the organizations' graphs presented throughout this paper are embedded as Scalable Vector Graphics (SVG) image, which enables the reader to zoom in and view each node in each graph.

Table 1: Organization Crawling Results

Org.	Crawler Version	#Total 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%

3.2.1 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 Figure 1). We also collected information on 84 employee positions inside the company. Out of these 84 employees, we identified 20 in management positions. Most of the discovered 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 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 Figure 2). We also collected information on the positions of 168 company employees. Out of these 168 individuals, 76 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 32,876 informal links among 1,429 Facebook users who, according to their Facebook profile page, worked for the company (see Figure 3). When we also collected information on the positions of 461 employees, we learned 227 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.

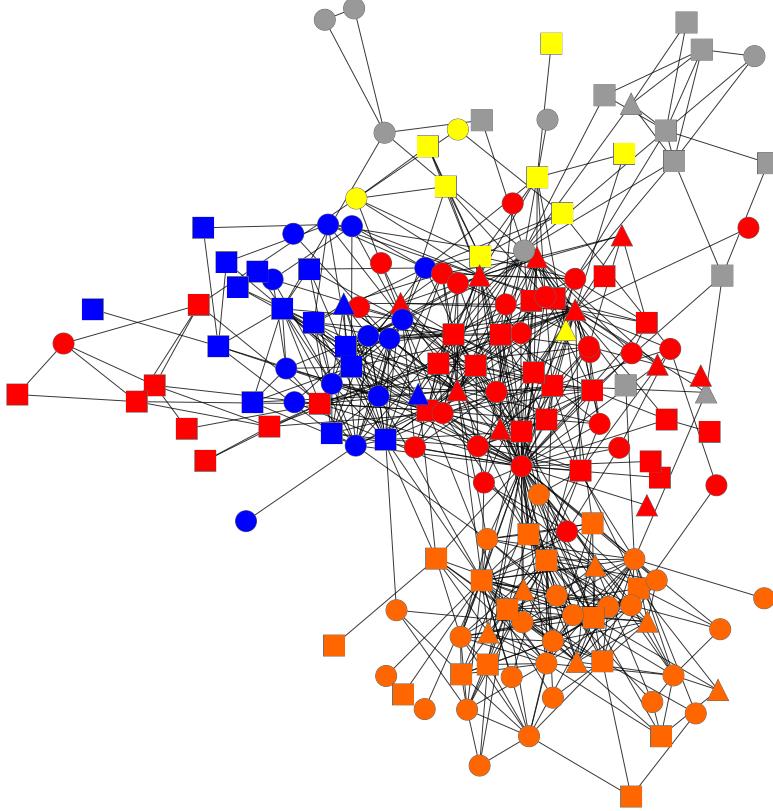


Figure 1: **S1 Company:** *Blue nodes - R&D and administration groups in Asia. Red nodes - primarily hardware verification engineers and chip designers in Asia. Yellow nodes - Hardware R&D. Orange nodes - acquired startup company. Gray nodes - R&D in Asia.*

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 Figure 4). We also succeeded in collecting information on the positions within the company for 1,511 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 1,511 employees, 230 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 45,266 informal links among 5,793 Facebook users who, according to their Facebook profile page,

