Inferring User Demographics and Social Strategies in Mobile Social Networks

Yuxiao Dong†, Yang Yang‡, Jie Tang‡, Yang Yang†, Nitesh V. Chawla†
†Department of Computer Science and Engineering, University of Notre Dame
†Interdisciplinary Center for Network Science and Applications, University of Notre Dame
‡Department of Computer Science and Technology, Tsinghua University
ydong1@nd.edu, yyang.thu@gmail.com, jietang@tsinghua.edu.cn, yyang1@nd.edu, nchawla@nd.edu

ABSTRACT

Demographics are widely used in marketing to characterize different types of customers. However, in practice, demographic information such as age, gender, and location is usually unavailable due to privacy and other reasons. In this paper, we aim to harness the power of big data to automatically infer users' demographics based on their daily mobile communication patterns.

Our study is based on a real-world large mobile network of more than 7,000,000 users and over 1,000,000,000 communication records (CALL and SMS). We discover several interesting *social strategies* that mobile users frequently use to maintain their social connections. First, young people are very active in broadening their social circles, while seniors tend to keep close but more stable connections. Second, female users put more attention on cross-generation interactions than male users, though interactions between male and female users are frequent. Third, a persistent same-gender triadic pattern over one's lifetime is discovered for the first time, while more complex opposite-gender triadic patterns are only exhibited among young people.

We further study to what extent users' demographics can be inferred from their mobile communications. As a special case, we formalize a problem of double dependent-variable prediction—inferring user gender and age simultaneously. We propose the *WhoAmI* method, a Double Dependent-Variable Factor Graph Model, to address this problem by considering not only the effects of features on gender/age, but also the interrelation between gender and age. Our experiments show that the proposed WhoAmI method significantly improves the prediction accuracy by up to 10% compared with several alternative methods.

Categories and Subject Descriptors

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

Keywords

Demographic prediction, Social strategy, Mobile social network, Human communication

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623703.

1. INTRODUCTION

From a recent report by the International Telecommunications Union (ITU) at the 2013 Mobile World Congress, the number of mobile-phone subscriptions reached 6.8 billion, corresponding to a global penetration of 96%. As of 2014, the number of mobile subscriptions across the globe has exceeded the world population. These mobile devices record huge amounts of user behavioral data, in particular users' daily communications with others. This provides us with an unprecedented opportunity to study how people behave differently and form different groups.

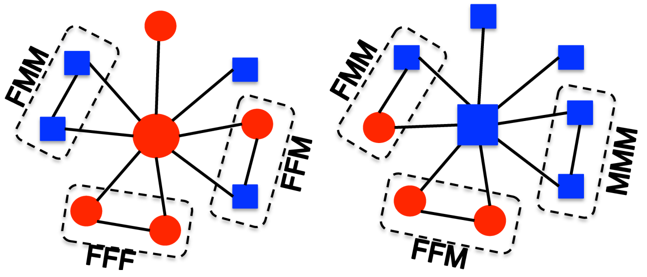
Previous work on mobile social networks mainly focuses on macro-level models, like network formation [26], scale free [5], duration distribution [4, 29], and mobility modeling [30, 33]. Recently, however, researchers have started to pay more attention to the micro-level analysis of the mobile networks. For example, Eagle et al. [6] studied the friendship network of 100 specific mobile users (students or faculties at MIT). They investigated human interactions (what people do, where they go, and with whom they communicate) based on the machine-sensed environmental data collected by mobile devices. However, they do not consider the user demographics. More recently, Nokia Research organized the 2012 Mobile Data Challenge to infer mobile user demographics by using communication records of 200 users [21, 35]. However, the scale of the network is very limited. In this paper, we aim to leverage a large-scale mobile network to study how users' communication behaviors correlate with their demographic information.

Demographic information is often available in the telecommunication industry. For example, a *postpaid* mobile user[1] is required to create an account by providing detailed demographic information (e.g., name, age, gender, etc.). However, a recent report[2] indicates that there is still a large portion of *prepaid* users[3] (also commonly referred to as pay-as-you-go) who are required to purchase credit in advance of service use. Statistics show that 95% of mobile users in India are prepaid, 80% in Latin America, 70% in China, 65% in Europe, and 33% in the United States. Even in the U.S., the switch to prepaid plans is accelerating during the economic recession from 2008. Prepaid services allow the users to be anonymous—no need to provide any user-specific information. However, building demographic profiles for all customers is critical to mobile service providers. This can help them make better marketing strategies (e.g., identify potential customers and prevent customer churns). Moreover, by using demographic information, service providers can supply users with more personalized services and focus on enhancing the communication experience. Therefore, one interesting but challenging question is the extent to which user demographic

[1] http://en.wikipedia.org/wiki/Postpaid_mobile_phone

[2] http://www.itu.int/

[3] http://en.wikipedia.org/wiki/Prepaid_mobile_phone



(a) Female Ego Network (b) Male Ego Network.

Figure 1: Ego Network. Red circle: female; blue square: male; MMM: Male-Male-Male; FFF: Female-Female-Female.

profiles can be inferred from their mobile communication interactions.

In this paper, we employ a real-world large mobile network comprised of more than 7,000,000 users and over 1,000,000,000 communication records (CALL and SMS) as the basis of our study, which we use to systematically investigate the interplay of user communication behaviors and demographic information. Through the study, we first reveal several intriguing *social strategies* that users of different ages and genders use to maintain their social connections. Based on the discoveries, we then develop a unified probabilistic model to predict users' demographics based on their communication behaviors. To the best of our knowledge, we are the first to study the problem of inferring user demographics and social strategies in such a real-large mobile network.

To highlight several of our key findings, we discovered that young people are very active in broadening their social circles, while seniors have the tendency to maintain small but close connections. We found that male users maintain broader social connections than female users when they are young. We also found that cross-gender communications are more frequent than those between same-gender users. We also observed frequent cross-generation interactions that are essential for bridging age gaps in family, workplace, education, and human society as a whole [20]. We discovered significant gender differences in human interactions throughout their lives, which reflects dynamic gender-biased social strategies. Figure 1 shows examples of both female and male ego networks. Human social behaviors reveal both same-gender relationships—'FFF' (Female-Female-Female) and 'MMM' (Male-Male-Male) triads—and opposite-gender social groups ('FMM' and 'FFM' triads) during the dating and reproductively active period (18-34 years old). Upon entering into middle-age, people's attention to opposite-gender triads quickly disappears. However, the insistence and social investment on same-gender social groups ('FFF' and 'MMM') lasts for a lifetime. Finally, our analysis shows strong interrelations between users' age and gender. For example, a 20-year-old female's behaviors are distinct from not only a 20-year-old male's, but also from a 50-year-old female's.

Based on these interesting discoveries, we further study to what extent users' demographic information can be inferred by mobile social networks. We formally define a double dependent-variable classification problem. The objective is to infer users' gender and age simultaneously by leveraging their interrelations. This problem is very different from traditional classification problems, where only the correlations between the dependent variable Y and feature vector $\mathbf{X}$ are considered. In this problem, we are given two dependent variables Y (gender) and Z (age), and feature vector $\mathbf{X}$. We aim to capture the correlations between $\mathbf{X}$ and Y, $\mathbf{X}$ and Z, and the interrelations between Y and Z to infer Y and Z simultaneously. To address this problem, we present the *WhoAmI* method, whereby the interrelations between two dependent variables are modeled. The WhoAmI method is able to infer user gender and age simultaneously. On both the CALL and SMS networks, the proposed method can achieve an accuracy of 80% for predicting users' age and gender according to their daily mobile communication patterns, significantly outperforming (by up to 10% in terms of F1-Measure shown in Figure 2) several alternative methods (Cf. §5 for details of the comparison methods). To scale up the proposed method to handle large-scale networks, we further develop a distributed learning algorithm, which can reduce the computational time to sub-linear speedup (9–10× with 16 cores) by leveraging parallel computing.

