

RESEARCH ARTICLE

# Inferring Social Status and Rich Club Effects in Enterprise Communication Networks

Yuxiao Dong<sup>1</sup>, Jie Tang<sup>2</sup>, Nitesh V. Chawla<sup>1\*</sup>, Tiancheng Lou<sup>3</sup>, Yang Yang<sup>1</sup>, Bai Wang<sup>4</sup>

**1** Interdisciplinary Center for Network Science and Applications, Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, United States of America, **2** Department of Computer Science and Technology, Tsinghua University, Beijing, P. R. China, **3** Google Inc, Mountain View, CA, United States of America, **4** Department of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing, P. R. China

\* [nchawla@nd.edu](mailto:nchawla@nd.edu)



 OPEN ACCESS

**Citation:** Dong Y, Tang J, Chawla NV, Lou T, Yang Y, Wang B (2015) Inferring Social Status and Rich Club Effects in Enterprise Communication Networks. PLoS ONE 10(3): e0119446. doi:10.1371/journal.pone.0119446

**Academic Editor:** Renaud Lambiotte, University of Namur, BELGIUM

**Received:** April 16, 2014

**Accepted:** January 19, 2015

**Published:** March 30, 2015

**Copyright:** © 2015 Dong et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** We believe our paper meets the data sharing requirements. Per the policies on sharing: If proprietary data are used, the manuscript must include an analysis of public data that validates the conclusions so that others can reproduce the analysis and build on the findings. We do use two data sets in the paper. One is proprietary and the other is publicly available (Enron Data), which validates the conclusions on the proprietary data, allowing others to reproduce the analysis and build on the findings. We provide these references on Enron data: Klimt B, Yang Y (2004) The enron corpus: A new dataset for email classification

## Abstract

Social status, defined as the relative rank or position that an individual holds in a social hierarchy, is known to be among the most important motivating forces in social behaviors. In this paper, we consider the notion of status from the perspective of a position or title held by a person in an enterprise. We study the intersection of social status and social networks in an enterprise. We study whether enterprise communication logs can help reveal how social interactions and individual status manifest themselves in social networks. To that end, we use two enterprise datasets with three communication channels — voice call, short message, and email — to demonstrate the social-behavioral differences among individuals with different status. We have several interesting findings and based on these findings we also develop a model to predict social status. On the individual level, high-status individuals are more likely to be spanned as structural holes by linking to people in parts of the enterprise networks that are otherwise not well connected to one another. On the community level, the principle of homophily, social balance and clique theory generally indicate a “rich club” maintained by high-status individuals, in the sense that this community is much more connected, balanced and dense. Our model can predict social status of individuals with 93% accuracy.

## Introduction

Social status refers to the relative rank (or position) that an individual holds in a society [1, 2]. For instance, an indicator of status in an enterprise setting could be the position that an individual holds in the company; and an indicator of status in a social media (such as Twitter, Facebook, etc.) could be the number of ‘followers’ ‘friends’ one has. Apart from economic incentive, social status is known to be among the important motivating forces in social behaviors [3], and achieving a higher status in a social network is often a motivating force in influencing an individual’s behavior [4, 5]. An individual with higher status is likely to have more influence or control on the message, mission, or vision within a social system.

research. Machine Learning: ECML 2004: 217{226. Lee C (2011). Communication networks part 1: the enron e-mail dataset. URL <http://sociograph.blogspot.com/2011/04/communication-networks-part-1-enron-e.html>. Our factor graph source code is also publicly available.

**Funding:** YD, NVC and YY are supported by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053, the United States Air Force Office of Scientific Research and the Defense Advanced Research Projects Agency grant #FA9550-12-1-0405; JT is supported by the National High-tech R & D Program (No. 2014AA015103), National Basic Research Program of China (No. 2014CB340500, No. 2012CB316006), Natural Science Foundation of China (No. 61222212), a research fund supported by Huawei Inc. and Beijing key lab of networked multimedia; BW is supported by National Basic Research Program of China (2013CB329603). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

**Competing Interests:** The authors claim that this work was done when TL was a Ph.D. student at Tsinghua University, and there are not any competing interests with Google. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

However, the relationship between social network structure and social status has not been widely recognized or studied, especially from the perspective of communication channels, such as voice call, short message, and email. Researchers have studied the interplay of social status and language signals [6, 7], cognitive politics [8], management science [9], race perception [10] and collectivist cultures [11]. Csermely et al. [12] find that the core and periphery network structure plays an important role in the development of a large variety of complex organisms and organizations. Tang et al. [13–15] illustrate how the opinion leaders and ordinary users correlate to social tie formation and type in social behaviors. Leskovec et al. [16, 17] develop a theory of status in online social networks, which provides a different organizing principle for directed networks of signed links. In addition, researchers have inferred demographic information such as age or gender from social networks [18–21]. Essentially, these studies derive an individual's status from computing node centralities or other node attributes such as demographics, rather than the social status of the nodes. In this paper, we consider social status from the perspective of position or title in an organization and its inter-relationship with social networks defined by communication channels.

The dynamics of social status can actively influence the strategies to make friends and maintain connections, which could re-structure the social circles of individuals. The social relationships of individuals can also rearrange the promotion or demotion of their social status and further reform the social hierarchy (or circles) in the organization. Furthermore, social status might also impact how information diffuses or cascades in a network via different communication channels. We discover that there are, indeed, different communications and social networking underpinnings as a consequence of individuals with different status in a social network. This can be especially compelling in a corporate organization that by design has individuals with different positions (or status)—managers and their direct reports (or subordinates).

To that end, we collected three communication networks from two different enterprises. Two mobile communication networks, i.e., voice call (CALL) and short message (SMS) networks, are extracted from two-month communication logs of an Asian telecommunication company with 50 managers and 182 subordinates. And an email communication (EMAIL) network is from Enron Inc. with 155 managers and 22,322 subordinates spanning over one year [22]. This allows us to consider nuances of not only the two different enterprises but also three different communication channels. Specifically, we posit and evaluate whether communication patterns in an organization vary depending on the social status of an individual, and inversely whether we can effectively build a model to infer social status using the communication patterns. We indicate the relative ranks of users in the company as their social status, i.e., managers (M) as high status and subordinates (S) as low status (Please see [Materials and Methods](#) for additional details about the data).

