Inferring Social Roles and Statuses in Social Networks

Yuchen Zhao[†] Guan Wang [†] Philip S. Yu[†*] Shaobo Liu[‡] Simon Zhang[‡]
[†]University of Illinois at Chicago, USA [‡]LinkedIn Corp. [*]King Abdulaziz University, Saudi Arabia
yzhao@cs.uic.edu, {gwang26, psyu}@uic.edu, {sliu, xzhang}@linkedin.com

ABSTRACT

Users in online social networks play a variety of social roles and statuses. For example, users in Twitter can be represented as advertiser, content contributor, information receiver, etc; users in Linkedin can be in different professional roles, such as engineer, salesperson and recruiter. Previous research work mainly focuses on using categorical and textual information to predict the attributes of users. However, it cannot be applied to a large number of users in real social networks, since much of such information is missing, outdated and non-standard. In this paper, we investigate the social roles and statuses that people act in online social networks in the perspective of network structures, since the uniqueness of social networks is connecting people. We quantitatively analyze a number of key social principles and theories that correlate with social roles and statuses. We systematically study how the network characteristics reflect the social situations of users in an online society. We discover patterns of homophily, the tendency of users to connect with users with similar social roles and statuses. In addition, we observe that different factors in social theories influence the social role/status of an individual user to various extent, since these social principles represent different aspects of the network. We then introduce an optimization framework based on *Factor Conditioning Symmetry*, and we propose a probabilistic model to integrate the optimization framework on local structural information as well as network influence to infer the unknown social roles and statuses of online users. We will present experiment results to show the effectiveness of the inference.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Data Mining; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Algorithms, Experimentation

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

Keywords

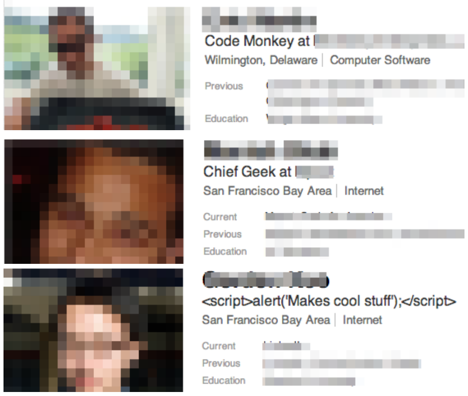
social networks; social role; social status; network inference; user modeling; LinkedIn

1. INTRODUCTION

As social networks become emerging platforms that people connect, communicate and share, there are tremendous knowledge on social networks and the online social structures reflect the social relations. Social role and status is one primary concept on individual users of a society. Social roles and statuses are defined as the part that people act as members in the society. They represent the degree of honor or prestige attached to the position of each individual [26].

In online social networks, people behave differently in social situations because they carry different latent social roles and statuses, which entail various expectations that society puts on them. There are diversified roles and statuses on different social network platforms. For example, the social roles in Twitter can be advertiser, company supporter, content contributor, information receiver, etc; the social roles in the professional network Linkedin can be engineer, salesperson, recruiter, manager, etc. Studying social roles and statuses is very helpful to gain the insights of the whole society as well as manage social resources at the individual level. Understanding social roles and statuses is crucial to many social network applications, including advertising targeting, marketing, personalization, recommendation, etc.

Conventional approaches [1][15][18][20][33] use mining techniques on textual or categorical information to predict user attributes in online social networks. Such information can be users' tweets, profiles and status updates. However, in a real social network, the textual and structured information is usually unavailable and noisy due to the following three reasons. (1) Missing Data: Previous research has shown that less than 1% of Twitter users produce 50% of the content [29]. A large number of online users view feed updates and make connections in online social networks. However, they do not include much textual and categorical information (e.g., work place, interests, geographic location, etc.) in their profiles since such information is mostly not mandatory in social networks. Thus, there is very little textual and categorical data that can be used to infer the social roles and statuses in this case. (2) Outdated Data: A large number of users do not actively update their profiles based on their latest states in a timely manner. Therefore, the inference based on such outdated data may lead to less meaningful predictions and even have adverse effects while applying misleading inference results into practice. However, such users may



[redacted]
Code Monkey at [redacted]
Wilmington, Delaware | Computer Software
Previous [redacted]
Education [redacted]

[redacted]
Chief Geek at [redacted]
San Francisco Bay Area | Internet
Current [redacted]
Previous [redacted]
Education [redacted]

[redacted]
<script>alert('Makes cool stuff');</script>
San Francisco Bay Area | Internet
Current [redacted]
Previous [redacted]
Education [redacted]

Figure 1: Three example profiles on a professional social network showing users use 'non-standard' and creative descriptions.

still use the social networks to connect with new friends and communicate with others. Thus, their network behaviors and characteristics might represent their latest states and can be useful to infer the true social roles and statuses. (3) Non-standard Data: Recently online social network users often use 'creative' and 'non-standard' descriptions and tags in their profiles. Figure 1 illustrates three users in a professional social network with creative profiles. Although all these three users belong to the same social role 'software engineer' in a professional network, they use 'code monkey', 'geek' and even a line of javascript code that prints 'make cool stuff' to describe themselves. Conventional mining approaches can not capture the real meaning in such case and therefore are unable to handle 'non-standard' data properly.

As missing, outdated and non-standard data largely exist among online users in social networks, the conventional methods cannot be applied to predict the social roles and statuses for such massive users. In the meantime, manually labeling the social roles and statuses is a time-consuming and error prone process, which can not be scalable for a real-world social network. Rather than addressing the data quality, we propose a framework that utilizes the network structures to infer user social roles and statuses while data is imperfect. Therefore, in this paper, we study the social role and status prediction problem in a semi-supervised setting. In other words, a portion of social roles and statuses are known either by mining methods with high precision and recall, or manually labeled by domain experts. The goal is to infer the roles and statuses of unlabeled users via exploring network structures and characteristics.

Despite its value and significance, the social role and status inference problem in social networks has not been fully investigated. The problem is non-trivial and brings some unique research challenges. First, what are the most essential factors and principles that can reflect social roles and statuses of users? Can we discover some fundamental patterns from the networks that can identify the role or status of online social network users? Second, since users in social networks are connected, can we quantitatively evaluate if connected users have the tendency to associate with people having similar social roles and statuses? Third, can we design a model to formalize this inference problem? How can we incorporate the network characteristics of individual users as well as the social relationships with neighbors/friends into the model in a principled way?