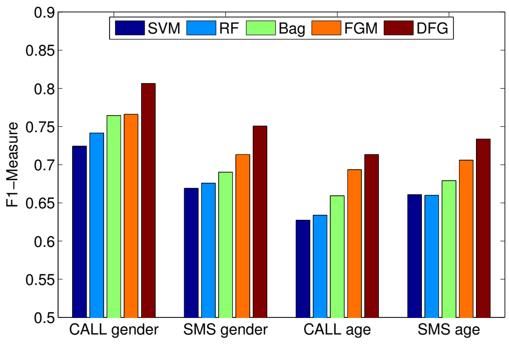


Figure 2: Demographic prediction performance of comparison methods (Cf. §5 for details of the comparison methods).

Organization. We introduce the mobile networks in Section 2. Followed by the interesting discoveries in Section 3, we formalize the double dependent-variable demographic classification problem and present our solution in Section 4. Prediction results are shown in Section 5. Finally, we summarize the related work in Section 6 and conclude this work in Section 7.

2. MOBILE NETWORK

Data. The dataset used in this paper is extracted from a collection of more than 1 billion (1,000,229,603) call and text-message events from an anonymous country [26, 17], which spans from Aug. 2008 to Sep. 2008. We construct two undirected and weighted mobile communication networks from the deidentified and anonymous data: call network (referred to as CALL) and messaging network (referred to as SMS). Specifically, we view each user as a node v_i and create an edge e_{ij} between two users v_i and v_j if and only if they made reciprocal calls (v_i called v_j and also v_j called v_i for at least one time during the observation period) or messages between each other. The strength w_{ij} of the edge is defined as the number of communications between v_i and v_j. Then we extract the largest connected components from each network as our experimental networks: CALL and SMS. The resultant CALL network consists of 7,440,123 nodes and 32,445,941 edges and the SMS network is composed of 4,505,958 nodes and 10,913,601 edges. The data also does not contain any communication content.

Demographics. In this dataset, around 45% of the users are female and 55% are male. We compare the demographic population distribution of mobile users in the dataset with the 2008 global population distribution. We found that both female and male users between the ages of 20 and 55 are strongly overrepresented in the mobile population compared to the global population. This is reasonable, because teenagers (under 18 years old) and the elderly (aged 80 or over) use mobile phones less frequently. Thus in our study, we focus on users aged between 18 and 80 years old. To simplify the notations, we use F and M to respectively denote the female and male users. Following [10, 3], we also split users into four groups according to ages: Young (18-24), Young-Adult (25-34), Middle-Age (35-49), and Senior (> 49).

Network Characteristics. We present a basic correlation analysis between network characteristics and user demographics to see how people of different gender and age maintain their mobile social networks. In particular, we consider the following network metrics:

- *Degree Centrality*: the number of edges incident upon a node in the network;
- *Neighbor Connectivity*: the average degree of neighbors of a specific user [37];
- *Triadic Closure*: the local clustering coefficient (cc) of each user [15];
- *Embeddedness*: the degree that people are enmeshed in networks [9]. More accurately, a user u's embeddedness is defined as

$$\frac{1}{|N_u|}\sum_{v\in N_u}\frac{|N_u\cap N_v|}{|N_u\cup N_v|} \tag{1}$$

where N_u is the neighbors of u.

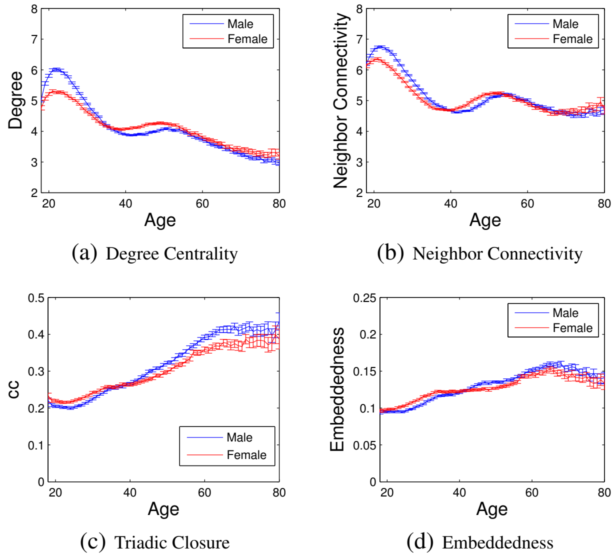
Figure 3 plots the correlations between the four network metrics and the user age. From sub-figures 3(a)-3(b), we observe that the degree and neighbor connectivity of both female and male users achieve peak values around 22 years old, then decrease with valleys around 38-40 years old. An interesting phenomenon is that before this valley, the males have clearly higher scores on both metrics (degree and neighbor connectivity), while the situation is reversed after this point.

From sub-figures 3(c)-3(d), we see that both triadic closure and embeddedness increase when users become older. Similar to the first two metrics, there is also a reversion phenomenon at age 38-40. The difference lies in that the male's triadic closure and embeddedness are at first smaller than the female's, and then become larger after the reversion point. All four network metrics are observed at a 95% confidence interval.

Social Strategies. From a sociological perspective, the analysis that results in Figure 3 can be also explained by different social strategies that people use to maintain their social connections. It seems that young people (who have higher degree scores) are very active in broadening their social circles, while seniors (who have higher triadic closure scores cc) tend to keep small but more stable connections.

3. COMMUNICATION DEMOGRAPHICS

Human communications form the structural backbone of human societies, in the shape of networks [26]. The mobile data provides rich information for understanding human communications in real-world daily life. In this section, we focus on the individual level to study how a user communicates with others in her/his *ego network*. To be precise, one's ego network is defined by viewing herself/himself as the central node and her/his directed friends as surrounding nodes. Clearly, the ego network is a sub-network of the original network. Figure 1 gives two examples of ego networks. Herein, we investigate the interplay of human communication interactions and demographic characteristics in ego networks. Three social strategies are revealed from the data, including homophily on gender and age, cross-generation communication, and demographic dynamics of social relationships. We present the discoveries in the CALL network due to the page limit, and the SMS network shows the similar results.



(a) Degree Centrality

(b) Neighbor Connectivity

(c) Triadic Closure

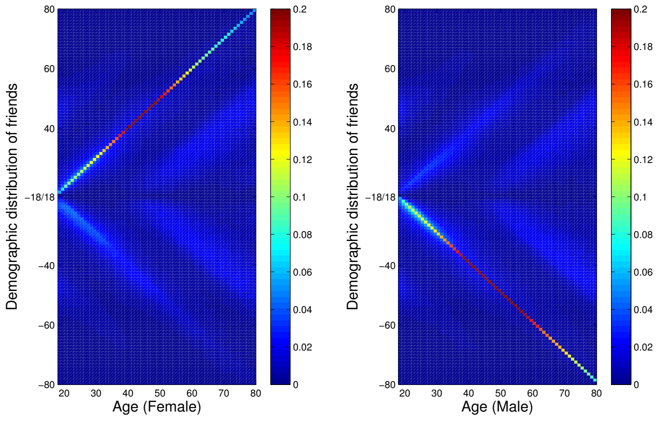
(d) Embeddedness

Figure 3: Correlations between demographics and network characteristics.

