Mining Fraudsters and Fraudulent Strategies in Large-Scale Mobile Social Networks

Yang Yang⋆, Yuhong Xu, Yizhou Sun, Yuxiao Dong, Fei Wu and Yueting Zhuang

Abstract—The rapid development of modern communication technologies—in particular, (mobile) phone communications—has largely facilitated human social interactions and information exchange. However, the emergence of telemarketing frauds can significantly dissipate individual fortune and social wealth, resulting in potential slow down or damage to economics. In this work, we propose to spot telemarketing frauds, with an emphasis on unveiling the "*precise fraud*" phenomenon and the strategies that are used by fraudsters to precisely select targets. To study this problem, we employ a one-month complete dataset of telecommunication metadata in Shanghai with 54 million users and 698 million call logs. Through our study, we find that user's information might has been seriously leaked, and fraudsters have preference over the target user's age and activity in mobile network. We further propose a novel semi-supervised learning framework to distinguish fraudsters from non-fraudsters. Experimental results on a real-world data show that our approach outperforms several state-of-the-art algorithms in accuracy of detecting fraudsters (e.g., $+0.278$ in terms of F1 on average). We believe that our study can potentially inform policymaking for government and mobile service providers.

Index Terms—Social network, Fraud detection, Fraudulent strategy.

1 Introduction

Fraudulent activities are increasing rapidly with the technology development of global communication in recent years. Millions of people suffer with frauds terribly. For instance, in China, phone fraud has been acknowledged as a significant problem. Estimations by both Qihoo[1] and Tencent[2] show that there are over 500 million phone frauds in 2016, which causes financial loses around 16.4 billion USD. Meanwhile, less than 3% of these cases are resolved. On August 29th of 2016, a college professor at Beijing was reported to have lost 2.67 million USD to a phone fraudster who claimed himself as a judicial officer. Besides the financial impact on individuals, the consequences of phone frauds have been even more tragic, even life-threatening.

Fraud detection has attracted lots of efforts. However, the data availability and high sensitivity have caused this domain to be largely untouched by academia. Most existing work on fraud detection [1], [2], [3], [4] construct experiments on synthetic data or real-world data with limited scale. In this paper, we study on a large-scale mobile social network in real world, which covers a *complete* set of call logs in Shanghai and spans 30 days from September 1st to 30th, 2016. For each call log, the anonymous phone numbers, along with the starting and ending time of the conversation, are recorded. We also obtain annotations of fraudsters made by crowd.

Still, there are many other challenges remained. The first challenge is caused by data sensitivity, which forbids us to access the content information of each call log. It would be easier to detect fraudsters by monitoring particular topics in calls' content such as financial transfer. Without content information, due to privacy issue, we are forced to use meta information to make the inference.

How could well educated people, like college professor in the above case, be swindled? Through our study, we show that **user's information might has been seriously leaked** and fraudsters select targets according to some strategy, instead of randomly (See details in Section 3). How to unveil fraudulent strategy to better understand fraud is the second challenge.

The third challenge is the label imbalance.Indeed, in our data, more than 95.2% of users are non-fraudsters. While the imbalance problem has been addressed by credit card fraud detection [5] and insurance fraud detection [6] technologies, to the best of our knowledge, it has not been well studied under the context of telecommunications.

To address the first and the second challenge, we design and construct several exploratory analysis on our real mobile network to study the behavior patterns of fraudsters. We disclose several fraudulent strategies based on our experiments. For example, we find that fraudsters have preference on young people, and ones who are active in phone communications. We also find that it is better for us to hang up the fraudulent phone call immediately, instead of spending time on slagging off the fraudster to avoid receiving more fraudulent calls.

Based on our discoveries, we design a novel factor-graph based model, *FFD*, to distinguish fraudsters. More

- *Yang Yang is with the College of Computer Science and Technology, Zhejiang University, China. E-mail: yangya@zju.edu.cn, corresponding author.*
- *Yuhong Xu is with the NetEase Fuxi Lab, Hangzhou, China. This work is done when the second author is a master student in Zhejiang University. E-mail: xuyhhh@gmail.com*
- *Yizhou Sun is with the Department of Computer Science, UCLA. E-mail: yzsun@cs.ucla.edu*
- *Yuxiao Dong is with the Microsoft Research Redmond. E-mail: ericdongyx@gmail.com*
- *Fei Wu is with the College of Computer Science and Technology, Zhejiang University, China. E-mail: wufei@zju.edu.cn*
- *Yueting Zhuang is with the College of Computer Science and Technology, Zhejiang University, China. E-mail: yzhuang@zju.edu.cn*

[1] http://www.360.com, a Chinese internet security company.

[2] http://www.tencent.com, one of the largest Internet companies in the world

specifically, our model incorporates fraudsters' structural information and preference on choosing targets. We further propose a semi-supervised learning framework to utilize both the known and unknown labels and address the label sparsity challenge. According to our experiments, we see that our model achieves an improvement on F1 of 0.278 comparing with several state-of-the-art methods.

It is worthwhile to highlight our contributions as follows:

- Based on a real phone-communication data, we disclose how fraudsters and non-fraudsters behave differently in mobile network.
- We study the "precise fraudulent strategy" and appeal to everyone to make sure the protection of personal information has been brought to the forefront.
- We propose a novel framework to distinguish fraudsters from others in a given mobile network.
- We validate the effectiveness of our model on a large-scale mobile network in real world.

2 Data and Problem

Data statistics. We use a mobile dataset that contains *complete* telecommunication records between users in Shanghai, spanning a month from September 1st, 2016, to September 30th, 2016 (1 month). The data is provided by China Telecom, the major mobile service providers in China, and consists of 698,811,827 call logs between 54,169,476 users. Each call log contains the caller's number, the callee's number, the starting time, and the ending time. Since personal identification is required to obtain a mobile number in China, we are also able to retrieve several personal attributes of each user, including age, sex, and birthplace. Our dataset was anonymized by China Telecom for privacy concerns. Throughout the paper, we report only overall statistics in this dataset without revealing any identifiable information of individuals. Please notice that since obtaining a phone number is nontrivial in China, it is uncommon for a people to obtain multiple phone numbers. We thereby regard each phone number as unique user. A similar dataset is also used in [7] and [8].

Data labeling. We then introduce how we obtain the ground truth data. In general, we collect the label data from Baidu[3] and Qihoo 360[4], which both set up report telephone to collect fraudsters phone calls. More specifically, given a user and her telephone number, we employ APIs of both Baidu and Qihoo 360 to see if the user is once reported as a fraudster or not. We label the user as a fraudster if she has been reported as a fraudster by someone either to Baidu or Qihoo 360. These ground truth data comes from a large number of complaints and therefore has a very high level of confidence. In this way, we obtain annotations of 340,550 users in total, among which there are 15,660 fraudsters (around 4.6%).To extend the artifact annotations and discover more fraudsters automatically, we next formulate the problem of fraudster mining.

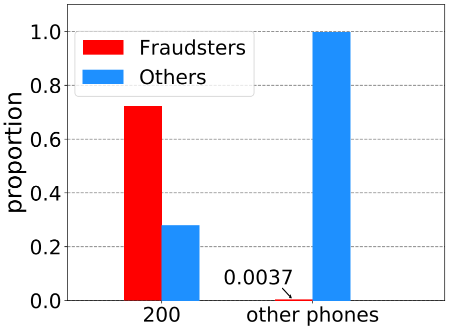


Fig. 1. Composition of the users dialing the "200" and others.