We consider the following questions in this paper: 1) What are the fundamental clues or patterns that may subtly reveal individuals' status in social networks? 2) How does the status of individuals influence the formation of network structure? 3) How accurately can we infer individuals' status from social network structure?

We observed that across the different communication channels (CALL/SMS/EMAIL), there exist consistent patterns arising from social status and its impact on network structure. We also discovered that various social theories and characteristics can be indicative of social status. The key findings, validated at a 95% statistical significance, include:

- High-status individuals are more likely to be spanned as “structural holes” in networks than their subordinates, which indicates that she or he is linked to individuals in different parts of

the network that are otherwise not well connected to one another [23]. Thus the “managers” act as the bridge across groups.

- At the neighborhood level, the number of common connections maintained by high-status individuals is over three times higher than the number of common connections maintained by the subordinates.
- The social community among high-status users is much more balanced and denser than subordinates, which further unveils the “rich club” [24] effects of high-status employees in enterprise networks.

Inspired by the observations around the social structure and characteristics, and their potential to infer social status in a network, we also developed a probabilistic graphical model to predict social status using the aforementioned characteristics as features. The proposed model, referred to as Factor Graph Model (FGM), associates a latent variable to each user in the communication network to represent her or his social status. We demonstrate that the presented model can accurately infer as many as 93% of social users’ status by leveraging the correlations between network structure and social status.

## Results

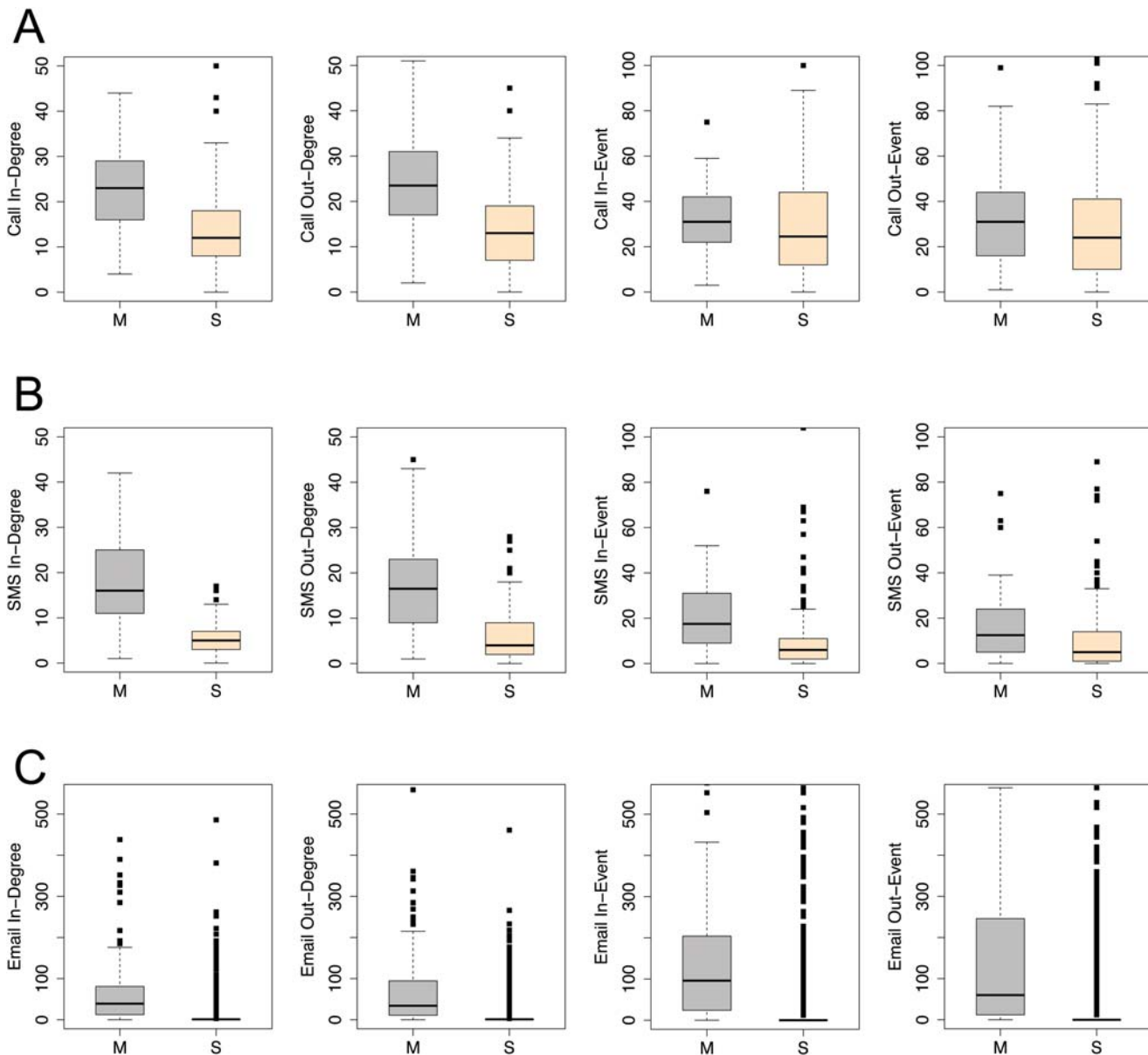
### Communication Behaviors

We first examine the communication patterns of the different staff (managers and subordinates) in the enterprise. Note, we are only focusing on the intra-company communication behaviors.

[Fig 1](#) shows the differences of four communication characteristics, including in-degree, out-degree, in-event, and out-event, between managers and subordinates from the three different channels—CALL, SMS, and EMAIL—across the two different companies. For each channel, one’s in- or out- degree is defined by the number of contacts who make or receive the communications, and the number of events is defined by the count of communications. We find that managers use mobile phones or emails more frequently than their subordinates. For example, on average, each manager makes about 60 calls (out-event) to 25 receivers (out-degree) in two months, while each subordinate only makes 40 calls to about 10 people. We can also see that both managers’ number of calls and number of receivers are around three to four times that of a subordinate. We also compare the difference among various communication channels. Clearly, there is a larger gap between manager and subordinate in SMS than CALL behaviors. As for EMAIL, the communication differences on frequency between managers and subordinates become more substantial than the mobile channels. Thus, the characteristics of interaction across different communication channels have the potential to reveal the social status of individuals in networked communications. We also studied duration of phone calls between different staff but found no significant difference between managers and subordinates in the actual duration of CALLs (See [S1 Text](#)).