In this paper, we first quantify the correlations between the network structures of users and their social roles/statuses. We systematically study the effects and patterns on five social principles and concepts that are related to the inference problem: *homophily*, *triadic closure*, *reach*, *embeddedness* and *structural holes*. These correspond to a variety of key aspects of the social network, including neighbor influence, tie density, centrality, tie strength & trust and connectivity. We find patterns of homophily, the tendency of users to connect with users with similar social roles and statuses. We also show that social principles and theories all provide different degrees of predictive powers of social role and status. However, most of them are weak signals and cannot independently infer effectively.

We introduce the concept of *Factor Conditioning Symmetry* based on the social factors and formulate an optimization framework to model the local effects upon social roles and statuses given the observed social factors. To integrate the optimization framework with the homophily property, we define and propose a factor graph based probabilistic model considering both the individual network structures and the social relationships via network influence.

Our results on real social network data sets show that we can reliably infer the unknown social roles and statuses with as few as 20% labeled users given, and our proposed model significantly outperforms baselines on a number of measures. The results further suggest that network reach is the most important social factor with regard to social roles/statuses. Structural hole is complementary with most additive effect, and triadic closure is also useful, while the effect from strength of tie is more marginal.

2. RELATED WORK

Traditionally, there have been efforts to study node classification and labeling in relational data. Most of these approaches can be grouped into two categories: (1) methods using some non-network features to train an traditional local classifier, e.g., Naive Bayes and decision trees. (2) methods using network propagation on weighted edges to determine the unknown labels. A survey can be found in [1].

Recently, some research work studied the user profile inference problem under some specific settings in the online network context. For example, editors in Wikipedia have been studied in [28]. Email users of Palin's email network have been analyzed in [11]. In [20], the authors studies user attribute inference in university social networks by applying community detection. However, all previous methods either focus on a specific network such as email network[15][11] and Wikipedia network[28], or have some strong assumptions of the data. Such strong implication does not exist in general social role and status inference in social networks and thus cannot be directly applied. For example, user attributes from university students[20], such as year and department, usually are identical within the same university network community. However, a closely connected community in a general social network may correspond to a division in a company, which include people from different roles, e.g., designer, engineer, tester, salesperson and manager. Wang et al.[27] proposed a framework to discover magnet communities in social networks.

There has been some previous work on social network inference problems in different contexts. For example, Henderson et al.[10] proposed a role discovery framework on networks. But it is unsupervised and essentially a clustering approach (using matrix factorization), which is not directly applicable in the semi-supervised setting of this paper. Nevertheless, we extract the same categories of the features used in [10] as a baseline in the experiment and present results in Section 5. Myers et al.[21] proposed a method to infer latent social networks based on convex programming. Social network relevance from interpersonal communication is studied in [7]. Tang et al.[24] proposed a predictive model on inferring social ties across heterogeneous networks. The privacy concerns related to public and private profiles have been explored in [34]. Zhao et al.[32] proposed a method to solve graph classification in the PU learning setting. In addition, graphical models have been applied on network data in many applications. Topical factor graph has been introduced in [25] to analyze social influence. Tan et al.[23] used a factor graph based model to perform sentiment analysis in social networks. All above previous work focuses on different dimensions from social roles and statuses, thus cannot be directly applied. Also recently a number of approaches have been proposed to infer and predict social links [4][16], which are different from our goal: inferring roles and statuses.

3. CORRELATING SOCIAL ROLES AND STATUSES WITH SOCIAL NETWORKS

In this section, we study a number of key sociology theories and quantitatively analyze the correlations between the social roles / statuses of online users in social networks and these fundamental social psychological concepts. One should note that we do not intend to enumerate all social factors, but rather use representative social principles that reflect different aspects of network structures for individual users.

3.1 Data

We first describe the data that we use in the analyses. We obtain a sample of network data of users in the IT industry from Linkedin[1] internally . We use the social network users in this specific industry because the readers are probably more familiar with the background in the IT industry. There are four social roles that we identify: Research & Development (**R&D**), Marketing & Sales (**M&S**), Human Resource (**HR**) and Executives (**EXE**). These four roles cover the majority of individuals in the IT industry. The Executive role is defined as users with an equivalent title of ‘Director’ or higher. To construct the data set, we obtain all the users from a variety of IT companies, including Microsoft, HP, IBM, Facebook, etc. There are 45,162 users in the data set and the average node degree is 214.86. We use a mixture of classification models built on available textural/categorical information and manual labeling to identify and verify the social role/status of each user. Although we only present the analyses and explanations based on four roles in the IT industry due to the space limit, we also explored the social principles in a wide range of industries and in various social roles, and the observations are similar. Besides the IT industry data set, we add two more data sets in Section 5 to test the performance of proposed framework.

3.2 Homophily

Homophily[19] refers to the tendency of users in social networks that have ties with similar other individuals, which is also known as “birds of a feather flock together”. Homophily is a fundamental characteristic of social networks. Singla et al. [22] discovered that people who chat with each other are more likely to share interests in the MSN Messenger network. Leskovec et al. [14] also found the tendency of ‘like to associate with like’ in viral marketing.

In order to study the homophily pattern on social roles and statuses, we show the probabilities of connected users that have the same social role/status in Figure 2(a). We also plot the probabilities where the social relations are created randomly as a baseline. One can observe that the random probabilities for R&D role is much higher than other three roles. The reason is that a large portion of professionals are in R&D role in the IT industry. From the figure, it is obvious that the similarities between friends with regard to social roles are significantly larger than that of random pairwise users. This clearly suggests that the homophily pattern exists on social roles and statuses. In other words, people have the tendency to have ties with others who carry similar social roles and statuses.

3.3 Triadic Closure

Triadic closure is one of the most basic principles in social network theory on social relationships [13][9]. It involves three individuals in a social network i, j, k where i is connected to both j and k. If j and k are later connected, it may somehow imply that the connection between j and k is resulted from their connections to i. Triadic closure has been widely used to study the strength and density of social ties in sociology theories.