Problem formulation. We construct a mobile communication network based on call logs in our dataset. Formally, we build a directed graph $G = (V, E)$, where V is the set of users, and each directed edge $e_{ij} \in E$ indicates that the user v_i calls v_j $(v_i, v_j \in V)$. Each user in V is associated with a label $y_i \in Y$ to denote if she is a fraudster $(y_i = 1)$, a normal user $(y_i = 0)$, or we do not have her identity yet $(y_i = ?)$.

During network construction, we find an interesting phenomenon that fraudsters dial the number "200" much more times than non-fraudsters (Figure 1). More than 70% of fraudulent phone calls are made though "200". By a study, we find that "200" is a transit number used by fraudsters to cover their true numbers and to save telephone fee. According to this discovery, for two calls made at the same time, one is from user A to "200", and another is from "200" to another user B, we merge them as a unified call log from A to B. We then formulate our problem below.

Definition 1. **Fraudster mining.** Given a mobile communication network $G = (V, E)$, and the identity vector $Y = \{Y^L, Y^U\}$ with missing values, where Y^L denotes labeled identity information of users in G and Y^U stands for unknown identities, our goal is to inference the missing values in Y, i.e., to detect fraudsters that be lurking among other users.

3 Exploratory Analysis

In this section, we convey several exploratory analysis with three purposes: (1) distinguish fraudsters from others; (2) disclose fraudulent strategy; and (3) disclose collaborating patterns of fraudsters.

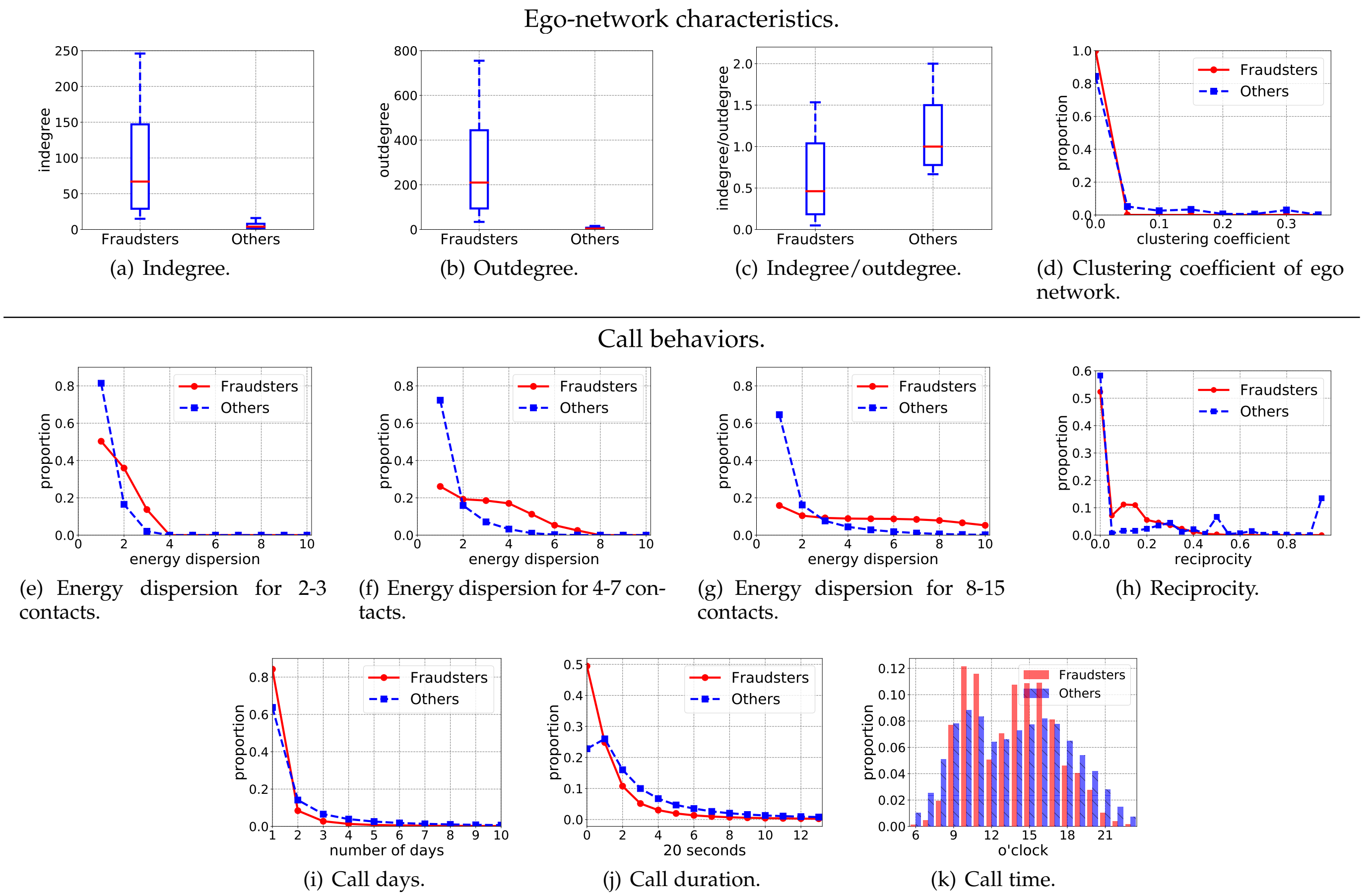
3.1 Distinguish Fraudsters from Others

To understand how fraudsters differ from non-fraudsters, we examine a wide range of features, which include ego-network characteristics and call behaviors, derived from users' mobile communication networks and personal attributes.

Ego-network characteristics. A user's degree reflects how active she is in telephone communications. Figure 2(a) and Figure 2(b) demonstrate the 0.15, 0.3, 0.5, 0.7, and 0.85 quantiles of users' indegree and outdegree respectively. As expected, fraudsters have much larger outdegree, as their

[3] http://www.baidu.com, one of the largest Internet companies in the world.

[4] http://www.360.com, a Chinese internet security company.



(a) Indegree.

(b) Outdegree.

(c) Indegree/outdegree.

(d) Clustering coefficient of ego network.

(e) Energy dispersion for 2-3 contacts.

(f) Energy dispersion for 4-7 contacts.

(g) Energy dispersion for 8-15 contacts.

(h) Reciprocity.

(i) Call days.

(j) Call duration.

(k) Call time.

Fig. 2. Feature comparison between fraudsters and normal users.

job is to make a lot of fraudulent phone calls. To our surprise, fraudster' indegree is also larger than non-fraudsters, which to some extent suggests that the success probability of fraudulent calls is higher than we expected (many of users will call back). In addition, the ratio of fraudsters' indegree to their outdegree is less than non-fraudsters (0.56 vs 1.13 as shown in Figure 2(c)), as there are more out-going calls for a fraudsters to make, comparing with in-coming calls she receives.