### Social Status vs. Social Theories

Social status and the resulting patterns of communication characteristics, across the three different networks, give us an opportunity to study the interplay between individuals’ status and their network influence via several well-known social theories, including structural hole, social balance, homophily and social clique. The observations from CALL, SMS and EMAIL channels unveil strong “rich club” phenomenon [25, 26] in enterprise networks. First, managers are much more likely to be spanned as “structural holes” in networks than subordinates. Second,



**Fig 1. Communication Attributes vs. Social Status.** (A). CALL attributes; (B). SMS attributes; (C). EMAIL attributes. Each box plot represents the distribution of the number of attributes per time unit (each month in A and B, and each year in C), including in-degree, out-degree, in-event and out-event.

doi:10.1371/journal.pone.0119446.g001

they also maintain 3-8 times more common neighbors compared to subordinates. Third, the managers form more balanced triads than those by the subordinates. Finally, the social community among managers is much closer and denser than subordinates. We use a null model on the randomized version of the empirical data and report the statistical significance of the results on structural hole and social balance. The details of the null model are introduced in *Materials and Methods*.

**Structural Hole.** The principle, that individuals can benefit from serving as intermediaries between others who are not directly connected, forms the underpinning for the theory of

**Table 1. Structural Holes vs. Social Status.** The probability that staffs in companies are spanned as structural holes (SH) extracted from communication network structures. M: Managers; S: Subordinates.

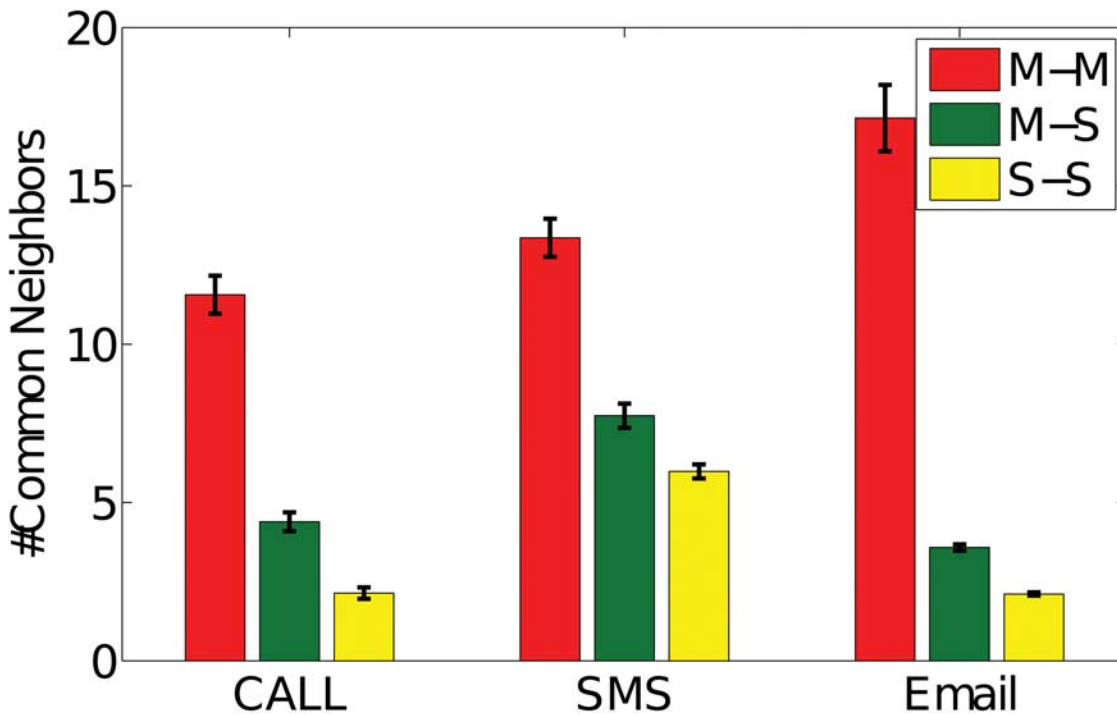
	CALL	SMS	EMAIL
M as SH	0.700 ****	0.550 ****	0.430 ****
M as SH (Random)	0.207	0.207	0.007
S as SH	0.300 ****	0.450 ****	0.570 ****
S as SH (Random)	0.793	0.793	0.993

\* $p < 0.05$ ;  
 \*\* $p < 0.01$ ;  
 \*\*\* $p < 0.001$ ;  
 \*\*\*\* $p < 0.0001$

doi:10.1371/journal.pone.0119446.t001

structural holes [23, 27–29]. For example, a manager in a department with a diverse range of connections can be considered as a structural hole, with a number of ties to staff in other different departments. The fundamental question is: do high-status individuals span as structural holes in networks? Here we consider the HIS algorithm [29] to estimate the likelihood of each node in the network to span as structural hole, and then categorize them into two groups based on the likelihood. We select the top 21% (the percentage of managers in mobile networks) of individuals in CALL and SMS networks and the top 0.67% individuals (the percentage of managers in EMAIL network) in EMAIL Network as structural holes based on their HIS scores and the rest as ordinary individuals. Our conjecture is that if the status does not correlate with structural holes, the probability that managers are structural holes should be the same as the ratio of managers (21% in mobile and 0.67% in EMAIL). However, our analysis in Table 1 clearly shows that managers are more likely (70% in CALL, 55% in SMS, and 43% in EMAIL) to be spanned as structural holes across the three networks. In other words, the structural holes extracted from enterprise communication network structure reveal the social status of staff in their company. This can be explained by the fact that managers usually need to operate the responsibility of correspondents and organizers within the company, especially for the experience for connecting different departments or groups to cooperate.

