How Long will She Call Me? Distribution, Social Theory and Duration Prediction

Yuxiao Dong[†§⋆], Jie Tang[‡], Tiancheng Lou[♯], Bin Wu[♮] and Nitesh V. Chawla[†§]

[†]Department of Computer Science and Engineering, University of Notre Dame
[‡]Department of Computer Science and Technology, Tsinghua University
[♯]Google Inc., USA
[♮]Beijing University of Posts and Telecommunications
[§]Interdisciplinary Center for Network Science and Applications, U. of Notre Dame
ydong1@nd.edu, jietang@tsinghua.edu.cn, acrush@google.com,
wubin@bupt.edu.cn, nchawla@nd.edu

Abstract. Call duration analysis is a key issue for understanding underlying patterns of (mobile) phone users. In this paper, we study to which extent the duration of a call between users can be predicted in a dynamic mobile network. We have collected a mobile phone call data from a mobile operating company, which results in a network of 272,345 users and 3.9 million call records during two months. We first examine the dynamic distribution properties of the mobile network including periodicity and demographics. Then we study several important social theories in the call network including strong/weak ties, link homophily, opinion leader and social balance. The study reveals several interesting phenomena such as people with strong ties tend to make shorter calls and young females tend to make long calls, in particular in the evening. Finally, we present a time-dependent factor graph model to model and infer the call duration between users, by incorporating our observations in the distribution analysis and the social theory analysis. Experiments show that the presented model can achieve much better predictive performance compared to several baseline methods. Our study offers evidences for social theories and also unveils several different patterns in the call network from online social networks.

1 Introduction

Analysis of mobility-based usage patterns can not only help understand users' requirements but also reveal underlying patterns behind user behaviors. The discovered patterns can be used to evaluate traffic demand and forecast call volumes, and also as a tool for infrastructure monitoring (such as switches and

⋆ This work was done when the first author was vising Tsinghua University. Jie Tang is supported by the Natural Science Foundation of China (No.61222212, 61073073). Yuxiao Dong and Nitesh V. Chawla are supported by the Army Research Laboratory (W911NF-09-2-0053), and the U.S. Air Force Office of Scientific Research and the Defense Advanced Research Projects Agency (FA9550-12-1-0405).

cables). There is a lot of work on mobile call network analysis, e.g., scaling properties analysis [22, 5], distribution analysis [21, 19, 1], behavior prediction [27, 28], social ties analysis [3, 2, 25], and link prediction [14, 20].

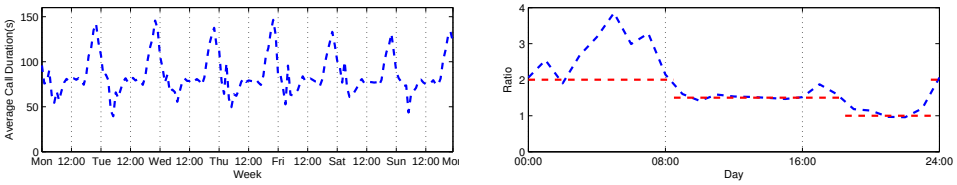
Vaz de Melo et al. [19] studied the call duration distributions of individual users in large mobile networks. They found that the call duration distribution of each user follows the log-logistic distribution, a power-law-like distribution and further designed a model for modeling the behavior of users based on their call duration distributions. The work has mainly focused on studying the call duration distributions of individual users. In [21], Seshadri et al. examined the distributions of the number of phone calls per customer; the total talk minutes per customer; and the distinct number of calling partners per customer. They found that the distributions significantly deviate from the expected power-law or lognormal distributions. However, both papers do not answer questions like what is the call duration distribution between two different users? How the distributions depend on the status (e.g., position, age, and gender) of the communicating users? And how the call duration reflects different properties of social ties between (or among) mobile users?

We focus on the call duration analysis. We understand and model the intricacies of social theory with the predictability of call duration between given two nodes in a network. What are the fundamental patterns underlying the call duration between people? What is the difference of call duration patterns between different groups of people? To which extent can we predict a call's duration between two users?

Contribution. We conduct systematic investigation of the call duration behaviors in mobile phone networks. Specifically, the paper makes the following contributions:

1. We first present a study on the call duration distributions. In particular, we focus on the dynamic properties of the duration distributions.
2. Second, we study the call duration network from a sociological point of view. We investigate a series of social theories in this network including social balance [4], homophily [13], two-step information flow [10], and strong/weak ties hypothesis [6, 11]. Interestingly, we have found several unique patterns from the call duration network. For example, different from the online instant messaging networks, where people with more interactions would stay longer in each communication, while in the mobile call network, it seems that people who are familiar with each other tend to make shorter calls.
3. Based on the discovered patterns, we develop a time-dependent factor graph model, which incorporates those patterns into a semi-supervised machine learning framework. The discovered patterns of social theories are defined as social correlation factors and the dynamic properties of call duration are defined as temporal correlation factors. The model integrates all the factors together and learns a predictive function for call duration forecast.