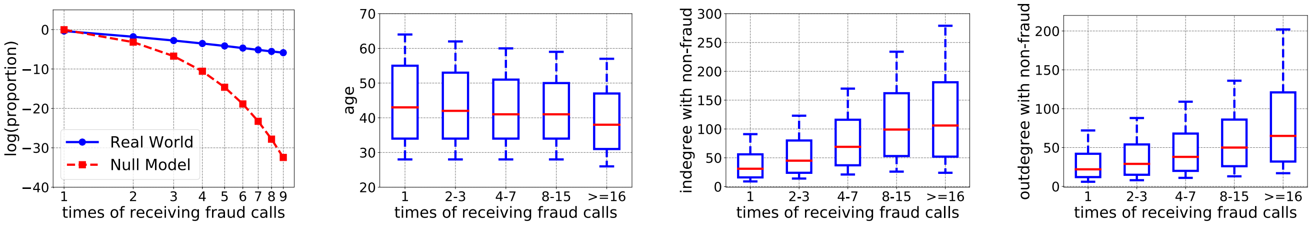
Clustering coefficient measures the fraction of triangles in the ego-networks, which is widely considered in social network analysis [9] [10] [11]. It roughly reflects how connected a user's contacts are to each other. Not surprisingly, according to Figure 2(d), fraudsters have low clustering coefficient that close to 0 (0.002 on average), as their contacts barley know each other. In addition, we find that many non-fraudsters have a value of 0 on this metric, which is due to the low degree of these users in our data.

Calling behavior. According to a user's calling logs, we define *energy dispersion*, the proportion of a user's energy invested in relationship with one of her contacts, as the proportion of times she calls a particular contact. Our data show that non-fraudsters tend to focus on spending their energy on fewer contacts comparing with fraudsters, who would like to expand their connections (potential targets) aggressively. Please notice that the result might be influenced by the number of contacts a user has. To validate if this observation is robust, we group users according to the number of their contacts, and examine the result in each group. It turns out that the observation is consistent in each group (Figure 2(e)-2(g)).

As expected, we find that contacts of fraudsters are less likely to call them back or have called them before, thus have a reciprocal relationship. Specifically, as Figure 2(h) shows, the fraction of fraudsters' reciprocal calls is lower than that of non-fraudsters. And since fraudsters have higher degree than others, the value diversity leads to a smoother distribution curve.

Given a user v_i and one of her contacts v_j, we examine the nature of their relations by two metrics: (1) the number of days that v_i calls v_j, and (2) the duration of each call. Making calls on many days and calls of long duration are more likely to involve intimate relations or are driven by substantial business. On the other hand, calls of short duration tend to be quick check-ins or frauds. As Figure 2(i) and 2(j) show, fraudsters are unlikely to call the same user on more than 2 days. Meanwhile, calls made by fraudsters last shorter than ones made by non-fraudsters.

Figure 2(k) present the distribution of users' phone calls at different time. Form the figure, we see that, comparing with non-fraudsters, fraudsters make more calls at working hours (e.g., 10:00 am - 12:00 pm) and make less calls at after hours (e.g., 12:00 pm - 13:00 pm and the night after 19:00 pm). It suggests that *fraudsters work like office staffs with labor disciplines*.



(a) Distribution of fraudulent phone calls.

(b) Age.

(c) Indegree.

(d) Outdegree.

Fig. 3. Basic analysis of fraudulent strategy. (a) examines if fraudsters choose targets randomly by conducting a null model. (b)-(c) analyze the correlation between the number of fraud phone calls a user receives and her age (indegree or outdegree).

3.2 Disclose Fraudulent Strategy

In this section, we aim to explore which users are more likely to be targeted by fraudsters, start from the following basic question.

Are you targeted by fraudsters randomly? To answer this basic question, we create a null model based on the assumption that fraudsters choose targets to call according to a uniform distribution (every user have the same probability to be called). We then compare the number of fraudulent phone calls each user receives in real data and that generated by the null model, to see if the assumption holds.

More specifically, for each fraudster with out-degree as d_o, we replace her d_o out-links by generating new links. Each new link is connected to a random selected user according to a uniform distribution. We simulate the above random process 100 times and obtain the distribution of the fraudulent calls each user receives averagely. As Figure 3(a) shows, the result of null model and our data is significantly different. For instance, the proportion of users that receive more than 3 fraudulent phone calls in real data is larger than that in null model (1.29% vs 0.01%).

The fact that some of the users will receive more fraudulent phone calls than others suggests that **fraudsters choose targets not randomly but according to some strategy**. In addition, it implies that **users' personal information might has been leaked**, so that fraudsters are able to make strategies. The next question is, who is more likely to be targeted by fraudsters?

Basic analysis. We make a hypothesis that fraudsters choose targets according to users' personal information and activities in the mobile network. We then study to see if these two aspects are correlated with the number of fraudulent phone calls a user receives. It turns out that 5 types of user information are commonly considered by fraudsters: age, indegree, outdegree, call duration with fraudsters and non-fraudsters. Figure 3(b)-3(d) demonstrate the results on the first three features and we omit the remaining due to space limitation. From the figures, we see that fraudsters prefer young people that lack of social experiences. Besides, users who make more calls are more likely to leak their personal information. Thus they are also preferred by fraudsters.

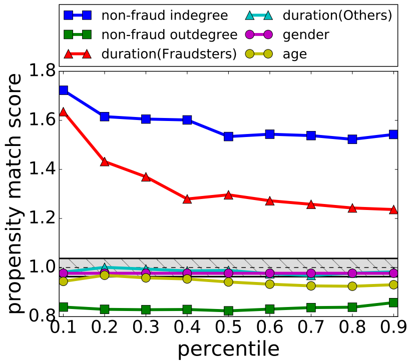


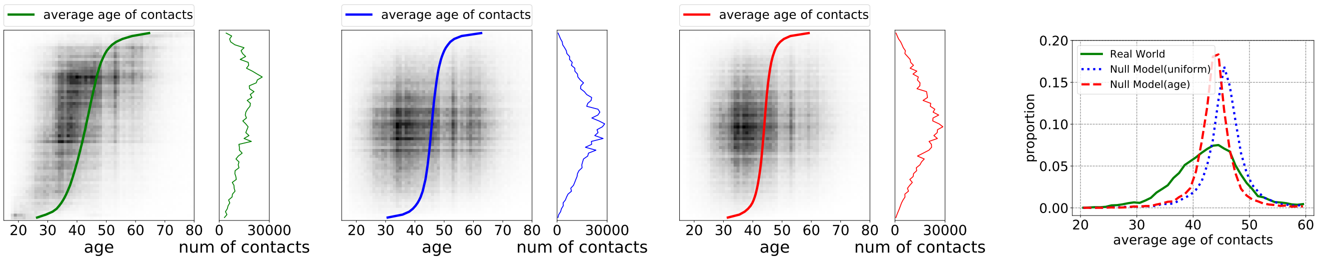
Fig. 4. Propensity match on analyzing the features that influence fraudulent strategy. Darker area indicates the confidence interval.

Propensity match. However, the above result may have bias as these 6 features may be correlated with each other. To further testify our assumption, we proceed the propensity score match analysis [12], a causal relation validation experiment. We show the result in Figure 4, where the x-axis indicates the percentile of a feature we use to divide users into treated group and untreated groups. The y-axis indicates the fraction of users, who receive more than one fraudulent calls, to others. If the fraction is greater than 1, it proves that the corresponding feature influences the frequency of a user being targeted positively (negatively if the fraction is less than 1, or has no influence if the faction lies within the confidence interval $[1 \pm 0.015]$).