In order to quantitatively measure the basic pattern of triadic closure in social networks and its relation with the social role and status of each user, we use **Local Clustering Coefficient**[13][9][7] for each individual user in online social networks to study the effects of triadic closure :

Definition 1. *(***Local Clustering Coefficient [LCC]***)*

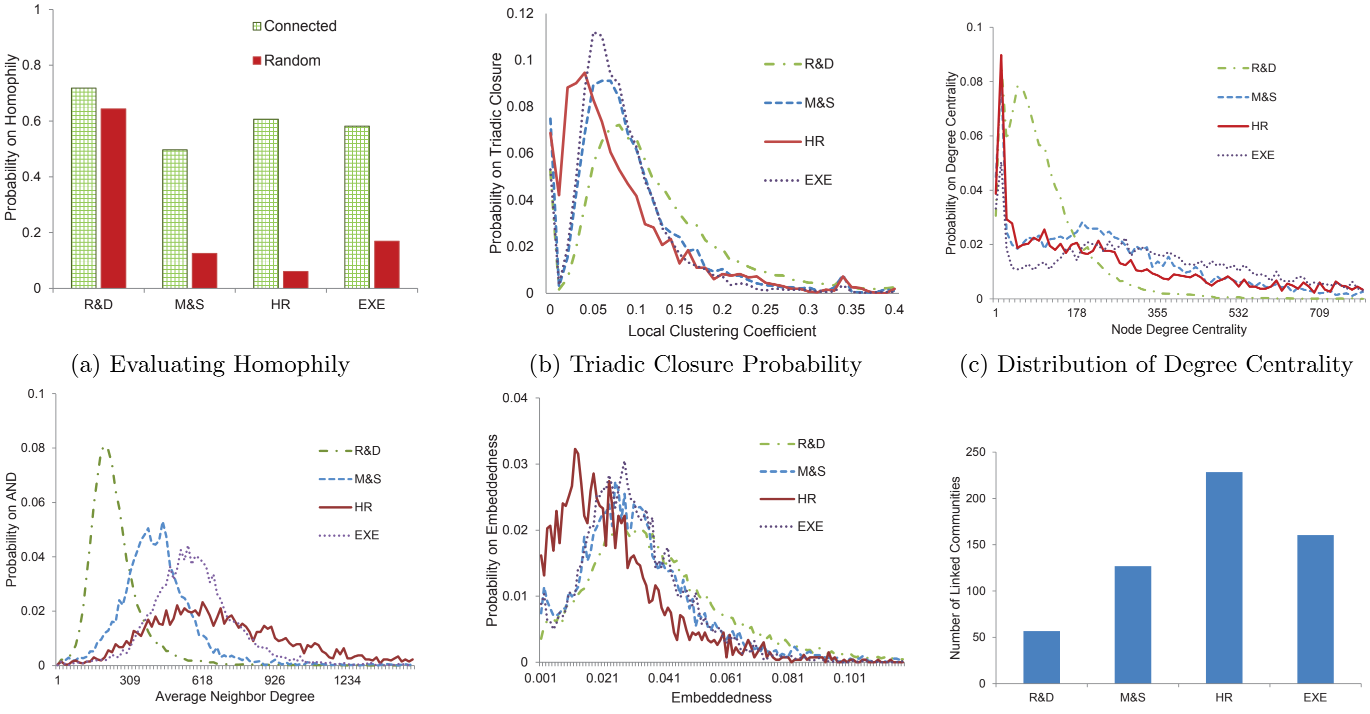
$$LCC_i = \frac{2 \cdot |\{e_{j,k} : j, k \in N_{v_i}|\}}{|N_{v_i}| \cdot (|N_{v_i}| - 1)} \tag{1}$$

where v_i is a given user node, N_{v_i} is the set of v_i's neighbors, $e_{j,k}$ is the edge connecting users j and k, and j, k are two neighbors of i.

The definition of Local Clustering Coefficient quantifies the closeness of neighbors to a clique. Intuitively, it counts the number of triangles of user i with neighbors and then is normalized by the total number of triangles if i is involved in a clique. The value of LCC should be in the range of $[0, 1]$.

Figure 2(b) shows the probability of closure in terms of local clustering coefficient for four different roles in the IT industry. One can observe that the LCC scores for the majority social network users are within the range of $[0.03, 0.1]$. The curve of users with R&D role is the flattest among that of all the roles. In addition, its peak has the largest LCC value at 0.0710. These indicate that the users in R&D have a relative dense social ties compared with users in other roles. The reason might be users in R&D usually only connect with co-workers and close friends, and their social roles as researchers or developers do not require them to actively explore new connections as other roles, e.g., marketing and sales. We further observe that the curve for HR role shifts left compared with other three curves. This observation fits our intuition quite well: users as recruiter and staffing agent connect with a large number of individuals from different backgrounds and communities. Thus, their social roles lead

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



(a) Evaluating Homophily (b) Triadic Closure Probability (c) Distribution of Degree Centrality

(d) Distribution of Average Neighbor Degree (e) Distribution of Embeddedness (f) Illustrating Structural Holes

Figure 2: Correlating Social Networks with Social Roles/Statuses

to relative low density of ties. Furthermore, the curve representing users in M&S is similar to that of users in Executive role. However, Executive role has a higher likelihood value at the peak, whereas M&S has a longer and heavier tail. All these observations clearly show that users in online social networks show diversified triadic closure patterns with different social roles because they function differently in the online society as they do offline. In the meantime, we also note that the likelihood curves in Figure 2(b) have some overlaps, thus using triadic closure alone is not effective enough to infer the social roles and statuses of online social network users.

3.4 Reach

Another important aspect of networks we study is the reach of individual users in online social networks. We first measure the reach of individuals in the network using **Degree Centrality**, which is defined as the number of ties that a user has. The distribution of degree centrality is shown in Figure 2(c). We observe that the R&D role has a distinct probability distribution compared with other three roles. The distribution of R&D role has a much steeper shape and 80% of users in R&D have node degrees which are less than 200. We further note that the distributions of M&S and EXE roles have longer tails (the tails are partially off the figure). This suggests that these two roles usually have to access more resources of the network because of the properties of their social functions. Furthermore, we notice that the EXE role has the lowest probability on low node degrees. This can also be explained by its social functions.

