User Modeling on Demographic Attributes in Big Mobile Social Networks

YUXIAO DONG and NITESH V. CHAWLA, University of Notre Dame
JIE TANG, Tsinghua University
YANG YANG†, Zhejiang University
YANG YANG‡, Northwestern University

Users with demographic profiles in social networks offer the potential to understand the social principles that underpin our highly connected world, from individuals, to groups, to societies. In this article, we harness the power of network and data sciences to model the interplay between user demographics and social behavior and further study to what extent users' demographic profiles can be inferred from their mobile communication patterns. By modeling over 7 million users and 1 billion mobile communication records, we find that during the active dating period (i.e., 18–35 years old), users are active in broadening social connections with males and females alike, while after reaching 35 years of age people tend to keep small, closed, and same-gender social circles. Further, we formalize the demographic prediction problem of inferring users' gender and age simultaneously. We propose a factor graph-based *WhoAmI* method to address the problem by leveraging not only the correlations between network features and users' gender/age, but also the interrelations between gender and age. In addition, we identify a new problem—coupled network demographic prediction across multiple mobile operators—and present a coupled variant of the *WhoAmI* method to address its unique challenges. Our extensive experiments demonstrate the effectiveness, scalability, and applicability of the *WhoAmI* methods. Finally, our study finds a greater than 80% potential predictability for inferring users' gender from phone call behavior and 73% for users' age from text messaging interactions.

CCS Concepts: • **Information systems** → **Data mining**; • **Human-centered computing** → **Collaborative and social computing**; **Social networks**; **Mobile computing**; • **Social and professional topics** → **User characteristics**; **Gender**; **Age**; • **Applied computing** → **Sociology;**

Additional Key Words and Phrases: Gender and age, demographic prediction, node attributes, ego networks, social tie and triad, mobile communication, mobile phone data, computational social science

ACM Reference Format:
Yuxiao Dong, Nitesh V. Chawla, Jie Tang, Yang Yang, and Yang Yang. 2017. User modeling on demographic attributes in big mobile social networks. ACM Trans. Inf. Syst. 35, 4, Article 35 (July 2017), 33 pages.
DOI: http://dx.doi.org/10.1145/3057278

Jie Tang and Yang Yang† are supported by the National High-tech R&D Program (2014AA015103, 2015AA124102) and National Basic Research Program of China (2014CB340506). 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 National Science Foundation (NSF) Grants BCS-1229450 and IIS-1447795. We sincerely thank Reid A. Johnson for his insightful comments.
Authors' addresses: Y. Dong and N. V. Chawla, Interdisciplinary Center for Network Science and Applications (iCeNSA) and Department of Computer Science and Engineering, University of Notre Dame, IN 46556 USA; emails: {ydong1, nchawla}@nd.edu; J. Tang, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P. R. China; email: jietang@tsinghua.edu.cn; Y. Yang†, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, P. R. China; email: yangya@zju.edu.cn; Y. Yang‡, Kellogg School of Management, Northwestern University, IL 60208 USA; email: yang.yang@kellogg.northwestern.edu. This work was done when Y. Yang† and Y. Yang‡ were Ph.D. students at Tsinghua University and University of Notre Dame, respectively.

DOI: http://dx.doi.org/10.1145/3057278

1. INTRODUCTION

As of 2016, the number of mobile users is 4.611 billion, corresponding to a global penetration of 62%; The number of mobile subscriptions across the globe reaches 7.377 billion in 2016, which is approximately the same with the world population, from a recent report by the International Telecommunications Union (ITU). On average, each mobile user makes, receives, or avoids 22 phone calls and sends or receives text messages 23 times, and checks their phones up to 150 times a day.[1] 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 build and maintain connections in mobile communication networks.

Previous work on mobile communication networks mainly focused on macro-level models, like network distributions [Onnela et al. 2007], scale free [Du et al. 2009], duration distributions [Dong et al. 2013; Seshadri et al. 2008], and mobility modeling [Gonzalez et al. 2008; Wang et al. 2011; Dong et al. 2015a]. Recently, researchers have also started to pay more attention to the micro-level analysis of the mobile networks. For example, Eagle et al. [2009] 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. Meng et al. [2016] studied the mobile communication networks of 200 students at the University of Notre Dame. They explored the interplay between individuals' evolving interaction patterns and traits. However, these works did not consider the interplay between user demographics and communication behavior. More recently, Nokia Research organized the 2012 Mobile Data Challenge to infer mobile user demographics by using communication records of 200 users [Ying et al. 2012; Mo et al. 2012]. However, the scale of the network is very limited. In this article, we leverage a large-scale mobile network to study how users' communication behaviors correlate with their demographic attributes.

Contributions. We employ a real-world large mobile network composed of more than 7,000,000 users and over 1,000,000,000 communication records (voice phone call and short text messaging) as the basis of our study, which we use to systematically investigate the interplay of user communication behavior and demographic information. Through the study, we first unveil several intriguing *social strategies* that users of different age and gender use to meet their social needs, that is, building new connections and maintaining existing relationships. Simultaneously, we examine the differences between people's phone call and text messaging behavior. Based on the discoveries, we then develop a unified probabilistic model—WhoAmI—to predict users' demographic profiles 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-world large mobile network.

This work expands on our previous work [Dong et al. 2014] in the following ways. First, we investigate social strategies from not only the voice phone call network but also the short text messaging network and further conclude the networking differences and similarities between human phone call and text messaging behaviors. Second, we propose to use a null model to validate the statistical significance of social strategies observed from network structures. Third, we generalize the previous prediction model, which can only handle two dependent variables, to support multiple dependent variables, enabling the simultaneous inference of any number of interrelated node attributes. Fourth, we identify a new problem—coupled network demographic prediction

[1] http://www.dailymail.co.uk/news/article-2276752/Mobile-users-leave-phone-minutes-check-150-times-day.html.

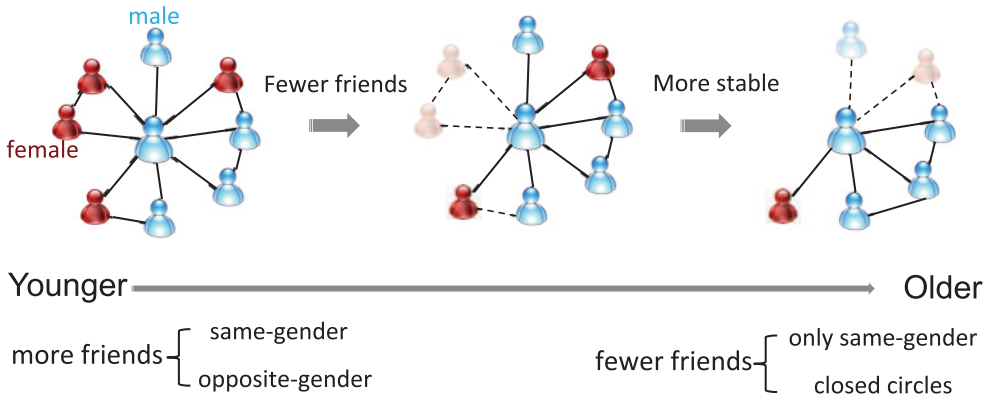


Fig. 1. Evolution of demographic-based social strategies in human communication.

across two mobile operators. To solve the unique challenges raised in coupled networks, we propose a variant version of the WhoAmI method—CoupledMFG. In order to handle large-scale (coupled) networks, we further present a distributed learning algorithm accompanying with the models. Finally, in addition to the prediction experiments in Dong et al. [2014], we demonstrate two real-world telecommunication applications: one for the normal demographic prediction problem, that is, the prediction of one mobile operator's prepaid users' demographics by the machine learning model trained on its postpaid users, and the other one for the coupled network demographic prediction, that is, the inference of competitors' user profiles by using the model trained on one operator's own users.

Key Findings. Our study unveils the significant social strategies and their evolution across the lifespan in human communication, which are highlighted in Figure 1. Specifically, we discover that younger people are very active in broadening their social circles, while older people tend to maintain smaller but more closed connections. We find that the communications between two younger opposite-gender users are more frequent than those between same-gender users. We also observe frequent cross-generation interactions that are essential for bridging age gaps in family, workplace, education, and human society as a whole [Mead 1970]. We unveil that people expand both same-gender and opposite-gender connections during their active dating period (18–34 years old), while they maintain only same-gender social groups in mobile communication after 35 years of age. Finally, our analysis shows strong interrelations between users' age and gender. For example, a 20-year-old female's social networking behavior is distinct from not only a 20-year-old male's, but also from a 50-year-old female's.

Demographic Prediction. 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-label classification problem. The objective is to simultaneously infer users' gender and age by leveraging their interrelations. This problem is 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 a 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 simultaneously infer Y and Z. To address this problem, we present the *WhoAmI* method, whereby the interrelations between multiple dependent variables can be modeled. As a result, the presented WhoAmI method is able to simultaneously infer users' gender and age. The experiments demonstrate

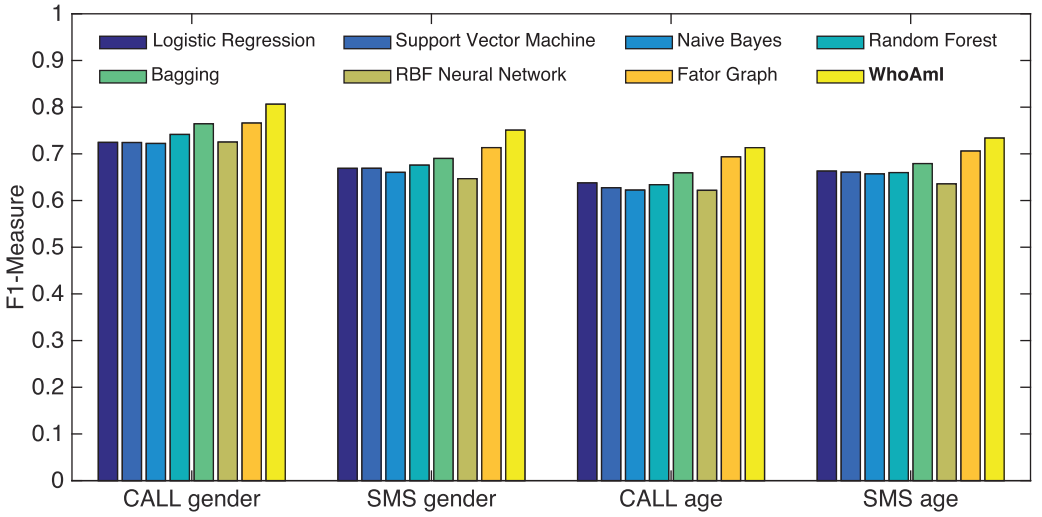


Fig. 2. Demographic prediction performance (cf. Section 7 for details of the comparison methods).

that the proposed method can achieve an accuracy of 80% for predicting users' gender and 73% for predicting users' age according to daily mobile communication patterns, significantly outperforming (by up to 10% in terms of F1-Measure shown in Figure 2) several alternative methods (cf. Section 7 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$\times$ with 16 CPU cores) by leveraging parallel computing.