From the figure, we see that the call duration with fraudsters has clear positive influence, while the duration of calls with normal users has negative influence. Thus once we realize that we are talking to a fraudster, it is better for us to *hang up the phone immediately, instead of spending time on slagging off the fraudster*. Besides, user age, outdegree and indegree with non-fraudsters have negative and positive influences respectively (p-values $\ll 10^{-9}$). Other features such as gender have no clear influence.

3.3 Disclose Collaborations between Fraudsters

Do different fraudsters use the same strategy? Taking user age as an example, we examine if fraudsters' preferences over target user's age is different. In particular, we define a vector ρ_i, where ρ_{ix} indicates the number of fraudster v_i's non-fraudster-contacts that are aged as x. In Figure 5(a), each row indicates ρ_i of a fraudster, where darker area indicates larger values. Please notice that fraudster are ranked



(a) Age preference of fraudsters in mobile network.
(b) Null model follows uniform distribution.
(c) Null model follows the age distribution of fraudsters' targets.
(d) Age distribution of users receiving fraudulent calls.

Fig. 5. Analyzing if different fraudsters use different strategies.

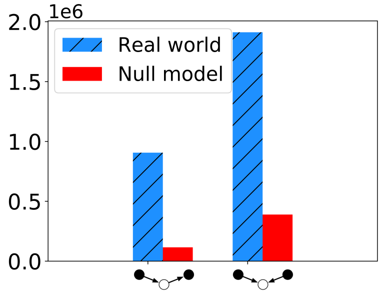


Fig. 6. Proportion of two potential collaborate patterns of fraudsters in real world and null model. Black vertexes indicate fraudsters while while vertexes denote non-fraudsters.

by the average value of ρ_i, so that bottom rows refer to fraudsters who prefer to target young people, while top rows refer to those prefer to cheat older people. The line in the figure denotes the average age of a fraudster's targets. Its sloping shape suggests that fraudsters take different strategies in respect of user age (e.g., some of them tend to contact with young people).

We next aim to disentangle the observation result due to the nature distribution of non-fraudsters' ages and fraudster's strategy. To this end, we design a null model, which assumes that fraudsters have no preference over age at all. In particular, we follow a random process to shuffle outlinks of each fraudster according to a uniform distribution. In turn, the average of ρ shall follow the natural distribution of non-fraudsters' ages. However, from the result in Figure 5(b), we see that the average age of fraudsters' contacts rises almost vertically. It suggests that there is less calls to younger or older people comparing with the result in real world.

To further testify, we conduct the second null model, which is based on the assumption that all fraudsters have the same preference over age. This time, we regenerate the age of each fraudster's contacts according to the distribution ϵ of the average age of fraudsters' targets in our data. More specifically, the x-th dimension of ϵ is defined as $\sum_i \rho_{ix}/Z$, where Z is a normalization term. Comparing Figure 5(c) and Figure 5(a), we see that the assumption of fraudsters having the same strategy over age does not hold. To demonstrate the results more clearly, we also present the distribution of average age of fraudsters' targets in real-world and two null models in Figure 5(d).

Do fraudsters work together? We consider two potential collaborate patterns of two fraudsters A and B: (1) A calls a user and ask her to call B to pay money for ID identification; and (2) A calls a user, who will receive another call from B after a while. We validate the existence of these two patterns by comparing the proportion of these two types of call logs among others in real world and the null model, which shuffles links according to a uniform distribution. Figure 6 shows the result, where the proportion of each pattern in our data is much higher than that in null model. Investigating more collaborate patterns is worth for our future work.

Summary. In this section, we convey several exploratory analysis and mainly obtain the following conclusions:

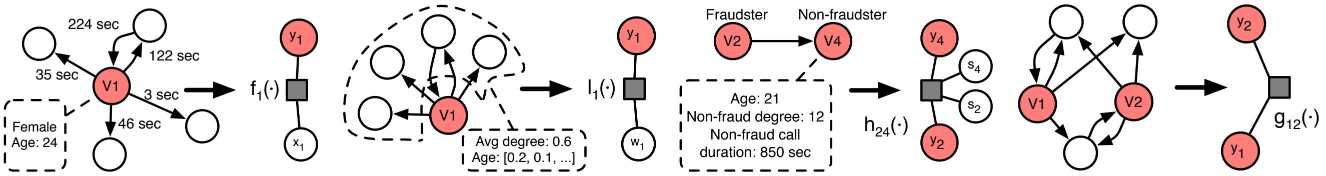
- Fraudsters and non-fraudsters behave differently on communicating with others.
- Fraudsters choose target not randomly, but have preference over users' age and activity in mobile network.
- Different fraudsters use different strategies, and we validate two potential collaborate patterns between them.

4 Fraudster and Fraudulent strategy Detector

Leveraging the insights gleaned from our analysis in Section 3, we develop a *semi-supervised* framework, *Fraudster and Fraudulent strategy Detector* (*FFD*), to distinguish fraudsters from non-fraudsters. In particular, we use a probabilistic factor graph to represent our framework, and aim to model the joint probability of a given mobile network G and user identity Y (i.e., fraudster or non-fraudster), which contains annotations Y^L and missing values Y^U, i.e., $P(Y, G)$ We then learn the model to find a configuration of parameters, which maximizes the joint probability.

Unfortunately, the inference of joint probability $P(Y, G)$ is often intractable. Factor graph factorizes the "global" probability as a product of "local" factor functions, each of which depends on a subset of variables in the graph [13]. In our model, we define four types of factor functions: *attribute factor*, *macro interactive factor*, *micro interactive factor*, and *group factor*. We introduce how we define each factor next, and give illustrations in Figure 7.

Attribute factor. Inspired by our analysis in Section 3.1, a user v_i's characteristics $\boldsymbol{x}_i$ in G, such as degree distribution, reflects her identity y_i. Thus, we define an attribute factor $f(\mathbf{x}_i, y_i)$ to represent the correlation between $\boldsymbol{x}_i$ and y_i. Please refer to Table 1 for details of features contained in $\boldsymbol{x}_i$.



(a) Attribute factor. (b) Macro interactove factor. (c) Micro interactive factor. (d) Group factor.

Fig. 7. Illustration of how we construct different types of factors.

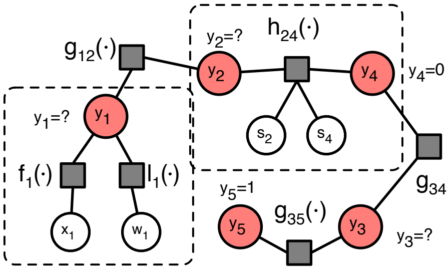


Fig. 8. Graphical representation of the proposed model. Factors between known labels (i.e., y_4 and y_5) and unknown labels (i.e., y_1, y_2, and y_3) reflect the advantage of our semi-supervised learning framework. We only show the attribute factors and macro interactive factors of y_1 and omit others due to space limitation.

Formally, we instantiate the factor by an exponential-linear function:

$$f(\mathbf{x}_i, y_i) = \frac{1}{Z_1} \exp\{\boldsymbol{\alpha}_{y_i} \cdot \mathbf{x}_i\}. \tag{1}$$

where $\boldsymbol{\alpha}$ is a vector of real valued parameters; and Z_1 is a normalization term to ensure that the the sum of the probabilities equals to 1. For each type of identity y_i, $\boldsymbol{\alpha}_i$ is an $|x_i|$-length vector, where the j-th dimension indicates how x_{ij} distributes over y_i. For instance, let us say x_i denotes the degree of v_i. The factor $f(\cdot)$ then captures the fact that fraudsters and non-fraudsters have different degree distributions as shown in Figure 2.