Homophily on Gender and Age. The principle of homophily suggests that people tend to be connected with those who are similar to them [14]. It has been extensively studied and verified in online social networks [16, 19] and mobile networks [4, 12]. With the ego network of each user, we study the demographic homophily on both gender and age. Figure 4 shows friends' demographic distribution for female and male users of different ages. The X-axis represents central users' age from 18 to 80 years old and the Y-axis represents the demographic distribution of users' friends, in which positive numbers denote female friends' ages and negative numbers denote male friends'. The spectrum color, which extends from dark blue (low) to red (high), represents the probability of one's friends belonging to the corresponding age (Y-axis) and gender (positive or negative). Interestingly, there exists a highlighted diagonal line in each figure, which suggests that people tend to communicate with others of similar age. In particular, the age homophily is much stronger for people aged between 35 to 55 years old. Simultaneously, the highlighted diagonals appear in the same gender range, i.e. females appear in the positive Y range (F) and males in the negative Y range (M), which shows the existence of a high degree of gender homophily.

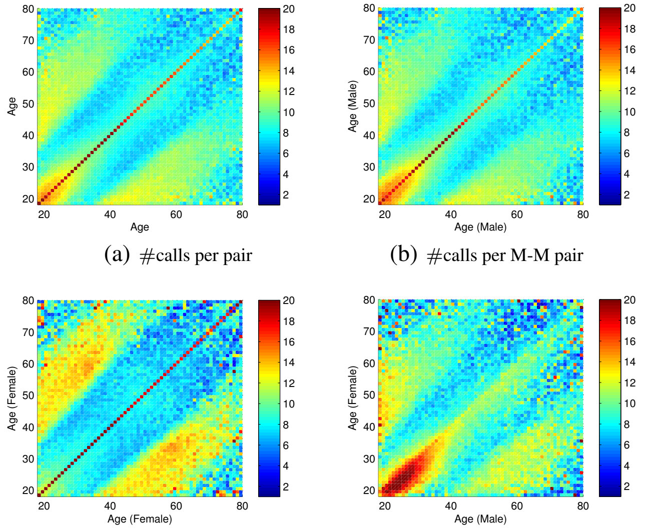
Cross-generation Communication. The analysis above confirms the existence of demographic homophily in mobile social networks. But what are the patterns underlying cross-generational communication, e.g., parents and kids? In fact, the cross-generational communication is a fundamental issue in sociology. The bulk of research [20, 28] has been conducted toward bridging age gaps in human society at large.

In Figure 5, we use heat maps to visualize the communication frequencies for different demographics. Figure 5(a) reports the average number of calls per month between people. Figures 5(b)-5(d) detail the analysis by reporting the average numbers of calls between two male users, two female users, and one male and one



(a) Demog. dist. of Female's friends (b) Demog. dist. of Male's friends

Figure 4: Friends' demographic distribution. X-axis: (a) female age; (b) male age. Y-axis: age of friends (positive: female friends, negative: male friends). The spectrum color represents the friends' demographic distribution.



(a) #calls per pair (b) #calls per M-M pair

(c) #calls per F-F pair (d) #calls per M-F pair

Figure 5: Strength of social tie. XY-axis: age of users with specific gender. The spectrum color represents the number of calls per month. (a), (b), and (c) are symmetric.

female user, respectively. Again, we discover highlighted diagonal lines in Figures 5(a)-5(c), which correspond to the gender and age homophily. We also notice that there are highlighted areas corresponding to cross-generational communications. In Figure 5(a), the color of cross-generation areas that extends from green to yellow indicates that on average 13 calls per month have been made between people aged 20-30 and those aged 40-50 years old. This corresponds to communications between parents and children, managers and subordinates, advisors and advisees, etc.

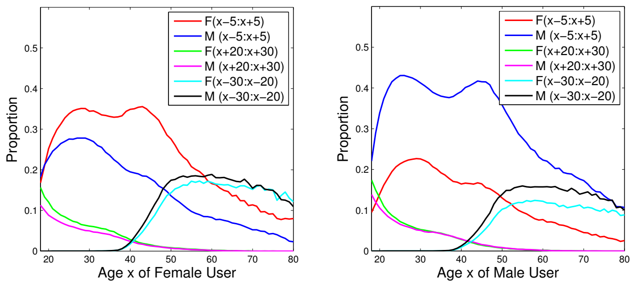
In addition, we observe that the communications between female users seem to be much more frequent than those between male users (Cf. Figures 5(b) and 5(c)). This is possibly because communications between mothers and daughters are more frequent than those between fathers and sons. Moreover, we find a "brothers" phenomenon among male users. In Figure 5(b), the diagonal highlights from 18 to 34 years (Young and Young-Adult) are much larger and denser than the corresponding area in Figure 5(c). This indicates that male users tend to maintain broader social connections than female users when they are young. Finally, from Figure 5(d), we observe a highlighted red area between people aged 18-34 years old, which means that cross-gender communications are more frequent than those between users of the same gender. A similar observation has also been reported in [16].

Demographic Dynamics. Human interactions between demographics reveal homophily or cross-generation phenomena not only in topologically but also in their dynamics. Evolutionary theory suggests that one's social strategy on her/his ego network vary and evolve across her/his lifetime as a function of the tradeoff between different social needs [27]. Herein, we focus our attention on the demographic dynamics in ego networks.

In Figure 6, we first report the friends' demographic distribution in one's ego network as a function of the central user's age. The X-axis denotes the central user's age x, and the Y-axis denotes the proportion of the age-groups her/his friends belong to, including the same generation ($x - 5$ to $x + 5$), older generation ($x + 20$ to $x + 30$), and younger generation ($x - 30$ to $x - 20$). From this, it is apparent that strong demographic dynamics exist in human interactions. First, the young and young-adult put increasing focus on the same generation ('blue' and 'red' lines) and maintain decreasing interactions with the older generation ('green' and 'pink' lines). Second, the middle-aged and seniors start to devote more attention to the new generation ('black' and 'cyanine' lines) even despite the sacrifice of homophily (the decreasing of 'blue' and 'red' lines). Third, during middle-age, the connections with the age-matched opposite-gender in ego networks ('blue' in Figure 6(a), 'red' in Figure 6(b)) decrease, and the proportion of same-gender friends ('red' in Figure 6(a) and 'blue' in Figure 6(b)) stay relatively stable.

In Figure 7, the heat map visualizes the distribution of minimum age (X-axis) and maximum age (Y-axis) of three users in a close social triad structure. Figures 7(a) and 7(d) show the same-gender triads: 'FFF' and 'MMM', and Figures 7(b) and 7(c) present the age distribution for users in opposite-gender triads: 'FFM' and 'FMM'. When people are young, the strong triadic relationships are revealed in all four kinds of gender-triads: 'FFF', 'MMM', 'FFM' and 'FMM' by highlighted red areas at the left-bottom corners. More excitingly, when entering middle-age, people only maintain the same-gender triadic relationships, which is revealed by the red diagonal lines in Figure 7(a) and Figure 7(d). However, the opposite-gender triadic relationships vanish when people pass 34 years old in Figure 7(b) and 7(c). The instability of opposite-gender triadic relationships and the persistence of same-gender triadic relationships across one's lifetime are novel discoveries and reveal human social strategy dynamics that are more complex and crucial than can be revealed by a static view.

Social Strategies. First, Figure 4 confirms that people have the tendency to interact with others with similar gender and age. Second, Figure 5 shows that the cross-generation interactions are maintained to pass the torch of family, workforce, and human knowledge from generation to generation in social society. Third, human interactions reveal striking gender differences in social triadic relationships across one's lifespan, which reflects dynamic gender-bias of human behaviors from young to old. Figure 7 shows that people tend to expand their social connections with both females and



(a) Proportion of Female's friends' age (b) Proportion of Male's friends' age

Figure 6: Dynamics of friends' demographic distribution in ego networks. X-axis: (a) female age; (b) male age. Y-axis: Proportion of friends' age groups.

males during the dating and reproductively active period, and put more social investment on maintaining same-gender social groups after entering into middle-age.