Experimental results show that the presented model incorporating the discovered social patterns and the dynamic distributions significantly improves the prediction performance (5-18%) by comparing with several baseline method-



(a) (b)

Fig. 1. Duration Periodicity. (a). X-axis: Time in one week. Y-axis: Average call duration. (b). X-axis: Time in one day. Y-axis: The ratio between call times (<60s) and call times (>60s).

s using Support Vector Machine, Logistic Regression, Bayesian Network, and Conditional Random Fields.

2 Mobile data and characteristics

The data set used in this paper is made of a large collection of non-America call records provided by a large mobile communication company[1]. The data set contains 3.9 million call records during two months (December 2007 and January 2008). Each call record contains the phone numbers of the caller and the callee, and the start and end time of the call. Based on this, we construct a social network by viewing each phone number as a user and creating a relationship between user A and B if A made a call to B. The weight of the relationship is quantified by the number of calls between the two related users. In this way, the resultant network contains around 272,345 nodes and 521,925 edges.

We first study the distribution of calls between users. Clearly, the distribution fits a "power-law" distribution in the network. We also found that some users had intensive communications (more than 10 calls in 8 weeks) with each other. About 20% of the pairs of users produce 80% of the call records, which satisfied the Pareto Principle (also known as 80-20 rule) [18, 16]. Thus in this work we mainly focus on the call duration between these pairs of users. For each user, we also extract her/his profile information such as age and gender information. A further statistic shows that there are about 60% calls which are less than 60 seconds (1 minute) and remaining 40% calls (>60s).

2.1 Periodicity

There exist periodic patterns for call duration between human beings. We reach this conclusion by tracking daily calls of mobile phone users. Figure 1(a) shows the average call duration curve on both weekdays and weekends. It clearly shows that there exists obvious week-period and day-period laws for the duration. From

[1] Data and codes are publicly available at http://arnetminer.org/mobile-duration.

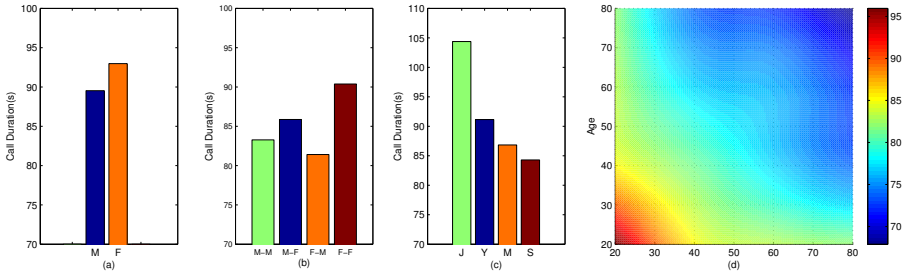


Fig. 2. Call duration of users by gender and age. (a). duration of different genders. (b). duration of different pairwise genders. (c). duration of different age(J: junior, Y: youth, M: middle-age, S: senior). (d). heat-map by plotting age *vs.* age and the color represents the duration of calls.

Monday to Sunday, we can see that the daily duration curves are very similar: 1) In work hours from 8:00 a.m. to 7:00 p.m., people call each other with stable average duration (75s); 2) After getting off work, the average duration between each other increases to 150 seconds gradually until mid-night; 3) From mid-night to early morning, the duration becomes shorter gradually and reaches to its lowest value (about 50s); 4) It ascends to 75 seconds at 8:00 a.m. Moreover, we perform a temporal analysis by tracking the hourly call duration in one day (see Figure 1). We observe that the curve of ratio between the number of calls (<60s) and calls (>60s) varies unevenly over hours: 1) From mid-night to 8:00 a.m., probability that people call each other with duration (<60s) is at least twice than duration which is greater than 60 seconds; 2) In work hours, the ratio is stable to 1.5. 3) From 18:00 p.m. to mid-night, the number of calls with a duration less than 60 seconds is almost the same as the number of calls with a duration larger than 60 seconds.

2.2 Demographics

How does the call duration distribution depend on the gender and age of callers? In this section, we examine the interplay of communication and user demographic attributes. First, we seek to understand how long males and females call. Figure 2 (a) and (b) represent the duration difference by different genders or between different gender-gender pairs. Figure 2 (a) shows that females tend to make longer calls than males. In Figure 2 (b), it shows that, in male-male calls, 84 seconds are taken per call which is lower than 91 seconds for female-female, whereas male-female calls, per call takes 86 seconds, whereas 81 seconds for female-male. Second, we report the analysis on different duration distribution based on age of users. Figure 2 (c) shows that, the average durations for juniors (0, 25], youths (25, 40], middle-aged people (40, 55], seniors (55, +) are 105, 91, 86, 84 seconds respectively, and they decrease as people get older. Figure 2 (d) uses a heat-map visualization to call duration for different age-age pairs. The rows and columns represent the age of both caller and receiver and the color at

