Predicting Social Ties in Mobile Phone Networks

Huiqi Zhang and Ram Dantu
Department of Computer Science and Engineering
University of North Texas
Denton, TX 76203 USA
huiqizhang@my.unt.edu, rdantu@unt.edu

Abstract— **A social network dynamically changes since the social relationships (social ties) change over time. The evolution of a social network mainly depends on the evolution of the social relationships. The social-tie strengths of person-to-person are different one another even though they are in the same group. In this paper we investigate the evolution of person-to-person social relationships, quantify and predict social tie strengths based on call-detail records of mobile phones. We propose an affinity model for quantifying social-tie strengths in which a *reciprocity index* is integrated to measure the level of reciprocity between users and their communication partners. Since human social relationships change over time, we map the call-log data to time series of the social-tie strengths by the affinity model. Then we use ARIMA model to predict social-tie strengths. For validation of our results, we used actual call logs of 81 users collected for a period of 8 months at MIT by the Reality Mining Project group and also used call logs of 20 users collected for a period of 6 months by UNT's Network Security team. These users have around 5000 communication partners. The experimental results show that our model is effective. We achieve prediction performance with accuracy of average 95.2% for socially close and near members. Among other applications, this work is useful for homeland security, detection of unwanted calls (e.g., spam), and marketing**.

Index Terms— **Social-tie strength, Reciprocity index, ARIMA, Prediction, Social groups, Social networks, Social relationships**

I. INTRODUCTION

The modern telecommunication and Internet technologies, such as mobile communications and social media, unite people around the world into a Wide Area Social Network (WASN). Of all, the mobile phone plays a significant role in everyone's daily life, becoming the most popular communication tool. Mobile phone is more convenient for people to communicate with others and accelerates information exchange by interpersonal contacts. As a result, it also accelerates changes and evolution of social relationships. Since the mobile phones have become an important tool of modern human daily life, calling patterns may reflect different human relationships and behaviors, and changes in calling pattern may expose signs of social relationships and behavior changes. For example, the calling patterns of a person with his/her friends differ from those with spammers.

This work is supported by the National Science Foundation under grants CNS- 0627754, CNS-0516807, CNS-0619871 and CNS-0551694. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Organizations or individual elements may be interested in different social network properties. For example, people in homeland security-related departments are interested in particular groups of persons such as terrorists, robbers, and other groups that present security risk. Business persons want to know which groups of people interested in their products. Network designers and operators want to know overall users' distributions and patterns to efficiently and effectively use and distribute resources and enhance quality of services. Unfortunately, almost all existing social network research has focused on overall social-network structures and properties such as clusters and communities. These research efforts lack analysis for one-to-one or one-to-many relationships and behaviors in the detail-necessary when interested in special groups or clusters of people. These detailed features of human relationships are more important for detecting terrorists, spam and user preferences.

A social network is defined as a set of actors (individuals) and the ties (relationships) among them [1]. These relational ties and actors compose the fundamental interests in social networks. Also, a social group can be defined as a set of people who have common interests, like the same subjects, share their experience, express similar ways of thinking and reacting, share the same opinions, do similar things, and have the same goals. They actively exchange information. In the presence of new events, they discuss with each other to decide what to do.

A social network dynamically changes since the social relationships (social ties) change over time. The evolution of a social network mainly depends on the evolution of the social relationships. The social-tie strengths of person-to-person are different one another even though they are in the same groups. In this paper we investigate the evolution of person-to-person social relationships, quantify and predict social tie strengths based on call-detail records of mobile phones.

The remainder of this paper is organized as follows: In Section 2, we briefly review the related work. In Section 3 we describe the methods for quantifying social-tie strengths and reciprocity index, *ARIMA* model for prediction of social-tie strengths. We perform the experiments and validation with actual call logs, and discuss the results in Section 4. Finally, we have the conclusions in Section 5.

II. RELATED WORK

A. Social Tie Strengths

The study of social networks has been applied in modern sociological studies for some time. The major applications focus on tasks such as measuring interpersonal relations in groups, describing properties of social structures and individual social environments [1].