We also explore the **Average Neighbor Degree (AND)** which represents the '2-hop' reach of individuals in networks. The distribution of average neighbor degree is shown in Figure 2(d). To our surprise, the distributions on all four roles have more obvious distinctions than that of degree centrality. Similarly, R&D role has the steepest probability curve which represents that the number of ties associated with R&D role is relatively small. H&R role has the flattest probability function and a heavier tail. This is because users in the H&R role such as recruiters are more dependent on social resources via their connections. In addition, the EXE role has a relative larger neighborhood spread than that of the M&S role. We also test on **Median Neighbor Degree (MND)** and the distribution patterns are similar. From these observations, we can infer that the reach of individual users in the network can potentially indicate their roles and statuses, because the society puts on different expectations for each role/status.

3.5 Tie Strength and Trust

Another social principle that we study is the strength of tie. To quantitatively measure the strength of tie associated with a user in a social network, we define **Embeddedness** of user v_i as:

Definition 2. *(***Embeddedness***)*

$$Emb_{v_i} = \frac{1}{|N_{v_i}|} \sum_{v_j \in N_{v_i}} \frac{|N_{v_i} \cap N_{v_j}|}{|N_{v_i} \cup N_{v_j}|} \qquad (2)$$

Embeddedness measures the degree that individuals are enmeshed in social networks [8]. The embeddedness score of a node is high if the node has a large overlap of neighborhoods with its neighbors. In sociology, a high embeddedness score also represents *trust* and *confidence*, since the presence of mutual friends reduces the chance of misbehavior [6]. We illustrate the probability on embeddedness related to the four roles in Figure 2(e). As we can see, R&D, M&S and EXE have similar embeddedness likelihood distributions, where the curve of R&D is shifted to the right slightly. The overall embeddedness score of HR is smaller than other roles. This indicate that the tie strength associated with users in the HR role is relatively weak, although their neighborhood spread is larger than other roles.

3.6 Structural Holes

The last social principle we review is structural holes [5]. In sociology, a structural hole represents a user who connects with multiple non-interacting parties [1]. The name comes from the intuition that an "empty space" will be left in the network if remove such a user. A user of structural hole property is structurally important because she connects diverse regions in the social network. We compute the **Number of Communities (NOC)** that each individual user connects to, and use it to represent the property of structural holes. In a professional network, we define each company/organization as a community, since different companies do not interact closely and can be approximately regarded as non-interacting parties.

We present the average number of connected communities for each role in Figure 2(f). Clearly, different roles represent diverse degrees to structural holes. The values of HR and EXE are about four and three times of the value of R&D, respectively. The high value of HR is because of their large number of connections, whereas the social functions of executives require them to interact and collaborate with multiple parties via their local bridges.

3.7 Summary

In summary, different social properties represent various patterns and can be utilized to differentiate social roles and statuses, because these properties measure various aspects of network, e.g., tie density, centrality, tie strength, etc. We also note that most social factors are weak signals and cannot independently infer social roles and statuses effectively.

4. MODELING SOCIAL ROLES AND STATUSES

In the previous section, we observe that a variety of social principles and concepts show different degrees of correlations with social roles and statuses. We also discover that the existence of homophily on social roles and statuses. In this section, we first introduce Factor Conditioning Symmetry (**FCS**) and use the equality to model the local influence of individual nodes from observed social factors. Then, combined with the homophily property, we propose a factor graph based model **SRS** to infer Social Roles and Statuses by integrating these social factors and neighbor effects in a meaningful manner, such that the model is capable to infer social roles and statuses effectively.

We first introduce some notations and definitions that we will use throughout the rest of the paper. Assume we have a partially labeled social network $G = (V^L, V^U, E, X)$, where V^L is the set of labeled users with social roles/statuses and V^U is the set of unlabeled users in the social network. We note that the set of all users in the network $V = \{v_i\} = V^L \cup V^U$ and $V^L \cap V^U = \emptyset$. E represents the set of all edges in the network. X is the set of five social factors of users we studied in Section 3, i.e., *LCC, degree centrality, AND, embeddedness* and *NOC*. Let v_i be a user in the network, y_{v_i} be the label for user v_i and X_{v_i} be a vector of network attributes of user v_i. Suppose the set of labels to be $R = \{1, ..., r\}$, which contains r different roles/statuses. The goal is to infer the labels of users with unknown social roles: $y_{v_i} \in R$ where $v_i \in V^U$. The above notations are summarized in Table 1.

Table 1: Notations

Symbol	Description
$G = (V^L, V^U, E, X)$	a partially labeled social network
V^L	the set of labeled users
V^U	the set of unlabeled users
E	the set of edges
X	the set of network attributes
v_i	a user in the social network
Y	a vector of labels for all users
y_{v_i}	the label for user v_i
X_{v_i}	a vector of network attributes of user v_i
N_{v_i}	the neighbors of user v_i
$R = \{1, ..., r\}$	r different social roles/statuses
$h_k(y_{v_i}, X_{v_i})$	node feature function of v_i with role k
$f_{k,l}(y_{v_i}, y_{v_j})$	edge feature function of the edge between v_i with role k and v_j with role l

We have observed that the social factors in X have predictive powers on social roles in Section 3, and the homophily property suggests that users tend to have similar roles to their connected neighbors. Therefore, the social role inference algorithm of each user should consider: *Step 1*: its own social factors of each user; *Step 2*: the network influence from neighbors; *Step 3*: the integration of the above two steps in a principled way. In the following, we first describe how to measure the effects of the first two steps, then we discuss how to integrate them using a factor graph model.

In order to infer the unknown social roles and statuses, we construct a factor graph on a given social network G based on the *Markov assumption* that (1) the social roles of y_{v_i} are influenced by the social factors X_{v_i} associated with users and (2) the social roles are also affected by their immediate neighbors N_{v_i}.

We define two types of feature functions in the factor graph which correspond to the above two assumptions:

- **Node Feature Function:** $h_k(y_{v_i}, X_{v_i})$ models the local influence upon the social roles and statuses given attributes X_{v_i}.
- **Edge Feature Function:** $f_{k,l}(y_{v_i}, y_{v_j})$ captures the homophily effects of connected nodes with regard to the social roles and statuses.

where k and l are two indices that specify the labels of nodes.