**Link Homophily.** Homophily is the tendency of individuals to associate and bond with similar others [30, 31]. The presence of homophily has been widely discovered in some form or another, including age, gender, class, and organizational role. Lazarsfeld and Merton [30] argued that individuals with similar social status are more likely to associate with each other, which is called status homophily. Particularly, following the theory of homophily [31], we consider the neighbors of one individual as her or his attributes and examine the correlation of the neighbors of different individuals. Then, the concept of link homophily [15] tests whether two individuals who share more common neighbors will have a tendency to have similar social status in the company. The average number of common neighbors by two managers ranges from 12 to 17 across the three networks from Fig 2. Surprisingly, the average number of common neighbors of pairs of subordinates only reaches around two in CALL or EMAIL networks and six in SMS network. As homophily phenomena gets more reflected among managers, we can contend that two individuals are much more likely to be two managers in the company if they share more common neighbors. Managers’ ability of creating and maintaining social connections in enterprise networks is more prominent than subordinates’. This could have the potential to further promote their status in companies, which further highlights the rich club effect.



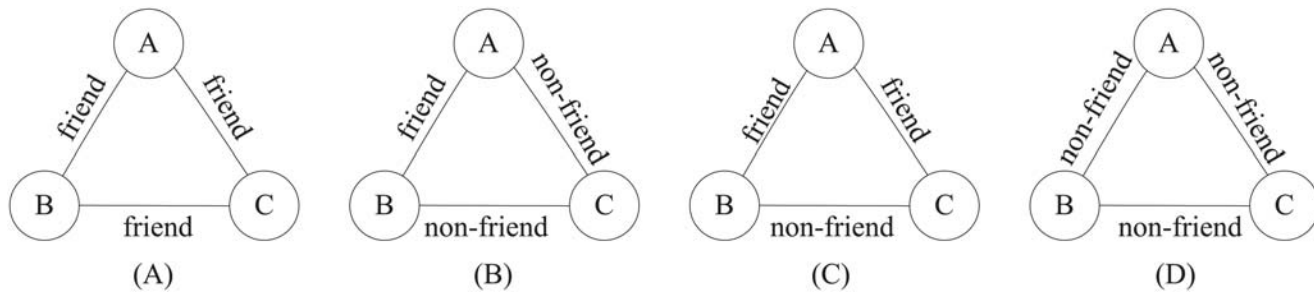
**Fig 2. Homophily vs. Social Status.** The number of common friends of each pair of employees are considered as the measurement of link homophily in social network, thus generating three different types of ties: M-M, M-S, S-S. Error bars show the confidence interval with significance level  $\alpha = 0.05$ .

doi:10.1371/journal.pone.0119446.g002

**Social Balance.** Triad is one of the simplest grouping of individuals that can be studied and is mostly investigated by microsociology [2]. Given a triad  $(A, B, C)$ , if  $A$  and  $B$  are friends and if  $A$  and  $C$  are friends as well, then social balance theory [32] suggests that  $B$  and  $C$  are also likely to be friends, which results in a balanced triad. Similarly it is also a balanced triad where  $A$  and  $B$  are friends, while  $B$  and  $C$ , and  $A$  and  $C$  are non-friends. The other structured triads are called unbalanced triads. For each node, the balance ratio is defined as the ratio of the number of balanced triads to the number of unbalanced triads. An illustrative example is shown in Fig 3. According to the social balance theory, a stable social network tends to be a balanced structure by containing densely connected triads [33]. Here, we aim to test whether the enterprise communication networks also have balanced structures with respect to social status. Given one individual and her/his ego network, we calculate three kinds of social balance ratios, i.e., the balance ratio among her/his manager-friends  $M-sb$ , the balance ratio among her/his subordinate-friends  $S-sb$ , and the overall balance ratio among all her/his friends  $sb$ . We find that the managers' overall balance ratios are larger than the subordinates' across all the three channels in Table 2. Moreover, the managers are more likely to form balanced structure among their manager-friends, and the subordinates with subordinates. In other words, the individuals in organizations have the tendency to create or maintain balanced relationships with people of the same status; this phenomenon coincides with the link homophily observed above. We conjecture that the relatively high status empowers the managers to connect with more people and maintain the relationships within the enterprise, enhancing the chance to promote their status.

**Social Clique.** Clique is a concept in both social sciences and graph theory. In social sciences, clique is used to describe a group of persons who interact with each other more regularly and intensely than others in the same setting [34]. In graph theory, a clique is defined as a





**Fig 3. Illustration of structural balance theory.** Triads (A) and (B) are balanced, while (C) and (D) are not balanced.

doi:10.1371/journal.pone.0119446.g003

subset of nodes such that for any two nodes, there exists an edge connecting them [35]. Interacting with people in one clique has the indication of close and strong relationships with each other. Here we aim to examine how managers and subordinates form cliques and to which extent they are connected. We build two sub-networks that only contain managers or subordinates respectively for each type of a network derived from each of the communication channels. Fig 4 shows the distributions of clique size, conditioned on the status of individuals (employees in the enterprise). For reference, we also plot the overall clique distribution in each full network. It is obvious that the distributions of managers and subordinates are quite different. The maximal cliques for 50 managers are 12, 20 in CALL and SMS networks, respectively. It is interesting that the clique sizes vary across the two different communication channels, albeit in the same company. The clique size for the 155 managers in EMAIL is 9. In comparison, the maximal clique sizes for 182 subordinates in the CALL and SMS networks are 9 and 10, respectively; and the clique size for the 22,232 subordinates in the EMAIL network is 9. We also find that the most frequent cliques in subordinates' sub-networks are 4/5/3-clique in CALL/SMS/EMAIL, respectively, which are much smaller than the 11/13/4-clique in the CALL/SMS/EMAIL for managers' sub-networks.