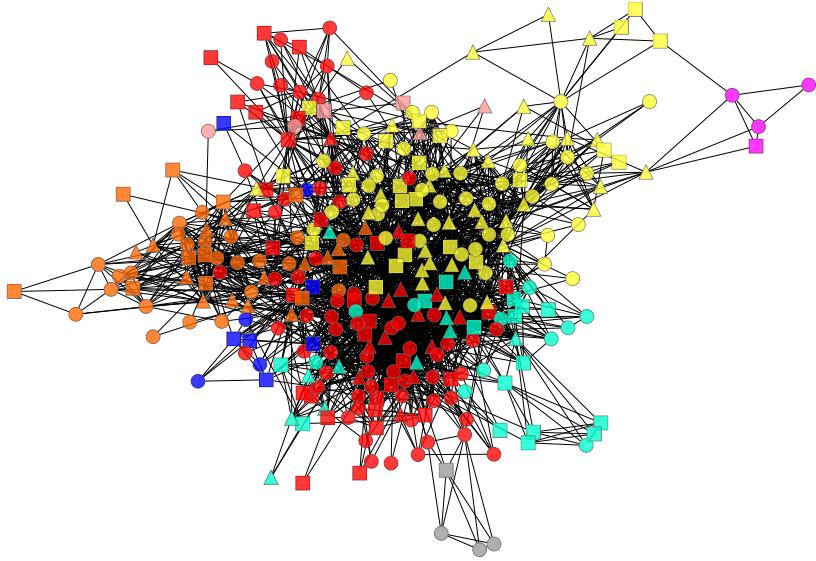


Figure 2: **S2 Company:** *Blue nodes - IT group in the Middle East. Red and Orange nodes - R&D groups in the Middle East. Purple nodes - North American group. Yellow nodes - managers and international project managers. Cyan nodes - R&D teams in Australia and the Middle East. Gray nodes - European group.*

worked for the corporation (see Figure 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 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 Figure 6). We also succeeded in collecting information on the company positions of 1,131 employees, out of which 461 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. In this section, we demonstrate that it is possible to pinpoint leadership roles solely by analyzing the structure of the informal

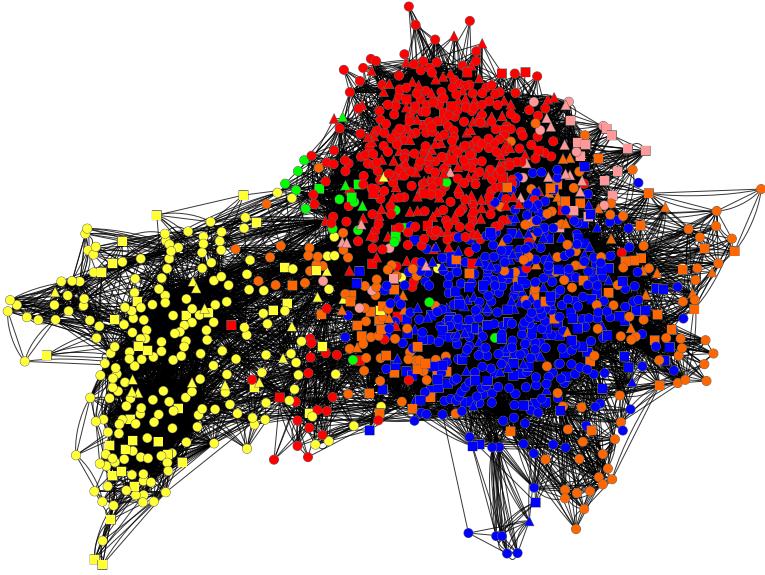


Figure 3: **M1 Company:** *Blue and Orange nodes - R&D divisions. Red nodes - senior management. Yellow nodes - international consultants and support engineers. Green nodes - North American headquarter employees.*

social network of an organization’s employees.

Let $G = \langle V, E \rangle$ 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, for each user $v \in V$ in the informal social network, we calculated eight centrality measures. Next, for each centrality measure, we examined the top 10 and the top 20 users who received the maximal score. By reviewing the selected employees’ Facebook and LinkedIn profile pages and checking the employment status declared by the individual, we manually classified whether or not the user held a management position (team leader, project manager, vice president, etc.). In many cases, however, the user’s profile information was not enough to reveal the user’s specific position inside the organization. To overcome this problem, we cross-referenced the user’s personal details with other publicly available online sources, such as Google search engines. By using these methods, in many cases we succeeded in revealing the user’s position within the organization.

Lastly, we used several Machine Learning algorithms to build classifiers that can automatically identify management roles inside an organization based on the different centrality measures of the vertices in the informal social network. 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.