Ogatha discusses the strength of social relations between two persons which they measured with email conversation [2]. The relation is strong if email between two persons is exchanged frequently, recently, and reciprocally and Ogatha uses a formula for the strength, which is a function of user-determined importance weights and the number of received and sent emails. In [3], [4] and [5] social network evolution is studied as its members' attributes change. Backstrom *et al.* perform a large-scale empirical analysis of social-network evolution in which interactions between people are inferred from time-stamped e-mail headers [6]. The social-network-evolution model of topics over time is proposed in [7].

The social network analysis and social clusters of the above work are mainly based on blogs, emails, or the World Wide Web [8], [9], [10]. In [11] [12] structure and tie strength of mobile telephone-call graphs are investigated. Kurucz *et al.* applied the spectral clustering method to telephone call graph partition [13]. Teng and Chou identify the communities of mobile users on call-detail records using a triangle approach [14]. Palla *et al.* has developed an algorithm based on clique percolation to investigate the time dependence of overlapping social-groups so as to discover relationships characterizing social group evolution and capturing the collaboration between colleagues and the calls between mobile phone users [15]. Eagle *et al.* present a method for measuring human behavior, based on contextualized proximity and mobile phone data, to study the dyadic data using the nonparametric multiple regression quadratic assignment procedure (MRQAP) [16]. The MRQAP is a standard technique to analyze social-network data, to discover behavioral characteristics of friendship using factor analysis and to predict satisfaction based on behavioral data. Hidalgo and Rodriguez-Sickert study the stability of social ties by defining and measuring the ties' persistence. They show that persistence of ties and perseverance of nodes depends on degree, clustering, reciprocity and topological overlap [17]. Dasgupta *et al.* propose a spreading activation-based technique to predict potential churners by examining a current set of churners and its underlying social network [18]. Candia *et al.* investigates the spatiotemporal anomalies of calls and patterns of calling activity using standard percolation theory tools [19]. They report that inter-event time of consecutive calls is heavy-tailed. Almost all of the above research focuses on general features of social networks and social groups.

Onnela *et al.* investigate the relationship between local network topology and the associated weights which represent the strength of social ties by the aggregate call duration and the cumulative number of calls between the individuals. They measure the number of common neighbors by the link overlaps for identifying the interconnectedness of communities. Their results provide quantitative evidence for the weak ties hypothesis [22]. Leskovec *et al.* propose a model of network evolution consisting of node arrivals, edge initiation, and edge destination selection processes based on four large online social networks. They find that nodes arrive at a prespecified rate, edge initiations follow a "gap" process and edge destination selections follow triangle-closing model [23]. Du *et al.* propose the first utility-driven graph generator for weighted time-evolving networks based on the patterns that cliques follow, like the *Clique-Degree Power-Law* and *Clique-Participation Law* and the observation of the weights on the edges of triangles followed power laws [24].

B. Prediction Applications

Time series prediction plays an important role in various applications. Several schemes have been widely deployed for predicting weather, environment, economics, stock, market, earthquakes, flooding, network traffic and call center traffic [28]. Companies use predictions of demands for making investments and efficient resource allocation. The call centers predict workload so that they can get the right number of staff in place to handle it. Network traffic prediction is used to access future network capacity requirements and to plan network development for optimum use network resources and improve quality of services.

Most of prediction applications in social networks are focused on link prediction [29]. Gilbert and Karahalios propose a predictive model that maps online social network data to tie strength which is a linear combination of the predictive variables [26]. This model classifies friends as strong and weak ties. Kahanda and Neville use models including logistic regression, bagged decision trees, and naive Bayesian classifiers to predict strong ties in online social network based on transactional information such as communication, file transfer, email and etc. [27].

The social-tie strength prediction in existing work are only considered to be static weights and not taken into account for their change over time. The existing approaches in link prediction and social network evolution are mainly based on the structure measurements such as centrality, density, clustering coefficient and clique in graph theory, in which the communication frequencies are not considered.

In our study, we focus on quantifying individual social-tie strengths using a probability model, *affinity* [20], in which the reciprocity index is integrated based on mobile phone call detail records. We use affinity to measure the similarity between probability distributions, and to quantify the social ties' strengths between actors in groups. Since social relationships change over time, we consider social-tie strength is a function of time. We map call-log data to a time series of the social-tie strengths. Then we apply *ARIMA* model to predict social-tie strengths. This study differs from previous work on measurements [1-19], [22-24], [26, 27] and [29] which focused on general structures of social networks and the social-tie strengths in existing work were not considered to be functions of time. In real world social relationship between two people changes over time.