4. DEMOGRAPHIC PREDICTION

All the observations above clearly demonstrate the strong demographic interrelations between users' gender and age. For example, a 20-year-old female's behaviors are distinct from not only a 20-year-old male's, but also from a 50-year-old female's. Based on this, we formalize the demographic prediction as a double dependent-variable classification problem, i.e., we infer gender and age simultaneously. The *WhoAmI* method, a Double Dependent-Variable Factor Graph Model, is proposed to solve this problem by leveraging not only the correlations between features and gender/age, but also the interrelations between gender and age.

4.1 Problem Definition

Let $G = (V, E, Y, Z)$ denote the undirected and weighted mobile network, where V is a set of $|V| = N$ users and $E \subseteq V \times V$ is a set of communication edges (CALL or SMS) between users. Each user $v_i \in V$ is associated with demographic information, i.e., Gender $y_i \in Y$ and Age $z_i \in Z$. $\mathbf{X}$ is the attribute matrix, where each row $\mathbf{x}_i$ represents an $|\mathbf{x}_i|$ dimensional feature vector for user v_i. Given this, we define our problem below.

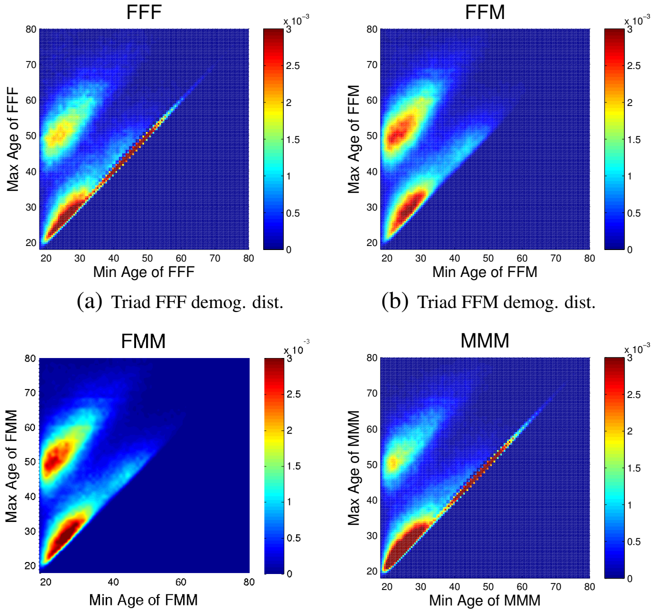
Problem 1. **Demographic Prediction:** Given a partially labeled network $G = (V^L, V^U, E, Y^L, Z^L)$ and the attribute matrix $\mathbf{X}$, where V^L is a set of users with labeled demographic information Y^L and Z^L, and V^U is a set of unlabeled users, the objective is to learn a function

$$f : G = (V^L, V^U, E, Y^L, Z^L), \mathbf{X} \rightarrow (Y^U, Z^U)$$

to predict users' gender and age simultaneously, where Y^U, Z^U are the demographic information for the unlabeled users V^U.

Different from previous work on demographic prediction [3, 10], where users' gender and age are inferred by modeling $P(Y|G, \mathbf{X})$ and $P(Z|G, \mathbf{X})$ separately, our problem here is to model $P(Y, Z|G, \mathbf{X})$ and predict users' gender and age simultaneously. We leverage not only the correlation between $\mathbf{X}$ and Y/Z but also the interrelations between gender Y and age Z. The motivation here comes from the fact that Section 3 reveals strong demographic interrelations in human communication behaviors.

In this work, we infer gender as a binary classification problem, i.e., Female or Male, and infer age as a four-class classification problem by splitting users' age into four groups: Young (18-24), Young-Adult (25-34), Middle-Age (35-49), and Senior (> 49).



(a) Triad FFF demog. dist. (b) Triad FFM demog. dist.

(c) Triad FMM demog. dist. (d) Triad MMM demog. dist.

Figure 7: Demographic distribution in social triad. X-axis: minimum age of three users in a triad. Y-axis: maximum age of three users. The spectrum color represents the distributions.

4.2 The WhoAmI Framework

Our goal is to design a unified model to capture not only users' attributes on demographics but also the interrelations between users' gender and age. We propose the WhoAmI method, a Double Dependent-Variable Factor Graph Model. To handle large-scale networks, we further develop a distributed learning algorithm.

4.2.1 Double Dependent-Variable Factor Graph

We define an objective function by maximizing the conditional probability of users' gender Y and age Z given their corresponding attributes and the input network structure, i.e., $P_\theta(Y, Z|G, \mathbf{X})$. The factor graph [13] provides a way to factorize the "global" probability as a product of "local" factor functions, which makes the maximization simple, i.e.,

$$P(Y, Z|G, \mathbf{X}) = \frac{P(\mathbf{X}, G|Y, Z)P(Y, Z)}{P(\mathbf{X}, G)} \propto P(Y, Z|G)P(\mathbf{X}|Y, Z)$$
$$\propto \prod_{v_i \in V} P(\mathbf{x}_i|y_i, z_i) \prod_{c \in G} P(Y_c, Z_c) \quad (2)$$

where $P(Y_c, Z_c)$ denotes the probability of labels given the structure c of the network and $P(\mathbf{x}_i|y_i, z_i)$ is the probability of generating attributes $\mathbf{x}_i$ given the label y_i and z_i.

In this model, we design three kinds of factors. The first is an attribute factor $f(y_i, z_i, \mathbf{x}_i)$ for capturing correlations between users' demographics and communication attributes. The second is a dyadic factor $g(\mathbf{y}_e, \mathbf{z}_e)$ for modeling correlations between users' demographics and their direct relationships in ego networks, where Y_c in Eq. 2 is represented as $\mathbf{y}_e$ (y_i, y_j), and Z_c is denoted by $\mathbf{z}_e$ (z_i and z_j) iff $e_{ij} \in E$. The third is a triadic factor $h(\mathbf{y}_c, \mathbf{z}_c)$ for correlating users' demographics and triadic relationships in their ego

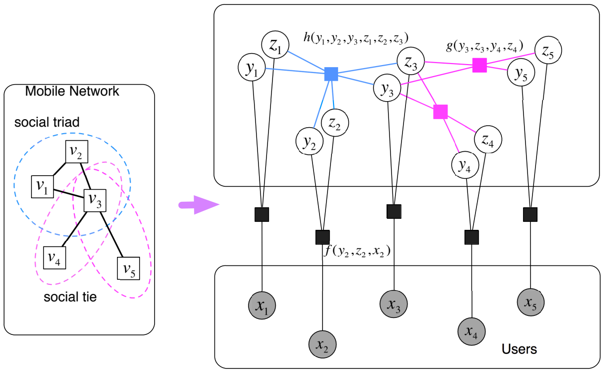


Figure 8: An illustration of the proposed model. y_i **and** z_i **indicate the gender and age of the user** v_i**.** x_i **denotes communication attributes of the user** v_i **extracted from the mobile network** G**.** $f(y_i, z_i, \mathbf{x}_i)$**,** $g(\mathbf{y}_e, \mathbf{z}_e)$**, and** $h(\mathbf{y}_c, \mathbf{z}_c)$ **respectively represent attribute factor, dyadic factor, and triadic factor in the proposed model.**

networks. Similarly, $\mathbf{y}_c$ means y_i, y_j, y_k and $\mathbf{z}_c$ is z_i, z_j, z_k when three users v_i, v_j, v_k form a close triangle structure.