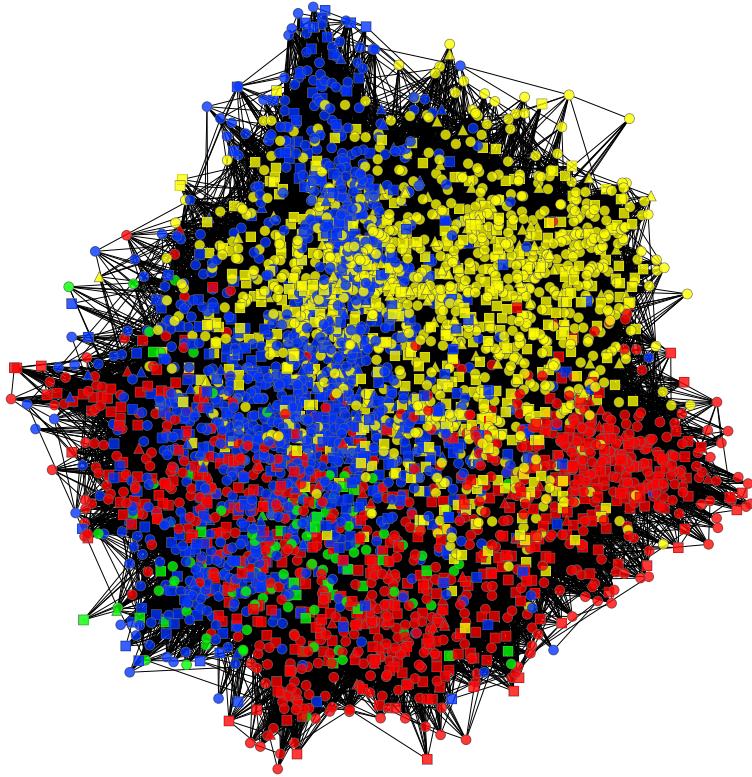


Figure 4: **M2 Company:** *Blue and Green nodes - R&D and Specific Domain Experts (SDE) connected to North American and Asia employees. Red nodes - R&D and SDE connected to Australia, Europe and North America. Yellow nodes - R&D and SDE connected to Africa, North America, and Asia.*

4.1 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:¹⁰ Degree centrality (DG), Closeness centrality (CL) [29], Betweenness centrality (BC) [15], HITS (H) [22], PageRank (PR) [32], Eigenvector centrality (EC) [30], Communicability centrality (CC) [13], and Load centrality (LC) [31]. We then sorted the crawled organization’s users’ list according to the different centrality measures.

We manually inspected the top 20 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, 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. We will refer to managers who do not report their position on Facebook as concealing their management position. 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.

¹⁰The centrality measures were calculated by using the Networkx [17] Python package.

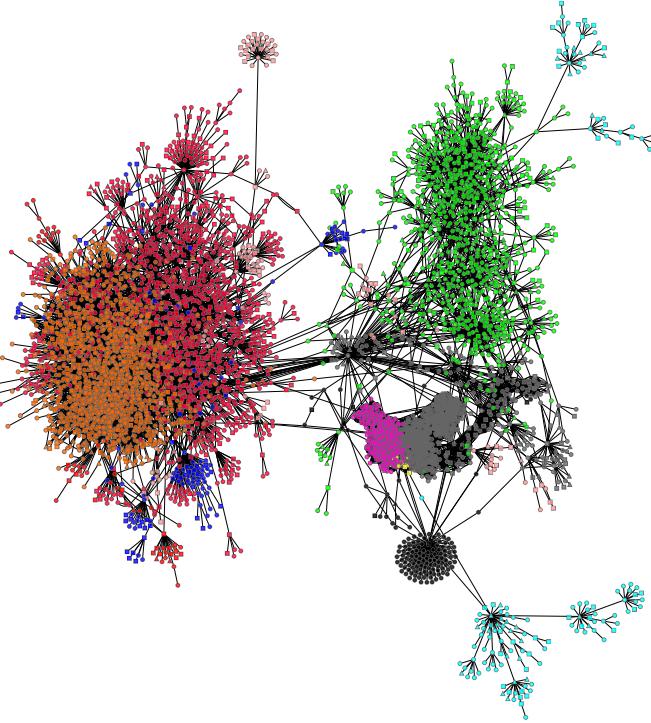


Figure 5: L1 Corporation: *Blue nodes* - South American support engineers. *Red nodes* - South American Branch (IT, support engineers, analysts, and PMs). *Orange nodes* - South American Branch (management, sales, marketing, project managers, support engineers, and administration). *Yellow nodes* - Eastern Europe Pricing Analysts. *Purple nodes* - Eastern European consultants (marketing, sales and pricing) and support engineers. *Black nodes* - North American Branch, and East Asia - R&D. *Green nodes* - Middle East R&D and North American headquarters (management and sales). *Gray nodes* - European consultants and sales and South Asian analysts.