We demonstrate a simple factorized network with four users in Figure 3 to illustrate the factorization. One can observe that the inference not only depends on the attributes of users (Step 1), but also is affected by neighbors (Step 2). For example, the prediction of v_2 is influenced by its own attributes X_{v_2} and neighbors v_1, v_3, v_4 via the edge feature function $f_{k,l}(y_{v_i}, y_{v_j})$. The task is to infer the roles and statuses of unlabeled users by maximizing the probability likelihood (Step 3). We first define the two feature functions formally as follows.

4.1 Node Feature Function

As the node feature function represents the local information of each user v_i and v_i can be either labeled or unlabeled, we define the node feature function based on whether v_i is labeled:

$$h_k(y_{v_i}, X_{v_i}) = \begin{cases} 1, & v_i \in V^L, y_{v_i} = k \\ 0, & v_i \in V^L, y_{v_i} \neq k \\ P_{v_i}^k, & v_i \in V^U \end{cases} \tag{3}$$

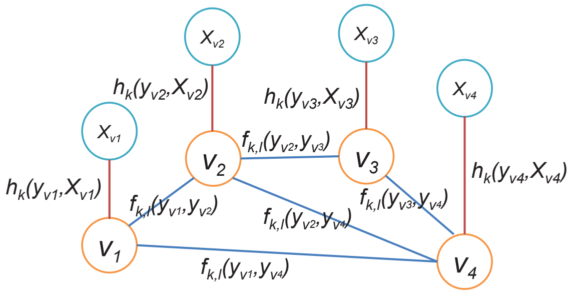


Figure 3: An example of factor graph with four users $\{v_1, v_2, v_3, v_4\}$. Each user v_i is associated with an attribute vector X_{v_i}. $h_k(y_{v_i}, X_{v_i})$ is the node feature function, whereas $f_{k,l}(y_{v_i}, y_{v_j})$ is the edge feature function defined on the edge between users v_i and v_j.

We note that if the user v_i has a label k, $h_k(y_{v_i}, X_{v_i})$ equals to 1 since the ground truth is known. P_{v_i} represents a vector of probabilistic estimates on the roles R of user v_i: $P_{v_i} = \{P^1_{v_i}, ..., P^r_{v_i}\}$.

Since usually a social network has multiple roles and statuses rather than binary labels, inspired by previous work on multi-class classifications [30], we compute the value of P_{v_i} from the **P**airwise **P**robabilities of **R**oles and Statuses (**PPR**). We define the conditional pairwise probability as:

Definition 3. *(**P**airwise **P**robabilities of **R**oles and Statuses (**PPR**))*

$$r_{k,l}(v_i, X_{v_i}) = P(y_{v_i} = k | y_{v_i} = k \ or \ y_{v_i} = l, X_{v_i}) \quad (4)$$

It is clear that Eq. 4 defines the probability of v_i being the role/status of k conditioned on v_i being either k or l given the attributes X_{v_i}. We further introduce the **Factor Conditioning Symmetry** on PPR:

Lemma 1. *(**F**actor **C**onditioning **S**ymmetry (**FCS**))*

$$r_{k,l}(v_i, X_{v_i}) \cdot P(y_{v_i} = l | X_{v_i}) = r_{l,k}(v_i, X_{v_i}) \cdot P(y_{v_i} = k | X_{v_i}) \quad (5)$$

Proof. Based on the definition of PPR, we have:

$$r_{k,l}(v_i, X_{v_i}) = \frac{P(y_{v_i} = k | X_{v_i})}{P(y_{v_i} = k \text{ or } y_{v_i} = l | X_{v_i})} \quad (6)$$

Similarly,

$$r_{l,k}(v_i, X_{v_i}) = \frac{P(y_{v_i} = l | X_{v_i})}{P(y_{v_i} = k \text{ or } y_{v_i} = l | X_{v_i})} \quad (7)$$

With Eq. 6 and 7, we have:

$$\frac{r_{k,l}(v_i, X_{v_i})}{r_{l,k}(v_i, X_{v_i})} = \frac{P(y_{v_i} = k | X_{v_i})}{P(y_{v_i} = l | X_{v_i})} \quad (8)$$

which can be rewritten as Eq. 5 in Lemma 1. □

We use Lin's method [17] to estimate *PPR* and we denote the estimated value as $\hat{r}_{k,l}(v_i, X_{v_i})$. From the *Factor Conditioning Symmetry*, an effective probability estimate on P_{v_i} should make both sides in Eq. 5 as close as possible. Therefore, we estimate the probability P_{v_i} by solving the following optimization problem:

$$\min_{P_{v_i}} \frac{1}{2} \sum_{k=1}^{r} \sum_{l=1}^{r} \left(\hat{r}_{k,l}(v_i, X_{v_i}) \cdot P^l_{v_i} - \hat{r}_{l,k}(v_i, X_{v_i}) \cdot P^k_{v_i}\right)^2$$

$$\text{s.t.} \quad P^k_{v_i} \geq 0, k = 1, ..., r; \sum_{k=1}^{r} P^k_{v_i} = 1. \quad (9)$$

Eq. 9 can be further converted to a quadratic programming form to solve:

Definition 4. *(**Factor Conditioning Optimization**)*

$$\min_{P_{v_i}} \frac{1}{2} P^T_{v_i} Q P_{v_i} \quad (10)$$

$$\textit{where} \quad Q_{kl} = \begin{cases} \sum_{m=1, m \neq k}^{r} \hat{r}^2_{m,k}(v_i, X_{v_i}), & k = l \\ -\hat{r}_{k,l}(v_i, X_{v_i}) \cdot \hat{r}_{l,k}(v_i, X_{v_i}), & k \neq l \end{cases}$$

Lemma 2. *Factor Conditioning Optimization in Eq. 10 defines a convex quadratic programming problem.*

Proof. For any non-negative vector z,

$$z^T Q z =$$

$$\frac{1}{2} \sum_{k=1}^{r} \sum_{l=1}^{r} \left(\hat{r}_{k,l}(v_i, X_{v_i}) \cdot z_l - \hat{r}_{l,k}(v_i, X_{v_i}) \cdot z_k\right)^2 \geq 0 \quad (11)$$