Macro interactive factor. According to our analysis in Section 3.2, fraudster have preferences over several characteristics of the targets (e.g., non-fraudulent outdegree, age, etc.). On the other hand, a user who uses fraudulent strategies to call others is more likely to be a fraudster. To model this, a straight-forward way is to represent v_i's strategy by the feature vector $\mathbf{w}_i$, such as the age distribution of v_i's contacts, and then use a factor $l(\mathbf{w}_i, y_i)$ to quantify the correlation between $\mathbf{w}_i$ and y_i. We call $l(\cdot)$ as the macro interactive factor as it represents the strategy that a user uses to interact with others based on statistics at macro level. Similar with the definition of attribute factor in Eq. 1, we define $l(\cdot)$ as

$$l(\mathbf{w}_i, y_i) = \frac{1}{Z_2} \exp\{\boldsymbol{\eta}_{y_i} \cdot \mathbf{w}_i\}. \tag{2}$$

where $\boldsymbol{\eta}$ is the model parameter and Z_2 is the normalization term. We define $f(\cdot)$ and $l(\cdot)$ as two factors to differentiate the contributions of network characteristics and fraudulent strategy in the fraudster mining task.

Micro interactive factor. To capture fraudulent strategy at micro level, we introduce another factor. According to our analysis, a user's age, non-fraud indegree/outdegree, and call duration with fraudsters are four important aspects that being considered in fraudulent strategies (Figure 4). Thus, we define $\boldsymbol{s}_i$ to indicate these features of user v_i. Next, for a particular pair of two users v_i and v_j, we define the micro interactive factor $h(\cdot)$ as

$$h(\boldsymbol{s}_i, \boldsymbol{s}_j, y_i, y_j) = \begin{cases} 1 & y_i = y_j \\ \frac{1}{Z_3} \exp\{\boldsymbol{\beta} \cdot I(y_i, y_j) \cdot \boldsymbol{s}_j\} & y_i = 1, y_j = 0 \\ \frac{1}{Z_3} \exp\{\boldsymbol{\beta} \cdot I(y_i, y_j) \cdot \boldsymbol{s}_i\} & y_i = 0, y_j = 1 \end{cases} \tag{3}$$

where $I(\cdot)$ is defined as a vector of indicator functions; β is a $4 \times |\boldsymbol{s}|$-length parameter vector (2^2, the combination number of two users' identities); Z_3 is the normalization term. The factor function equals to 1 when v_i and v_j have the same identity as the fraudulent strategy only appears in calls between fraudsters and normal users.

Group factor. According to Section 3.3, we see that fraudsters may work together as a group. In addition, they may contact with the same user to cheat on him cooperatively (Figure 6). Thus, for two users who have more than ϵ common contacts, we define a group factor $g(y_i, y_j)$, which represents the correlation between user v_i and v_j's identities, as follows:

$$g(y_i, y_j) = \frac{1}{Z_4} \exp\{\gamma \cdot I(y_i, y_j)\}. \tag{4}$$

where γ is the model parameter; and Z_4 is a normalization term. Another intuition behind this factor is that users have similar contacts tend to have the same identity.

By putting every factor together, we finally obtain the complete structure of our model, which is shown in Figure 8.

Objective function. By integrating all the factor functions together, and according to the Hammersley-Clifford theorem [14] we obtain the following log-likelihood objective function.

$$\begin{aligned} \mathcal{O}(\theta) &= \log P_\theta(Y, G) \propto \log P_\theta(Y|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{w}) \\ &= \sum_i \alpha_{y_i} \boldsymbol{x}_i + \sum_i \eta_{y_i} w_i + \sum \boldsymbol{\gamma} \cdot I(y_i, y_j) \\ &+ \sum_{i,j} \sum_{y_i \neq y_j} \boldsymbol{\beta} \cdot I(y_i, y_j) \cdot [\delta(y_i = 1)\boldsymbol{s}_j + \delta(y_j = 1)\boldsymbol{s}_i] - \log Z \end{aligned} \tag{5}$$

where $\theta = \{\alpha, \beta, \gamma, \eta\}$ is a parameter configuration of the proposed model; and Z is a normalization term.

Model learning. Learning the model is to find a configuration for the free parameters $\theta = \{\alpha, \beta, \gamma, \eta\}$ that maximizes

ALGORITHM 1: Learning algorithm of the proposed model.

```
Input: a mobile network $G$, a partially labeled user identity
       vector $Y$, the learning rate $\lambda$, and the sampling number
       $K$.
Output: estimated parameter $\theta$
Initialize $\theta$ and $\epsilon$ randomly;
repeat
    for $k = 1$ to $K$ do
        Sample a user $v_i \propto \epsilon_i$;
        Sample a user $v_j \propto 1 - \epsilon_j$;
        Construct factor function $h(\boldsymbol{s}_i, \boldsymbol{s}_j, y_i, y_j)$;
    end
    Call LBP to calculate $\mathbb{E}_{P_\mu(Y^U|G,\theta)}\mathbf{Q}(Y^U)$;
    Call LBP to calculate $\mathbb{E}_{P_\mu(Y|G,\theta)}\mathbf{Q}(Y)$;
    if $y_i \in Y^L$ then
        $\epsilon_i = 1$;
    else
        $\epsilon_i = P(y_i = 1|Y_{\neg i}, G, \theta)$;
    end
    Calculate $\nabla_\theta$ with Eq. (6);
    Update $\theta_{new} = \theta_{old} - \lambda \cdot \nabla_\theta$;
until Convergence;
```

TABLE 1
List of features that used in fraudster mining. † are ones used in macro interactive strategy factor of our model, while others are used in attribute factor.

Feature	Description
Demographics of v_i and her neighbors	
demographics	Age and gender of v_i.
neighbor-age†	Age distribution of v_i's neighbors.
neighbor-sex†	Sex distribution of v_i's neighbors.
province diversity†	Entropy of the distribution of the provinces that v_i's neighbors come from.
Ego-network characteristics of user v_i	
indegree	The number of v_i's neighbors that have made calls to v_i.
outdegree	The number of v_i's neighbors that have received calls from v_i.
neighbor degree†	Average degree of v_i's neighbors.
clustering coefficient	$\frac{\|e_{jk}:v_j,v_k \in V, e_{jk} \in E\|}{d_v(d_{v_i}-1)}$, where v_j and v_k are v_i's neighbors, and d_{v_i} is v_i's degree.
Call behavior	
call duration	v_i's average call duration.
neighbor call duration†	Call duration of v_i's neighbors.
duration variance	Root mean square of v_i's call duration
energy dispersion	v_i's energy dispersion, see more details in Section 3.1.
reciprocal call	The probability of v_i calling another user who will call back.
call time	The distribution of v_i makes calls at different time, which varies from 0:00 am to 23:59 pm.