Table 3 presents the leadership identification *precision at* the top 10 (T-10) and top 20 (T-20) user profiles for the various centrality measures. The results indicate that each of the calculated centrality measures can assist in identifying managers inside the organizations. Closeness demonstrated the highest average precision at 20 (0.76), while PageRank received the lowest score (0.70).

Table 4 reports the number of concealed management roles that can be detected by using the closeness measure. Out of 85 managers detected by focusing on the top 20 Facebook users with the highest Closeness centrality within the informal social network of their organization, 40% did not report their positions on Facebook.

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, therefore, will not be identified using this technique. To overcome the problem complex centrality criteria, we used state-of-the art Machine Learning algorithms to

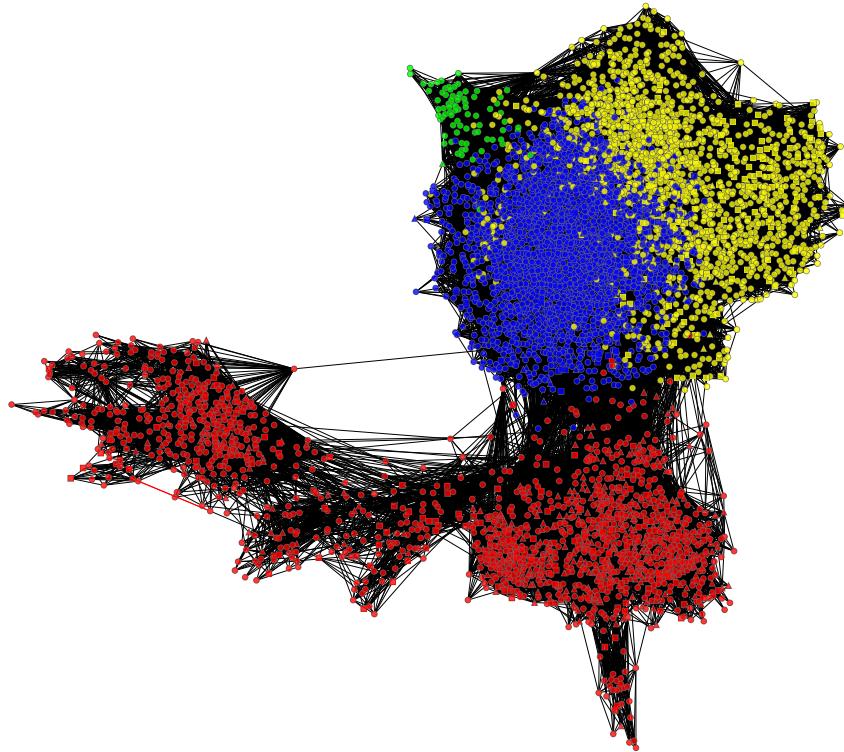


Figure 6: **L2 Corporation:** *Blue nodes* - East Asia Headquarter (management and consultants). *Red nodes* - international senior management and researchers. *Yellow nodes* - East Asian headquarters (R&Ds and consultants). *Green nodes* - the company’s amateur sports team.

classify management roles in each organization (see Section 4.2).

4.2 Machine Learning

Using state-of-the-art 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 and 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 user in the organization. The target attribute is a binary attribute which indicates whether or not the user held a management role inside the organization, while the instance features are the different extracted centrality measures. We created a sufficient number of training instances by quickly reviewing the users’ data extracted from their Facebook profiles. By analyzing the crawled organizations’ user Facebook profiles, we discovered that an average of 28.1% of the collected users had inserted at least partial information about their previous and current employment positions into their Facebook 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

Table 2: Collected Organization Datasets

Org.	Size	Discovered Employees	Links	Employees Disclosing Positions on Facebook
S1	500-1K	165	726	54(32.7%)
S2	1K-2K	320	2,369	104(32.5%)
M1	2K-10K	1,429	32,876	383(26.8%)
M2	10K-20K	3,862	87,324	1,531(39.6%)
L1	50K+	5,793	45,266	1,601(27.6%)
L2	50K+	5,524	94,219	1,131(20.5%)
Total	-	17,093	262,780	4,804(28.1%)