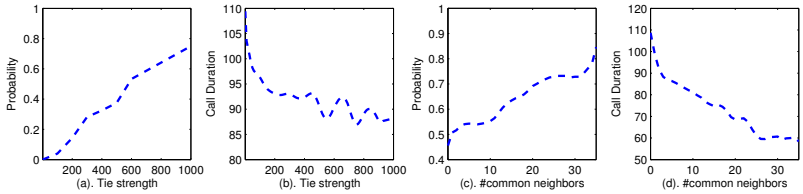


Fig. 3. Tie strength and Link homophily. X-axis: (a)(b). Tie strength as the increase of call times; (c)(d). The number of common neighbors between two callers. Y-axis: (a)(c). Probability that the duration is less than 60s, conditioned on tie strength or #common neighbors; (b)(d). Average call duration.

each age-age cell captures the duration of this pair. The color spectrum extends from blue (short duration) through green, yellow, and onto red (long duration). In this Figure, it is evident that older people tend to have shorter conversations than young users. This trend of obviously descending-order duration in pairwise age fits individual case, as age increases.

3 Social Theory

Besides the dynamic properties of duration distribution, we investigate the interplay of human call behaviors and social theory, and try to answer the question: how social theories, i.e., social tie, homophily, social balance theory etc. are satisfied in the mobile social network? More specifically, we connect the call duration to four classical social psychological theories and focus our analysis on the network based correlation via the following statistics:

1. Strong/weak Ties [6, 11]. How long do people with a strong or weak tie call?
2. Link homophily [13]. Do similar users tend to call each other with long or short duration?
3. Opinion leader [10]. How different (or how similar) are the calling behavior patterns between opinion leaders and ordinary users?
4. Social balance [4]. How does the duration-based network satisfy the social balance theory? To which extent?

Social Tie. Interpersonal ties, generally, come in two varieties: strong and weak. It is argued that weak ties are responsible for the majority of the structure of social networks and the transmission of information through the networks [6], but strong ties make people move to the same circles [11]. The strength of tie represents the extent of closeness of social contacts [4]. In mobile network, we define strong ties, representing frequent calls between two users, and weak ties, representing more casual social contacts with less calls between two users. Such a definition suggests a way of thinking about and answering the following question: How long do people call each other with a strong or weak tie? Figure 3 (a) illustrates our interesting finding: weak ties have a lower probability that their duration is less than 60 seconds. The stronger the tie between two users is, the larger the probability that their duration is less than 60 seconds is. When their tie

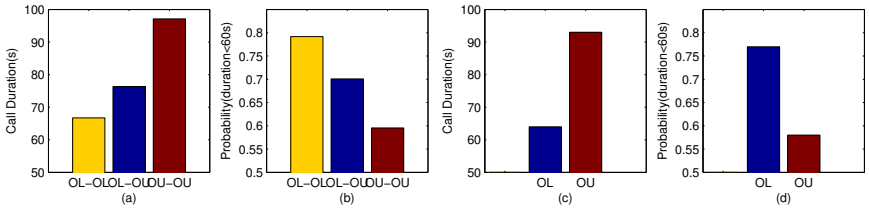


Fig. 4. Opinion leader. OL-Opinion leader; OU-Ordinary user. X-axis: (a)(b). calls between two users; (c)(d). calls made by OL or OU. Y-axis: (b)(d). average call duration; (b)(d). probability that the duration is less than 60s.

strength reaches to 1,000 (calls before), there is a high probability (approximately 80%) that the duration of their future call is less than 60 seconds. In Figure 3 (b), we can see that the average call duration between strong ties is shorter than the calls between weak ties. This finding from both figures seems to be different from the situation in online instant messaging networks, where people with more interactions would stay longer in each communication.

Link homophily. The principle of homophily [13] suggests that users with similar characteristics tend to associate with each other. Particularly, we study link homophily and test whether two users who share common links (caller or receiver) will have a tendency to call each other with longer or shorter duration. In Figure 3 (c), we can see clearly that the probability that the duration is less than 60 seconds when people have more common neighbors becomes higher gradually. Figure 3 (d) shows that the average duration between pairwise users becomes shorter and shorter when they have more and more common neighbors. Intuitively, in mobile communication, more homophily (more common neighbors) and stronger ties (more call times) between two people means that they are familiar with each other. In the point of human behaviors, thus, we can say that the call duration between acquaintances has larger probability to make a short call.

Opinion leader. Opinion leadership is a concept that arises out of the theory of two-step flow of communication propounded by Lazarsfeld [12] and Katz [9, 10], which suggests that innovation (idea) usually flows first to *opinion leaders*, and spreads to more people from them. There are several considerable algorithms to detect opinion leaders in social networks. We apply PageRank [17] to sort all users in our mobile phone data, then top 1% users are labeled as opinion leader according to their PageRank score and the others as ordinary users [8]. Figure 4 clearly shows that the calls between two opinion leaders have 30% shorter duration than the calls between two ordinary users in Figure 4 (a), and the average duration made by opinion leaders is also approximately 30% lower than ordinary users in Figure 4 (c). Figure 4 (b) shows that there is 80% possibility that the duration is less than 60 seconds, when an opinion leader calls another opinion leader, and the possibility is 60% when an ordinary user calls an ordinary user. As to individuals, there are the similar patterns in Figure 4 (d).