III. The Proposed Approach

The approach proposed here for quantifying social-tie strengths relies first on a computation of a reciprocity index. We then compute the affinity including the reciprocity index between two users. We use this affinity model to map the social-tie strengths into a time series. Finally we apply *ARIMA* model to predict social-tie strengths. These steps are presented in the next Sections.

A. Reciprocity Index:

In social networks, one of the important relationships between people is reciprocity. Reciprocity can be defined as

the action of returning similar acts [1]. To investigate how people interactively construct their social relationships, we focus on the reciprocity of actions that take place in a telecommunication environment.

In a mobile-phone social networks, actor *i* and actor *j* may call each other multiple times. Reciprocity reflects their relationship in a period of time. The existing mutual indices cannot measure this kind of relationship. The existing mutual (reciprocity) indices measure the tendency of mutual choices for actors (nodes) in a graph [1]. They do not deal with communication frequency and response time. We propose a reciprocity index, $\rho_{a\leftrightarrow b}$ which does measure the tendency of reciprocity for actors *a* and *b* in a group. This was described in detail in [31].

Suppose that the number of phone call arrivals is a Poisson process; this has been shown by Bregni *et al.* [25]. Then the probability of no arrivals in the interval [0, t] is given by

$$P(\tau > t) = e^{-\lambda t}$$

where λ is the arrival rate and τ is interarrival time. The occurrence of at least one arrival between 0 and t is given by

$$P(\tau \le t) = 1 - e^{-\lambda t}$$

Considering actor *a* calls actor *b* at time t_i with rate $\lambda_a t$, the probability of actor *b* calling actor *a* back (reciprocity) at a time t_j with rate $\lambda_b t$ can be computed by

$$P(a \to b \,\&\, b \to a) = P(a \to b)P(b \to a \mid a \to b)$$
$$= P(a \to b)[P(b \to a) + \rho_{a\leftrightarrow b}P(b \xrightarrow{not} a)]$$
$$= (1 - e^{-\lambda_a t_i})[(1 - e^{-\lambda_b (t_j - t_i)}) + \rho_{a\leftrightarrow b} e^{-\lambda_b (t_j - t_i)}]$$

The expected value, $E(R \mid \rho_{a\leftrightarrow b})$, of number of reciprocity from *b* to *a* is the total number of calls, S, from *a* to *b* times this probability, i. e.

$$E(R \mid \rho_{a\leftrightarrow b}) = S(1 - e^{-\lambda_a t_i})[(1 - e^{-\lambda_b (t_j - t_i)}) + \rho_{a\leftrightarrow b} e^{-\lambda_b (t_j - t_i)}]$$

After rearranging the terms, we have

$$\rho_{a\leftrightarrow b} = [R - S(1 - e^{-\lambda_a t_i})(1 - e^{-\lambda_b (t_j - t_i)})] / S(1 - e^{-\lambda_a t_i}) e^{-\lambda_b (t_j - t_i)} \quad (1)$$

where R is observed number of reciprocity.

The $\rho_{a\leftrightarrow b}$ is 0 if there is no tendency toward reciprocity and 1 if there is a maximal tendency toward reciprocity.

We compute the reciprocity indices by Eq. (1) for the call-log data.

B. Quantifying Social Tie Strengths

We can represent the observed data in a bi-dimensional matrix, where rows describe data units and columns describe categorical variables. In the first step we apply a probabilistic similarity coefficient, i.e., *affinity*, measured in a probability scale. In the second step we define an aggregation criterion for merging similar clusters of elements. In the third step we use internal validation to assess the validity of the clustering results, *i.e.*, similarity coefficients comparing a classification with the original data sets.

Affinity measures the similarity between probability distributions. Because our problem belongs to discrete events, we only consider finite event spaces. Let

$$S_N = \{P = (p_1, p_2, ...p_N) \mid p_i \ge 0, \sum_{i=1}^{N} p_i = 1\}$$

be the set of all complete finite discrete probability distributions and $P, Q \in S_N$. The Hellinger distance between P and Q is defined as

$$d_H^2(P,Q) = \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i} - \sqrt{q_i})^2 \quad (2)$$

$d_H^2(P,Q) \in [0,1]$, $d_H^2(P,Q) = 0$ if P = Q and $d_H^2(P,Q) = 1$ if P and Q are disjoint [20].