Therefore, the joint distribution can be further factorized as:

$$P(Y, Z|G, \mathbf{X}) = \prod_{v_i \in V} f(y_i, z_i, \mathbf{x}_i) \times \prod_{e_{ij} \in E} [g(\mathbf{y}_e, \mathbf{z}_e)] \prod_{c_{ijk} \in G} [h(\mathbf{y}_c, \mathbf{z}_c)] \quad (3)$$

Figure 8 shows an illustration of our proposed model, which consists of two layers of nodes. The bottom layer contains random variables and the upper layer contains the three kinds of factors introduced above. The joint distribution over the whole set of random variables can be factorized as the product of all factors. Specifically, we instantiate the three factors as follows.

Attribute factor. We use this factor $f(y_i, z_i, \mathbf{x}_i)$ to represent the correlation between user v_i's demographics and her/his network characteristics $\mathbf{x}_i$. More specifically, we instantiate the factor by an exponential-linear function:

$$f(y_i, z_i, \mathbf{x}_i) = \frac{1}{W_v} \exp\{\alpha_{y_i z_i} \cdot \mathbf{x}_i\} \quad (4)$$

where α is one parameter of the proposed model, and W_v is a normalization term. For each pair of (y_i, z_i), $\alpha_{y_i z_i}$ is an $|\mathbf{x}|$-length vector, where the k-th dimension indicates how x_{ik} distributes over (y_i, z_i). For example, let's say x_{ik} represents the degree of user v_i. This factor can capture the fact that people with different gender and age have the different network properties shown in Figure 3. Traditional probabilistic graphical models can only model the correlations between features and one single type of dependent variable, while our proposed model captures how the features distribute over two types of dependent variables jointly.

Dyadic factor. We next define the dyadic factor $g(\mathbf{y}_e, \mathbf{z}_e)$, where $e_{ij} \in E$, to represent the correlation between user v_i and v_j's demographic information. Specifically, we have

$$g(\mathbf{y}_e, \mathbf{z}_e) = \begin{cases} \frac{1}{W_{e_1}} \exp\{\beta_1 \cdot g'_1(y_i, y_j)\} \\ \frac{1}{W_{e_2}} \exp\{\beta_2 \cdot g'_3(y_i, z_i)\} \\ \cdots \\ \frac{1}{W_{e_6}} \exp\{\beta_6 \cdot g'_6(z_i, z_j)\} \end{cases} \quad (5)$$

where β_p is the model parameters for this type of factor, $g'_p(\cdot)$ is defined as a vector of indicator functions, and W_{e_p} is the normalization term. We can enumerate in total 6 different combinations of each pair of demographic variables from (y_i, y_j, z_i, z_j). The intuition behind is that v_i's friends' demographics distribute differently by varying either v_i's own age or gender, as Figure 6 suggests.

Triadic factor. We finally define the triadic factor $h(\mathbf{y}_c, \mathbf{z}_c)$ to represent the correlation among the demographics of social triads, where $c = \{v_i, v_j, v_k | e_{ij}, e_{jk}, e_{ik} \in E\}$ indicates the triangle structure in G. More specifically, we have

$$h(\mathbf{y}_c, \mathbf{z}_c) = \begin{cases} \frac{1}{W_{c_1}} \exp\{\gamma_1 \cdot h'_1(y_i, y_j, y_k)\} \\ \frac{1}{W_{c_2}} \exp\{\gamma_2 \cdot h'_2(y_i, y_j, z_i)\} \\ \cdots \\ \frac{1}{W_{c_{20}}} \exp\{\gamma_{20} \cdot h'_{20}(z_i, z_j, z_k)\} \end{cases} \quad (6)$$

where $h'_q(\cdot)$ is the vector of indicator functions and W_{c_q} is the normalization term similar with W_{e_p}. There are 20 different kinds of three-variable enumerations from $(y_i, y_j, y_k, z_i, z_j, z_k)$. We use these triadic factors to model the distributions of users' age and gender within a social triangle. See details in Figure 7.

Finally, combing Eq. 4, 5, 6 into Eq. 3, we define the objective function as the log-likelihood of the proposed model as:

$$\mathcal{O}(\alpha, \beta, \gamma) = \sum_{v_i \in V} \alpha_{y_i z_i} \mathbf{x}_i + \sum_{e_{ij} \in E} \sum_{p=1}^{6} \beta_p g'_p(\cdot) + \sum_{c_{ijk} \in G} \sum_{q=1}^{20} \gamma_q h'_q(\cdot) - \log W \quad (7)$$

where $W = W_v W_e W_c$ is the global normalization term, $W_e = \prod_{e_p=1}^{6} W_{e_p=1}$, and $W_c = \prod_{c_q=1}^{20} W_{c_q}$.

The technical novelty of the proposed model is that it considers two types of labels in a unified framework, which differentiates our model from traditional classification models. By considering two types of labels, the main advantage is that our model can characterize the correlations between gender/age and features more precisely.

4.2.2 Feature Definition

Given a network with labeled and unlabeled users, the goal is to infer unlabeled users' demographic information. This scenario satisfies a real-world application for mobile operators. There are three kinds of features designed in our experiments, namely attribute features, friend features, and circle features. Specifically, given an ego network with one central user v and her/his direct friends, we extract three kinds of features for this central user v as follows:

Individual feature. Characteristic attributes are extracted based on the network topological properties discussed in Section 2. It includes the degree, neighbor connectivity, clustering coefficient, embeddedness, and weighted degree (#calls or #messages).

Friend feature. There are two types of friend features, including friend attributes and dyadic factor. First, friend attributes are used to model the demographic distribution of v's direct friends in her/his ego network, including the number of connections to female, male, young, young-adult, middle-age, and senior friends. In the prediction scenario, not all friends of the central user v are labeled with gender or age information, so we extract the friend attributes only based on her/his labeled friends. Second, to further model the distribution of unlabeled and labeled friends together, dyadic factor is used as the other type of friend feature (Cf. Eq. 5).

Circle feature. Similarly, it also contains two kinds of features: circle attributes and triadic factor. First, circle attributes refer to the triadic demographic distribution of v's ego network. Because we aim to infer the central user v's demographics, we count the numbers of different gender triads, i.e., 'FF-v', 'FM-v', 'MM-v', and different age-group triads. Let A/B/C/D denote the young/young-adult/middle-age/senior age-groups, respectively. There are in total ten kinds of triads based on age-groups: 'AA-v', 'AB-v', 'AC-v', 'AD-v', 'BB-v' ,'BC-v' ,'BD-v', 'CC-v', 'CD-v', 'DD-v'. Second, triadic factor is used to model the demographic distributions over triangles with both unlabeled and labeled users (Cf. Eq. 6).

The individual, friend, and circle attributes are captured by the attribute factor in our DFG (Cf. Eq. 4). There are in total 24 attribute features used in our models.

4.2.3 DFG Learning and Inference

The goal of learning the DFG model is to find a configuration for the free parameters $\theta = \{\alpha, \beta, \gamma\}$ that maximize the log-likelihood of the objective function $\mathcal{O}(\theta)$ in Eq. 7 given by the training set, i.e.,

$$\theta^\star = \arg\max \mathcal{O}(\theta)$$

Learning. We first introduce how we learn the model in a single-processor configuration, and then explain how to extend the learning algorithm to a distributed one for handling large-scale networks.