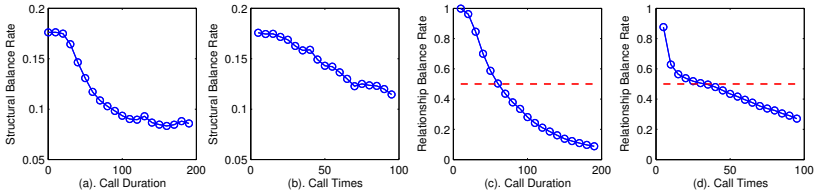


Fig. 5. Social Balance. X-axis: Whether a link is a non-friend(negative) one based on call duration(a)(c) or call times(b)(d). Y-axis: (a)(b) structural balance rate. (c)(d) relationship balance rate.

Social balance. Now, we connect our work to the social balance theory [4]. For each triad (a group of three users), structural balance property implies that either all three of these users are friends or only one pair of them are friends. We assume two users are friends if they call each other at least once. In Figure 5 (a) and (b), it clearly shows that the mobile call network does not satisfy the structural balance theory and the balance rate decreases when the average duration or call times increases. As to relationship balance, the balance rate is the percentage of triangles with even number of negative ties. To adapt the theory to our problem, we assume whether a tie is a negative one based on either average call duration or call times, where the premise is that there exists at least one call between any two users in the triangle. Figure 5 (c) and (d) show that it is much more likely (more than 50%) for users to be connected with a balanced relationship when their duration is less than 60 seconds or they call each other less than 40 times. It represents that mobile network satisfies relationship balance in lower call times or shorter duration.

4 Duration Prediction

4.1 Problem Definition

Now, we study how to design a machine learning model to infer the call duration in the mobile call network based on the discovered patterns from the analysis of data distribution and social theory. We first give necessary definitions and then present a formal definition of the duration prediction problem. We assume that each user is associated with a number of attributes and thus have the following definition.

Definition 1. ***Attributes Matrix****: Let* $\boldsymbol{X}$ *be an* $N \times d$ *attribute matrix of people in which every row* X_i *corresponds to a user, each column an attribute, and an element* x_{ij} *denotes the* j^{th} *attribute value of user* v_i*.*

The attributes matrix describes user-specific characteristics and can be defined in different ways. In the call network, an attribute can be defined as night call ratio and the value of an attribute can be defined as the frequency of calls occurring at night. Then, we define a dynamic call network with node attributes and call duration logs, as the input of our problem.

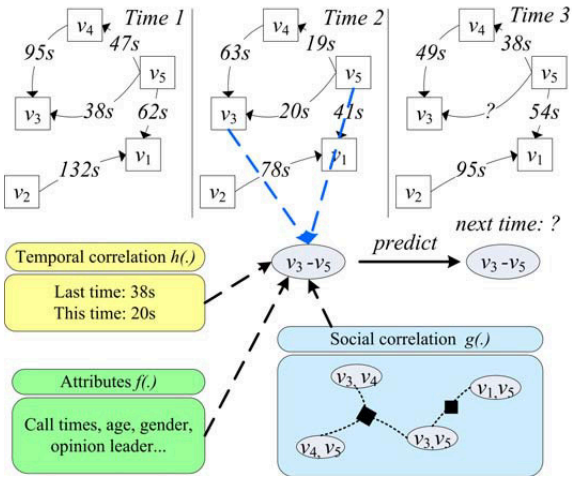


Fig. 6. Model illustration of duration prediction in a dynamic mobile call network.

Definition 2. ***Dynamic Call Network****: A network at time t can be denoted as* $G^t = (V, E, \boldsymbol{X}, Y^t)$*, where* V *is the set of users and* E *is the set of call links between users, and* $\boldsymbol{X}$ *represents the attribute matrix of all users in the network, and* Y^t *is a set of the call duration score between two users at time* t*. Then we can define the dynamic call network* $\boldsymbol{G} = \{V, E, \boldsymbol{X}, \boldsymbol{Y}\}$ *and* $\boldsymbol{Y} = Y^1 \cup Y^2 \cup \cdots \cup Y^T$*.*

Based on the above concepts, we can define the problem of call duration prediction.

Problem 1. **Call Duration Prediction.** Given a dynamic call network $\mathbf{G} = \{V, E, \mathbf{X}, \mathbf{Y}\}$, the goal of the prediction is to learn a mapping function $f : (\mathbf{G}, \mathbf{Y}) \rightarrow Y^{T+1}$to predict the call duration in the next time stamp.

Future call duration can be defined as two cases, first, we use the past call duration to predict the duration of next call; and the second case is to predict the average duration of several calls in a future period (one user can call the other user more than one time) based on the historic call records. Furthermore, there might be more than one call between two users in the next time stamp. We consider two kinds of test cases. The first one is to predict the first call duration in next time stamp, which is called next call duration prediction; The other case is to predict the average call duration in the next time stamp, which is called average call duration prediction. We consider two different scenarios: a binary classification task by setting a threshold $T_{threshold}$ in call duration.