The affinity between probability measures P and Q is defined as

$$A(P,Q) = 1 - d_H^2(P,Q) = \sum_{i=1}^{N}\sqrt{p_i q_i} \quad (3)$$

$A(P,Q) \in [0,1]$, $A(P,Q) = 1$ if P = Q and $A(P,Q) = 0$ if P and Q are disjoint [20].

In this paper, we use three attributes incoming (*in*), outgoing (*out*) and reciprocity (*reci*) of calls.

Let m_i, n_i be the number of calls, where $i \in \{in, out, reci\}$. $P = (p_{in}, p_{out}, p_{reci})$ is a vector of normalized frequencies over the training period. $Q = (q_{in}, q_{out}, q_{reci})$ is a vector of normalized frequencies of the same attributes observed over the testing period. Then $p_i = m_i / \sum_i m_i$ where $i \in \{in, out, reci\}$ and $q_i = n_i / \sum_i n_i$ where $i \in \{in, out, reci\}$.

The reciprocity part is computed by Eq. (1).

We compute affinity between P and Q is as follows:

$$A(P,Q) = \sum_i \sqrt{p_i q_i} \text{ where } i \in \{in, out, reci\} \quad (4)$$

We map call-log data into time series of the affinity values (social-tie strengths) by Eq. (4) and apply Seasonal Auto Regressive Integrated Moving Average (*SARIMA*) models for predicting the future values.

C. SARIMA Prediction Models

Seasonal AutoRegressive Integrated Moving Average (*SARIMA*) models integrate Seasonal (periodic), AutoRegressive (*AR*), Integrated (I), and Moving Average (*MA*) into a general comprehensive time series model [30].

Let $\{X_t\}$ be a stationary time series.

A $AR(p)$ represents that each observation is a function of the previous *p* observations, which defined as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + e_t$$

where X_t is an observation, ϕ_i(*i=1, ..., p*)are coefficients to be estimated and $e_t \sim (0, \sigma^2)$ (white noise).

A $MA(q)$ describes that each observation is a function of the previous *q* errors, which defined as

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q}$$

where θ_i(*i=1, ..., q*)are coefficients to be estimated and $e_t \sim (0, \sigma^2)$.

$AR(p)$ and $MA(q)$ can be combined in an $ARMA(p,q)$ model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q}$$

Considering the backward linear operator *B* defined by $BX_t = X_{t-1}$, $B^i X_t = X_{t-i}$ for any integer *i*.

By using the backshift linear operator, an $AR(p)$ can be written as

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = \phi_p(B) X_t = e_t$$

where $\phi_p(B)$ is defined by

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

Similarly, $MA(q)$ can be written as

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} = \theta_q(B) e_t$$

where $\theta(B)$ is defined by

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$$

A $ARMA(p,q)$ model can be written in backshift linear operator form

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) X_t = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) e_t$$

That is,

$$\phi_p(B) X_t = \theta_q(B) e_t \quad (5)$$

If $\{Y_t\}$ is an nonstationary time series, $ARMA(p,q)$ can be extended to $ARIMA(p,d,q)$.

$I(d)$ component removes trend by conducting d differencing operations between consecutive observations making time series stationary.

Considering the difference linear operator Δ defined by

$$\Delta Y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1-B) Y_t$$

The stationary time series $\{X_t\}$ can be obtained by the d^{th} difference Δ^d of nonstationary time series $\{Y_t\}$

$$X_t = \Delta^d Y_t = (1-B)^d Y_t \quad (6)$$

By substituting X_t in (1) with (2), we have $ARIMA(p,d,q)$ model

$$\phi_p(B) \Delta^d Y_t = \theta_q(B) e_t \quad (7)$$

or

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B) e_t \quad (8)$$

Applying a similar idea to seasonal time series, we use seasonal difference and backshift operators.

Let $\{Z_t\}$ be a seasonal time series.

The seasonal difference linear operator Δ_s defined by

$$\Delta_s Z_t = Z_t - Z_{t-s} = Z_t - B^s Z_t = (1 - B^s) Z_t$$

where s is the length of the seasonal variation (period).