Table 3: Management Positions Percentage Based on Centrality Measures (Precision at 10/20)

Org.	Cat.	DG	CL	BC	H	PR	EC	CC	LC
S1	T-10	0.50	0.40	0.60	0.30	0.50	0.30	0.30	0.60
	T-20	0.35	0.30	0.30	0.30	0.25	0.30	0.30	0.30
S2	T-10	0.80	0.90	0.80	0.90	0.70	0.90	0.90	0.80
	T-20	0.70	0.75	0.75	0.70	0.75	0.70	0.75	0.75
M1	T-10	1	1	0.80	1	1	1	1	0.80
	T-20	1	0.95	0.85	1	0.85	1	1	0.85
M2	T-10	0.83	0.71	0.86	0.83	0.86	0.83	0.83	0.88
	T-20	0.73	0.82	0.69	0.80	0.71	0.80	0.80	0.69
L1	T-10	0.55	0.80	0.80	0.78	0.60	0.78	0.78	0.80
	T-20	0.65	0.75	0.70	0.56	0.65	0.56	0.56	0.70
L2	T-10	1	1	1	1	1	1	1	1
	T-20	0.92	1	1	1	1	1	1	1
Avg.	T-10	0.78	0.8	0.81	0.80	0.78	0.80	0.80	0.81
	T-20	0.725	0.76	0.715	0.73	0.70	0.73	0.735	0.715

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 4,767 users’ profiles. Out of the 4,767 manually classified positions, we identified 1,470 users who held management positions (see Table 4). All these profiles were fed into WEKA [18], a popular suite of Machine Learning software, as training instances. Using WEKA, we tested many different Machine Learning algorithms, such as *OneR* (OR), *K-Nearest-Neighbors* (IBk) with $K \in \{1, 3, 10\}$, *Naive-Bayes* (NB), *Decision tree* (J48), *Logistic* (LG), and *RandomForest* (RF). Lastly, we evaluated each classifier by using the 10-folds cross validation method and calculating the *Accuracy*, *F-measure*, and *AUC* (Area Under the ROC curve) (see Table 5). We used T-tests with a significance of 0.05 to compare the different classifiers. According to the T-test results, for every organization except S1, all the classifiers returned better accuracy results than the naive *ZeroR* (ZR) classifier. Moreover, in most cases the simple OneR classifier is sufficient enough to obtain a near maximum accuracy. However, better AUC and F-measure results were obtained using more advanced classifiers, such as Logistic, and RandomForest classifiers.

Table 4: Organization’s Hidden Management Positions

Org.	Classified Employee Positions	Classified Management Positions	Closeness T20 Management Positions	Hidden T20 Management Positions
S1	84	20 (23.80%)	6	5 (83.33%)
S2	168	76 (45.20%)	15	4 (26.66%)
M1	461	227 (49.20%)	19	10 (47.40%)
M2	1,511	223 (14.76%)	14	2 (14.30%)
L1	1,619	463 (28.60%)	15	3 (20%)
L2	924	461 (49.90%)	16	10 (62.50%)
Total	4,767	1,470 (30.80%)	85	34 (40%)

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 [6] to separate each informal social network into disjointed communities. Each community is marked with a different color in Figures 1-6. 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 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. During this process, we reviewed several thousand employees’ profiles and identified the organizational positions of 4,767 users. 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 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 social network into five main communities (see Figure 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