the log-likelihood objective function $\mathcal{O}(\theta)$. In this work, we use gradient descent method to solve the function. The gradient for each parameter θ is calculated as:

$$\begin{aligned}\nabla_\theta &= \frac{\partial \log P(Y|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{w}, \theta)}{\partial \theta} \\ &= \frac{\log \sum_{Y^U} \exp\{\theta^T \boldsymbol{Q}(Y^U)\} - \log \sum_Y \exp\{\theta^T \boldsymbol{Q}(Y)\}}{\partial \theta} \\ &= \mathbb{E}_{P_\theta(Y^U|\boldsymbol{x},\boldsymbol{s},\boldsymbol{w},\theta)}\mathbf{Q}(Y^U) - \mathbb{E}_{P_\theta(Y|\boldsymbol{x},\boldsymbol{s},\boldsymbol{w},\theta)}\mathbf{Q}(Y) \qquad (6)\end{aligned}$$

where $\boldsymbol{Q}(Y) = ((\sum_i f(\boldsymbol{x}_i, y_i))^T, (\sum_i l(\boldsymbol{w}_i, y_i))^T, \sum_{i,j} h(\boldsymbol{s}_i, \boldsymbol{s}_j, y_i, y_j), \sum_{i,j} g(y_i, y_j))^T$. One challenge here is to calculate the two expectations. The graphical structure of our model may be arbitrary and contain cycles. Thus, we adopt Loopy Belief Propagation (LBP) [15] to approximately compute the marginal probabilities of Y and Y^U. We are then able to obtain the gradient by summing over all the label nodes. An important point here is that the LBP process needs to be proceeded twice during the learning procedure, one for estimating $P(Y|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{w}, \theta)$ and again for $p(Y^U|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{w}, \theta)$. We update each parameter with a learning rate λ as follows:

$$\theta_{new} = \theta_{old} - \lambda \cdot \nabla_\theta. \qquad (7)$$

Implementation notes. The computation of micro-level strategy factor requires to numerate all user pairs, which is time consuming ($O(|V|^2)$). To handle this, we propose a sampling algorithm: we sample K pairs of users, who tend to have different identities, to build the micro-level strategy factor. Specifically, during each iteration, we first select a user v_i proportionally to the marginal probability of she being a fraudster (i.e., $P(y_i = 1|Y_{\neg i}, G)$); we then select another user v_j according to the marginal probability of v_j being a non-fraudster (i.e., $P(y_i = 0|Y_{\neg i}, G)$). For each iteration, we sample K times. Our sampling algorithm reflects the advantage of semi-supervised learning framework as it utilizes both annotated labels and unknown labels.See details in Algorithm 1.

Time complexity. It takes $O(T|E|)$ to execute LBP in our algorithm, where T is the number of iterations in LBP, and $|E|$ is the number of edges in graphical model. The gradient computation step takes $O(|E| + |V|)$, where $|V|$ is the number of random variables in graphical model. Therefore, our model has a time comlexity of $O(RT|E|)$ in its inference and learning procedure, where R is the number of iterations in outer loop.

5 Experimental Results

In this section, we present the results from a series of experiments to evaluate the effectiveness of our proposed approach, by using the dataset introduced in Section 2. All the experiments are implemented in C++ on a 1.2GHz Intel Cores server with 56 CPUs and 396GB RAM, running Ubuntu 14.04.5.

5.1 Experimental Setup

Given the mobile network G and the identity vector Y, the task in our experiment is to determine the missing values

TABLE 2
Performance of detecting fraudsters.

Method	Precision	Recall	F1-score
LR	0.802	0.486	0.605
CRF	0.761	0.501	0.604
SVM	0.840	0.369	0.512
FrauDetector	0.157	0.116	0.133
FRAUDAR	0.646	0.041	0.077
GCN	0.230	0.493	0.314
FFD	**0.808**	**0.547**	**0.652**

in Y. In particular, we construct a 5-fold cross validation to train and test our approach. For evaluation, we consider the following performance metrics: Precision, Recall, and F1-score. The ratio of fraudsters to non-fraudsters is around $1:21$, which is imbalanced.

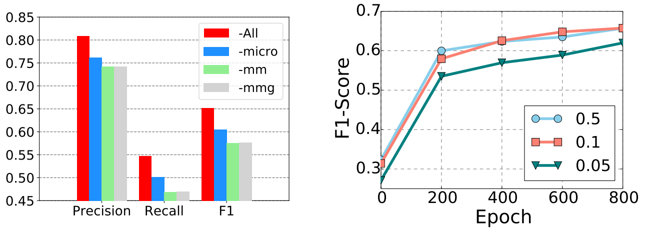
Baseline methods. We consider the following approaches as baselines in our experiments:

- LR: it uses all features listed in Table 1 to train a Logistic Regression, and then applies it to determine whether a user is a fraudster.
- CRF: it is a graphical model based on Conditional Random Field (CRF). This method uses the same features with LR. We further use support vector machine (SVM) as classifier to test the generality of our feature.
- FrauDetector: it is an unsupervised graph-mining-based fraudulent phone call detection framework proposed in [2]. To some extend, this algorithm could be regarded as a weighted HITS [16].
- FRAUDAR: it is also an unsupervised graph-based algorithm proposed in [1]. As it only supports bipartite graph, we reconstruct our mobile network as follows: we define two duplicated sets V^1 and V^2, which both contain all users. For a calling log that the user v_j calls v_j, we create a link from $v_i^1 \in V^1$ to $v_j^2 \in V^2$.
- GCN: it is a commonly used network embedding algorithm proposed in [17]. This algorithm can learn the embedding of these nodes that resembles the local context in a semi-supervised way. We use a 2-layer graph convolution structure and apply all features listed in Table 1 as input.
- *FFD*: it is our proposed model. We empirically set the model parameters as: $K = 10000$ and $\lambda = 0.1$.

5.2 Quantitative Results

Model comparison. Table 2 lists the fraudster detection performance of different methods. Overall, our model, *FFD*, consistently achieves better performance than baseline methods in all metrics. For instance, *FFD* improves F1-score of 0.278 on average. We produced sign tests for each result, which confirms that all the improvements of our proposed models over the other methods are statistically significant ($p \ll 0.01$).

More specifically, we find that the graph-based methods (i.e., FrauDetector, FRAUDAR) only incorporate limited structural information. For example, FrauDetector is mainly based on the call frequency and call duration, which is similar to our call behavior features to some extent, while FRAUDAR mainly considers the density of the given graph at macro level. Another limitation of graph-based methods is that they are unsupervised and can only utilize label information to a limited extent. Besides call behavior features, *FFD* further incorporates user demographic information and characterizes the ego-network of each user into a semi-supervised learning framework. Therefore our approach outperforms graph-based methods.



(a) Factor comparison. (b) Parameter comparision for λ.

Fig. 9. Factor and parameter comparision. In factor comparison, "-All" stands for our model that uses all factors; "-micro" does not consider micro interactive factor; "-mm" further removes macro interactive factor; "-mmg" only considers attribute factor.

One thing worth to mention is that classifier-based methods (i.e., LR, CRF and SVM), which use the same features with our approach, outperform graph-based methods. It suggests that our proposed features are general enough to be utilized in several different machine learning models and are effective in the fraudster detection task. However, these classifier-based methods heavily rely on labeled data. Their performance will be hurt when the labels of fraudsters are sparse, whereas our approach, under the help of micro-level strategy factor and group factor, further utilize unknown labels and estimates its parameters under a semi-supervised learning algorithm.

Surprisingly, the performance of GCN does not reach expectation which might be caused by the imbalance of the label. This further illustrates the effectiveness of the framework we proposed.