The D^{th} difference Δ_s^D of seasonal time series $\{Z_t\}$ is

$$\Delta_s^D Z_t = (1 - B^s)^D Z_t$$

A seasonal $AR(P)$ operator of order P is defined as

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}$$

A seasonal $MA(Q)$ operator of order Q is defined as

$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}$$

The seasonal time series can be modeled as

$$\Phi_P(B^s) \Delta_s^D Z_t = \Theta_Q(B^s) Y_t \quad (9)$$

or

$$\Phi_P(B^s)(1 - B^s)^D Z_t = \Theta_Q(B^s) Y_t \quad (10)$$

Substituting (7) and (8) into (9) and (10) respectively, yield the seasonal $ARIMA(p,d,q)$ model with period *s*

$$\phi_p(B) \Phi_P(B^s) \Delta^d \Delta_s^D Z_t = \theta_q(B) \Theta_Q(B^s) e_t \quad (11)$$

or

$$\phi_p(B) \Phi_P(B^s)(1-B)^d (1 - B^s)^D Z_t = \theta_q(B) \Theta_Q(B^s) e_t \quad (12)$$

which are denoted by $SARIMA(p,d,q) \times (P,D,Q)_s$.

The Box-Jenkins method [30] uses an iterative approach of identifying a possible model from a general class model. The chosen model is then checked against the historical data to see whether it accurately describes the series. The model fits well if the residuals are generally small. It comprises the following steps.

Step 1: Model Identification

The first step in model identification is to determine whether the series is seasonal (periodic) and stationary and whether the time series requires differencing to remove seasonality and trend. The initial parameters of an *ARIMA* model are based on an examination of a plot of the time series. Stationary series appears to vary about a fixed level. It is useful to look at a plot of the series along with the sample autocorrelation function (*ACF*). The autocorrelation function (*ACF*) and partial autocorrelation function (*PACF*) are used to identify the best model. If the *ACF* of the dependant variable approaches zero as the number of lags increases, the series is stationary. A nonstationary time series show little tendency for the *ACF*'s to decrease in size as the number of lags increase. A nonstationary time series is indicated if the series appears to grow or decline over time and the sample autocorrelations fail to die out rapidly.

If the series is not stationary, it can often be converted to a stationary series by *d* differencing. If differencing, the original series is replaced by a series of differences. A $ARIMA(p,d,q)$ model is then specified for the differenced series. The $ARIMA(p,d,q)$ models reduce to the $ARIMA(p,q)$ models.

Once a stationary series has been obtained, the form of the model to be used must be identified. This involves selecting the most appropriate lags (values for *p* and *q*) for the *AR* and *MA* parts. These are chosen by finding the lowest *p* and *q* for each residual of the estimated equation.

Step 2: Model Estimation

Once a tentative model has been selected, the parameters for that model must be estimated. The parameters in $ARIMA(p,d,q)$ models are estimated by minimizing the sum of squares of the fitting errors. This usually involves the use of a least squares estimation process. The residual mean square error is useful for assessing fit and comparing different models.

After initial *p* and *q* are chosen, the residuals from this estimate are analyzed with *ACF* and *PACF*. The *PACF* for the k^{th} lag is the correlation coefficient between e_t and e_{t-k}. In particular, the last lag before the *PACF* moves toward zero with an exponential decay is typically a good value for *p*, and the last lag before the *ACF* moves toward zero with an exponential decay is typically a good value for *q*.

After the initial *p, d* and *q* are estimated, the *ACF* and *PACF* of the residuals are calculated and inspected. If these *ACF* and *PACF* are all significantly different from zero, the equation can be considered as the final one since the residuals are free from autoregressive and moving average components. If one

or more of the *ACF* or *PACF* are significantly different from zero, increase p if *PACF* is significant or q if *ACF* is significant by one and reestimate the model. This process will be performed until the residuals have no autoregressive and moving average components.

Step 3: Model Validation

As a final check, it is suggested to compare the variance of $ARIMA(p,d,q)$ with those of $ARIMA(p+1,d,q)$ and $ARIMA(p,d,q+1)$. If $ARIMA(p,d,q)$ has the lowest variance, it should be considered the final equation.