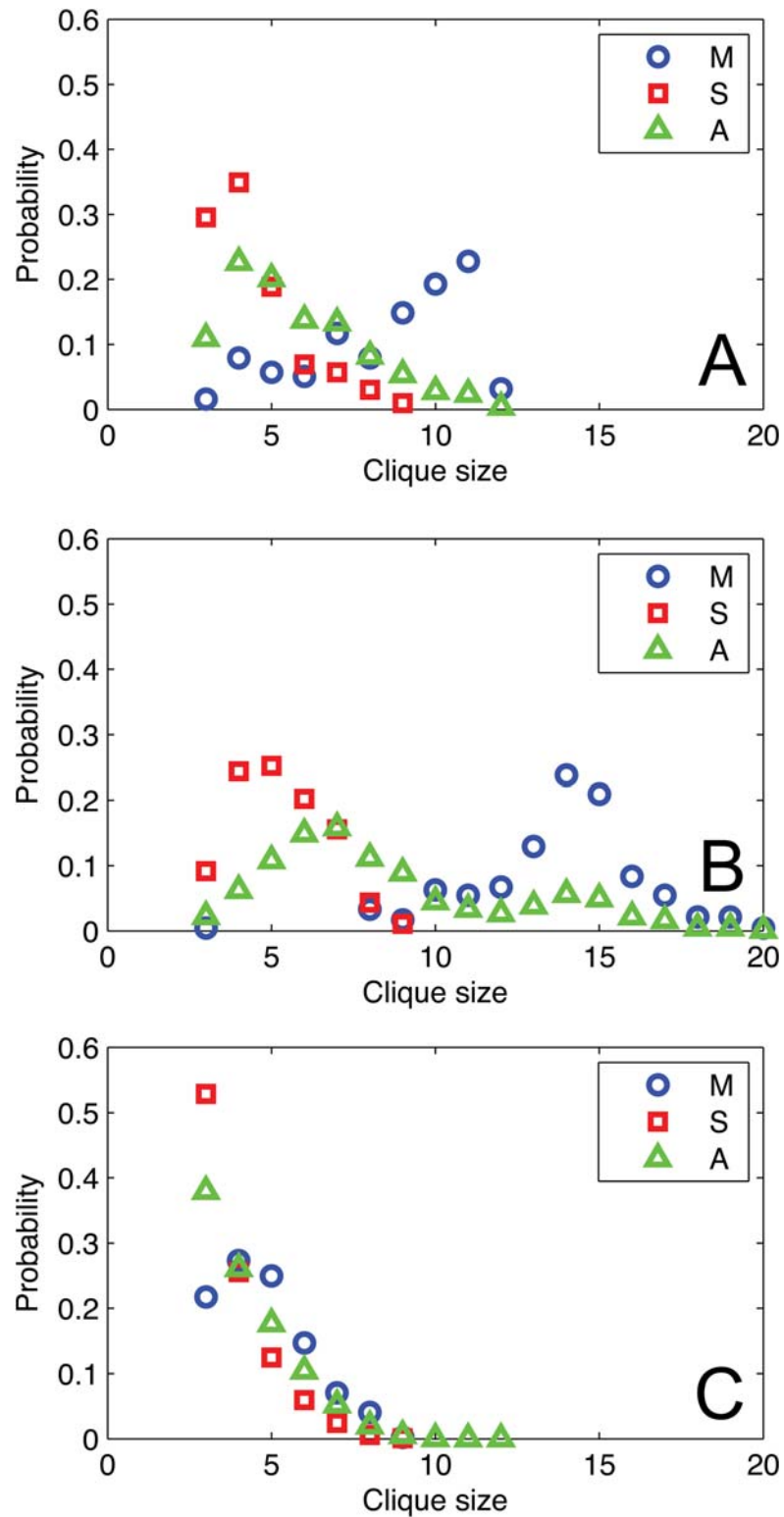
In addition, we note the difference in the network properties between the two sub-networks in Table 3, even though they belong to the same enterprise mobile networks. Specifically, the clustering coefficient for sub-network (M) is two times higher than sub-network (S), which coincides with our observations above that the managers share more dense connections than the subordinates in an organization. The correlations between social status and several social theories provide the evidence of “rich club” maintained by high-status individuals.

**Table 2. Social Balance vs. Social Status.** The social balance ratio of staffs with different status in companies in three communication networks. M: Managers; S: Subordinates. M-sb: social balance ratio of manager-friends; S-sb: social balance ratio of subordinate-friends; sb: social balance ratio of all friends.

	M (CALL)	S (CALL)	M (SMS)	S (SMS)	M (EMAIL)	S (EMAIL)
M-sb	0.569***	0.348****	0.546****	0.468****	0.455****	0.047****
S-sb	0.174***	0.254**	0.289	0.299	0.066**	0.082****
sb	0.340	0.312	0.325	0.311	0.165	0.124

\* $p < 0.05$ ;  
 \*\* $p < 0.01$ ;  
 \*\*\* $p < 0.001$ ;  
 \*\*\*\* $p < 0.0001$

doi:10.1371/journal.pone.0119446.t002



**Fig 4. Social Clique vs. Social Status.** Distribution of social clique in enterprise communication networks. M: Managers; S: Subordinates; A: All employees.

doi:10.1371/journal.pone.0119446.g004



**Table 3. Topology characteristics of sub-networks.** M: Managers; S: Subordinates.

attributes	M(CALL)	S(CALL)	M(SMS)	S(SMS)	M(EMAIL)	S(EMAIL)
#nodes	50	182	50	182	155	22322
#edges	511	1858	693	1627	964	29912
clustering coefficient	0.63	0.31	0.78	0.45	0.46	0.10
associative coefficient	-0.04	0.02	-0.29	-0.12	0.03	-0.14

doi:10.1371/journal.pone.0119446.t003

### Social Status Prediction

We now consider the core problem: can we leverage the observations from communication behaviors to infer an individual’s social status? We use several classical data mining models to infer social status, including Naive Bayes (NB), Bayes Network (BNET), Logistic Regression Classification (LRC) and Conditional Random Fields (CRF) [36]. We also propose a Factor Graph Model (FGM) to leverage the social theories to help status prediction. We use Weka [37] for NB, BNET and LRC methods. NB, BNET and LRC use communication attributes to train classification models and apply them to predict individual status. For CRF and FGM models, both communication and social features are used to infer the labels of individuals. We quantitatively evaluate the performance of inferring individual status in terms of weighted Precision, Recall, F1-Measure and Accuracy. The setup of status prediction is introduced in [S1 Text](#).