To solve the optimization problem, we adopt a gradient decent method (or a Newton-Raphson method). Specifically, we derive the objective function with respect to each parameter with regard to our objective function in Eq. 7.

$$\begin{aligned}
\frac{\partial \mathcal{O}(\theta)}{\partial \alpha} &= \boldsymbol{E}[\sum_{v_i \in V} f(y_i, z_i, \mathbf{x}_i)] - \boldsymbol{E}_{P_\alpha(Y,Z|X)}[\sum_{v_i \in V} f(y_i, z_i, \mathbf{x}_i)] \\
\frac{\partial \mathcal{O}(\theta)}{\partial \beta} &= \boldsymbol{E}[\sum_{e_{ij} \in E} g(\mathbf{y}_e, \mathbf{z}_e)] - \boldsymbol{E}_{P_\beta(Y,Z|\mathbf{X},G)}[\sum_{e_{ij} \in E} g(\mathbf{y}_e, \mathbf{z}_e)] \\
\frac{\partial \mathcal{O}(\theta)}{\partial \gamma} &= \boldsymbol{E}[\sum_{c_{ijk} \in G} h(\mathbf{y}_c, \mathbf{z}_c)] - \boldsymbol{E}_{P_\gamma(Y,Z|\mathbf{X},G)}[\sum_{c_{ijk} \in G} h(\mathbf{y}_c, \mathbf{z}_c)]
\end{aligned} \tag{8}$$

where in the first Equation of Eq. 8, $\boldsymbol{E}[\sum_{v_i \in V} f(y_i, z_i, \mathbf{x}_i)]$ is the expectation of the summation of the attribute factor functions given the data distribution over Y, Z, and X in the training set, and $\boldsymbol{E}_{P_\alpha(Y,Z|X)}[\sum_{v_i \in V} f(y_i, z_i, \mathbf{x}_i)]$ is the expectation of the summation of the attribute factor functions given by the estimated model. The other expectation terms have similar meanings in the other two equations. As the network structure in the real-world may contain cycles, it is intractable to estimate the marginal probability in the second terms of Eq. 8. In this work, we adopt Loopy Belief Propagation (LBP) [23] to calculate the marginal probability of $P(Y, Z)$ and compute the expectation terms.

The learning process then can be described as an iterative algorithm. Each iteration contains two steps: First, we call LBP to calculate marginal distributions of unknown variables $P_\alpha(Y, Z|X)$. Second, we update α, β, and γ with the learning rate η by Eq. 9. The learning algorithm terminates when it reaches convergence.

$$\theta_{new} = \theta_{old} + \eta \cdot \frac{\partial \mathcal{O}(\theta)}{\partial \theta} \tag{9}$$

Distributed learning. We further develop a distributed algorithm to scale up our model to handle these large-scale networks. Our distributed learning algorithm utilizes a Message Passing Interface (MPI) framework, by which we can split the network into small parts and learn the parameters on different processors. As most computing time is consumed in the first step of our learning algorithm introduced above, we speed up this learning processes by distributing multiple 'slave' computing processors for this step. The second step is calculated in the 'master' processor by collecting the results from all 'slave' processors on the first step.

Specifically, our distributed learning algorithm based on the master-slave framework can be described in two phases. At the first phase, the large-scale network G is partitioned into K sub-networks $G_1, \cdots, G_k, \cdots, G_K$, and the K sub-networks are distributed to K 'slave' processors. At the second phase, we iteratively learn the parameters in two steps. First, each processor can compute the local marginal probability on its sub-network G_k. Second, the 'master' processor collects all gradients obtained from different subgraphs and updates all parameters. The second phase is repeated until convergence.

There are two notes for our model implementation. At the first phase, different from partitioning the large network based on communities in previous work [31], we partition the country-wide mobile network based on different administrative regions, i.e., provinces or states. The mobile operator can obtain this information when its users first register their mobile phones. The other one is that we first extract all features for each user from the original large network. We then split it into sub-networks that are handled by each 'slave' processor.

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

$$(Y^*, Z^*) = \arg\max\ \mathcal{O}(Y, Z|G, \mathbf{X}, \theta)$$

In this paper, we use the max-sum algorithm [18] to solve the above problem.

5. EXPERIMENTS

We present the effectiveness and efficiency of our proposed DFG model on demographic prediction by various experiments. The code used in the experiment is publicly available.[4]

5.1 Experiment Setup

Data and Evaluation. We use two large-scale mobile networks, CALL and SMS, to infer users' gender and age. Detailed data information is introduced in Section 2. To infer user demographics effectively for mobile operators, we only consider active users who have at least five contacts in two months. After filtering, there are 1.09 million and 304,000 active users in CALL and SMS networks, respectively. We repeat the prediction experiments ten times, and report the average performance in terms of weighted Precision, Recall, and F1-Measure. We consider weighted evaluation metrics because every class in female/male or young/young-adult/middle-age/senior is as important as each other. It is worth noting that the weighted Recall has the same score as Accuracy.

All code is implemented in C++, and experiments are performed in a server with four 16-core 2.4 GHz AMD Opteron processors with 256GB RAM. We use the speedup metric with different numbers of computing cores (1-16) to evaluate the scalability of our distributed learning algorithm.

Comparison methods. We compare our proposed DFG model that can capture the interrelation between two dependent variables

[4] http://arnetminer.org/demographic

(gender and age) with different classification algorithms, including Logistic Regression (**LRC**), Support Vector Machine (**SVM**), Naive Bayes (**NB**), Random Forest (**RF**), Bagging (**Bag**), Gaussian Radial Basis Function Neural Network (**RBF**), and Factor Graph Model (**FGM**). For LRC, NB, RF, Bag, RBF, we employ Weka[5] and use the default setting and parameters. For SVM, we use liblinear[6]. For FGM, the model proposed in [19] is used. Note that our proposed DFG model is equal to FGM if we do not consider the interrelations between gender and age.

For all comparison methods, we use the same unstructured features (characteristic, friend, and circle attributes) introduced in Section 4.2.2. For the graphical models, FGM and DFG , the structure features (dyadic and triadic factors) are further used to model user demographics on network structure. The major difference between our DFG and FGM model is that DFG can capture not only the correlation between different users, but also the interrelation between two dependent variables of each user, i.e., gender and age.

5.2 Experiment Results

We report the demographic prediction performance for different methods in the CALL and SMS networks as follows. In prediction experiments, we use 50% of the labeled data in each network as a training set and the remaining 50% for testing.

Predictive performance. Table 1 shows the prediction results of different algorithms on the four prediction cases, i.e., gender and age predictions in the CALL and SMS networks, respectively. Clearly our DFG model yields better performance than the other alternative methods in the four cases. The Bag algorithm achieves the best prediction results among all non-graphical methods. The FGM model outperforms a series of non-graphical algorithms by modeling the correlations among structured nodes via dyadic and triadic factors. The DFG model outperforms FGM by further leveraging the interrelations between user gender and age. In terms of weighted Precision, Recall, and F1-Measure, DFG achieves up to 10% improvements compared with the baselines for the prediction of users' gender and age. As for Accuracy, the DFG model can infer 80% of the users' gender in the CALL network and 73% of the users' age in the SMS network correctly. Finally, we observe that the CALL network can reveal more users' gender information than the SMS network, as the overall performance of gender prediction in CALL is about 5% higher than that in SMS. However, predicting age from SMS behaviors is relatively easier than predicting it from CALL communications.