4.2 Prediction Model

Social Time-dependent Factor Graph Model. Tang et. al. [25] first proposed a partially labeled factor graph model to infer social tie. Hopcroft et

al. [8] also proposed a triad based factor graph model for reciprocal relationship prediction in the Twitter network. Tan et al. [23] proposed a noise tolerant time-varying factor graph model for predicting users' behavior in social networks. In this work, we come up with a dynamic factor graph model based on previous partially labeled and triad factor graph model. The dynamic factor graph model incorporates both the correlations among latent variables in different timestamps and other social or attribute features for modeling and prediction. We take next call duration prediction as an example to formalize it in a dynamic factor graph model referred as Social Time-dependent Factor Graph Model (STFG) and propose an approach to learn the model for predicting call duration of pairwise callers. The name is derived from the idea that we incorporate social theory into the factor graph model.

Figure 6 illustrates the graphical illustration of STFG model. The top figure shows the dynamic call network of five users with duration and the bottom figure shows the proposed STFG model. The arrows indicate calls between two users and weight indicates the duration. In bottom figure, the model incorporates three different types of information including social theory (social correlation), user attributes and user's historic duration records (temporal correlation).

Now, we explain the proposed STFG model in details. Given a dynamic call network $\mathbf{G} = \{V, E, \mathbf{X}, \mathbf{Y}\}$, we can define the joint distribution over the durations Y^{T+1} given $\mathbf{G}$ as

$$p(Y^{T+1}|\mathbf{G}) = \prod f(x_i, y_i)g(X_c, Y_c)h(\mathbf{Y}, y_i^{T+1}) \tag{1}$$

The joint probability has three kinds of factor function, corresponding to the illustration in Figure 6. Specifically,

1. **Attribute factor:** $f(x_i, y_i)$. It represents the influence of an attribute of user v_i.
2. **Social correlation factor:** $g(X_c, Y_c)$. It denotes the influence of social relation Y_c.
3. **Temporal correlation factor:** $h(\mathbf{Y}, y_i^{T+1})$. It represents the dependency of one's duration at time $T+1$ on its durations at time t ($t \in \{1, \cdots, T\}$), which denotes the difference between our dynamic model with others [25, 8].

In principle, the three factors can be instantiated in different ways. In this work, we model them by the Hammersley-Clifford theorem [7] in a Markov random field. For the attribute factor, we accumulate all the attributes and obtain a local entropy for all users:

$$\frac{1}{Z_\alpha} \exp\{\sum_{i=1}^{|E|} \sum_{j=1}^{d} \alpha_j f_j(x_{ij}, y_i)\} \tag{2}$$

where α is the weight of function f_j and Z_α is a normalization factor. It can be defined as either a binary function or a real-value function. For example, for the user's social tie feature, we simply define it as a binary feature, that is if the link

between user v_i and v_j is a strong tie and v_i calls v_j with duration (>60s), then a feature $f_j(x_{ij} = 1, y_i = 1)$ is defined and its values is 1, otherwise 0. For social correlation factor, we define a set of correlation feature functions $g_k(X_c, Y_c)$ over each triad Y_c in the network. Then we define a social correlation factor function as follows:

$$\frac{1}{Z_\beta} \exp\{\sum_c \sum_k \beta_k g_k(X_c, Y_c)\} \tag{3}$$

where β_k is the weight of the function, representing the influence degree of k^{th} factor function on Y. We take opinion leader feature as the example to explain social correlation factor. It is defined as a binary function, that is, if a triad contains an opinion leader, then the value of a corresponding triad factor function is 1, otherwise 0.

For temporal correlation factor, we try to use it to model dynamic properties of duration distribution define it as:

$$\frac{1}{Z_\gamma} \exp\{\sum_{i=1}^{|E|} \sum_{t=1}^{T} \sum_m \gamma_m h_m(\mathbf{Y}, y_i^{T+1})\} \tag{4}$$

where $\mathbf{Y}$ is the past durations of the i^{th} pair callers; γ_m represents how strongly the periodicity of the m^{th} pair is. In reality, some users may call each other with similar durations in approximately same time every day or every week. For example, if user v_i and v_j call each other more than ten minutes in every everything, we can define a temporal function with value 1, otherwise 0.

Finally, a factor graph model is constructed by combining Eqs. 2-4 together into Eq. 1, *i.e.*,

$$\begin{aligned} p(Y^{T+1}|\mathbf{G}) = \frac{1}{Z} \exp\{&\sum_{i=1}^{|E|} \sum_{j=1}^{d} \alpha_j f_j(x_{ij}, y_i) \\ &+ \sum_c \sum_k \beta_k g_k(X_c, Y_c) + \sum_{i=1}^{|E|} \sum_{t=1}^{T} \sum_m \gamma_m h_m(\mathbf{Y}, y_i^{T+1})\} \end{aligned} \tag{5}$$

where $Z = Z_\alpha Z_\beta Z_\gamma$ is a normalization factor. Based on Eq. 5, we define the following log-likelihood objective function $\mathcal{O}(\theta) = \log p(Y^{T+1}|\mathbf{G})$:

$$\begin{aligned} \mathcal{O}(\theta) = &\sum_{i=1}^{|E|} \sum_{j=1}^{d} \alpha_j f_j(x_{ij}, y_i) + \sum_c \sum_k \beta_k g_k(X_c, Y_c) \\ &+ \sum_{i=1}^{|E|} \sum_{t=1}^{T} \sum_m \gamma_m h_m(\mathbf{Y}, y_i^{T+1}) - \log Z \end{aligned} \tag{6}$$

where $\theta = (\{\alpha\}, \{\beta\}, \{\gamma\})$ indicates a parameter configuration.