Step 4: Prediction

Prediction can involve either in-sample or out-of-sample. For the out-of-sample predictions, the data used is not included in the estimation of the model. When assessing how well a model predicts, we need to compare it to the actual data, this then produces a prediction error (difference between prediction and actual values) for each individual observation used for the prediction. Then the accuracy of the prediction needs to be measured.

IV. Experimental Results and Discussions

A. Real-life Data Sets and Parameters

***Real-life traffic profile*:** In this paper, actual call logs are used for analysis. These call logs of 81 users were collected for a period of 8 months at MIT [21] by the Reality Mining Project group. Additionally, the call logs of 20 users were collected for a period of 6 months by the Network Security team at UNT. These users have around 5000 communication partners.

The Reality Mining Project group collected data on mobile phone usage of 81 users, including user ID (unique number representing a mobile phone user), time of call, call direction (incoming or outgoing), incoming call description (missed or accepted), talk time, and tower ID (location of phone user). These 81 phone users were students, professors, and staff members. The collection of the call logs was followed by a survey to gather feedback from participating phone users about behavior patterns such as favorite hangout places; service providers; talk-time minutes; and phone users' friends, relatives and parents. More information about the Reality Mining Project can be found in [21].

B. Quantifying and Predicting Social Tie Strengths

We first compute affinity values by the model described the above section and map them in a time series in which the time unit is biweekly.

Fig. 1 shows both observed and predicted values of affinity including confidential intervals for the last 2 predicted values, where the x-axis indicates the time and the y-axis indicates the affinity values for user 60 with his/her partner 2538.

By the model creation and prediction procedure described in above section, we obtained the model in the form $ARIMA(2,0,3)$ and

$$\hat{X}_t = 0.9173 - 0.0580X_{t-1} - 0.9823X_{t-2} + 0.8866e_{t-1} + 0.8935e_{t-2} - 0.1070e_{t-3} + e_t$$

for user60 with his partner2538.

The square root of mean square error for prediction is 0.013. The partner 2538 is a socially close member of user 60.

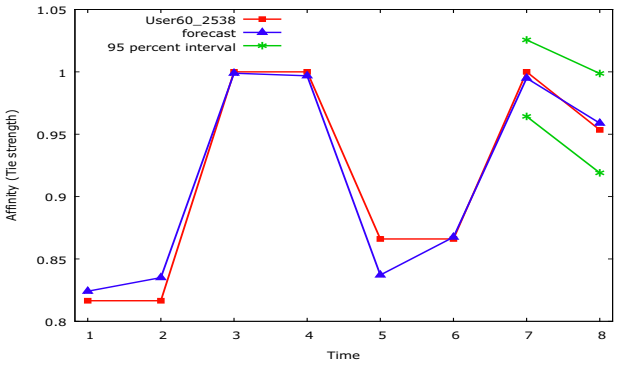

Fig. 1 The affinity predicted (triangle-line) and observed (block-line) values including confidential intervals (star-line) for the last 2 predicted values for user 60 with his/her partner 2538. The x-axis indicates the time (biweekly) and the y-axis indicates the affinity values.

Fig. 2 shows both observed and predicted values of affinity including confidential intervals for the last 2 predicted values, where the x-axis indicates the time biweekly and the y-axis indicates the affinity values for user 86 with his partner 929.

For user 86 with his partner 929, the model of affinity prediction is in the form $ARIMA(4,0,4)$ and

$$\hat{X}_t = 0.8272 + 0.3103X_{t-1} - 0.1687X_{t-2} - 0.4026X_{t-3} + 0.4228X_{t-4} - 0.4895e_{t-1} - 0.2072e_{t-2} - 0.4899e_{t-3} - 0.9996e_{t-4} + e_t$$

The square root of mean square error for prediction is 0.076. The partner 929 is a socially near member of user 86.