Effects of demographic interrelation. We evaluate the effects of demographic interrelation on the predictions. Without modeling the interrelation between gender and age, our proposed DFG model degenerates to a basic factor graph model (FGM/DFG-d). From Table 1, we clearly observe the 2% to 4% improvements achieved by DFG to FGM on weighted F1-Measure. We further analyze feature contributions for demographic prediction. Recall that in Section 4.2.2, besides the individual features, we introduced the friend features (friend attributes and edge factor) and circle features (circle attributes and triad factor). By removing either friend or circle features, we evaluate the decrease in predictive performance in terms of weighted F1-Measure, plotted in Figure 9. DFG-df, DFG-dc and DFG-dfc stand for the removing of friend features, circle features, and both, conditioned on DFG-d without modeling gender and age interrelations. Clearly, we can see that for inferring gender, the performance when removing circle features drops more than when removing friend features, which indicates a stronger contribution of circle features to gender prediction than friend features. However, for inferring user age, friend features contribute more than circle features. The feature contribution analysis further validates our observations of demographic-based social strategies, and demonstrates that the proposed model works well by capturing the observed phenomena.

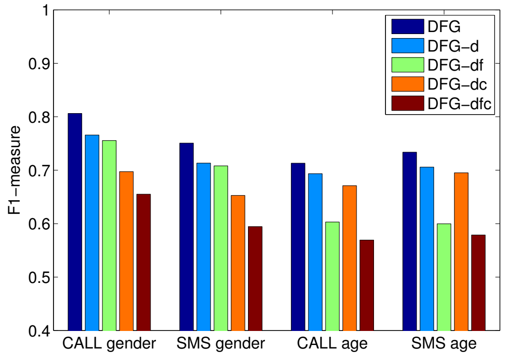


Figure 9: Feature Contribution Analysis. DFG is the proposed model. TFG-d is the basic version of DFG without modeling the correlation between gender and age. DFG-df stands for further ignoring friend features. DFG-dc stands for further ignoring circle features. DFG-dcf stands for ignoring both friend and circle features.

Training/test ratio. We provide further analysis on the effects of training ratio on predictive performance. Figure 10 shows the prediction results when varying the percentage of labeled users in the training set. Clearly, we can see rising trends as the training set increases in Figure 10(a) and 10(b). This indicates the positive effects of training data size on predicting the gender of mobile users. The smooth lines in Figure 10(c) and 10(d) reveal the limited contributions of training data size on predicting age. We can see that in all cases, obvious improvements can be obtained by our proposed DFG model with different sizes of training data.

Scalability. We verify the distributed learning algorithm by partitioning the original large-scale network into multiple sub-networks based on different administrative areas. We determine users' areas by their postal code information during registration. Each subnetwork in one area is used as the input for a given core. By utilizing MPI, our distributed algorithm can achieve 9-10× speedup with 16 cores with $<2\%$ drop in performance. Basically, our learning algorithm can converge in 100 iterations, and each iteration costs about 2-5 minutes for one single processor. By leveraging a distributed learning algorithm, our DFG model is efficient even for large-scale networks with millions of nodes.

6. RELATED WORK

The availability of mobile phone communication records has offered researchers many ways to analyze mobile networks, greatly enhancing our understanding of human mobile behaviors.

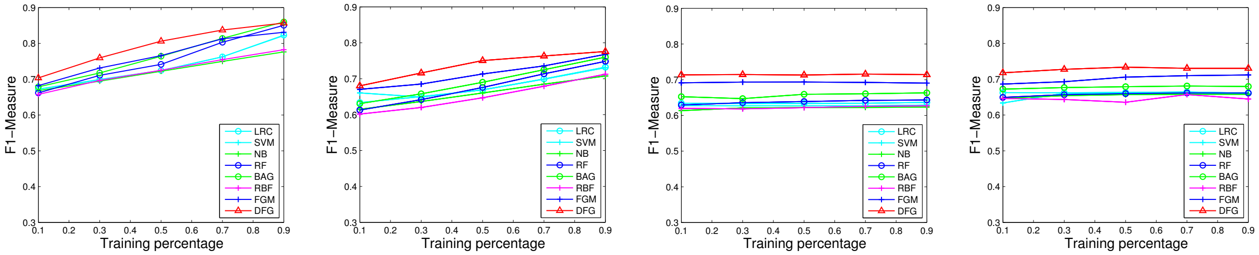
To better model the macro properties of mobile communication networks, Onnela [26] and Nanavati et al. [24] examine the local and global structure of a society-wide mobile communication network. Faloutsos et al. [29] first propose the double pareto-lognormal distribution to model the macro properties in call networks, which is beyond power-law and lognormal distributions. They further discover that not only the node properties but also

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

Table 1: Demographic Prediction performance by weighted Precision, Recall, and F1-Measure. The weighted Recall score is equal to the Accuracy score. The number in parentheses is the standard deviation.

Network	Method	Gender			Age		
		wPrecision	wRecall/Accu	wF1-Measure	wPrecision	wRecall/Accu	wF1-Measure
CALL	LRC	0.7327 (0.0003)	0.7289 (0.0003)	0.7245 (0.0005)	0.6350 (0.0005)	0.6466 (0.0003)	0.6337 (0.0005)
	SVM	0.7327 (0.0004)	0.7287 (0.0003)	0.7242 (0.0003)	0.6369 (0.0004)	0.6463 (0.0005)	0.6273 (0.0005)
	NB	0.7222 (0.0004)	0.7227 (0.0003)	0.7222 (0.0004)	0.6246 (0.0011)	0.6224 (0.0002)	0.6223 (0.0002)
	RF	0.7437 (0.0003)	0.7310 (0.0002)	0.7415 (0.0003)	0.6382 (0.0010)	0.6482 (0.0008)	0.6388 (0.0009)
	Bag	0.7644 (0.0005)	0.7648 (0.0004)	0.7643 (0.0005)	0.6607 (0.0010)	0.6688 (0.0004)	0.6592 (0.0005)
	RBF	0.7283 (0.0015)	0.7275 (0.0005)	0.7252 (0.0017)	0.6194 (0.0062)	0.6272 (0.0068)	0.6218 (0.0068)
	FGM	0.7658 (0.0096)	0.7662 (0.0115)	0.7659 (0.0113)	0.6998 (0.0094)	0.6989 (0.0087)	0.6935 (0.0089)
	DFG	0.8088 (0.0139)	0.8076 (0.0148)	0.8063 (0.0131)	0.7266 (0.0097)	0.7140 (0.0094)	0.7132 (0.0091)
SMS	LRC	0.6766 (0.0013)	0.6758 (0.0006)	0.6689 (0.0014)	0.6702 (0.0011)	0.6890 (0.0008)	0.6630 (0.0008)
	SVM	0.6749 (0.0006)	0.6750 (0.0005)	0.6690 (0.0007)	0.6654 (0.0163)	0.6884 (0.0006)	0.6607 (0.0006)
	NB	0.6231 (0.0003)	0.6655 (0.0011)	0.6603 (0.0021)	0.6563 (0.0014)	0.6588 (0.0015)	0.6570 (0.0012)
	RF	0.6399 (0.0009)	0.6749 (0.0009)	0.6757 (0.0009)	0.6623 (0.0013)	0.6775 (0.0008)	0.6598 (0.0011)
	Bag	0.6905 (0.0005)	0.6918 (0.0009)	0.6901 (0.0009)	0.6907 (0.0008)	0.6987 (0.0009)	0.6791 (0.0009)
	RBF	0.6712 (0.0006)	0.6592 (0.0131)	0.6468 (0.0139)	0.6295 (0.0062)	0.6640 (0.0051)	0.6356 (0.0042)
	FGM	0.7132 (0.0040)	0.7138 (0.0050)	0.7133 (0.0057)	0.7154 (0.0046)	0.7154 (0.0046)	0.7059 (0.0058)
	DFG	0.7589 (0.0187)	0.7549 (0.0159)	0.7507 (0.0178)	0.7409 (0.0199)	0.7303 (0.0208)	0.7337 (0.0198)



(a) CALL Gender Prediction (b) SMS Gender Prediction (c) CALL Age Prediction (d) SMS Age Prediction

Figure 10: Performance of demographic prediction with different percentages of labeled data.

clique structures follow the power-law distribution in mobile networks [5]. Dong et al. [4] investigate the mobile call duration behaviors in mobile social networks and find that people who are familiar with each other tend to make short calls. Recently, the emergence of work on mobility [30, 33, 8, 36] and location-based mobile networks [7, 2, 1], where human movements or locations are tracked by mobile phones, provides us a means of understanding and predicting mobile social behaviors. Eagle et al. [6] try to infer the friendship network in mobile phone data. As for mobility applications, mobile patterns are applied to predict the future location of users in [22, 25, 34]. However, most previous work focuses on scaling the macroscopic properties of mobile networks, while our work incorporates the micro-network structure to model human communication behaviors in mobile networks.