Therefore, the matrix Q is positive semidefinite and Eq. 10 is a convex function. □

4.2 Edge Feature Function

For the edge feature function, it models the influence from neighbors. We define it to be a function with an input of two users v_i and v_j who are connected in the social network:

$$f_{k,l}(y_{v_i}, y_{v_j}) = \frac{|e_{m,n} \in E : y_{v_m} = k, y_{v_n} = l|}{|v_m : y_{v_m} = k| \cdot |v_m : y_{v_m} = l|} \quad (12)$$

Intuitively, users with social roles/statuses k and l are more likely to be friends if these two roles are frequently connected in the observed data. Thus, $|e_{m,n} \in E, y_{v_m} = k, y_{v_n} = l|$ models the frequency of k and l being connected in the observed data. Then it is normalized by $|v_m : y_{v_m} = k| \cdot |v_m : y_{v_m} = l|$, which defines the number of connections if k and l roles are fully connected.

4.3 Global Optimization

Let Y be a vector of labels of all users. With the node feature function and edge feature function, we define the **S**ocial **R**oles and **S**tatuses Inference Model (**SRS**) as follows:

Definition 5. *(**S**ocial **R**oles and **S**tatuses Inference Model [**SRS**]) The factor graph based social roles and statuses inference model is:*

$$P(Y) = \frac{1}{Z} \left(\prod_{v_i \in V, k} h_k(y_{v_i}, X_{v_i}) \right)$$

$$\cdot \left(\prod_{v_i \in V} \prod_{v_j \in N(v_i), k, l} f_{k,l}(y_{v_i}, y_{v_j}) \right) \quad (13)$$

where Z is a normalization factor and k, l are the labels of users v_i and v_j.

The above definition defines a factorized probabilistic model with joint distribution. It is desired that the model can fit the data well, which is usually achieved by maximizing the likelihood of the given data.

We derive an iterative algorithm to maximize the joint probability distribution in Eq. 13 based on the loopy belief propagation[12]. We omit the details due to the space limit. We note that after the iterative propagation, all unlabeled nodes are assigned with social roles and statuses such that the marginal probabilities are maximized.

5. EXPERIMENT RESULTS

Here we present the effectiveness of the proposed SRS model on social role and status inference. We evaluate performance using precision, recall and F-score on each social role/status, as well as the overall accuracy. We also conduct sensitivity analyses with the fraction of labeled users and study the importance of social factors.

5.1 Data Sets

Beside the social network data set in the IT industry that we study in Section 3, we also extract a social network data set in the Finance industry from Linkedin to evaluate the proposed model. There are five social roles that we obtain in the finance industry which correspond to different social functions: *Finance*, *Sales*, *IT*, *Support* and *Operation*. The social network users in our data set cover diverse major companies in the finance industry, such as Goldman Sachs, Citi Group, Bank of America, Morgan Stanley, JP Morgan, etc. We have 76,186 users in the finance industry data set and the avenge node degree is 74.51. Similar to the IT industry data set, we employ a mixture of classification models built on available textural/categorical information and human manual labeling to label and validate the roles and statuses of the users.

To have a test on non-Linkedin data, we also use the Internet Movie Database (IMDB)[2] in the experiment. The IMDB data set has been used in previous work focusing on different problems[2][3][31]. We obtained five-year movie data from 2001 to 2005 and there are three roles in the data set: actor/actress, director and producer. The users in the data set are connected if they collaborate in a same movie.

5.2 Baselines

In order to demonstrate the effectiveness of the proposed model, we compare SRS against a number of baseline approaches. Since the SRS model considers both the local network structures of individual users and the effects from neighbor influence, we use the following approaches to show the performance of SRS from different perspectives:

(1) **BSVM:** We apply the SVM classifier on the social factors X as the first baseline. It evaluates the performance with only the local structural information of individual nodes. We refer to this method as **B**asic **SVM**.

(2) **Homophily:** In Section 3, we have seen the property of homophily associated with social roles and statuses. Previous studies on sociology[19] also suggest user attributes, such as ages, occupations and interests, may be inferred from neighbors. Therefore, we employ a baseline that applies majority votes on the labels of neighbors to infer social roles/statuses. We refer to this method as *Homophily*, which evaluates the network influence of social roles and statuses.

(3) **Community Detection:** Previous work [20] applies community detection on social network users to infer user attributes. We evaluate the performance of adopting community detection approaches to infer social roles.

(4) **RolX:** Henderson et al.[10] proposed an unsupervised role discovery approach that obtains structural feature vectors from networks and uses matrix factorization methods to cluster nodes, while each node cluster represents a role. We extract the feature vector in the same category as RolX, which includes local features, neighbor features and recursive features. Then we use SVM to train and test. We denote this method as RolX.

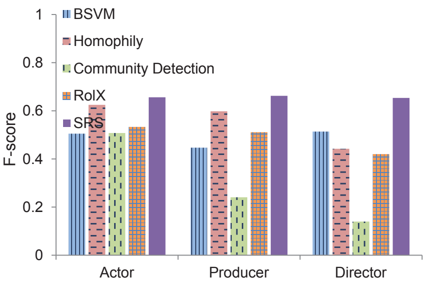


Figure 6: Results on the IMDB Data Set

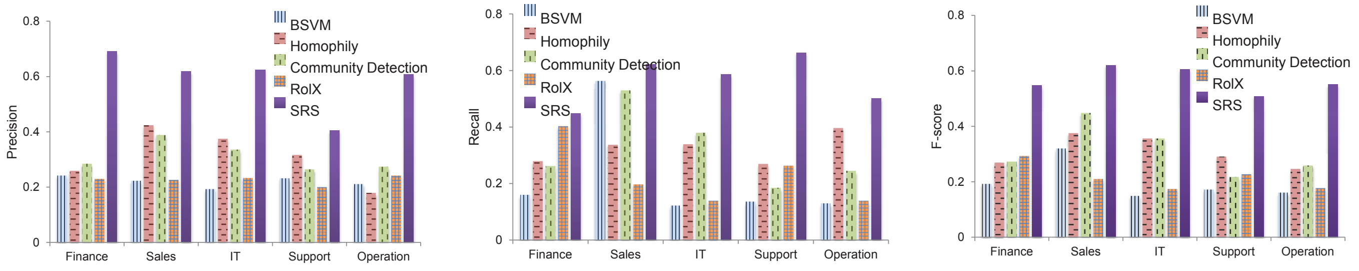
5.3 Performance in Different Roles/Statuses