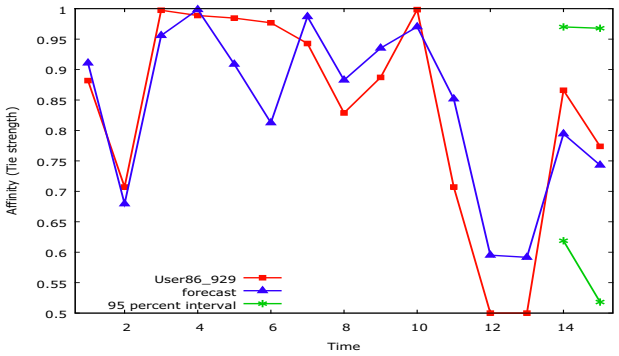

Fig. 2 The affinity predicted (triangle-line) and observed (block-line) values including confidential intervals (star-line) for the last 2 predicted values for phone user 86 with partner 929. The x-axis indicates the time (biweekly) and the y-axis indicates the affinity values.

Table 1 describes the prediction errors for randomly selected phone users, where "close" and "near" indicate that their relations are socially close or near. We achieved performance with accuracy of average 95.2% for prediction.

TABLE I

Prediction Errors for Phone Users

User ID	Relation	Mean Error	Mean Squared Error	Root Mean Squared Error
User60-2538	Close	0.0001	0.0001	0.0130
User60-3	Close	0.0079	0.0001	0.0100
User60-2518	Close	0.0245	0.0006	0.0248
User60-2519	Close	-0.0124	0.0002	0.0144
User60-2524	Close	-0.0011	0.00004	0.0070
User86-341	Close	0.00002	0.00015	0.0125
User86-911	Close	0.0093	0.00009	0.0098
User86-3514	Close	-0.0009	0.0000009	0.0009
User70-1661	Close	0.0000007	0.00002	0.0052
User70-1662	Close	0.00006	0.000002	0.0016
User50-3268	Close	0.0106	0.0014	0.0386

User39-316	Close	0.0014	0.000002	0.0015
User49-2003	Close	0.0027	0.000007	0.0027
User49-2005	Near	-0.028	0.0148	0.1216
User86-929	Near	-0.0088	0.0058	0.0767
User70-1664	Near	0.0599	0.0182	0.1352
User70-1667	Near	0.0082	0.0061	0.0770
User70-1680	Near	0.1145	0.0152	0.1233
User70-1788	Near	0.0759	0.0113	0.1063
User74-2906	Near	0.0040	0.0355	0.1884

V. CONCLUSION

In this paper we propose an affinity model for quantifying social-tie strengths in which a *reciprocity index* is integrated to measure the level of reciprocity between users and their communication partners based on mobile phone-call detail records. Since human social relationships change over time, we map the call-log data to time series the social-tie strengths by the affinity model. Then we use *ARIMA* model to predict social-tie strengths. Because of the diversities and complexities of human-social behavior, one technique cannot detect different features among human social behaviors. We integrate multiple probability and statistical methods for quantifying and predicting social-tie strengths, relation and communication patterns.

We may quantify relationships for a short-term period, say a month, or a long-term period, say, a year or more, using our model by adjusting the parameters. Errors occurred when the number of calls is few.

This work is useful for homeland security and for detecting unwanted calls, *e.g.*, spam and marketing. The experimental results show that our model is effective. In our future work we plan to analyze and study social-network dynamics and evolution.

ACKNOWLEDGMENT

We would like to thank Nathan Eagle and Massachusetts Institute of Technology for providing the call logs of the Reality Mining dataset.

REFERENCES

[1] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[2] Ogatha. "Computer supported social networking for augmenting cooperation" *Computer Supported Cooperative Work* 10, 2001, pp. 189-209, Kluwer Academic Publishers.

[3] Borgs, J. Chayes, M. Mahdian and A. Saberi. "Exploring the community structure of newsgroups," in *Proceedings of 10th ACM International. Conference on Knowledge Discovery and Data Mining*, 2004.

[4] P. Holme, M. Newman. "Nonequilibrium phase transition in the coevolution of networks and opinions," *arXiv physics/0603023*, March 2006.

[5] P. Sarkar, A. Moore. "Dynamic Social network analysis using latent space models," in *Proceedings of SIGKDD Explorations: Special Edition on Link Mining*, 2005.

[6] L. Backstrom, D. Huttenlocher, J. Kleinberg. "Group formation in large social networks: membership, growth, and evolution," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* pp. 44-54, 2006.

[7] Kossinets, D. Watts. "Empirical analysis of an evolving social network," *Science*, vol. 311, pp. 88 – 90, 2006.