Factor analysis. We further analyze the contribution of each factor by removing them one by one, and show the result in Figure 9(a). In the first removal step, we find that removing micro interactive factors (-m) hurts the performance the most (i.e., F1 drops 7%). When removing a second type of factors, macro interactive factors (-mm) is the most influential. After that, F1 drops around 12.6% and decreases little by further removing group factors (-mmg).

Parameter analysis. We analyze the convegence and performance of the model under different λ. As Figure 9(b) shows, our model can quickly reach convergence within 1000 iterations, and the learning process remains stable in all settings.

5.3 Qualitative Results

Fraudulent group detection. We finally study how fraudsters work together as different groups. To do this, we first use our model to infer the identity of each user in our dataset. We then extract a sub-network G' which only consists of fraudsters. For each pair of fraudsters in G', in

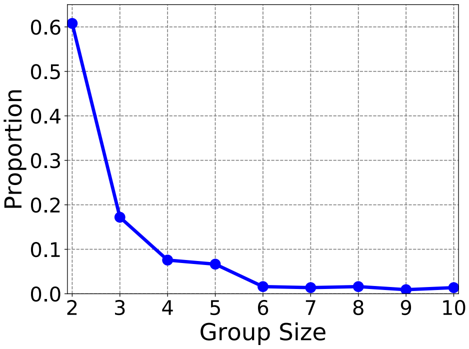


Fig. 10. Distribution of community scale.

addition to their original links, if they have more than 10 common neighbors in the complete graph G, we create a new link between them. Next, we extract fraudulent groups by applying the Girvan-Newman algorithm community detection algorithm [18] on G'. Figure 10 shows the distribution of the number of fraudsters each group contains. The clear heavy tail revealed in the figure suggests that *most fraudsters work along or within a small group*. For example, over 60% of groups only contain 2 fraudsters.

Case study. We give a case study to demonstrate how fraudsters identified by our model work together. We pick up a fraudulent group with the scale of 2 discovered by our method, and denote the two fraudsters in that group as user A and B respectively. We then present their call logs to a non-fraudster X in Table 3. We see that A and B take turns to call X, who will also call back sometime, and it lasts for two days. Interestingly, during the whole process of fraud, A and B have never called each other.

6 Related Work

Fraud detection has attracted significant research efforts recently due to its high impact on telecommunication [19], insurance [20], credit card [21], and health care [22]. In this section, we review related studies in the following aspects.

Classifier-based methods. In many literatures, fraud detection is formulated as a binary classification problem. That is, given a set of phone numbers, predict whether each number is normal or fraudulent. For example, Weatherford et al. [23] utilize user profiles that store long-term information and train neural networks to differentiate fraud behavior and normal one. Instead of neural networks, Yusoff [4] propose a model based on Gaussian mixed model (GMM) as the classifier. Dominik uses a threshold-type classification algorithm [24]. The major limitation of classifier-based methods is that, its performance is heavily influenced by annotations, and will be hurt when the label is sparse. In this work, we propose a semi-supervised learning framework to further utilize unknown labels and improve the performance.

Graph-based methods. This type of approaches mainly detect fraudsters by identifying unexpectedly dense regions of a network. For example, Hooi et al. [1] focus on spotting fraudsters in the presence of camouflage and propose the FRAUDAR algorithm. Tseng et al. [2] build a network with weighted edges to represent each pair of users' call duration and call frequency. They then perform a weighted HITS algorithm [16] on that network to learn the trust value of a phone number and detect fraudulent phone calls according to the trust value. Similar ideas are imposed in several other existing work [25], [26], [27], [28], [29]. Some other work adopt outlier detection techniques to identify unusual user profiles [30].

TABLE 3
Call logs between two fraudsters and a non-fraudster. User A and user B are two fraudsters, who are grouped together, detected by our model and X is a non-fraudster. A → X indicates that A calls X.

Users	Call time	Users	Call time
A→X	Sep. 5th, 15:21:05-15:22:23	A→X	Sep. 5th, 15:52:36-15:52:56
X→B	Sep. 5th, 15:23:36-15:23:53	X→B	Sep. 6th, 13:52:11-13:52:28
B→X	Sep. 5th, 15:32:50-15:33:13	A→X	Sep. 6th, 14:54:32-14:54:53
A→X	Sep. 5th, 15:46:48-15:47:08	B→X	Sep. 6th, 15:28:25-15:30:14
B→X	Sep. 5th, 15:51:22-15:51:50	A→X	Sep 6th, 15:57:40-15:58:33

The proposed method is based on the probabilistic graphical model, which has also been applied in fraud reviews detection [31], social event extraction [32] , signal processing [33], etc.

Many graph-based methods only consider limited types of features, which are mostly call frequency and call duration. In this work, by studying and distinguishing fraudsters from non-fraudsters, we propose several general and effective features. We believe that our features can benefit other work on fraud detection. Another difference between our work and others is that, to the best of our knowledge, we are the first to study fraudulent strategy.

Decomposition-based methods. These kind of approaches detect fraudsters by applying matrix decomposition. For example, Akoglu and Faloutsos [34] propose a SVD decomposition based method to detect anomalous nodes in a time-evolving graph, where a node is considered to be anomaly if its pattern is significantly different from its previous pattern. A similar work is introduced by Ide and Kashima [35], which focus on the problem of monitoring multi-tier web-based system. Sun et al. [36] propose a method for anomaly detection in dynamic graph, which uses the low-rank approximations as summaries of sparse graph. The reconstruction error is then used to evaluate the level of anomalous. Rossi et al. [37] build an none negative matrix factorization based algorithm to recursively extract node feature and determine the nodes' roles. Other works like [38], [39], [40] also use similar decomposition-based techniques to solve the fraud detection problem.

Outlier detection. Our work is also relevant to studies on identifying outlier, whose local structure or attributes are greatly deviate from other members in a social community. For example, Gao et al. [41] and Perozzi et al. [42] propose methods that simultaneously finds communities and outliers. Muller et al. [43] introduce a model that uses a outlier ranking technique in attributed graph. There are also many works that focus on dynamic graphs. For example, Peel and Clauset [44] use the generalized hierarchical random graphs(GHRG) to model community structures in a graph. Sun et al. [45] propose a method called GraphScope, which is a parameter-free algorithm that checks the changes of node partitions in network over time.

7 Conclusion

In this paper, we study the problem of mining fraudsters and fraudulent strategies in a large-scale mobile network. By analyzing a one-month complete dataset of telecommunication metadata in Shanghai with 698 million call logs between 54 million users, we find that fraudsters and non-fraudsters behave differently on communicating with others. In addition, fraudsters have preferences over users' age and activity in phone communications when they choose targets. Inspired by our exploratory analysis, we then propose a novel semi-supervised model to distinguish fraudsters from non-fraudsters.Experimental results demonstrate that our model achieves a significant improvement comparing with several state-of-the-art baseline methods.

As for the future work, it is interesting to think about how to discover a fraud group, instead of an individual fraudster, consists of fraudsters with different roles and duties. Based on this, the collaboration patterns of different fraud groups can be disclosed. Besides, one can extend our work by further considering geographical information of users, studying offline behaviors of fraudsters like how they move around the city.

Our work is limited by the data that we have access to. Although China Telecom is a major service provider and Shanghai is an important global city, the selection bias in our data may limit the generalizability of our work.

Acknowledgments. We acknowledge China Telecom for providing data and assistance in research. The work is supported by NSFC (61702447) and the Fundamental Research Funds for the Central Universities.