We further demonstrate one application scenario of demographic prediction in telecommunication industry. In the real world, there are two kinds of mobile subscriptions of a mobile operator: postpaid[2] and prepaid.[3] Specifically, a *postpaid* mobile user is required to create an account by providing detailed demographic information (e.g., name, age, gender, etc.). However, a recent ITU report indicates that there is still a large portion of *prepaid* users (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 was accelerating during the economic recession from 2008. Prepaid services allow the users to be anonymous—no need to provide any user-specific information. In this sense, mobile operators are highly motivated to infer their prepaid users' demographic profiles. We take one case study to demonstrate the effectiveness of our discoveries and methodologies on this real-world application of demographic prediction for prepaid users.

Coupled Network Demographic Prediction. In addition to its prepaid users, a mobile operator also does not have the demographic information of users of another operator. For example, in Figure 3 a mobile operator O_1 (e.g., AT&T) could have the communication logs of two O_1 users, and one O_1 user and one user of another operator O_2 (Verizon) [Dong et al. 2015]. In the real world, O_1 does not have the access to the demographic profiles of its competitor O_2's users. However, it is critical for mobile service providers to build the demographic profiles of its competitors' customers. This can

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

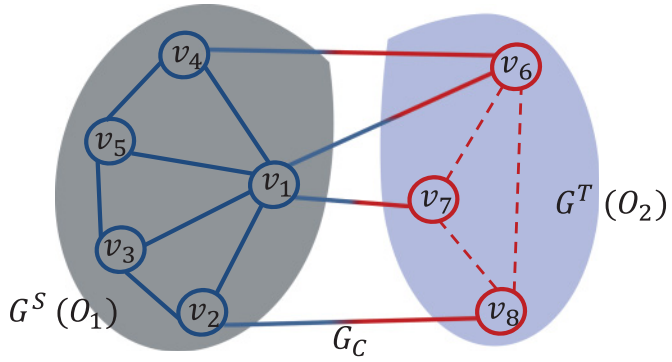


Fig. 3. An illustrative example of coupled networks across two mobile operators. The source network is mobile operator O_1's network. O_1 could also have the demographic information of its own users (postpaid). The objective is to predict the demographic profiles of users in its competitor O_2's network.

help them make better marketing strategies (e.g., identifying potential customers and preventing customer churning). Moreover, by using demographic information, service providers can supply users with more personalized services and focus on enhancing the communication experience.

In light of the real scenario in telecommunication, we formalize the coupled network demographic prediction problem, where we have the structure and user demographic information of one (source) network G^S (e.g., O_1) and the interactions between this network and another (target) network G^T (e.g., O_2). The goal is to predict the demographic attributes of users in the target network. This problem faces several unique challenges, including the cold start of the target network structure and, as a result, the asymmetry of source and target users' graph-based features. To address them, we present a coupled version of the WhoAmI method. Our experiments over six pairs of mobile operators demonstrate the predictability of competitors' user demographics, enabling the potential for business intelligence across mobile operators.

Organization. We introduce the mobile networks in Section 2. We report the social strategies that are discovered from human mobile communication networks in Section 3 and propose a null model to validate their statistical significance in Section 4. We formalize the demographic prediction problems in Section 5. We present our solutions for inferring user demographics in Section 6. Prediction results are demonstrated in Section 7. Finally, we summarize the related work in Section 8 and conclude this work in Section 9.

2. MOBILE NETWORK DATA WITH DEMOGRAPHICS

The dataset used in this article is extracted from a collection of more than 1 billion (1,000,229,603) phone call and text messaging events from an anonymous country [Gonzalez et al. 2008; Ercsey-Ravasz et al. 2012; Dong et al. 2014, 2015], which spans from August 2008 to September 2008. Notice that we only consider the communications that were made between users within this country. We construct two undirected and weighted mobile communication networks from the de-identified and anonymous data: a phone call network (referred to as CALL) and a text 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 or text messages (v_i called v_j and also v_j called v_i for at least one time during the observation period). The strength w_{ij} of the edge is defined as the number of communications between v_i and v_j per month. Then, we extract the largest connected component from each network as our

Table I. The Statistics of Mobile Networks

networks	#nodes	#edges
CALL network with user demographics ($CALL_d$)	7,440,123	32,445,941
SMS network with user demographics (SMS_d)	4,505,958	10,913,601
Reciprocal CALL network ($CALL_r$)	4,927,095	16,674,164
Reciprocal SMS network (SMS_r)	3,104,853	7,602,830
Largest Connected Component of reciprocal CALL network ($CALL_{rl}$)	4,295,638	15,787,538
Largest Connected Component of reciprocal SMS network (SMS_{rl})	2,369,078	6,660,172
$CALL_{rl}$ with user demographics ($CALL_{rld}$ / CALL)	4,292,227	15,765,196
SMS_{rl} with user demographics (SMS_{rld} / SMS)	2,064,898	5,689,696

Table II. The Distribution of Mobile Users' Gender and Age

	Young (18–24)	Young-Adult (25–34)	Middle-Age (35–49)	Senior (> 49)
female	4.77%	13.52%	16.16%	10.84%
male	5.23%	15.96%	19.73%	13.66%

experimental networks. We also generate the networks by filtering out the nodes that don't have demographic information. Table I lists the order and size of the resultant CALL and SMS networks. The data does not contain any communication content.

In this dataset, around 45% of the users are female and 55% are male. We compare the demographic population distribution of mobile users with the 2008 world population distribution, which was released by the U.S. Census Bureau international database.[4] We find 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, while teenagers (under 18 years old) and the elderly (aged 80 or over) are underrepresented. 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 denote the female and male users, respectively. Following [Hu et al. 2007; Bi et al. 2013], we also split users into four groups according to their ages: Young (18–24), Young-Adult (25–34), Middle-Age (35–49), and Senior (>49). The distribution of users' gender and age is listed in Table II.

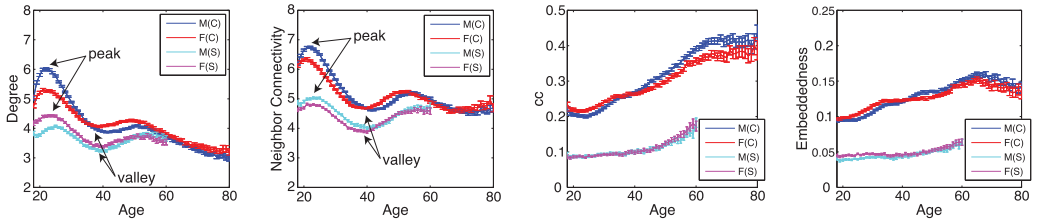
3. SOCIAL STRATEGIES IN MOBILE COMMUNICATION

Social strategies are used by people to meet their social needs, which is, together with being, having, and doing, considered among the basic human needs [Max-Neef et al. 1992]. Meeting with new people and strengthening existing relationships belong to the category of social needs. The mobile communication data provides rich information for discovering and characterizing human social strategies by which people build and maintain social connections. Previous studies [Palchykov et al. 2012] show that the strategies by which social needs are satisfied change over time, although the needs are constant across one's lifetime. In this section, we show how people communicate with each other across their respective lifetime. Specifically, we investigate the interplay of human communication interactions and demographic characteristics in the perspective of micro-level network structures, including ego networks, social ties, and social triads. We also use a null model to simulate the observations by randomly shuffling users' demographic profiles and report the statistical significance of the results in Section 4.

3.1. Social Strategies on Ego Networks

An ego network of one person is defined by viewing himself or herself as the central node and his or her one-degree friends as surrounding nodes [Freeman 1982]. Clearly, one's

[4]http://www.census.gov/idb/worldpopinfo.html.



(a) Degree Centrality (b) Neighbor Connectivity (c) Triadic Closure (d) Embeddedness

Fig. 4. Correlations between demographics and network characteristics. C means attributes observed from the CALL network and S means the SMS network. F denotes female and M denotes male.

ego network is a sub-network of the original network. Figure 1 presents an illustrative example of the evolution of one's ego network. We first examine the characteristics of the central node (ego) and then the distributions of this ego's friends (ego network) with respect to their demographic profiles.

Ego. We present a basic correlation analysis between network characteristics and user demographics to examine how an individual's gender and age influence her or his ego social networks. In particular, we consider the following network metrics:

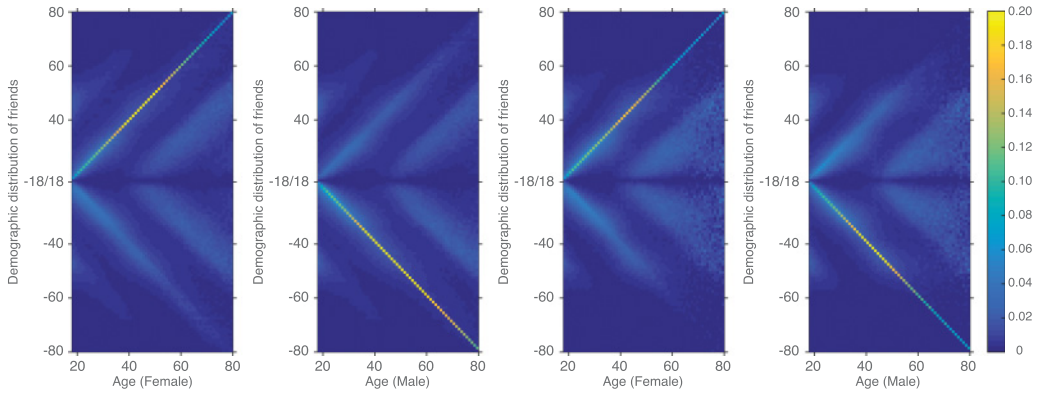
—*Degree Centrality*: the number of edges incident upon a node in the network;
—*Neighbor Connectivity*: the average degree of neighbors of a specific user;
—*Triadic Closure*: the local clustering coefficient (cc) of each user;
—*Embeddedness*: the degree that people are enmeshed in networks [Granovetter 1985]. More accurately, a user u's embeddedness is defined as $\frac{1}{|N_u|}\sum_{v\in N_u}\frac{|N_u\cap N_v|}{|N_u\cup N_v|}$, where N_u is the neighbors of u.

Figure 4 plots the correlations between the four network metrics and the users' age. From sub-figures 4(a)–4(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 4(c)–4(d), we see that both triadic closure and embeddedness increase when users become older. Similar to the first two metrics, there is also a reverse 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.

Ego Networks. With the ego network of each user, we study the demographic homophily on both gender and age. The principle of homophily suggests that people tend to be connected with those who are similar to them [Lazarsfeld and Merton 1954]. It has been extensively studied and verified in both online social networks [Leskovec and Horvitz 2008; Lou et al. 2013] and mobile networks [Dong et al. 2013; Kovanen et al. 2013].