Table 5: Machine Learning Classifiers Results

Org.	Measure	ZR	OR	J48	NB	IBK K=1	IBK K=3	IBK K=10	LG	RF
S1	Accuracy	76.11	68.36	71.93	72.96	65.28	74.32	75.17	73.63	67.67
	F-measure	0	0.11	0.01	0.29	0.25	0.27	0.15	0.08	0.24
	AUC	0.50	0.49	0.46	0.57	0.53	0.64	0.61	0.37	0.57
S2	Accuracy	54.78	60.45	62.2	63.03	63.33	61.34	65.13	60.99	58.75
	F-measure	0	0.55	0.50	0.42	0.57	0.55	0.65	0.48	0.55
	AUC	0.50	0.60	0.60	0.66	0.63	0.64	0.66	0.64	0.60
M1	Accuracy	50.76	65.73	66.47	63.3	61.63	67.34	65.09	70.72	64.67
	F-measure	0	0.63	0.59	0.47	0.60	0.64	0.64	0.66	0.64
	AUC	0.50	0.66	0.71	0.74	0.62	0.69	0.72	0.76	0.69
M2	Accuracy	85.24	85.13	85.96	82.24	79.14	82.69	86.45	87	83.46
	F-measure	0	0.22	0.22	0.32	0.26	0.26	0.29	0.33	0.30
	AUC	0.50	0.24	0.40	0.43	0.61	0.43	0.30	0.7	0.58
L1	Accuracy	71.4	69.15	71.61	70.79	64.40	67.36	68.43	72.2	66.28
	F-measure	0	0.28	0.27	0.19	0.37	0.34	0.33	0.11	0.37
	AUC	0.50	0.49	0.53	0.55	0.58	0.60	0.65	0.59	0.61
L2	Accuracy	50.22	49.91	52.92	53.9	57.11	58.53	58.66	58.88	55.71
	F-measure	0	0.48	0.38	0.23	0.57	0.58	0.61	0.55	0.57
	AUC	0.50	0.33	0.44	0.29	0.39	0.28	0.24	0.50	0.57

communities consisting of employees with different sets of skills. While three communities included mainly software developers (blue, red, and gray communities in Figure 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 (see Figure 2). By reviewing the S2 employees’ positions within the organization and user residence locations, we discovered that S2 has one headquarter office in the Middle East (blue, red, and orange communities in Figure 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 five well-connected communities (Figure 3). We discovered two of the company’s headquarters, both located in North America (green community in Figure 3), and also two large R&D divisions (blue and orange communities). Moreover, we succeeded in detecting the

company's senior management community (yellow 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 (Figure 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 21 communities (Figure 5). Fourteen of these communities represented nine different roles inside the organization. By examining only the residence and position information of these communities, it is possible to pinpoint the group of support engineers in South America (blue community in Figure 5). We also succeeded in detecting the company's marketing and sales division in Eastern Europe (yellow and purple communities) and the company's R&D divisions in North America and East Asia (black community). Moreover, we discovered part the company's R&D group in the Middle East and part of the North American management and sales group (green community).

5.2.6 L2 Communities

Our community detection algorithm separated the L2 social network into four communities (Figure 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 Figure 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.

6 Conclusions

This paper presents methods and algorithms that can be used to collect data from publicly available sources and analyze organizations' social networks. In order to collect organization datasets, we utilized crawling algorithms based on the homophily principle (see Algorithm 1) which can collect organizational data from online social networks like Facebook in matter of hours. Using these crawling algorithms, we collected data from the Facebook profiles of employees who worked at six different organizations. In contrast to the BFS social network crawler, which inefficiently collected organization data, the organization social network crawler presented in this paper succeeded in collecting data from

17,096 social networks users from six organizations with an average precision rate of 22.1% (Table 1).

We then used the collected organizational data to construct and analyze the informal social network among each organization's employees. By calculating eight centrality measures for each selected employee (i.e., Facebook user) we could uncover leadership roles inside the organization (Section 4). 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. Furthermore, the closeness centrality measure presented the best precision at 20 results, with an average precision at 20 of 76% (see Table 3). Also using the closeness centrality measure, we identified 85 management positions where 40% of these positions were hidden management roles and did not appear in the individuals' Facebook profiles (see Table 4).

In Section 4.2, we presented a more sophisticated method for identifying organization management positions by applying Machine Learning algorithms. Using WEKA software, we tested and evaluated several algorithms on the datasets which were based solely on calculated centrality measures. All the evaluated classifiers returned better accuracy results than the trivial ZeroR classifier. Moreover, better AUC and F-measure results were obtained using more advanced classifiers, such as Logistic, RandomForest, and IBk classifiers (see Table 5). We believe that these classification results can be improved by adding more features, such as an employee's age and gender as well as the employee's seniority, to the classification algorithm. Moreover, similar Machine Learning techniques can be applied to identify specific positions inside the organization, such as developers, sales representatives, support engineers, or senior managers.