We first show the performance of the proposed model SRS as well as the baselines. We use 50% users as the labeled nodes and the task is to infer the roles/statuses of the other 50% users. This setting corresponds to a natural assumption of a real-world social network. The effectiveness results in precision/recall/F-score on the Finance industry data set are illustrated in Figures 4(a), (b) and (c), respectively. F-score represents a harmonic mean between precision and recall, where F-score has a high value only if that both precision and recall are high. It is clear that SRS significantly outperforms the baselines on all the measures. Specifically, SRS improves the F-score by 0.3 - 0.45 compared with the BSVM scheme. This is in accordance with our previous observations that the social factors have predictive powers regarding social roles but they are not discriminative enough to predict effectively. This also suggests that the proposed SRS model can indeed integrate the social factors and network effects of connected users to improve the effectiveness. In addition, SRS generally gives a 0.2 - 0.3 improvement on F-score over the Homophily and Community Detection approaches. We further note that Homophily and Community Detection approaches have similar performance across different roles. This is natural since both methods only consider the property that similar users should be grouped together. Similarly, our proposed method SRS outperforms RolX in all the figures. We also note that the improvement of SRS on three measures remains consistent throughout all the five social roles in the Finance industry. This demonstrates that the proposed SRS model is consistently superior to the different baselines, irrespective of the social roles and statuses that are inferred.

We also test the SRS model and baseline approaches on the IT industry data set with four social roles. We present the results in Figures 5(a), (b) and (c) on precision, recall and F-score, respectively. They once again show that the SRS model generally outperforms four baseline models in terms of effectiveness. The above results on two real social network data sets clearly show that the SRS model is able to use the social factors of individual users and network influence through neighbors in a robust and consistent way over a variety of social roles/statuses from diverse social network contexts.

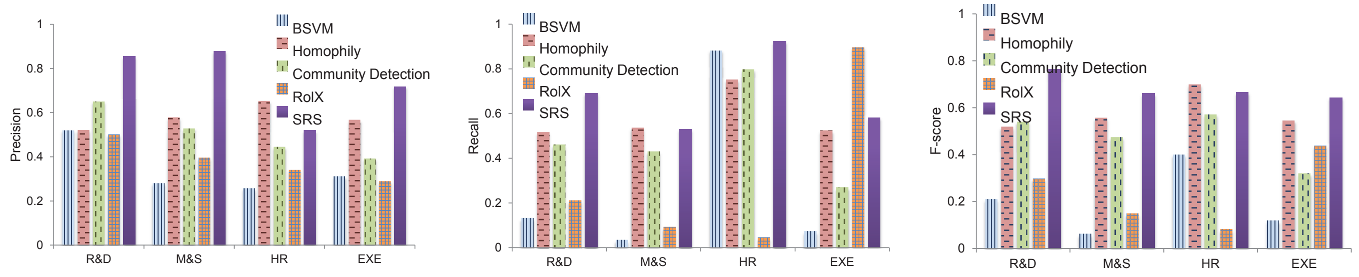
We present the results of IMDB data set in Figure 6. Due to the space limitation, we only show the F-score results. From the figure, one can observe that the proposed method SRS again outperforms all the baseline approaches. This also indicates that the proposed model is robust and consistent over different data sources.

[2] http://www.imdb.com/interfaces



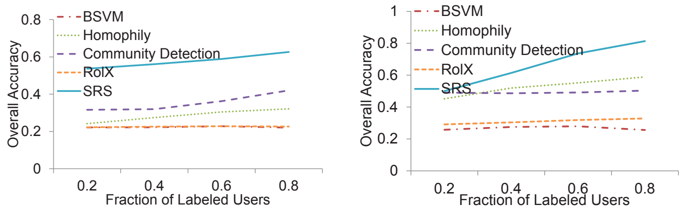
(a) Precision (b) Recall (c) F-score

Figure 4: Results on the Finance Industry Data Set



(a) Precision (b) Recall (c) F-score

Figure 5: Results on the IT Industry Data Set



(a) Finance Industry (b) IT Industry

Figure 7: Sensitivity Analysis over the Fraction of Labeled Users

5.4 Sensitivity Analysis Results

It is also valuable to test the effectiveness of the proposed model over different settings of labeled users. Therefore, we show the performance of SRS and baselines by varying the fractions that users are labeled in the data sets in Figure 7(a) and Figure 7(b) for the Finance industry data set and the IT industry data set, respectively. In both figures, the proportion of labeled users varies from 20% to 80% and it is illustrated on the X-axis. This creates a wide variety of scenarios of unknown social roles in social networks. The overall accuracy over all social roles is illustrated on the Y-axis. From both figures, it is evident that the performance generally improves with more users being labeled. This is quite natural since larger number of observed users provide more useful insights on inferring social roles and statuses. We also note that the improvement of SRS compared with baselines is consistent for all settings on the fraction of labeled users. This means the improvement of performance is not sensitive to the percentage that users are labeled. The robustness of the SRS model over large ranges of labeled users shows that the proposed approach can effectively infer social roles and statuses under different settings of observed data in social networks.

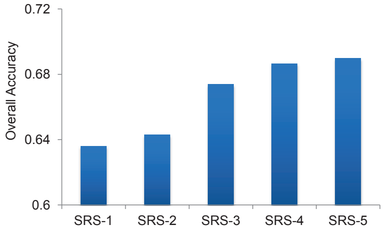
5.5 Social Factor Analysis

As we have shown incorporating social factors of individual users and neighbor influence effects in the proposed model SRS can effectively infer social roles and statuses, we further study the importance of different social factors used in SRS. We first compute the information gain of each social factor for the IT industry and present the results in Figure 8(a). Similar trend is also observed in the Financie industry which is not shown due to space limitation. The single most important social factor is the average node degree. This is line with our observations on investigating the distribution of average neighbor degree in Figure 2(d). The second most informative social factor is the degree centrality, which also measures the reach of networks. This is interesting that the average neighbor degree (representing the '2-hop' reach) is even more useful than the degree centrality (representing the '1-hop' reach). We suspect the reason is that the direct degree of individual users is sometimes noisy, e.g. an engineer can connect with over 1,000 users but a recruiter can also have less than 200 connections. However, the overall 'friends of friends' may capture the local structures more effectively since the above noises can be balanced out to a certain degree. In addition, the social factor NOC which illustrates the concept of structural holes only has a slightly lower information gain score than that of the degree centrality. This demonstrates that the connectivity and bridging effects can well reveal the roles and statuses of social network users.