Figure 5 shows friends' demographic distribution for female and male users of different age in the CALL and SMS networks. The X-axis represents a central user' age from 18 to 80 years old and the Y-axis represents the demographic distribution of that central user' friends, in which positive numbers denote female friends' age and negative numbers denote male friends'. The spectrum color, which extends from dark blue (low) to yellow (high), represents the probability of one's friends belonging to the corresponding age (Y-axis) and gender (positive or negative). Interestingly, there exist highlighted



(a) F ego net in CALL (b) M ego net in CALL (c) F ego net in SMS (d) M ego net in SMS

Fig. 5. Friends' demographic distribution in ego networks. X-axis: (a) the age of a female ego in CALL; (b) the age of male ego in CALL; (c) the age of a female ego in SMS; (d) the age of a male ego in SMS. Y-axis: the age of the ego's friends (positive, female friends; negative, male friends). The spectrum color represents the friends' demographic distribution.

diagonal lines in each sub-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 in the CALL network, and 40 to 50 years old in the SMS network. Simultaneously, the highlighted diagonals appear in the same gender range in both networks, that is, females appear in the positive Y range (F) in Figures 5(a) and 5(c) and males in the negative Y range (M) in Figures 5(b) and 5(d), which shows the existence of a high degree of gender homophily in mobile phone behavior.

Social Strategies. From a sociological perspective, the results in Figures 4 and 5 can be also explained by different social strategies that people use to maintain their social connections. First, younger people (who have higher degree centrality) are very active in broadening their social circles, while older people (who have higher triadic closure centrality *cc*) tend to keep smaller but more stable connections. This finding from large-scale networks coincides with previous survey studies that older people have lower rates of contact than young people [Marsden 1987; Cornwell 2011]. Second, people tend to communicate with others of similar gender and age, that is, gender and age homophily in mobile communications. Third, young people put increasing focus on the same generation and decreasing focus on the older generation, and the middle-age people devote more attention on the younger generation even at the cost of age homophily.

3.2. Social Strategies on Interpersonal Ties

An interpersonal tie is viewed as the connection between two people, and its strength represents the extent of closeness of social contacts [Onnela et al. 2007], such as strong ties [Krackhardt 1992; Shi et al. 2007] and weak ties [Granovetter 1973]. In mobile communication networks, tie strength is defined as the frequency of communications between each pair of users [Onnela et al. 2007; Palchykov et al. 2012].

In Figure 6, we use heat maps to visualize the communication frequencies for different demographics. Figures 6(a) and 6(e) report the average number of calls/messages per month between two users. Figures 6(b)–6(d) and 6(f)–6(h) detail the analysis by reporting the average numbers of calls/messages between two male users, two female users, and one male and one female, respectively. Again, we discover highlighted diagonal lines in Figures 6(a)–6(c), which correspond to the gender and age homophily. We

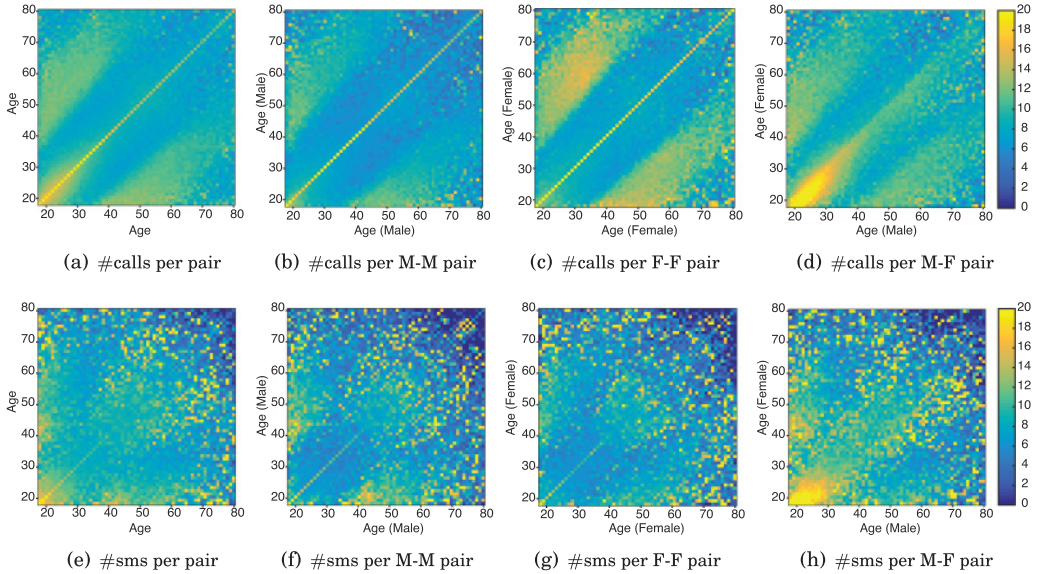


Fig. 6. Strength of social ties in the CALL and SMS networks. XY-axis: age of users with specific gender. The spectrum color represents the number of phone calls (text messages) per month. (a), (b), (c), (e), (f), and (g) are symmetric.

also notice that there are highlighted areas corresponding to cross-generation communications. In Figure 6(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 potentially corresponds to phone calls between parents and children, managers and subordinates, and advisors and advisees, and so on. These two discoveries can also be observed in Figures 6(e)–6(g) in the SMS network but not as obvious as in the CALL network.

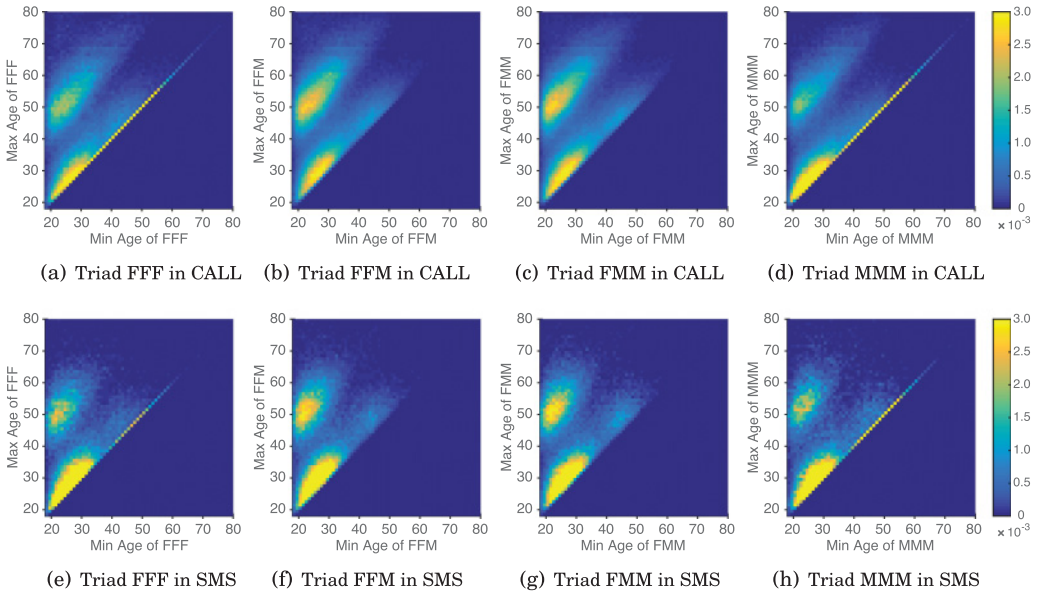
In addition, we observe that the cross-generation phone call communications between female users seem to be much more frequent than those between male users (cf. Figures 6(b) and 6(c)). Moreover, from Figures 6(d) and 6(h), we observe a highlighted yellow area between people aged 18–34 years old, which means that cross-gender communications are more frequent than those between users of the same gender. A similar observation has also been reported in the MSN network [Leskovec and Horvitz 2008].

Social Strategies. The social strategies unveiled from Figure 6 can be summarized as follows. First, frequent cross-generation interactions are maintained to bridge age gaps in both phone call and text messaging channels. Second, opposite-gender communication interactions among younger people are much more frequent than those between same-gender users. However, when people reach the 35 years of age, reversely, same-gender interactions are more frequent than those between opposite-gender users.

3.3. Social Strategies on Triads

A triad is one of the simplest groupings of individuals in social networks [Easley and Kleinberg 2010]. Three individuals form a triad if and only if each pair of them are friends. Herein, we investigate how male and female users maintain their social triadic relationships across their lifetime.

In Figure 7, the heat map visualizes the distribution of the minimum age (X-axis) and maximum age (Y-axis) of three users in a closed social triad structure. Figures 7(a)/7(e)



(a) Triad FFF in CALL (b) Triad FFM in CALL (c) Triad FMM in CALL (d) Triad MMM in CALL

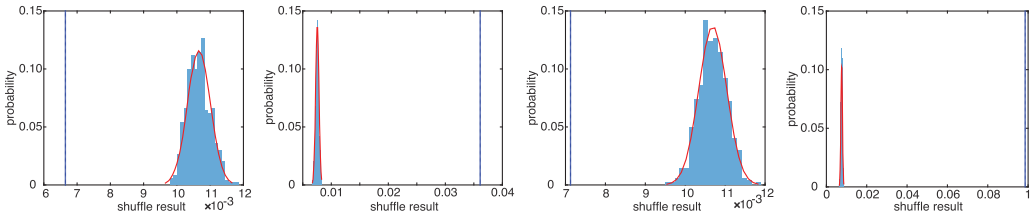
(e) Triad FFF in SMS (f) Triad FFM in SMS (g) Triad FMM in SMS (h) Triad MMM in SMS

Fig. 7. Social triad distribution in the CALL and SMS networks. X-axis: the minimum age of three users in a triad. Y-axis: the maximum age of three users. The spectrum color represents the distributions.

and 7(d)/7(h) show the same-gender triads: "FFF" and "MMM," and Figures 7(b)/7(f) and 7(c)/7(g) present the age distribution for users in opposite-gender triads: "FFM" and "FMM." Clearly, the triadic relationships are observed in all four kinds of gender-triads (i.e., "FFF," "MMM," "FFM," and "FMM") among young people by highlighted yellow areas at the left-bottom corners of each sub-figure. When entering middle-age ($>$ 35 years old), people only maintain the same-gender triadic relationships in mobile communications, which is revealed by the yellow diagonal lines in Figures 7(a)/7(e) and 7(d)/7(h). The opposite-gender triadic relationships vanish when people pass 35 years old observed in Figures 7(b)/7(f) and 7(c)/7(g). The instability of opposite-gender triadic relationships and the persistence of same-gender triadic relationships across one's lifetime are novel discoveries and reveal the dynamics of human social strategies across their lifespan.