[Table 4](#) shows the results of four methods for inferring individual status in these communication networks. Clearly, our model FGM yields better performance than other alternative methods. In CALL network, we can find that FGM achieves about 85% F1-Measure scores and Accuracy. In the SMS network, both F1-Measure and Accuracy of FGM reach 92%. However, the text-messaging network seems to reveal more status differentials than the calling network.

**Table 4. Social status prediction in enterprise communication networks.** NB: Naive Bayes, BNET: Bayes Network, LRC: Logistic Regression Classification, CRF: Conditional Random Fields, FGM: Factor Graph Model.

Status	Method	Precision	Recall	F1	Accuracy
CALL	NB	0.7334	0.7625	0.7416	0.7625
CALL	BNET	0.7409	0.6934	0.7110	0.6934
CALL	LRC	0.7065	0.6795	0.6904	0.6795
CALL	CRF	0.8078	0.8095	0.8086	0.8095
CALL	FGM	0.8514	0.8508	0.8511	0.8508
SMS	NB	0.8693	0.8734	0.8648	0.8734
SMS	BNET	0.8497	0.8512	0.8483	0.8512
SMS	LRC	0.8129	0.7850	0.7935	0.7850
SMS	CRF	0.8720	0.8761	0.8740	0.8760
SMS	FGM	0.9321	0.9276	0.9298	0.9276
EMAIL	NB	0.8847	0.8993	0.8847	0.8598
EMAIL	BNET	0.8936	0.9054	0.8164	0.8755
EMAIL	LRC	0.8761	0.8772	0.7653	0.8483
EMAIL	CRF	0.9033	0.8902	0.8967	0.8902
EMAIL	FGM	0.9319	0.9383	0.9373	0.9383

doi:10.1371/journal.pone.0119446.t004

By combining the CALL and SMS networks together, the predictive performance can be improved by 2%-3% compared with the results from the SMS network. The prediction performance from EMAIL channel outperforms the other two mobile channels by 1-8% in terms of Accuracy. In summary, the social status of 85% – 93% of individuals can be inferred from their communication interactions among their colleagues. This prediction results further confirm our observations on communication behaviors and social theories are general across different companies, even with different communication channels (CALL vs. SMS vs. EMAIL).

## Discussion

Interactions within a corporate enterprise are representative of the artifacts of the status of individuals in the enterprise. To that end, we analyzed the communication interactions (call, message, email) of managers and subordinates in a corporation by network properties. We find that the managers (or the high-status individuals) in a corporate hierarchy congregate as a “rich club”, maintaining denser and closer connections than the subordinates (or the low-status individuals) in the same enterprise. This phenomenon stands out with different social characteristics across the networks derived from different communication channels. This result also becomes evident from the larger social cliques of the managers.

The structure of communication networks collected from different channels also suggests that the managers (high-status) are more likely to be spanned as structural holes and maintain more balanced social circles than the subordinates (low-status). The number of common neighbors also indicates the stronger homophily between the high-status individuals. The social circles maintained by high-status individuals are much larger and closer than subordinates, simultaneously, the social capital, namely the collective benefits from social circles, that managers own are much greater than subordinates based on their social circles.

Finally, we study to what extent the social status of individuals can be inferred from their communication network structure. That is, can the observed communication and social characteristics in networks be used to develop a model for inferring the actual social status? This is an important experiment for lending an insight into predicting the status of an individual when the only observed information is the social network and the patterns of communication behavior. We propose a factor-graph based model, and demonstrate that our model is able to achieve about 85% of predictive accuracy using CALL network, about 92% of predictive accuracy using SMS, and over 93% of predictive accuracy using EMAIL. The performance trends clearly show that we are able to capture essential properties of social theories, which are general across different communication channels in different companies.

Social status characterizes the strategies that people organize their social connections. It offers a great potential to understand the underlying principles that drive human social activities and behaviors. Awareness of individual status can provide new perspectives in network science problems, such as link prediction, influence propagation, and community detection, which were considered in black-and-white network structure before.

## Materials and Methods

### Communication Network Data

The mobile dataset used in this paper is extracted from a large collection of call and text-message records, which span over two months. We construct mobile communication networks for all employees in a telecommunication company with 15 departments, where there are 232 staff, which include one CEO, four Vice Presidents, 45 department managers (each department has three managers) and 182 subordinates. However, the dataset provided for the study contains only two levels of status (manager or subordinate) for each individual, resulting into two

**Table 5. Network topology characteristics.** M: Managers; S: Subordinate.

attributes	CALL	SMS	EMAIL
#nodes	232	232	22477
#edges	3340	3406	44728
clustering coefficient	0.3326	0.4761	0.1241
associative coefficient	0.1195	-0.0894	-0.2153

doi:10.1371/journal.pone.0119446.t005

groups of managers (high-status) and subordinates (low-status). We construct two sub-networks for the mobile enterprise—one using the voice calls (CALL) and the other using the text messaging service (SMS).

The EMAIL network is extracted from the Email communication logs of Enron Inc. [22, 38]. It consists of 164,080 emails between 22,477 Enron employees, including 155 managers and 22,322 subordinates. Table 5 lists statistics of the three networks. *cc* is the average clustering coefficient, *ac* is the associative coefficient and *cn* denotes the number of components in the network.

### Null Model for Different Social Status of Staff

We use a null model [21] to validate the statistical significance of our social observations. A straightforward way to measure this is to compare the real values to the null model where the status of people is randomly assigned. We compare the real data to 10,000 randomized cases where managers and subordinates are randomly shuffled. First, we simulate the random process of allocating status to individuals with the same ratio as in the real data (50 managers and 182 subordinates of mobile networks, and 155 managers and 22,322 subordinates of EMAIL network) 10,000 times for the underlying network structure. The difference on social observation between empirical data  $x$  and the null model  $\tilde{x}$  can provide the interpretation for the deviation. The  $z$  score can examine whether the null model is true, i.e., there are no distinctions between individual status given the underlying communication network structure.

$$z(x) = \frac{x - \mu(\tilde{x})}{\sigma(\tilde{x})}$$

where  $\mu(\tilde{x})$  and  $\sigma(\tilde{x})$  are the mean and standard deviation of the observations on the null model. The null hypothesis is rejected at 2 sigmas (corresponding to the  $p$ -value  $< 0.01$ ).