Model Learning. Learning STFG is to estimate the remaining free parameters θ, which maximizes the log-likelihood objective function $\mathcal{O}(\theta)$, *i.e.*, $\theta^* = \arg\max \mathcal{O}(\theta)$

We use a gradient decent method (or a Newton-Raphson method) to optimize the objective function. We adopt α as the example to explain how we learn the parameters. Specifically, we first write the gradient of each α_j with regard to the objective function:

$$\frac{\partial \mathcal{O}(\theta)}{\partial \alpha_j} = \mathbb{E}[f_j(y_i, x_{ij})] - \mathbb{E}_{P_{\alpha_j}(y_i|x_{ij},G)}[f_j(y_i, x_{ij})] \tag{7}$$

where $\mathbb{E}[f_j(y_i, x_{ij})]$ is the expectation of feature function $f_j(y_i, x_{ij})$ given the data distribution and $\mathbb{E}_{P_{\alpha_j}(y_i|x_{ij},G)}[f_j(y_i, x_{ij})]$ is the expectation of feature function $f_j(y_i, x_{ij})$ under the distribution $P_{\alpha_j}(y_i|x_{ij}, G)$ given by the estimated model. Similar gradients can be derived for parameter β_k and γ_m.

Here, there is a challenge that the graphical structure in STFG model can be arbitrary and may contain circles, which makes it intractable to directly calculate the marginal distribution $P_{\alpha_j}(y_i|x_{ij}, G)$. Several approximate algorithms have been proposed, such as Loopy Belief Propagation (LBP) [15] and Mean field [26]. Due to the ease of implementation and effectiveness of LBP, in this work, we use LBP to approximate the marginal distribution $P_{\theta_k}(y_i|x_{ij}, G)$. We are then able to obtain the gradient by summing over all the factor graph nodes with the marginal probabilities. It is worth noting that we need to perform the LBP twice in each iteration, one time for estimating the marginal distribution of unknown variables $y_i =?$ and the other time for marginal distribution over all features. Finally, we update each parameter with a learning rate η with the gradient. Related algorithms can be found in [25, 24].

Prediction. With the estimated parameter θ, we can predict the future call durations. Specifically, the prediction problem can be cast as assigning the value of unknown call durations Y^{T+1} which maximizes the objective function given the learned parameters and network data.

$$Y^* = \arg\max \ \mathcal{O}(Y^{T+1}|\mathbf{G}, \mathbf{X}, \mathbf{Y}, \theta) \tag{8}$$

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

5 Experiments

Our goal here is to predict the next call duration and the average duration of calls in next time stamp based on historic call detail records. We use the first 7 week call detail records as historic data, the first call in the 8th week as next

Table 1. Binary duration prediction performance of different methods. Case 1: Next Call Duration Prediction; Case 2: Average Call Duration Prediction

	Method	Precision	Recall	F1-Measure
Case 1.	SVM	0.5057	0.5021	0.5042
	LRC	0.6184	0.5548	0.5173
	BNet	0.5812	0.5705	0.5692
	CRF	0.5865	0.5886	0.5871
	STFG	**0.6501**	**0.6375**	**0.6393**
Case 2.	SVM	0.4869	0.4875	0.4847
	LRC	0.6143	0.6044	0.5996
	BNet	0.5943	0.5902	0.5873
	CRF	0.6085	0.6054	0.6049
	STFG	**0.6695**	**0.6707**	**0.6692**

call duration and the average duration of calls in the 8th week as average call duration. For binary duration prediction, we present the results with $T_{threshold} = 60s$.

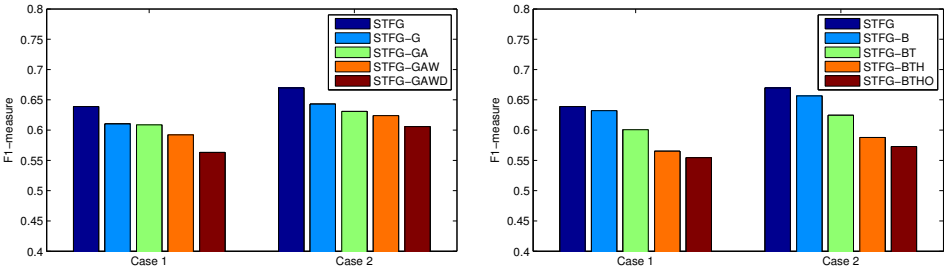
Baseline Methods. We compare our proposed model with four methods.
SVM: it uses the same attributes associated with each edge or node as features to train a classification model and then apply it to predict the call duration label in the test data. For SVM, we use SVM-light [2].
LRC: it uses the same attributes in SVM as features to train a logistic regression classification model and them apply it to predict the label in the test data.
BNet: the method uses the same features as that in SVM. The only difference is that it uses the Naive Bayes classifier.
CRF: it trains a Conditional Random Field model with attributes associated each edge. The difference of this method from our model is that it does not consider structural balance factors.
STFG: our proposed model, which trains a factor graph model with unlabeled data.