In this study we also used the community detection algorithm to separate each organization's social network into disjointed communities (see Figures 1- 6). 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. For each organization, we discerned 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 [4]. In the case of company S2, we found this organization had many project managers who worked in different countries across the world. In companies M1 and L2, we uncovered the senior management community and their informal friendship connections. Detecting an organization's senior management community can assist in identifying undisclosed management 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 the Asian branch consisted of R&D and Specific Domain Expert employees who worked with company's employees in different continents. We discovered the L1 company's support divisions in South America and the company's sales and marketing division in Eastern Europe.

We believe this study has several future research directions. 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 [42]. These multi-label social networks can provide valuable insights and assist in better understanding the organization as a whole. 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. 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 Greg Lindsay [24], is to use the collected organizational social network to identify isolated teams inside an organization.

A research direction we have already started to pursue is to examine the implications of malicious users utilizing the collected organizational social network by collect additional information about the organization. Such users might perform a series of friend requests to company employees [11] or attack a specific employee inside a targeted organization [12].

7 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

- [1] A. Acquisti and R. Gross. *Imagined communities: Awareness, Information sharing, and privacy on the Facebook*. 2006.
- [2] T. Allen and S. Cohen. Information flow in research and development laboratories. *Administrative Science Quarterly*, pages 12–19, 1969.
- [3] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: When bots socialize for fame and money. 2011.
- [4] R. Burt. *Structural holes: the social structure of competition*. Harvard Univ Pr, 1995.
- [5] C. Campbell, P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531. ACM, 2003.
- [6] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- [7] J. Constine. Facebook’s growth since ipo in 12 big numbers. *TechCrunch*, 2013.
- [8] C. P. Diehl, G. Namata, and L. Getoor. Relationship identification for social network discovery. In *AAAI*, volume 22, pages 546–552, 2007.
- [9] J. Diesner, T. L. Frantz, and K. M. Carley. Communication networks from the enron email corpus “it’s always about the people. enron is no different”. *Computational & Mathematical Organization Theory*, 11(3):201–228, 2005.

- [10] C. Dwyer, S. Hiltz, and K. Passerini. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. In *Proceedings of AMCIS*, pages 1–12. Citeseer, 2007.
- [11] A. Elishar, M. Fire, D. Kagan, and Y. Elovici. Organizational intrusion. *ASE Cyber Security Conference (CyberSecurity)*, 2012.
- [12] A. Elishar, M. Fire, D. Kagan, and Y. Elovici. Homing socialbots: Intrusion on a specific organization’s employee using socialbots. *International Workshop on Social Network Analysis in Applications (SNAA)*, 2013.
- [13] E. Estrada and J. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.
- [14] M. Fire, O. Tenenboim, L. Puzis, R. Lesser, L. Rokach, and Y. Elovici. Computationally efficient link prediction in variety of social networks. 2012.
- [15] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.
- [16] M. Gjoka, C. Butts, M. Kurant, and A. Markopoulou. Multigraph sampling of online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1893–1905, 2011.
- [17] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 2008.
- [18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.
- [19] E. Jacobson and S. Seashore. Communication practices in complex organizations. *Journal of Social Issues*, 7(3):28–40, 1951.
- [20] M. Kilduff and D. Brass. Organizational social network research: Core ideas and key debates. *The Academy of Management Annals*, 4(1):317–357, 2010.
- [21] M. Kilduff and W. Tsai. *Social networks and organizations*. Sage Publications Ltd, 2003.
- [22] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [23] V. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.
- [24] G. Lindsay. Engineering serendipity. *New York Times*, 2013.
- [25] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, page 3, 2005.
- [26] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