Furthermore, there are several works on user demographic and profile modeling. Existing works try to infer user demographics based on their online browsing [10] and search [3] behaviors. Leskovec and Horvitz [16] examine the interplay of the MSN network and user demographic attributes. Tang et al. extract and model the researcher profiles in large-scale collaboration networks [32]. Additionally, researchers have used network information to identify user status differences in email [11] and LinkedIn networks [37]. Nokia research organized the 2012 Mobile Data Challenge to infer mobile user demographics by using 200 individual communication records without network information [21, 35]. Kovanen et al. [12] utilize temporal motifs to reveal demographic homophily in dynamic communication networks. The main difference between existing work and our efforts lies in that existing work mainly analyzes demographics (gender, age, status, etc.) separately, while our analysis and model consider the correlation among different demographic attributes.

7. CONCLUSION

In this paper, we study human interactions on demographics by investigating a country-wide mobile communication network. From this, we discover a set of social strategies stemming from human communications. First, young people put more focus on enlarging their social circles; as they age, they have the tendency to maintain small but close social relationships. Second, we observe a strong homophily of human interactions on gender and age simultaneously. Third, beyond these observations, we find that the frequent cross-generation interactions are maintained to pass the torch of family, workforce, and human knowledge from generation to generation in social society. Finally, we observe striking gender differences in social triadic relationships across individuals' lifespans, which reflects dynamic gender-bias of human behaviors from young to old.

Through these observations, we engage in answering the question of to what extent individual demographic can be revealed from mobile communication interactions. We formalize a demographic prediction problem to infer user gender and age simultaneously, and further propose the WhoAmI method to solve this problem by modeling not only the correlations between gender/age and features, but also the interrelations between gender and age. Experimental results in the CALL and SMS networks demonstrate the effectiveness and efficiency of our proposed model.

Detecting user demographics makes social networks more colorful and closer to our real human networks. For future work, some other social theories and strategies can be explored and validated for modeling human mobile social interactions. In addition, examining how the inferred demographic can help other topics in social network analysis, such as influence propagation, community detection, and network evolution, would also be very meaningful.

Acknowledgements. We sincerely thank Dr. Albert-L*á*szl*ó* Barab*á*si for sharing the mobile phone data with the University of Notre Dame. We also thank anonymous reviewers for their useful comments, and Reid Johnson at Notre Dame for his insightful discussion. Jie Tang and Yang Yang are supported by the National High-Tech R&D Program of China (No. 2014AA015103), Natural Science Foundation of China (No. 61222212, 61073073), and National Basic Research Program of China (2013CB329603) and a research fund supported by Huawei Inc. Nitesh V. Chawla, Yuxiao Dong, and Yang Yang are supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, and the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) grant #FA9550-12-1-0405.

8. REFERENCES

[1] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel. A survey on recommendations in location-based social networks. *ACM TIST*, 2014.

[2] M. Berlingerio, F. Calabrese, G. D. Lorenzo, R. Nair, F. Pinelli, and M. L. Sbodio. Allaboard: A system for exploring urban mobility and optimizing public transport using cellphone data. In *ECML/PKDD (3)*, pages 663–666. Springer, 2013.

[3] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: Social data meets search queries. In *WWW '13*, pages 131–140, 2013.

[4] Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla. How long will she call me? distribution, social theory and duration prediction. In *ECML/PKDD (2)*, pages 16–31, 2013.

[5] N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: Patterns and a utility-driven generator. In *KDD '09*, pages 269–278. ACM, 2009.

[6] N. Eagle, A. S. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *PNAS*, 106(36), 2009.

[7] H. Gao, J. Tang, X. Hu, and H. Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *CIKM '13*, pages 1673–1678, 2013.

[8] F. Giannotti and D. Pedreschi. *Mobility, data mining and privacy: Geographic knowledge discovery*. Springer, 2008.

[9] M. Granovetter. Economic action and social structure: The problem of embeddedness. *The American Journal of Sociology*, 1985.

[10] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *WWW '07*, pages 151–160, 2007.

[11] X. Hu and H. Liu. Social status and role analysis of Palin's email network. In *WWW '12 Companion*, pages 531–532. ACM, 2012.

[12] L. Kovanen, K. Kaski, J. Kertész, and J. Saramäki. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *PNAS*, 2013.

[13] F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE TOIT*, 47:498–519, 2001.

[14] P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society, New York: Van Nostrand*, pages 8–66, 1954.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08*, pages 462–470, 2008.

[16] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08*, pages 915–924. ACM, 2008.

[17] 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.

[18] H.-A. Loeliger. An introduction to factor graphs. *Signal Processing Magazine, IEEE*, 21(1):28–41, 2004.

[19] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to predict reciprocity and triadic closure in social networks. *ACM TKDD*, 7(2):5:1–5:25, 2013.

[20] M. Mead. *Culture and commitment: a study of the generation gap*. Natural History Press, 1970.

[21] K. Mo, B. Tan, E. Zhong, and Q. Yang. Your phone understands you. In *Nokia MDC '12*, 2012.

[22] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: A location predictor on trajectory pattern mining. In *KDD '09*, pages 637–646, 2009.

[23] 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.

[24] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: Findings and implications. In *CIKM '06*, pages 435–444, 2006.

[25] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *ICDM '12*, pages 1038–1043, 2012.

[26] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *PNAS*, 2007.

[27] V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. M. Dunbar. Sex differences in intimate relationships. *Scientific Reports*, 2:370, 2012.

[28] R. Prasad. *Generation Gap: A study of intergenerational sociological conflict*. Mittal Publications, 1992.

[29] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec. Mobile call graphs: beyond power-law and lognormal distributions. In *KDD '08*, pages 596–604. ACM, 2008.

[30] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 2010.

[31] J. Tang, S. Wu, and J. Sun. Confluence: Conformity influence in large social networks. In *KDD '13*, pages 347–355. ACM, 2013.

[32] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD '08*, pages 990–998, 2008.

[33] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *KDD '11*, pages 1100–1108. ACM, 2011.

[34] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *ICDE '13*, pages 254–265, 2013.

[35] J. Ying, Y.-J. Chang, C.-M. Huang, and V. S. Tseng. Demographic prediction based on user's mobile behaviors. In *Nokia MDC '12*, 2012.

[36] N. J. Yuan, Y. Wang, F. Zhang, X. Xie, and G. Sun. Reconstructing individual mobility from smart card transactions: A space alignment approach. In *ICDM'13*, pages 877–886, 2013.

[37] Y. Zhao, G. Wang, P. S. Yu, S. Liu, and S. Zhang. Inferring social roles and statuses in social networks. In *KDD '13*, pages 695–703, 2013.