5.1 Prediction Performance

We quantitatively evaluate the performance of inferring call duration in terms of *Precision, Recall and F1-Measure*.

Table 1 shows the results for binary duration prediction next call duration prediction and average call duration prediction, set under the threshold. From Table 1, we see that our method clearly outperforms the baseline methods on both cases. For next call duration prediction, the STFG achieves a 5-13% improvement compared with SVM, LRC, BNet, CRF methods in terms of *F1-Measure*. Now, we further validate the effectiveness of our STFG model in the

[2] http://svmlight.joachims.org/



(a) Distribution factor contribution analysis (b) Social factor contribution analysis

Fig. 7. Factor contribution analysis.

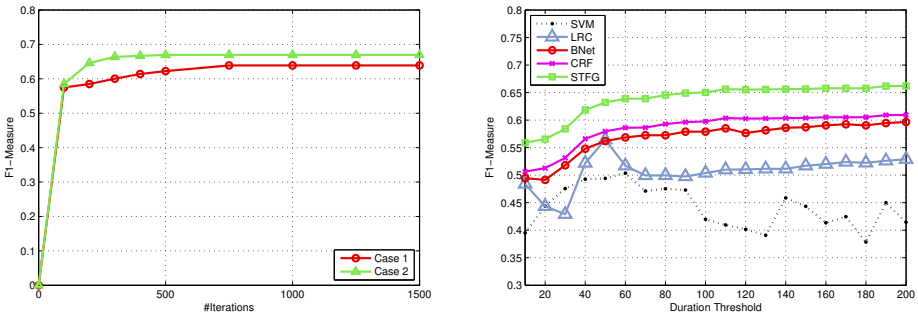
following three aspects: (1) contributions of distribution and social factors in the model; (2) convergence of the learning algorithm; (3) effect of different settings for the duration threshold.

Distribution factor contribution analysis. Now, we analyze how different distribution factors: gender (G), age (A), week periodicity (W), day periodicity (D), can help infer future call duration. We first remove the gender factor (denoted as STFG-G), followed by further removing the age factor (STFG-GA), week periodicity (STFG-GAW), and finally removing day periodicity (STFG-GAWD). Figure 7 (a) shows the *F1-Measure* of the different STFG models. Obviously, we can observe clear drops on the performance when removing each of the factors. The result indicates that our model works well by combining the dynamic properties of data distribution, and each factor in our model contributes improvement in the performance.

Social factor contribution analysis. In STFG, we also consider five different social factors: social balance theory (B), social tie (T), link homophily (H) and opinion leaders (O). Here, we take the analysis to evaluate the contribution of different social factors for the prediction performance. With the same removing operations, we also can see clear drops in *F1-Measure* score in Figure 7 (b). For both two cases, we can find that there is a quick drop when ignoring social tie or link homophily factors. Figure 7 (b) also shows that the other social factors contribute to the prediction of call duration in two cases.

Convergence analysis. We conduct an experiment to analyze the convergence property of the loopy belief propagation based learning algorithm. Figure 8(a) illustrates the convergence analysis results of the learning algorithm. For case 1, the LBP-based learning algorithm converges in about 300 iterations. For case 2, the learning algorithm reach to convergence in about 750 iterations. This suggests that the learning algorithm is able to reach convergence and its efficiency is acceptable.

Threshold analysis. Finally, we analyze how different settings for the parameter $T_{threshold}$ influence the performance of call duration prediction. Figure 8 (b)



(a) Convergence analysis. (b) Parameter $T_{threshold}$ analysis.

Fig. 8. Convergence and parameter analysis.

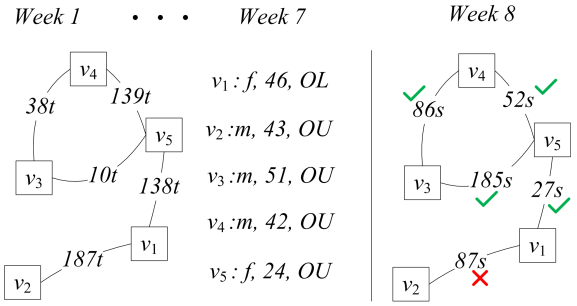


Fig. 9. Case study. Portion of the dynamic call network. The numbers associated with each link in left figure are the number of calls in first 7 weeks. $v_1 : f, 46, OL$ means v_1 is a 46-year female opinion leader. The right figure shows the real average call duration in the 8th week.

lists the average prediction performance of all methods in case 1 with $T_{threshold}$ varied. There are similar patterns in case 2. In general, most methods have similar patterns with different parameter settings, except SVM which is unstable as $T_{threshold}$ varies. It shows that when setting $T_{threshold}$ less 60 seconds, the prediction performance of all models is not very acceptable, while when setting it more than 60 seconds, the performance varies very little and has a slightly increasing trend. However, when $T_{threshold}$ comes to 180s, the number of calls (<180s) is about 9 times to the number of calls (>180s). It means that only one of ten calls in our daily life tends to be greater than 3 minutes.