Furthermore, the cross-generation triadic relationships are found in the left-middle light areas in each sub-figure. These left-middle light areas are almost isolated with other highlighted areas in each sub-figure, then we are curious about the distribution of the middle age of three users in one social triad. Our further study shows that the middle age in these triads are similar to either the minimum age (60%) or the maximum age (40%) among them, which means there are around 60% cross-generation communication triads are composed of two youths and one middle-age people, for example, 25-25-45 years old, respectively, in a triad, the remaining 40% are two middle-age and one young people, for example, 20-40-40 years old, and no triads like 20-30-40 years old are observed in this nationwide communication networks.

Social Strategies. The dynamics of gender differences on social decisions indicate the evolution of social strategies used by people to meet their social needs. People expand both the same-gender and opposite-gender social circles during the dating active period. However, people's attention to opposite-gender groups quickly disappears after entering into middle-age, and the insistence and social investment on same-gender social groups lasts for a lifetime.



(a) F-20 ego, F-60 friend (b) F-20 ego, M-20 friend (c) M-20 ego, F-60 friend (d) M-20 eog, M-20 friend

Fig. 8. Illustrative cases of shuffled results and true value in CALL. We select two points from Figure 5(a) and two from Figure 5(b) to show the shuffled results. Blue line represents the true values from the data (Figure 5); blue histograms plot the shuffled results; red line represents the fitted normal density curve.

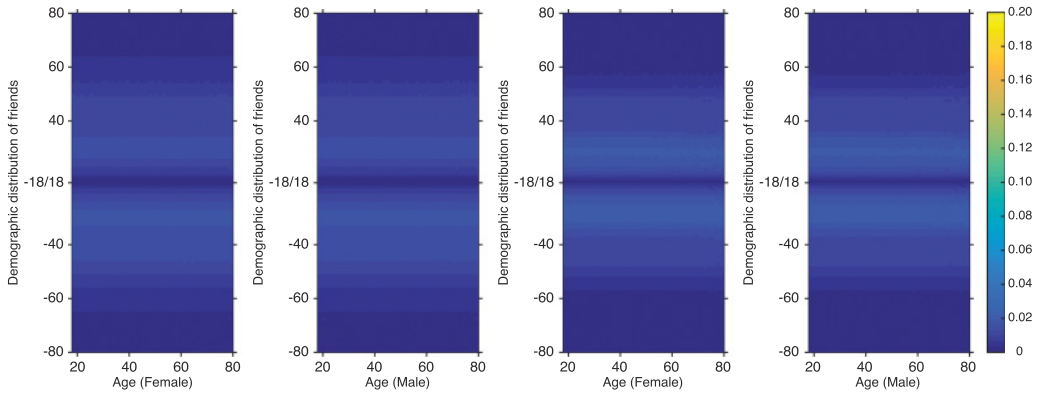
3.4. Summary

According to our comprehensive analysis on the interplay of demographic profiles and mobile communications, we unveil striking gender- and age-based networking differences, which reflect the dynamic social strategies that evolve as a function of the balance between different social needs across lifespans. In summary, we provide the following social phenomena relating to mobile communications:

—Figure 4 demonstrates that younger people are active in broadening their social connections, while older people have the tendency to maintain smaller but more closed connections.
—Figure 5 confirms demographic homophily; that being said, people tend to interact with others with similar gender and age in both phone call and text messaging channels.
—Figure 6 shows that cross-gender social relationships exhibit more frequent communications than the same-gender ones, and the cross-generation interactions are maintained to pass the torch of family, workforce, and human knowledge from generation to generation in social society.
—Figure 7 unveils that people tend to expand their social connections with females and males alike during younger and more dating-active period and put more social investment on maintaining same-gender social groups after entering into middle-age.
—In addition, the gap between the younger and older people in text-messaging channel (e.g., Figure 7(e)) is larger than that in phone calls (Figure 7(a)), while the difference between males and females (e.g., Figure 6(b) versus 6(c)) in phone-call channel are more significant than that in messaging communications (Figures 6(f) versus 6(g)).

4. THE NULL MODEL IN NETWORKS

We validate the statistical significance of the social strategies observed in the CALL and SMS networks in Section 3 by using a null model. The idea of the statistical test is to compare the demographic-based observations x from empirical data to those $\{\tilde{x}\}$ provided by the null model, wherein the demographic profiles of users are randomly shuffled [Kovanen et al. 2013; Dong et al. 2015b]. On the null model, we first randomly assign the demographic profiles of the users on the underlying communication networks and then observe the social strategies that are derived from the randomly allocated user demographics. We simulate the random process 10,000 times and get the mean $\mu(\tilde{x})$ and standard deviation $\sigma(\tilde{x})$ of the observations $\{\tilde{x}\}$ on the null model. For example, we use four data points selected from Figure 5 to illustrate the statistical test, that is, two points ($X = 20$, $Y = 60$) and ($X = 20$, $Y = -20$) from Figures 5(a) and 5(b), respectively. Figure 8 reports the histograms of shuffled results $\{\tilde{x}\}$ of the four points.



(a) F ego net in CALL (b) M ego net in CALL (c) F ego net in SMS (d) M ego net in SMS

Fig. 9. Friends' demographic distribution (shuffle). X-axis: (a) the age of a female ego in CALL; (b) the age of male ego in CALL; (c) the age of a female ego in SMS; (d) the age of a male ego in SMS. Y-axis: the age of the ego's friends (positive: female friends, negative: male friends). The spectrum color represents the friends' demographic distribution.

First, it is clear that the true values x (blue lines) observed from Figure 5 largely fall out of the shuffled distributions (histogram plots). Further, we can see that the shuffled distributions are close to the fitted normal distributions (red lines). Accordingly, we use z-score to examine the numerical gap between the empirical data x and the randomly shuffled results $\{\tilde{x}\}$ on the null model [Sprinthall 2011]:

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

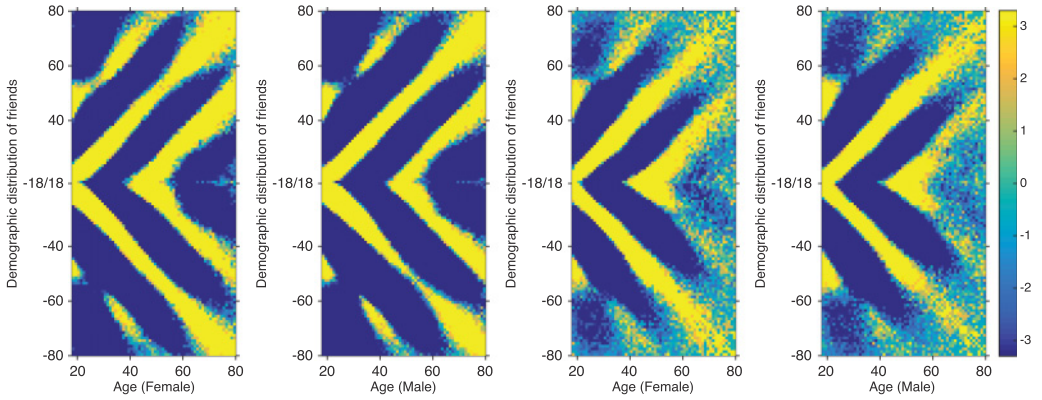
A z-score of 0 indicates that there exists no difference between empirical data and the null model. A positive (negative) z-score represents that the empirical data is over-(under-) represented than expected by chance. In specific, $|z(x)| \geq 3.3$ (corresponding to p-value ≤ 0.001) indicates that the observation from the empirical data is extremely statistically significant.

The statistical tests are conducted for all the social strategies observed on ego networks, social ties, and social triads in mobile phone call and text messaging behavior. We associate each observation figure of the social strategies presented in Section 3 with the shuffled results and z-score plots. Specifically, the results on ego networks are shown in Figures 9 and 10. The shuffled results and z-scores on social ties in the CALL and SMS networks can be found in Figures 11 and 12, respectively. Figures 13 and 14 present the values of shuffled means and z-scores of the social strategies on social triad observed in both the CALL and SMS networks, respectively.

From the figures, we can see that there are large differences between the heatmaps of the observations (data) and those of the means of 10,000 simulating results (shuffle). Moreover, we find that the color of the areas we are interested in from each z-score plot tells that $|z(x)| \geq 3.3$. That being said, each social strategy we observed in the mobile communication networks is (extremely) statistically significant.

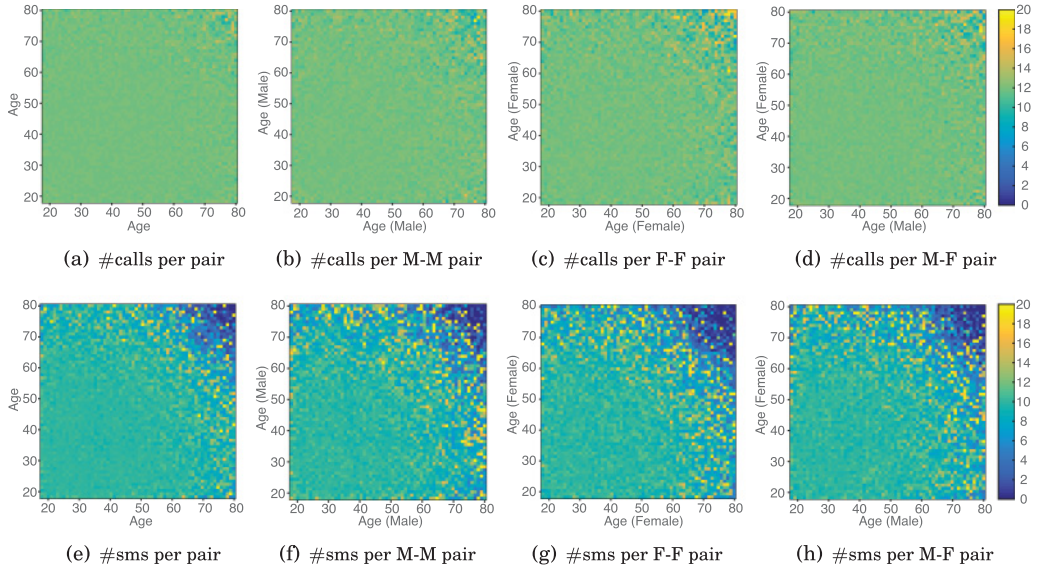
5. DEMOGRAPHIC PREDICTION PROBLEMS

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, that is, gender $y_i \in Y$ and age $z_i \in Z$. We further define an attribute matrix $\mathbf{X}$, where each row



(a) F ego net in CALL (b) M ego net in CALL (c) F ego net in SMS (d) M ego net in SMS

Fig. 10. Friends' demographic distribution (z-score). X-axis: (a) the age of a female ego in CALL; (b) the age of male ego in CALL; (c) the age of a female ego in SMS; (d) the age of a male ego in SMS. Y-axis: the age of the ego's 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