We further show the overall accuracy by adding social factors one by one to SRS according to the importance of social factors in Figure 8(a), i.e., first add AND, followed by Degree, then NOC, etc. The results are illustrated in Figure 8(b), where SRS-k denotes an SRS model with the top k social features. We note that adding NOC to the model has a large gain on the overall accuracy, while adding Degree only achieves a fairly small gain. This is due to the fact that structural hole is a complementary concept not captured by reach via AND and Degree, while Degree and AND are both reach measures, i.e., degree measures of the node itself and its neighbors, respectively. LCC is also helpful indicating the usefulness of capturing the concept of triadic closures, while the effect of the strength of tie, i.e., embeddness, is

Social Factor	IG
AND	0.2607
Degree	0.1881
NOC	0.1777
LCC	0.0872
Emebeddness	0.0663

(a) Information Gain



(b) Contribution of Social Factors

Figure 8: Relative Importance of Social Factors

more marginal. We also observe that there is a clear improvement on accuracy with more social factors included. This demonstrates that each social factor we obtained has its own contribution to the performance, since the social factors measure different aspects of the network structures.

6. CONCLUSION

In this paper, we study inferring social roles and statuses in online social networks, where the categorical and textural information is often missing, outdated and non-standard. We explored five social principles and concepts that represent a variety of network characteristics and quantify their relations with social roles and statuses. We propose a novel probabilistic model SRS, which can integrate both the local social factors of individual users and network influence via neighbors in a principled way. The experiment results on two real social network data sets show that the proposed model greatly outperforms a number of baseline models and is able to effectively infer in a wide range of scenarios. In this study, we discover the patterns of homophily associated with social roles and statuses. Among the social factors, we find that network reach is the most important with regard to social roles/statuses. Structural hole is complementary with most additive effect, and triadic closure is also useful, while the effect from strength of tie is more marginal. We believe that our results provide a promising step towards understanding social behaviors and social situations at the individual level and have many potential applications in social networks.

Acknowledgement

This work is supported in part by NSF through grants IIS-0905215, CNS-1115234, IIS-0914934, DBI-0960443, and OISE-1129076, and US Department of Army through grant W911NF-12-1-0066.

7. REFERENCES

[1] C. C. Aggarwal. *Social Network Data Analytics.* Springer Publishing Company, Incorporated, 1st edition, 2011.

[2] C. C. Aggarwal, Y. Zhao, and P. S. Yu. On text clustering with side information. ICDE '12, pages 894–904.

[3] C. C. Aggarwal, Y. Zhao, and P. S. Yu. Outlier detection in graph streams. ICDE '11, pages 399–409, 2011.

[4] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM '11*, pages 635–644, New York, NY, USA.

[5] R. Burt. *Structural holes: The social structure of competition.* Harvard University Press, 1995.

[6] E. David and K. Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press, New York, NY, USA, 2010.

[7] M. De Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts. Inferring relevant social networks from interpersonal communication. In *WWW '10*, pages 301–310, New York, NY, USA.

[8] M. Granovetter. Economic Action and Social Structure: The Problem of Embeddedness. *The American Journal of Sociology*, 91(3):481–510, 1985.

[9] S. Guo, M. Wang, and J. Leskovec. The role of social networks in online shopping: information passing, price of trust, and consumer choice. In *EC '11*, pages 157–166, New York, NY, USA.

[10] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. Rolx: structural role extraction & mining in large graphs. KDD '12, pages 1231–1239. ACM, 2012.

[11] X. Hu and H. Liu. Social status and role analysis of palin's email network. In *WWW '12 Companion*, pages 531–532, New York, NY, USA.

[12] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning.* The MIT Press, 2009.

[13] G. Kossinets and D. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[14] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

[15] A. Leuski. Email is a stage: discovering people roles from email archives. In *SIGIR '04*, pages 502–503, New York, NY, USA.

[16] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03*, pages 556–559, New York, NY, USA.

[17] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. *Mach. Learn.*, 68(3):267–276, Oct. 2007.

[18] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30(1):249–272, Oct. 2007.

[19] M. McPherson, L. S. Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[20] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *WSDM '10*, pages 251–260, New York, NY, USA.

[21] S. A. Myers and J. Leskovec. On the convexity of latent social network inference. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *NIPS '10*, pages 1741–1749. Curran Associates, Inc., 2010.

[22] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW '08*, pages 655–664, New York, NY, USA.

[23] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *KDD '11*, pages 1397–1405, New York, NY, USA.

[24] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM '12*, pages 743–752, New York, NY, USA.

[25] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD '09*, pages 807–816, New York, NY, USA.

[26] H. Tischler. *Introduction to sociology.* Wadsworth Publishing Company, 2010.

[27] G. Wang, Y. Zhao, X. Shi, and P. S. Yu. Magnet community identification on social networks. KDD '12, pages 588–596.

[28] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in wikipedia. In *iConference '11*, pages 122–129, New York, NY, USA.

[29] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *WWW '11*, pages 705–714, New York, NY, USA.

[30] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.*, 5:975–1005, Dec. 2004.

[31] Y. Zhao, C. C. Aggarwal, and P. S. Yu. On graph stream clustering with side information. In *SDM*, 2013.

[32] Y. Zhao, X. Kong, and P. S. Yu. Positive and unlabeled learning for graph classification. ICDM '11, pages 962–971.

[33] Y. Zhao, N. Sundaresan, Z. Shen, and P. S. Yu. Anatomy of a web-scale resale market: a data mining approach. WWW '13, pages 1533–1544.

[34] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *WWW '09*, pages 531–540, New York, NY, USA.