5.2 Qualitative Case Study

We present a case study to demonstrate the effectiveness of STFG model, see Figure 9. The left figure shows a portion of the dynamic call network from 1st to 7th week. Green colored sign indicates that our model predicts correctly whether the label of duration (<60s or >60s) between respective users or not.

Red colored sign means that our model did not infer the real duration label. In left figure, there exists stronger ties in $(v_1, v_2), (v_1, v_5)$ and (v_4, v_5) than (v_3, v_4) and (v_3, v_5). STFG model predicted $(v_1, v_2), (v_1, v_5), (v_4, v_5)$ as short calls (<60s) and (v_3, v_4), (v_3, v_5) as long calls (>60s) based on social tie theory. Our model predicted four of five labels successfully. User v_5 as a young female tends to make short calls with others, so STFG predicted $(v_1, v_5), (v_4, v_5)$ correctly. As to (v_3, v_5), our model made a compromise between gender factor and social tie factor, and finally predicted it as a long call (>60s) because of the weak tie between v_3 and v_5. STFG missed the prediction between v_1 and v_2, as it was misguided by the strong tie and opinion leader status of v_1.

6 Conclusion

In this paper, we systematically investigated a large mobile call duration network. We first identified and studied the dynamic properties of mobile calling patterns and characteristics, and how they relate to the social network attributes. We discover some interesting social patterns — stronger the ties, lower the probability of call duration; average duration between pairwise users becomes shorter and shorter when they have more and more common neighbors; opinion leaders tend to have shorter call durations; and social balance tends to emerge with shorter call durations. Inspired by these observations, we combined them in to a feature vector to learn a time-dependent factor graph model. Experimental results show that the presented model incorporating the discovered social patterns and the dynamic distributions significantly improves the predictive performance (5-18%) by comparing it with several baseline methods.

Our work has a significant impact in studying the usage patterns of cell phone communication, which can then impact the capacity planning of the communication networks, as well as informing social attitudes and behaviors. Our study can inform cascading effect of information and behavior through a cell phone network, and how duration of phone calls and social topology are closely intertwined.

References

1. C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLOS ONE*, 5, 07 2010.
2. K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *EDBT '08*, pages 668–677. ACM, 2008.
3. N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106, 2009.
4. D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.
5. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 2008.

6. M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.
7. J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. *Unpublished manuscript*, 1971.
8. J. E. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *CIKM'11*, pages 1137–1146, 2011.
9. E. Katz. The two-step flow of communication: an up-to-date report of an hypothesis. *In Enis and Cox(eds.), Marketing Classics*, pages 175–193, 1973.
10. E. Katz and P. F. Lazarsfeld. *Personal Influence*. The Free Press, 1955.
11. D. Krackhardt. *The Strength of Strong ties: the importance of philos in networks and organization*. Cambridge, Harvard Business School Press, Hershey, USA, 1992.
12. P. F. Lazarsfeld, B. Berelson, and H. Gaudet. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press, New York, USA, 1944.
13. P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, pages 8–66, 1954.
14. R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD '10*, pages 243–252. ACM, 2010.
15. K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI'99*, pages 467–475, 1999.
16. M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 2005.
17. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.
18. V. Pareto. *Manuale di economia politica*. Societa Editrice, 1906.
19. C. F. Pedro O.S. Vaz de Melo, Leman. Akoglu and A. A.F.Loureiro. Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. In *PKDD '10*, pages 354–369. Springer, 2010.
20. A. Scellato, Salvatore. Noulas and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD '11*, pages 1046–1054. ACM, 2011.
21. M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *KDD '08*, pages 596–604. ACM, 2008.
22. C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science*, 2010.
23. C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *KDD'10*, pages 1049–1058, 2010.
24. J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM'12*, pages 743–752, 2012.
25. W. Tang, H. Zhuang, and J. Tang. Learning to infer social ties in large networks. In *ECML/PKDD'11*, pages 381–397, 2011.
26. E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *UAI'03*, pages 583–591, 2003.
27. Y. Zhang, J. Tang, J. Sun, Y. Chen, and J. Rao. Moodcast: Emotion prediction via dynamic continuous factor graph model. In *ICDM'10*, pages 1193–1198, 2010.
28. H. Zhuang, A. Chin, S. Wu, W. Wang, X. Wang, and J. Tang. Inferring geographic coincidence in ephemeral social networks. In *ECML/PKDD'12*, pages 613–628, 2012.