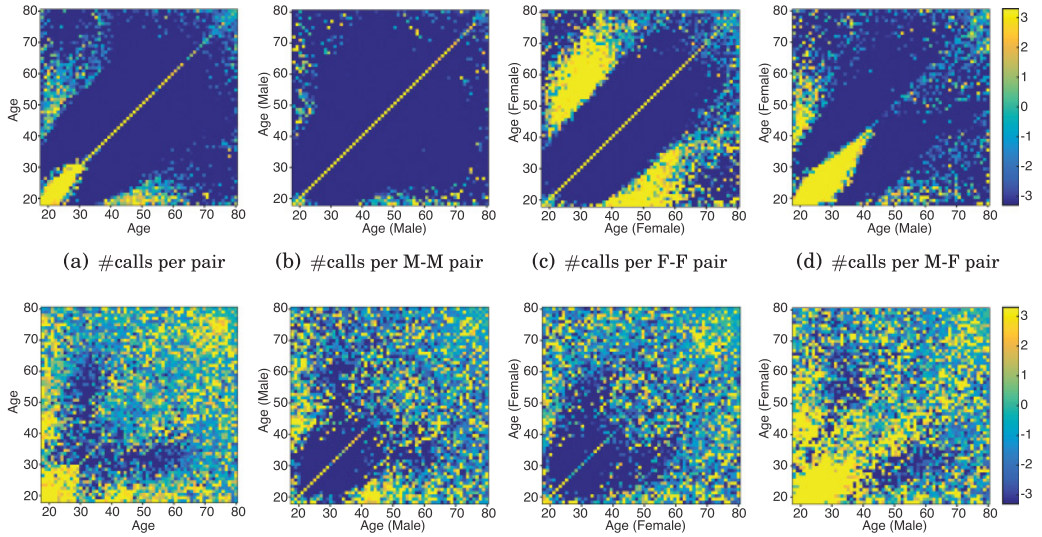
(e) #sms per pair (f) #sms per M-M pair (g) #sms per F-F pair (h) #sms per M-F pair

Fig. 11. Strength of social ties in the CALL and SMS networks (shuffle). XY-axis: age of users with specific gender. The spectrum color represents the number of calls (messages) per month. (a), (b), (c), (e), (f), and (g) are symmetric.

$\mathbf{x}_i$ represents an $|\mathbf{x}_i|$ dimensional feature vector for user v_i. Given this, we formalize our problem as follows.

PROBLEM 1. ***Demographic Prediction:*** *Given a partially labeled network* $G = (V^L, V^U, \boldsymbol{E}, Y^L, Z^L)$ *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*

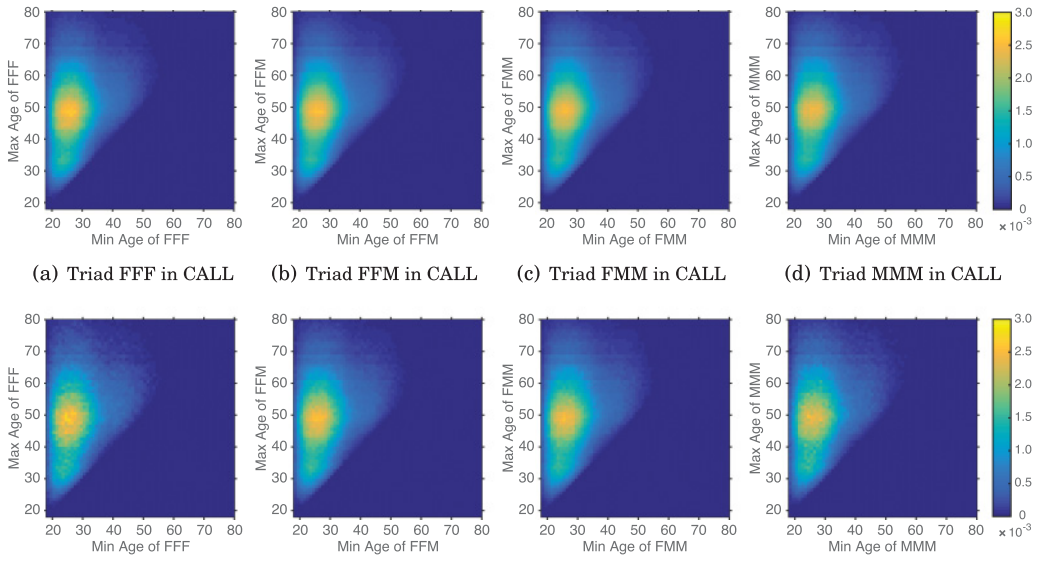
$$f : G = (V^L, V^U, E, Y^L, Z^L), \mathbf{X} \rightarrow (Y^U, Z^U)$$



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

(e) #sms per pair (f) #sms per M-M pair (g) #sms per F-F pair (h) #sms per M-F pair

Fig. 12. Strength of social ties in the CALL and SMS networks (z-score). XY-axis: age of users with specific gender. The spectrum color represents the number of calls (messages) per month. (a), (b), (c), (e), (f), and (g) are symmetric.



(a) Triad FFF in CALL (b) Triad FFM in CALL (c) Triad FMM in CALL (d) Triad MMM in CALL

(e) Triad FFF in SMS (f) Triad FFM in SMS (g) Triad FMM in SMS (h) Triad MMM in SMS

Fig. 13. Social triad distribution in the CALL and SMS networks (shuffle). X-axis: minimum age of three users in a triad. Y-axis: maximum age of three users. The spectrum color represents the distributions.

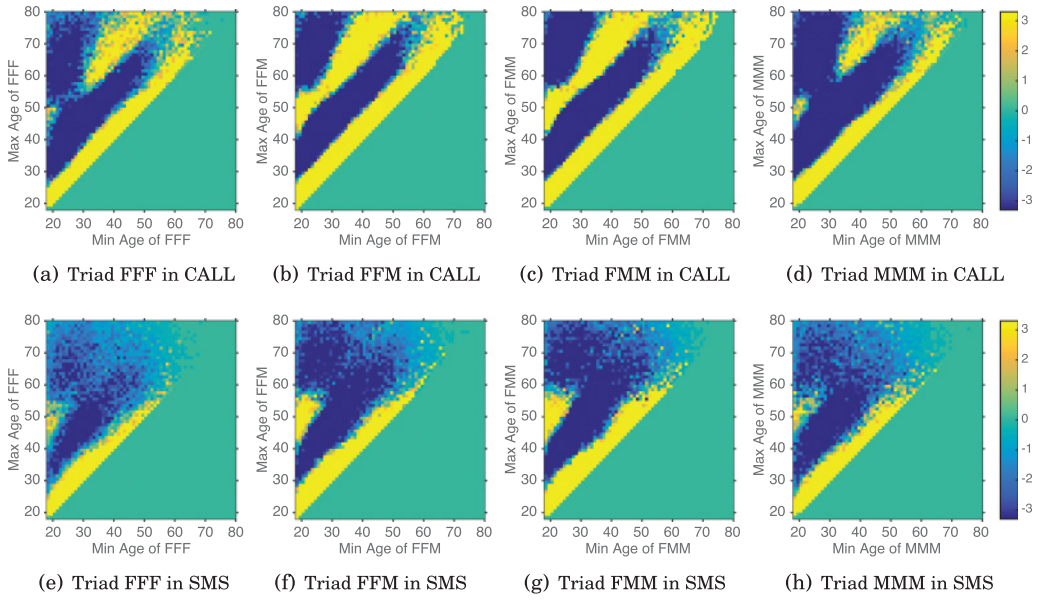


Fig. 14. Social triad distribution in the CALL and SMS networks (z-score). X-axis: minimum age of three users in a triad. Y-axis: maximum age of three users. The spectrum color represents the distributions.

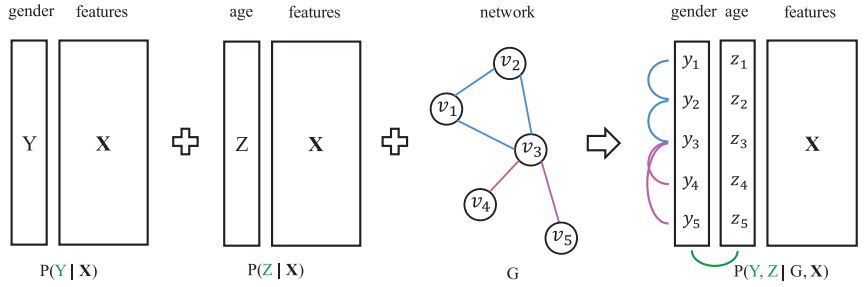


Fig. 15. An illustration of the proposed demographic prediction problem. In addition to model the correlations between labels (Y or Z) and features ($\mathbf{X}$) of each node, we propose to further model the structural correlations among different nodes (G) as well as the interrelations between one node's two labels, that is, Y and Z.

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

Different from previous work on demographic prediction [Bi et al. 2013; Hu et al. 2007], where users' gender and age are inferred by modeling $P(Y|\mathbf{X})$ and $P(Z|\mathbf{X})$ separately (see Figure 15), our problem here is to model $P(Y, Z|G, \mathbf{X})$ for the joint inference of users' gender and age. Specifically, we leverage not only the correlations between $\mathbf{X}$ and Y/Z but also the structural correlations among nodes and interrelations between gender Y and age Z. The motivation here comes from the fact that there exist strong network effects and demographic interrelations in human communication behavior, which was demonstrated in Section 3. For example, a 20-year-old female's behavior is distinct from not only a 20-year-old male's, but also from a 50-year-old female's.

In addition, there are usually multiple mobile operators in telecommunication market—for example, the two mobile operators in Figure 3. A mobile operator O_1

(e.g., AT&T) could have the communication records of its users and also the communication logs between its users and users of another operator O_2 (e.g., Verizon) [Dong et al. 2015]. It would be very useful for the operator O_1 to have the demographic profiles of users of its competitor O_2 for business intelligence and precision marketing, such as acquiring new users from and preventing customer churning to competitors.

To solve this problem, we define the concept of coupled networks and formulate the problem of coupled network demographic prediction across multiple operators in mobile communication.

Definition 5.1. **Coupled Networks:** Given a source network $G^S = (V^S, E^S)$ and a target network $G^T = (V^T, E^T)$, they compose coupled networks if there exists a cross link e_{ij} with one node $v_i \in V^S$ and the other node $v_j \in V^T$. The cross network $G^C = (V^C, E^C)$ is a bipartite network containing all the cross links in the coupled networks.

Figure 3 shows a typical example of coupled networks with the left network of mobile operator O_1 as the source network G^S and the right network of another mobile operator O_2 as the target network G^T. The links between these two networks represent the communications between users belonging to these two different mobile operators, which, with their linked nodes in G^S and G^T, constitute the cross network G^C.