References

[1] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, "Fraudar: Bounding graph fraud in the face of camouflage," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 895–904.

[2] V. S. Tseng, J. Ying, C. Huang, Y. Kao, and K. Chen, "Fraudetector: A graph-mining-based framework for fraudulent phone call detection," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 2157–2166.

[3] J. Xu, A. H. Sung, and Q. Liu, "Behaviour mining for fraud detection." *Journal of Research and Practice in Information Technology*, 2007.

[4] M. I. M. Yusoff, I. Mohamed, and M. R. A. Bakar, "Fraud detection in telecommunication industry using gaussian mixed model," in *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on*, 2013, pp. 27–32.

[5] P. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *IEEE Intelligent Systems & Their Applications*, vol. 14, no. 6, pp. 67–74, 1999.

[6] T. Ormerod, N. Morley, L. Ball, C. Langley, and C. Spenser, "Using ethnography to design a mass detection tool (mdt) for the early discovery of insurance fraud," in *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, 2003, pp. 650–651.

[7] Y. Yang, C. Tan, Z. Liu, F. Wu, and Y. Zhuang, "Urban dreams of migrants: A case study of migrant integration in shanghai," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018, pp. 507–514.

[8] Y. Yang, Z. Liu, C. Tan, F. Wu, Y. Zhuang, and Y. Li, "To stay or to leave: Churn prediction for urban migrants in the initial period," in *Proceedings of the Twenty-Seventh World Wide Web Conference*, 2018, pp. 967–976.

[9] Y. Yang, J. Tang, and J. Li, "Learning to infer competitive relationships in heterogeneous networks," *ACM Transactions on Knowledge Discovery from Data*, pp. 1432–1441, 2017.

[10] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *KDD '14*. ACM, 2014, pp. 15–24.

[11] Y. Yang, J. Tang, J. Keomany, Y. Zhao, J. Li, Y. Ding, T. Li, and L. Wang, "Mining competitive relationships by learning across heterogeneous networks," pp. 1432–1441, 2012.

[12] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 544–21 549, 2009.

[13] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp. 498–519, 2001.

[14] J. M. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," *Unpublished manuscript*, 1971.

[15] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *UAI'99*, 1999, pp. 467–475.

[16] J. Kleinberg, "Hubs, authorities, and communities," *ACM Computing Surveys*, vol. 31, p. 5, 1999.

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[18] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 06111, 2004.

[19] R. A. Becker, C. Volinsky, and A. R. Wilks, "Fraud detection in telecommunications: History and lessons learned," *Technometrics*, vol. 52, no. 1, pp. 20–33, 2010.

[20] P. Picard, "Economic analysis of insurance fraud," in *Handbook of insurance*, 2000, pp. 315–362.

[21] S. B. E. Raj and A. A. Portia, "Analysis on credit card fraud detection methods," in *Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on*, 2011, pp. 152–156.

[22] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health care management science*, vol. 11, no. 3, pp. 275–287, 2008.

[23] M. Weatherford, "Mining for fraud," *IEEE Intelligent Systems*, vol. 17, no. 4, pp. 4–6, 2002.

[24] D. Olszewski, "A probabilistic approach to fraud detection in telecommunications," *Knowledge-Based Systems*, vol. 26, pp. 246–258, 2012.

[25] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, 2012, pp. 15–15.

[26] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 61–70.

[27] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 576–587.

[28] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "Catchsync: catching synchronized behavior in large directed graphs," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 941–950.

[29] B. Wu, V. Goel, and B. D. Davison, "Propagating trust and distrust to demote web spam." *MTW*, vol. 190, 2006.

[30] M. Onderwater, "Detecting unusual user profiles with outlier detection techniques," *VU University Amsterdam*, 2010.

[31] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in *ICWSM*, 2013.

[32] X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, "Exacting social events for tweets using a factor graph," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[33] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.

[34] L. Akoglu and C. Faloutsos, "Event detection in time series of mobile communication graphs," in *Army science conference*, 2010, pp. 77–79.

[35] T. Idé and H. Kashima, "Eigenspace-based anomaly detection in computer systems," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 440–449.

[36] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, "Less is more: Sparse graph mining with compact matrix decomposition," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 1, no. 1, pp. 6–22, 2008.

[37] R. A. Rossi, B. Gallagher, J. Neville, and K. Henderson, "Modeling dynamic behavior in large evolving graphs," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 667–676.

[38] H. Bunke, P. J. Dickinson, M. Kraetzl, and W. D. Wallis, *A graph-theoretic approach to enterprise network dynamics*. Springer Science & Business Media, 2007, vol. 24.

[39] M. A. Peabody, "Finding groups of graphs in databases," Ph.D. dissertation, Drexel University, 2002.

[40] P. Shoubridge, M. Kraetzl, W. Wallis, and H. Bunke, "Detection of abnormal change in a time series of graphs," *Journal of Interconnection Networks*, vol. 3, no. 01n02, pp. 85–101, 2002.

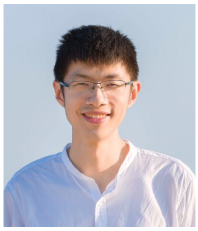
[41] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han, "On community outliers and their efficient detection in information networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 813–822.

[42] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller, "Focused clustering and outlier detection in large attributed graphs," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1346–1355.

[43] E. Muller, P. I. Sánchez, Y. Mulle, and K. Bohm, "Ranking outlier nodes in subspaces of attributed graphs," in *Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on*. IEEE, 2013, pp. 216–222.

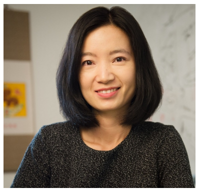
[44] L. Peel and A. Clauset, "Detecting change points in the large-scale structure of evolving networks." in *AAAI*, 2015, pp. 2914–2920.

[45] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "Graphscope: parameter-free mining of large time-evolving graphs," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 687–696.



Yang Yang received his Ph.D. degree from Tsinghua University in 2016. He is an associate professor in the College of Computer Science and Technology, Zhejiang University. His main research interests include data mining and social network analysis. He has been visiting scholar at Cornell University and Leuven University. He has published over 20 research papers in major international journals and conferences including: KDD, WWW, AAAI, and TOIS.

Yuhong Xu received his bachelor's and master's degrees from the College of Computer Science and Technology, Zhejiang University. He is currently an assistant researcher of artificial intelligence at NetEase Fuxi Lab. His main research interests include data mining and social network analysis.



Yizhou Sun received her Ph.D. degree from Computer Science Department, University of Illinois at Urbana Champaign in December 2012. She is an associate professor at Computer Science, UCLA. Her principal research interest is in large-scale information and social networks, and more generally in data mining.

Yuxiao Dong received his Ph.D. in Computer Science from University of Notre Dame in 2017. He is an applied scientist at Microsoft Research Redmond. His research focuses on social networks, data mining, and computational social science, with an emphasis on applying computational models to addressing problems in large-scale networked systems.

Fei Wu received his PhD degree in Computer Science from Zhejiang University in 2002. He is currently a full professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include artificial intelligence, multimedia retrieval and machine learning.

Yueting Zhuang received his PhD degree from the College of Computer Science at Zhejiang University. Now he is the professor and dean of the College of Computer Science, Zhejiang University. His main research interests include multimedia information analysis and retrieval, digital library.