- [27] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.
- [28] Y. Naddafa and S. Mutyalab. Social network analysis and community mining in organizations based on email records. 2010.
- [29] M. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.
- [30] M. Newman. The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2, 2008.
- [31] M. Newman et al. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review Series E-*, 64(1; PART 2):16132–16132, 2001.
- [32] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.
- [33] K. Provan, A. Fish, and J. Sydow. Interorganizational networks at the network level: A review of the empirical literature on whole networks. *Journal of management*, 33(3):479–516, 2007.
- [34] D. Pugh, D. Hickson, C. Hinings, and C. Turner. Dimensions of organization structure. *Administrative science quarterly*, pages 65–105, 1968.
- [35] J. Rooksby, A. Kahn, J. Keen, I. Sommerville, and J. Rooksby. Social networking and the workplace. *The UK Large Scale Complex IT Systems Initiative*, pages 1–39, 2009.
- [36] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.
- [37] J. Shetty and J. Adibi. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4, 2004.
- [38] J. Shetty and J. Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, pages 74–81. ACM, 2005.
- [39] M. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, 13(3):251–274, 1991.
- [40] C. Steinfield, J. DiMicco, N. Ellison, and C. Lampe. Bowling online: Social networking and social capital within the organization. In *Proceedings of the fourth international conference on Communities and technologies*, pages 245–254. ACM, 2009.
- [41] N. Tichy, M. Tushman, and C. Fombrun. Social network analysis for organizations. *Academy of Management Review*, pages 507–519, 1979.

- [42] J. Tyler, D. Wilkinson, and B. Huberman. E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, 21(2):143–153, 2005.
- [43] D. Wilkinson and B. Huberman. A method for finding communities of related genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5241, 2004.
- [44] G. Wilson and W. Banzhaf. 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.*, pages 3256–3263. IEEE, 2009.

Table 6: Organizations' Communities

Org.	Comm. Color	#Users	#Links	Number of Facebook Profiles with Positions	Number of Classified Users' Positions	Description
S1	Blue	30	96	6	16	R&D and administration groups in Asia
	Red	62	234	24	37	Mainly hardware verification engineers and chip designers in Asia
	Yellow	10	13	3	8	Hardware R&D
	Orange	46	197	13	21	Acquired startup company
	Gray	17	29	5	11	R&D in Asia
S2	Blue	10	16	5	6	IT group in the Middle East
	Red	109	645	25	45	R&D groups in the Middle East
	Orange	48	230	16	26	R&D groups in the Middle East
	Yellow	100	575	39	58	Managers and international PM
	Purple	4	5	1	1	Group in North America
	Gray	4	6	1	1	European group
	Cyan	39	155	15	27	R&D teams in Australia and the Middle East
M1	Blue	467	7,685	100	163	R&D division
	Red	425	11,706	86	129	Senior management
	Orange	217	2,526	46	75	R&D divisions
	Yellow	254	3,023	47	51	International consultants and support engineers
	Green	23	95	4	7	North American Headquarter
M2	Blue	1,329	23,549	504	498	R&D and SDE connected to North American and Asian employees
	Red	1,071	16,637	437	430	R&D and SDE connected to Australia, Europe and North America
	Yellow	1,348	24,080	556	551	R&D and SDE connected to Africa, North America, and Asia
	Green	921	1,058	33	32	R&D and SDE connected to North America and Asia employees
L1	Blue	141	148	45	50	South America support engineers
	Red	1,461	1,934	471	473	South American Branch (IT, PM, Support engineers, and Analysts)
	Yellow	15	172	6	7	Eastern European Pricing Analysts
	Orange	1,613	7,407	448	422	South American Branch (Management, Sales, Marketing, PM, Support engineers, and Administration)
	Purple	443	13,837	100	110	Eastern European (Marketing, Sales and Pricing) consultants and support engineers
	Green	921	1,482	243	246	Middle East R&D and North American Headquarters (Management and Sales)
	Gray	774	17,247	201	175	European Consultants and Sales and South Asian Analysts.
	Cyan	154	151	46	67	East Asian - R&D
L2	Black	143	146	9	9	North American Branch, East Asian - R&D
	Blue	2,285	42,230	220	140	East Asian Headquarter (management and consultants)
	Red	1,573	19,841	605	449	International Senior management (Senior management, Senior researchers)
	Yellow	1,588	19,023	264	218	East Asian Headquarter (R&Ds and consultants)
	Green	78	1,478	4	1	The company's amateur sports team