PROBLEM 2. ***Coupled Network Demographic Prediction:*** *Given the source network G^S with its users' demographic profiles Y^S, Z^S and the cross network G^C in coupled networks $G = (G^S, G^T, G^C)$, the task is to find a predictive function:*

$$f : G^S = (V^S, E^S, Y^S, Z^S), G^C = (V^S, V^T, E^C) \rightarrow (Y^T, Z^T),$$

where Y^T and Z^T are the set of demographic labels—gender and age—of users V^T in the target network G^T.

The difference between the coupled network demographic prediction and Problem 1 lies in the cold start of network structures between target users in Problem 2. For example, in Figure 3, the triangle structures (v_6, v_7, v_8), (v_1, v_6, v_7) cannot be observed by the operator O_1, making it impossible to leverage the correlations built upon these structures in the prediction task. The real-world and yet challenging setting of the coupled network demographic prediction can be directly applied by a mobile operator to infer the demographic profiles of its competitors' users, facilitating the acquirement of new users from competitor operators.

We treat users' gender as a **binary** random variable, that is, Female or Male, and users' age as a **four-class** variable by splitting users' age into the four groups mentioned above [Hu et al. 2007; Bi et al. 2013], that is, Young (18–24), Young-Adult (25–34), Middle-Age (35–49), and Senior (> 49). The distribution of users' gender and age is listed in Table II.

6. THE WHOAMI FRAMEWORK

Leveraging the insights gleaned from our network analysis in Section 3, we develop a unified model to capture not only the correlations between users' communication behaviors and demographic profiles but also the interrelations among users' different demographic attributes. In our previous work [Dong et al. 2014], the proposed *DFG* (Double Label Factor Graph) model is only capable of handling the interrelations between two dependent variables (e.g., gender Y and age Z). In this extension, we generalize the WhoAmI method to a Multiple Label Factor Graph Model (MFG). The MFG is general to model the interrelations among multiple (more than two) dependent variables. To illustrate the way that MFG captures the interrelations between multiple

(> 2) labels, we assume that in addition to one's gender Y and age Z, each user is also associated with another demographic attribute S (e.g., income). However, notice that in the mobile data only two demographic attributes—gender and age—are available. Therefore, in Section 7, we use the proposed approach to predict these two attributes.

To infer users' demographic attributes in coupled networks, we propose a variant of the Multiple Label Factor Graph—CoupledMFG—that is able to address the unique challenges presented in this task. To handle large-scale networks, we further develop a distributed learning algorithm.

6.1. Multiple Label Factor Graph

We define an objective function by maximizing the conditional probability of users' gender Y, age Z, and S given their corresponding attributes $\mathbf{X}$ and the input network structure G, that is, $P_\theta(Y, Z, S|G, \mathbf{X})$. The factor graph [Kschischang et al. 2001] provides a way to factorize the "global" probability as a product of "local" factor functions, which makes the maximization simple, that is,

$$P(Y, Z, S|G, \mathbf{X}) = \frac{P(\mathbf{X}, G|Y, Z, S)P(Y, Z, S)}{P(\mathbf{X}, G)} \propto P(Y, Z, S|G)P(\mathbf{X}|Y, Z, S) \tag{1}$$
$$\propto \prod_{v_i \in V} P(\mathbf{x}_i|y_i, z_i, s_i) \prod_{c \in G} P(Y_c, Z_c, S_c),$$

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

Our proposed model consists of three kinds of factor functions. The first one is an attribute factor $f(y_i, z_i, s_i, \mathbf{x}_i)$ for capturing correlations between users' demographics and communication attributes. The second one is a dyadic factor $g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)$ for modeling correlations between users' demographics and their direct relationships in ego networks, where Y_c in Equation (1) is represented as $\mathbf{y}_e$ (y_i and y_j), Z_c is denoted by $\mathbf{z}_e$ (z_i and z_j), and S_c by $\mathbf{s}_e$ (s_i and s_j) iff $e_{ij} \in E$. The third one is a triadic factor $h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)$ for correlating users' demographics and triadic relationships in their ego networks. Similarly, $\mathbf{y}_c$ refers to y_i, y_j, y_k, while $\mathbf{z}_c$ refers to z_i, z_j, z_k, and $\mathbf{s}_c$ is s_i, s_j, s_k when three users v_i, v_j, v_k form a closed triangle structure c_{ijk}, that is, $e_{ij}, e_{ik}, e_{jk} \in E$.

Therefore, the joint distribution can be further factorized as

$$P(Y, Z, S|G, \mathbf{X}) = \prod_{v_i \in V} f(y_i, z_i, s_i, \mathbf{x}_i) \times \prod_{e_{ij} \in E} [g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)] \times \prod_{c_{ijk} \in G} [h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)]. \tag{2}$$

Figure 16 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 Factors. We design the factor $f(y_i, z_i, s_i, \mathbf{x}_i)$ 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:

$$f(y_i, z_i, s_i, \mathbf{x}_i) = \frac{1}{W_v} \exp\{\alpha_{y_i z_i s_i} \cdot \mathbf{x}_i\}, \tag{3}$$

where α is one parameter of the proposed model, and W_v is a normalization term. For each (y_i, z_i, s_i), $\alpha_{y_i z_i s_i}$ is an $|\mathbf{x}|$-length vector, where the kth dimension indicates how x_{ik} distributes over (y_i, z_i, s_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 demographic

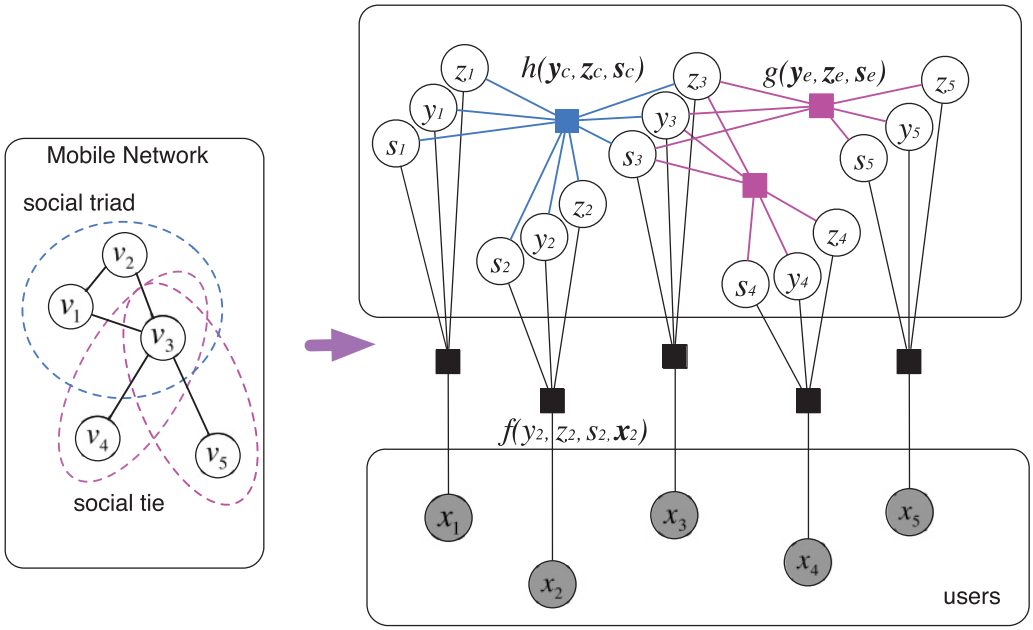


Fig. 16. An illustration of the proposed model. $y_\cdot$, $z_\cdot$, and $s_\cdot$ indicate the gender, age, and newly added label of the user v_i., x_i denotes communication attributes of the user v_i extracted from the mobile network G. $f(y_i, z_i, s_i, \mathbf{x}_i)$, $g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)$, and $h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)$, respectively, represent attribute factor, dyadic factor, and triadic factor in the proposed model.

profiles have different network properties shown in Figure 4. 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 jointly distribute over multiple dependent variables.

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

$$g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e) = \begin{cases} \frac{1}{W_{e_1}} \exp\{\beta_1 \cdot g'_1(y_i, y_j)\} \\ \frac{1}{W_{e_2}} \exp\{\beta_2 \cdot g'_2(y_i, z_i)\} \\ \cdots \\ \frac{1}{W_{e_{14}}} \exp\{\beta_{14} \cdot g'_{14}(z_j, s_i)\} \\ \frac{1}{W_{e_{15}}} \exp\{\beta_{15} \cdot g'_{15}(s_i, s_j)\} \end{cases}, \tag{4}$$

where β_p is the model parameter 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 $C_6^2 = 15$ different combinations of each pair of demographic variables from $(y_i, y_j, z_i, z_j, s_i, s_j)$. The intuition behind this is that v_i's friends' demographics distribute differently by varying either v_i's own age or gender or income, as Figure 5 suggests.

Triadic Factors. We finally define the triadic factor $h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)$ 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 closed triangle structure in G. More specifically, we have

$$h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c) = \begin{cases} \frac{1}{W_{c_1}} \exp\{\gamma_1 \cdot h'_1(y_i, y_j, y_k)\} \\ \frac{1}{W_{c_2}} \exp\{\gamma_2 \cdot h'_2(y_i, y_j, z_i)\} \\ \cdots \\ \frac{1}{W_{c_{83}}} \exp\{\gamma_{83} \cdot h'_{83}(s_i, s_j, z_k)\} \\ \frac{1}{W_{c_{84}}} \exp\{\gamma_{84} \cdot h'_{84}(s_i, s_j, s_k)\} \end{cases}, \tag{5}$$

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 C_9^3 different kinds of three-variable enumerations from $(y_i, y_j, y_k, z_i, z_j, z_k, s_i, s_j, s_k)$. We use these triadic factors to model the distributions of user demographics within a closed social triangle (see details in Figure 7).

Finally, incorporating Equations (3), (4), and (5) into Equation (2), we define the objective function as the log-likelihood of the proposed model as

$$\mathcal{O}(\alpha, \beta, \gamma) = \sum_{v_i \in V} \alpha_{y_i z_i s_i} \mathbf{x}_i + \sum_{e_{ij} \in E} \sum_{p=1}^{15} \beta_p g'_p(\cdot) + \sum_{c_{ijk} \in G} \sum_{q=1}^{84} \gamma_q h'_q(\cdot) - \log W, \tag{6}$$

where $W = W_v W_e W_c$ is the global normalization term, $W_e = \prod_{e_p=1}^{15} W_{e_p}$, and $W_c = \prod_{c_q=1}^{84} W_{c_q}$.

The technical novelty of the proposed model is that it considers different types of labels in a unified framework, which differentiates our model from traditional classification models. By considering three types of labels in this special case, the main advantage is that our model can characterize the interrelations between different demographic labels and the structural correlations between different users as well as correlations between labels and features.

6.2. Feature Definition

Given a network with labeled and unlabeled users, the goal is to infer unlabeled users' demographic information, which is in accordance with the real-world application scenarios. There are two types of features designed in our experiments, namely *nonstructural attribute features* and *structural features*. Specifically, given an ego network with one central user v and her/his direct friends, we extract three kinds of attribute features for this central user v as follows:

Individual attributes are extracted based on the network topological properties discussed in Section 3.1. It includes the degree, neighbor connectivity, clustering coefficient, embeddedness, and weighted degree (#calls or #messages) of each node.