### Factor Graph Model for Status Prediction

From the machine learning perspective, if we consider each individual as an instance in a learning model, we will speak of each individual as a relatively ranking position (such as manager and subordinate). If we assume that all data instances (individual-based instances) are independent, then we can leverage standard machine learning algorithm to learn a classifier. However, the instances are not necessarily independent. To that end, we also leverage a probabilistic graphical model [39] to build our models. We consider factor graph [40] that is able to model the correlations among variables. We have also used The Factor Graph Model (FGM) in our previous works [15, 18].

Let  $G = (V, E)$  denote the communication network, where  $V$  is a set of individuals and  $E \subseteq V \times V$  is a set of edges. Each edge  $e_{uv}$  is created if individual  $u \in V$  and individual  $v \in V$  have communication logs between each other. In the FGM model, the communication network  $G$  is

directly transformed as a factor graph with each node as an individual and each edge as communication relationship between two individuals. For each individual node  $v \in V$ , a hidden variable  $y_v$  is introduced to represent the relative rank (social status) of the corresponding individual. For example, in our mobile network, we use each individual's position in a mobile company as the status, thus we can define two ranks for  $y$  to respectively represent manager and employee. Given some labeled training data  $(G, \mathbf{X}, Y)$ , where  $\mathbf{X}$  is the individual attribute matrix, the objective function can be defined as a log-likelihood function

$$\mathcal{O}(\theta) = \log P(Y|G, \mathbf{X}, \theta)$$

where  $\theta$  are parameters to learn from the training data. If we consider  $P(\cdot)$  as an exponential distribution over various available features in the social network, we can formally define the log-likelihood objective function as:

$$\mathcal{O}(\theta) = \sum_v \sum_k^K \theta_k f_k(x_v, y_v) + \sum_c^C \theta_c f_c(y_u, y_v, y_w) - \log Z \tag{1}$$

where  $f_k(x_v, y_v)$  is the  $k$ -th feature defined over node  $v$ ,  $K$  is the number of features for each node,  $f_c(y_u, y_v, y_w)$  is the  $c$ -th correlation feature defined over each triangle  $c = \{u, v, w | e_{uv}, e_{uw}, e_{vw} \in E\}$ ,  $C$  is the set of all closed triangles in the graph  $G$ , and  $Z$  is a normalization factor. The details of the features used for factor graph model are defined in [S1 Text](#).

For model learning, the task is to find a parameter configuration  $\{\theta\}$  to maximize the log-likelihood objective function [Eq 1](#), i.e.,

$$\theta^* = \operatorname{argmax} \mathcal{O}(\theta) \tag{2}$$

In this work, we use a gradient descent method (or a Newton-Raphson method) to optimize the convex objective function [\[36, 39\]](#). Specifically, we first write the gradient of each  $\theta_k$  with regard to the objective function:

$$\frac{\partial \mathcal{O}(\theta)}{\partial \theta_k} = \mathbb{E}[f_k(x_v, y_v)] - \mathbb{E}_{P_{\theta_k}(y_v|x_v, G)}[f_k(x_v, y_v)] \tag{3}$$

where  $\mathbb{E}[f_k(x_v, y_v)]$  is the expectation of feature function  $f_k(x_v, y_v)$  given the data distribution and  $\mathbb{E}_{P_{\theta_k}(y_v|x_v, G)}[f_k(x_v, y_v)]$  is the expectation of feature function  $f_k(x_v, y_v)$  under the distribution  $P_{\theta_k}(y_v|x_v, G)$  given by the estimated model.

The graphical structure in the above model can be arbitrary and may contain circles, which makes it intractable to directly calculate the marginal distribution  $P_{\theta_k}(y_v|x_v, G)$ . To solve this challenge, we use Loopy Belief Propagation, due to its ease of implementation and effectiveness, to approximate the marginal distribution  $P_{\theta_k}(y_v|x_v, G)$ . Then we are able to obtain the gradient by summing up all the factor graph nodes. Finally, we update each parameter with a learning rate  $\eta$  with the gradient. Related algorithms can be found in [\[15, 41\]](#).

With the estimated parameter  $\theta$ , we can now assign the value of unknown labels  $Y$  by looking for a label configuration that will maximize the objective function, i.e.

$$Y^* = \operatorname{argmax} \mathcal{O}(Y|G, \mathbf{X}, \theta) \tag{4}$$

Obtaining exact solution is again intractable. The LBP is utilized to calculate the marginal probability for each node in the factor graph. Finally, labels that produce the maximal probability will be assigned to each factor graph node.

## Supporting Information

**S1 Text. Inferring Social Status and Rich Club Effects in Enterprise Communication Networks.**

(PDF)

## Author Contributions

Conceived and designed the experiments: YD JT NVC TL YY BW. Performed the experiments: YD TL YY. Analyzed the data: YD TL YY BW. Wrote the paper: YD JT NVC BW.