[8] X. Wang, A, McCallum. Topics over Time: "A Non-Markov continuous-time model of topical trends," in *Proceeding of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[9] R. Kumar, J. Novak, O. Raghavan, A. Tomkins. "Structure and evolution of blogspace," *Communications of ACM,* 47(12): 35-39, 2004.

[10] R. Kumar, J. Novak, A. Tomkins. "Structure and evolution of on line social networks," in *Proceedings of the twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2006.

[11] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. "On the structural properties of massive telecom graphs: Findings and implications," in *Proceeding of the Fifteenth ACM CIKM Conference on Information and Knowledge Management,* 2006.

[12] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. "Structure and tie strengths in mobile communication networks," in *Proceedings of the National Academy of Sciences of united State of America*, vol. 104, no. 18, pp. 7332-7336, May 1, 2007.

[13] M. Kurucz, A. Benczur, K. Csalogany, L. Lukacs. "Spectral clustering in telephone call graphs," in *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop,* 2007.

[14] W. Teng, M. Chou. "Mining communities of acquainted mobile users on call detail records," in *Proceedings of the 22nd Annual ACM Symposium on Applied Computing,* 2007.

[15] G. Palla, A. Barabasi, T. Vicsek. "Quantifying social group evolution," *Nature*. vol. 446, pp. 664-667, 2007.

[16] N. Eagle, A. Pentland, and D. Lazer. "Inferring friendship network structure by using mobile phone data," in *Proceedings of the National Academy of Sciences*. vol. 106, no. 36, pp. 15274-15278, 2009.

[17] A. C. Hidalgo, C. Rodriguez-Sickert, "The dynamics of a mobile phone network," *Physica* A 387, 3017-3024, 2008.

[18] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea A. Nanavati. "Social ties and their relevance to churn in mobile telecom networks," in *Proceedings of the 11th ACM international conference on extending database technology: Advances in database technology*, 2008.

[19] J. Candia, M. C. Gonzalezm, P. Wang, T. Schoenharl, G. Madey, A. Barabasi. "Uncovering individual and collective human dynamics from mobile phone records". *Journal of Physics A: Mathematical and Theoretical*. 41, 224015, 2008.

[20] M. Fannes and P. Spincemaile. "The mutual affinity of random measures". *Periodica Mathematica Hungarica,* vol. 47, pp. 51–71, 2003.

[21] Massachusetts Institute of Technology: Reality Mining. http://reality.media.mit.edu/ 2009.

[22] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, M. A. Menezes, K. Kaski, A.-L. Barabasi and J. Kertesz. "Analysis of a large-scale weighted network of one-to-one human communication," *New Journal of Physics*, vol. 9, no. 6, 179, 2007.

[23] J. Leskovec, L. Backstrom, R. Kumar and A. Tomkins. "Microscopic evolution of social networks," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2008.

[24] N. Du, C. Faloutsos, B. Wang and L. Akoglu. "Large human communication networks: patterns and a utility-driven generator," in *Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2009.

[25] S. Bregni, R. Cioffi and M. Decina. "An empirical study on statistical properties of GSM telephone call arrivals," in *Proceeding of IEEE Global Telecommunications Conference,* 2006.

[26] E. Gilbert and K. Karahalios, "Predicting Tie Strength with Social Media", in *Proceedings of the 27th international conference on Human factors in computing systems*, pp. 211-220, 2009.

[27] I. Kahanda and J. Neville, "Using Transactional Information to Predict Link Strength in Online Social Networks", in *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, pp. 74-81, 2009.

[28] J. Gooijer and R. J. Hyndman. "25 years of time series forecasting", *International Journal of Forecasting*, vol. 22, issue 3: 442- 473, 2006.

[29] D. Liben-Nowell and J. Kleinberg. "The link prediction problem for social networks", *Journal of the American Society for Information Science and Technology*, vol. 58, issue 7, pp. 1019-1031, 2007.

[30] G. E. P. Box and G. M. Jenkins and G. C. Reinsel. *Time series analysis: Forecasting and control*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1994.

[31] H. Zhang, R. Dantu and J. Cangussu. "Quantifying Reciprocity in Social Networks". In *Proceedings of the IEEE InternationalWorkshop on Social Mobile Web (SMW09), in conjunction with SocialCom-09,* Vancouver, Canada, August, 2009.