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 setting, 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.

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, that is, "*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*."

Table III. Definition of Nonstructural Attribute Features Modeled by the Attribute Factor in Equation (3)

Attribute Type	Name	Description
Individual attributes	degree	number of contacts
	neighbor connectivity	average degree of neighbors
	clustering coefficient	local clustering coefficient
	embeddedness	the degree that people are embedded in networks
	weighted degree	number of communications (#calls or #messages)
Friend attributes	#female-friends	number of female contacts
	#male-friends	number of male contacts
	#young-friends	number of young contacts
	#young-adult-friends	number of young-adult contacts
	#middle-age-friends	number of middle-age contacts
	#senior-friends	number of senior contacts
Circle attributes A: young B: young-adult C: middle-age D: senior	#v-FF-triangles	number of *FF*-v triangles in v's ego network
	#v-FM-triangles	number of *FM*-v triangles in v's ego network
	#v-MM-triangles	number of *MM*-v triangles in v's ego network
	#v-AA-triangles	number of *AA*-v triangles in v's ego network
	#v-AB-triangles	number of *AB*-v triangles in v's ego network
	#v-AC-triangles	number of *AC*-v triangles in v's ego network
	#v-AD-triangles	number of *AD*-v triangles in v's ego network
	#v-BB-triangles	number of *BB*-v triangles in v's ego network
	#v-BC-triangles	number of *BC*-v triangles in v's ego network
	#v-BD-triangles	number of *BD*-v triangles in v's ego network
	#v-CC-triangles	number of *CC*-v triangles in v's ego network
	#v-CD-triangles	number of *CD*-v triangles in v's ego network
	#v-DD-triangles	number of *DD*-v triangles in v's ego network

Table III lists 24 nonstructural attribute features used in our models. Notice that friend and circle attributes can only be extracted from v's labeled friends. These three types of attribute features—individual, friend, and circle attributes—are captured by the attribute factor in our MFG model (cf. Equation (3)).

In addition, the structural features, captured by the dyadic factor (cf. Equation (4)) and triadic factor (cf. Equation (5)), are designed to model the demographic distributions over edges and triangles with both labeled and unlabeled users, which forms one of the advantages of the proposed factor graph-based model. Together with nonstructural friend attributes, structural features covered by dyadic factors form *friend features*. Similarly, *circle features* are composed of nonstructural circle attributes and triadic structural features.

6.3. WhoAmI Learning and Inference

The goal of learning the WhoAmI method 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 Equation (6) given by the training set, that is, $\theta^\star = \arg\max \mathcal{O}(\theta)$.

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 Equation (6).

$$\frac{\partial \mathcal{O}(\theta)}{\partial \alpha} = \boldsymbol{E}\left[\sum_{v_i \in V} \mathbf{x}_i\right] - \boldsymbol{E}_{P_\alpha(Y,Z,S|\mathbf{X})}\left[\sum_{v_i \in V} \mathbf{x}_i\right],$$
$$\frac{\partial \mathcal{O}(\theta)}{\partial \beta} = \boldsymbol{E}\left[\sum_{e_{ij} \in E} g'(\cdot)\right] - \boldsymbol{E}_{P_\beta(Y,Z,S|\mathbf{X},G)}\left[\sum_{e_{ij} \in E} g'(\cdot)\right], \tag{7}$$
$$\frac{\partial \mathcal{O}(\theta)}{\partial \gamma} = \boldsymbol{E}\left[\sum_{c_{ijk} \in G} h'(\cdot)\right] - \boldsymbol{E}_{P_\gamma(Y,Z,S|\mathbf{X},G)}\left[\sum_{c_{ijk} \in G} h'(\cdot)\right],$$

where in the first equation of Equations (7), $\boldsymbol{E}[\sum_{v_i \in V} \mathbf{x}_i]$ is the expectation of the summation of the attribute factor functions given the data distribution over Y, Z, S, and $\mathbf{X}$ in the training set, and $\boldsymbol{E}_{P_\alpha(Y,Z,S|X)}[\sum_{v_i \in V} \mathbf{x}_i]$ 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 Equation (7). In this work, we adopt Loopy Belief Propagation (LBP) [Murphy et al. 1999] to calculate the marginal probability of $P(Y, Z, S)$ 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 $P_\alpha(Y, Z, S|X)$. Second, we update α, β, and γ with the learning rate η by Equation (8). The learning algorithm terminates when it reaches convergence:

$$\theta_{new} = \theta_{old} + \eta \cdot \frac{\partial \mathcal{O}(\theta)}{\partial \theta}. \tag{8}$$

Prediction. With the estimated parameter θ, we can now assign the value of unknown labels Y, Z, S by looking for a label configuration that will maximize the objective function, that is, $(Y^*, Z^*, S^*) = \arg\max\ \mathcal{O}(Y, Z, S|G, \mathbf{X}, \theta)$. In this article, we use the max-sum algorithm [Kschischang et al. 2001] to solve the above problem.

Complexity. The complexity of the learning algorithm at each iteration is $O(|V| \cdot Q + |E| \cdot Q^2 + |C| \cdot Q^3)$, where $|V|, |E|, |C|$ are the numbers of users, edges, and triads in the graph, respectively, and Q is the number of classes of multiple labels. Specifically, $Q = |Y| \times |Z| \times |S|$ in the presented model, where $|Y| = 2$ is the number of gender labels—male and female, $|Z| = 4$ is the number of age labels—young, young-adult, middle-age, and senior, and $|S|$ is the number of income labels. Therefore, when learning over only gender and age in our prediction experiments, Q is equal to $|Y| \times |Z|$, that is 8.

6.4. Distributed Learning

We further leverage a distributed framework [Tang et al. 2013, 2016] to scale up our model to handle these large-scale mobile 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 process 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. An illustrative flow of the two steps can be found in Figure 17.

Specifically, the master-slave-based distributed learning framework [Tang et al. 2013, 2016] can be described in two phases. At the first phase, the large-scale

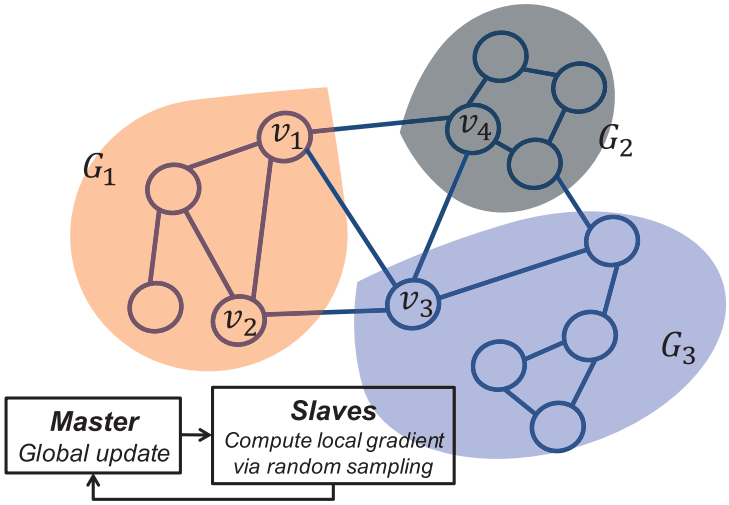


Fig. 17. An illustration of the master-slave learning scheme.

network G is partitioned into K sub-networks $G_1, \ldots, G_k, \ldots, G_K$ of balanced size, and the K sub-networks are distributed to K "slave" processors. At the second phase, we iteratively learn the parameters in two steps. At each iteration i, first, each processor can compute the local belief on its sub-network G_k according to Equation (9):

$$M_t^{k,i}(\chi_t) \propto f^k(\chi_t, \cdot) \prod_{u \in \Gamma(t)} m_{u \to t}^{k,i}(\chi_t), \tag{9}$$

where χ_t denotes the nodes in the local factor graph, $\Gamma(t)$ denotes χ_t's neighbors, and $m_{u \to t}^{k,i}$ denotes the belief (message) propagated from node χ_u to node χ_t, which is defined as the following equation:

$$m_{u \to t}^{k,i}(\chi_t) \propto \sum_{\chi_u} f^k(\chi_u, \cdot) g^k(\chi_u, \chi_t) h^k(\chi_u, \chi_t, \cdot) \prod_{s \in \Gamma(u) \setminus t} m_{s \to u}^{k,i-1}(\chi_u), \tag{10}$$

wherein the message will be normalized. Second, the "master" processor collects all local results obtained from different subgraphs and computes the marginal probability $P(\chi_t|\cdot)$ according to Equation (11) and updates the parameters according to Equations (7) and (8):

$$P^i(\chi_t|\cdot) = \sigma \sum_{k=1}^{K} M_t^{k,i}(\chi_t), \tag{11}$$

where σ is the normalization constant. This phase is repeated until convergence.

There are three notes for our model implementation. In order to achieve the balance among different slaves, we partition the nationwide mobile network into K subgraphs of roughly equal size. The second one is that we first extract all features for each user from the original full network. We then split it into subgraphs that are handled by each "slave" processor.

The third point worth noting is that a structural factor has to be eliminated in the distributed learning framework if it is defined over several nodes that belong to different subgraphs—for example, the triangle structures (v_1, v_2, v_3) and (v_1, v_3, v_4) in Figure 17. To address this issue, we propose to use virtual nodes [Tang et al. 2013, 2016] to construct the broken structural factors. For example, to complete the triad

factor over the triangle (v_1, v_2, v_3) that would be ignored in G_1 in Figure 17, we design a virtual node v_3' in G_1. In doing so, the factor graph over G_1 will capture the structural correlations of the three users' demographic information. As the completion of the triangle (v_1, v_2, v_3) in G_1, it will not be constructed in the other subgraph, that is, G_3. With that said, if three nodes of a triangle are distributed into three subgraphs, such as (v_1, v_3, v_4), one of the three involved subgraphs will be randomly selected to complete the triangle and leave the other two ignored.

6.5. Coupled Network Learning

Finally, we design a variant of the WhoAmI method to address the challenges in coupled network demographic prediction. As illustrated in Section 5, the problem faces two unique challenges. First, the missing of the target network structure makes it impossible to define triadic factors $h(\cdot)$ over three target users, such as the triangle structure (v_6, v_7, v_8) in Figure 3. Second, users' individual features across different mobile operators are asymmetric, due to the sparsity of the target network. For example, the connections between user v_1 and users from both the same operator O_1 (v_2, v_3, v_4, v_5) and the other operator O_2 (v_6, v_7) are observed for counting v_1's degree centrality, while for user v_6 in O_2, the associations with O_1's users (v_1, v_4) can be observed, and those with target users (v_7, v_8) are not observable. In this context, the individual features of source and target users follow different distributions, making it infeasible for a supervised learning framework.