## References

1. Rege M. Why do people care about social status. *Journal of Economic Behavior and Organization*. 2008; 66:233–242 doi: [10.1016/j.jebo.2006.04.005](https://doi.org/10.1016/j.jebo.2006.04.005)
2. Easley D, Kleinberg J. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press; 2010.
3. Milroy L, Milroy J. Social network and social class: Toward an integrated sociolinguistic model. *Language in Society*. 1992 3; 21:1–26 doi: [10.1017/S0047404500015013](https://doi.org/10.1017/S0047404500015013)
4. Max W. *Class, Status, Party*. *Essays in Sociology*. 1946;p. 180–195.
5. Rodkin PC, Berger C. Who bullies whom? Social status asymmetries by victim gender. *International Journal of Behavioral Development*. 2008; 22:473–485 doi: [10.1177/0165025408093667](https://doi.org/10.1177/0165025408093667)
6. Danescu C, Lee L, Pang B, Kleinberg J. Echoes of power: Language effects and power differences in social interaction. In: *WWW'12*; 2012. p. 699–708.
7. Gilbert E. Phrases that signal workplace hierarchy. In: *CSCW'12*; 2012. p. 1037–1046.
8. Kraus MW, Callaghan B. Noblesse Oblige? Social Status and Economic Inequality Maintenance among Politicians. *PLoS ONE*. 2014 01;.
9. Tichy NM, Tushman ML, Fombrun C. Social Network Analysis For Organizations. *Academy of Management Review*. 1979; 4(4):507–519 doi: [10.5465/AMR.1979.4498309](https://doi.org/10.5465/AMR.1979.4498309)
10. Freeman JB, Penner AM, Saperstein A, Scheutz M, Ambady N. Looking the Part: Social Status Cues Shape Race Perception. *PLoS ONE*. 2011 09; 6(9):e25107 doi: [10.1371/journal.pone.0025107](https://doi.org/10.1371/journal.pone.0025107) PMID: [21977227](https://pubmed.ncbi.nlm.nih.gov/21977227/).
11. Brown C, Efstratiou C, Leontiadis I, Quercia D, Mascolo C. Tracking serendipitous interactions: How individual cultures shape the office. In: *ACM CSCW'14*; 2014. p. 1072–1081.
12. Csermely P, London A, Wu LY, Uzzi B. Structure and dynamics of core-periphery networks. *Journal of Complex Networks*. 2013 12; 1:93–123 doi: [10.1093/comnet/cnt016](https://doi.org/10.1093/comnet/cnt016)
13. Tang J, Lou T, Kleinberg J. Inferring Social Ties across Heterogeneous Networks. In: *WSDM'12*; 2012. p. 743–752.
14. Dong Y, Tang J, Wu S, Tian J, Chawla NV, Rao J, et al. Link Prediction and Recommendation across Heterogeneous Social Networks. In: *IEEE ICDM'12*; 2012 p. 181–190.
15. Lou T, Tang J, Hopcroft J, Fang Z, Ding X. Learning to predict reciprocity and triadic closure in social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2013; 7(2):5 doi: [10.1145/2499907.2499908](https://doi.org/10.1145/2499907.2499908)
16. Leskovec J, Huttenlocher D, Kleinberg J. Signed networks in social media. In: *CHI'10*; 2010. p. 1361–1370.
17. Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks. In: *WWW'10*; 2010. p. 641–650.
18. Dong Y, Yang Y, Tang J, Yang Y, Chawla NV. Inferring User Demographics and Social Strategies in Mobile Social Networks. In: *KDD'14*. ACM; 2014. p. 15–24.
19. Palchikov V, Kaski K, Kertész J, Barabási AL, Dunbar RIM. Sex differences in intimate relationships. *Scientific Reports*. 2012;2:370.
20. Szell M, Thurner S. How women organize social networks different from men. *Scientific Reports*. 2013 Jul;3.
21. Kovanen L, Kaski K, Kertész J, Saramäki J. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences*. 2013;.

22. Klimt B, Yang Y. The enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004*. 2004;p. 217–226.
23. Burt RS. *Structural Holes: The Social Structure of Competition*. Cambridge, Mass.: Harvard University Press; 1995.
24. Colizza V, Flammini A, Serrano MA, Vespignani A. Detecting rich-club ordering in complex networks. *Nat Phys*. 2006 Feb; 2(2):110–115 doi: [10.1038/nphys209](https://doi.org/10.1038/nphys209)
25. Zhou S, Mondragon RJ. The rich-club phenomenon in the Internet topology. *Communications Letters, IEEE*. 2004 March; 8(3):180–182 doi: [10.1109/LCOMM.2004.823426](https://doi.org/10.1109/LCOMM.2004.823426)
26. McAuley J, da Fontoura Costa L, Caetano T. Rich-club phenomenon across complex network hierarchies. *Applied Physics Letters*. 2007 August; 91(8) doi: [10.1063/1.2773951](https://doi.org/10.1063/1.2773951)
27. Burt RS. Structural Holes and Good Ideas. *The American Journal of Sociology*. 2004; 110(2):349–399 doi: [10.1086/421787](https://doi.org/10.1086/421787)
28. Kleinberg J, Suri S, Tardos E, Wexler T. Strategic Network Formation with Structural Holes. In: *ACM EC'08*. New York, NY, USA: ACM; 2008. p. 284–293.
29. Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks. In: *WWW'13*; 2013. p. 825–836.
30. Lazarsfeld PF, Merton RK. Friendship as a social process: A substantive and methodological analysis. In: Berger M, Abel T, Page C, editors. *Freedom and Control in Modern Society*. New York: Van Nostrand; 1954. p. 18–66.
31. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. *Annual review of sociology*. 2001;p. 415–444.
32. Heider F. *The psychology of interpersonal relations*. Wiley; 1958.
33. Traag VA, Van Dooren P, De Leenheer P. Dynamical Models Explaining Social Balance and Evolution of Cooperation. *PLoS ONE*. 2013 04; 8(4):e60063 doi: [10.1371/journal.pone.0060063](https://doi.org/10.1371/journal.pone.0060063) PMID: [23634204](https://pubmed.ncbi.nlm.nih.gov/23634204/).
34. Salkind N. *Encyclopedia of Educational Psychology*. Sage Publications; 2008.
35. Alba RD. A graph theoretic definition of a sociometric clique. *The Journal of Mathematical Sociology*. 1973; 3(1):113–126 doi: [10.1080/0022250X.1973.9989826](https://doi.org/10.1080/0022250X.1973.9989826)
36. Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *ICML'01*; 2001 p. 282–289.
37. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. 2009; 11(1):10–18 doi: [10.1145/1656274.1656278](https://doi.org/10.1145/1656274.1656278)
38. Lee C. Communication networks part 1: the Enron e-mail dataset; 2011. Available from: <http://sociograph.blogspot.com/2011/04/communication-networks-part-1-enron-e.html>
39. Wainwright MJ, Jordan MI. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*. 2008; 1(1–2):1–305.
40. Kschischang FR, Frey BJ, andrea Loeliger H. Factor graphs and the sum-product algorithm. *IEEE TOIT*. 2001; 47:498–519.
41. Zhao Y, Wang G, Yu PS, Liu S, Zhang S. Inferring Social Roles and Statuses in Social Networks. In: *KDD'13*. ACM; 2013. p. 695–703.