In light of these issues and our previous work on coupled link prediction [Dong et al. 2015], we propose the coupled version of the WhoAmI method—CoupledMFG. By taking the coupled mobile networks as the input of a factor graph model, we have the following joint distribution:

$$\begin{aligned} P(Y, Z, S|G^S, G^C, \mathbf{X}) = & \prod_{v_i \in V^S} f^S(y_i, z_i, s_i, \mathbf{x}_i) \times \prod_{v_i \in V^T} f^T(y_i, z_i, s_i, \mathbf{x}_i) \\ & \times \prod_{e_{ij} \in E^S} [g^S(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)] \times \prod_{e_{ij} \in E^C} [g^C(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)] \\ & \times \prod_{c_{ijk} \in G^S} [h^S(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)] \times \prod_{c_{ijk} \in G^C} [h^C(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)]. \end{aligned} \tag{12}$$

This joint distribution factorizes all factors over the available structures in coupled networks. The first two terms model the attribute factors for users in source and target networks, respectively. Recall that one of the challenges is the asymmetry of users' individual attributes across these two networks, making it desired to separately model these two groups of attribute factors $f^S(\cdot)$ and $f^T(\cdot)$. The remaining four terms capture the structural correlations in coupled networks. Specifically, the third and fourth terms model the dyadic correlations, and the fifth and sixth terms model the triadic correlations in the source and cross networks, respectively. Further, all the latent variables in $g^S(\cdot)$ and $h^S(\cdot)$ are labeled, while only partial of latent variables in $g^C(\cdot)$ and $h^C(\cdot)$ are known to the model. Take the triadic factor $h^C(\cdot)$ over the triangle (v_1, v_4, v_6) in Figure 3 as an example, user v_6's demographic attributes are not available—in fact, they are the objective of the prediction model—and the demographics of users v_1 and v_4 are labeled for the learning algorithm.

One necessary question arises: Do the demographic correlations over edges $g(\cdot)$ and triangles $h(\cdot)$ follow the same distribution in source and cross networks? Our examination shows that there exists no significant distinction on the demographic distributions between source and cross networks. With that said, the semi-supervised nature of the proposed WhoAmI method enables the joint modeling of structural factors ($g(\cdot)$ and $h(\cdot)$) across source and target networks. To do so, we model the structural factors into

ALGORITHM 1: Distributed CoupledMFG Learning Algorithm.

Input: The source network G^S, the cross network G^C, the node set V^T of the target network G^T, and the learning rate η
Output: Parameters $\theta = (\alpha^S, \alpha^T, \beta, \gamma)$

Master initializes $\theta \leftarrow 0$;
Master constructs the coupled factor graph according to Equation (12) with G^S, G^C, V^T;
Master partitions the input mobile network into K subgraphs of relatively equal size;
Master completes the broken structural factors with virtual nodes;
Master forwards all subgraphs to slaves [Communication];
repeat
  Master broadcasts θ to Slaves [Communication];
  for $k = 1 \rightarrow K$ **do**
    Slave k computes local belief according to Equations (9) and (10);
    Slave k sends the local belief to Master [Communication];
  end
  Master calculates the marginal distribution for each variable according to Equation (11);
  Master calculates the gradient for each parameter according to Equation (7);
  Master updates the parameters according to Equation (8);
until *Convergence*;

the same parameter space. Specifically, we have the following log-likelihood objective function for the CoupledMFG model.

$$\mathcal{O}(\alpha, \beta, \gamma) = \sum_{v_i \in V^S} \alpha^S_{y_i z_i s_i} \mathbf{x}_i^S + \sum_{v_i \in V^T} \alpha^T_{y_i z_i s_i} \mathbf{x}_i^T + \sum_{p=1}^{15} \beta_p \sum_{e_{ij} \in E^S \cup E^C} g'_p(\cdot) + \sum_{q=1}^{84} \sum_{c_{ijk} \in G^S \cup G^C} \gamma_q h'_q(\cdot) - \log W, \tag{13}$$

where the two different parameters α^S and α^T are designed to separately model the attribute factors in source and target networks, and on the other hand, both the parameters β and γ are used to simultaneously model the dyadic and triadic factors across source and cross networks. In doing so, the CoupledMFG model is enabled to handle the two challenges in coupled network demographic prediction—the sparseness of the target network and as a result, the asymmetry of individual features in source and target networks.

The distributed learning algorithm for CoupledMFG is presented in Algorithm 1. In the algorithm, we also mark the communications between Master and Slaves. The learning algorithm will assign the target users (unlabeled) with demographic labels that maximize the marginal probabilities.

7. EXPERIMENTS

We present the effectiveness and efficiency of our proposed WhoAmI method on demographic prediction by various experiments. The code used in the experiment is publicly available.[5]

7.1. 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

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

infer user demographics effectively for mobile operators, we only consider active users who have at least five contacts in two months. After filtering out non-active users, there are 1.09 million and 304,000 active users in the 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.

All code is implemented in C++, and prediction 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 WhoAmI method that can capture the interrelation between two types of labels (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[6] and use the default setting and parameters. For SVM, we use liblinear.[7] For FGM, the model proposed in Lou et al. [2013] is used. Note that our proposed WhoAmI method is equal to FGM if we do not consider the interrelations between gender and age. In addition, other types of models have been used for capturing interaction effects from data, such as hierarchical multi-level models [Gelman and Hill 2006; Raudenbush and Bryk 2002]. However, rather than detecting and modeling the nested structures, the goal of this work is to demonstrate the effects of dyadic and triadic correlations between users' demographic attributes. Therefore, those models are not considered in the experiments.

For all comparison methods, we use the same unstructured features (individual, friend, and circle attributes) introduced in Feature Definition of Section 6.2. For the graphical models, FGM and WhoAmI, the structural features (dyadic and triadic factors) are further used to model user demographics on network structure. The major difference between our WhoAmI method and the FGM model is that WhoAmI can capture not only the structural correlations between different users but also the interrelations between two dependent variables of each user, that is, gender and age.

7.2. Experiment Results

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

Predictive Performance. Table IV shows the prediction results of different algorithms on the four prediction cases, that is, gender and age predictions in the CALL and SMS networks, respectively. Clearly our WhoAmI method yields better performance than the other alternative methods in all four cases. The Bag method 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 WhoAmI method outperforms FGM by further leveraging the interrelations between users' gender and age. In terms of weighted Precision, Recall, and F1-Measure, WhoAmI achieves up to 10% improvements, compared with the baselines for the prediction of users' gender and age. As for Accuracy, the WhoAmI method can infer 80% of the users' gender in the CALL network and 73% of the users'

[6]http://www.cs.waikato.ac.nz/ml/weka/.
[7]http://www.csie.ntu.edu.tw/cjlin/liblinear/.

Table IV. Demographic Prediction Performance by Weighted Precision, Recall, and F1-Measure

Network	Method	Gender			Age		
		wPrecision	wRecall	wF1-Measure	wPrecision	wRecall	wF1-Measure
CALL	LRC	0.7327	0.7289	0.7245	0.6350	0.6466	0.6337
	SVM	0.7327	0.7287	0.7242	0.6369	0.6463	0.6273
	NB	0.7222	0.7227	0.7222	0.6246	0.6224	0.6223
	RF	0.7437	0.7310	0.7415	0.6382	0.6482	0.6388
	Bag	0.7644	0.7648	0.7643	0.6607	0.6688	0.6592
	RBF	0.7283	0.7275	0.7252	0.6194	0.6272	0.6218
	FGM	0.7658	0.7662	0.7659	0.6998	0.6989	0.6935
	WhoAmI	0.8088	0.8076	0.8063	0.7266	0.7140	0.7132
SMS	LRC	0.6766	0.6758	0.6689	0.6702	0.6890	0.6630
	SVM	0.6749	0.6750	0.6690	0.6654	0.6884	0.6607
	NB	0.6231	0.6655	0.6603	0.6563	0.6588	0.6570
	RF	0.6399	0.6749	0.6757	0.6623	0.6775	0.6598
	Bag	0.6905	0.6918	0.6901	0.6907	0.6987	0.6791
	RBF	0.6712	0.6592	0.6468	0.6295	0.6640	0.6356
	FGM	0.7132	0.7138	0.7133	0.7154	0.7154	0.7059
	WhoAmI	0.7589	0.7549	0.7507	0.7409	0.7303	0.7337

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 text messaging behavior is relatively easier than predicting it from phone call communications. The reason can be reasoned from the discoveries in Section 3, where we find that the difference on the usage of text messages between the young and senior people is more strong than that in phone call usage, resulting the better performance in age prediction in SMS than CALL, while the gender homophily in phone calls is more obvious than in messages, leading to the advantage when predicting gender from the CALL network.

Effects of Demographic Interrelations. We evaluate the effects of demographic interrelation on the predictions. Without modeling the interrelation between gender and age, our proposed WhoAmI method degenerates to a basic factor graph model (FGM/WhoAmI-d). From Table IV, we clearly observe the 2% to 4% improvements achieved by WhoAmI to FGM on weighted F1-Measure. We further analyze feature contributions for demographic prediction. Recall that in Feature Definition of Section 6.2, besides the individual features, we introduced the friend features (nonstructural friend attributes and dyadic factors) and circle features (nonstructural circle attributes and triadic factors). By removing either friend or circle features, we evaluate the decrease in predictive performance in terms of weighted F1-Measure, plotted in Figure 18. WhoAmI-df, WhoAmI-dc, and WhoAmI-dfc stand for the removing of friend features, circle features, and both of them, conditioned on WhoAmI-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 users' age, friend features are more telling 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.

Scalability. We verify the distributed learning algorithm by partitioning the original large-scale networks into multiple sub-networks based on users' administrative areas.

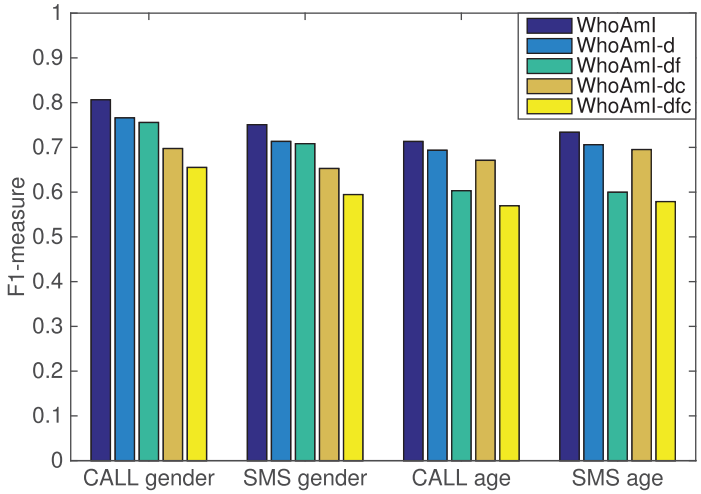


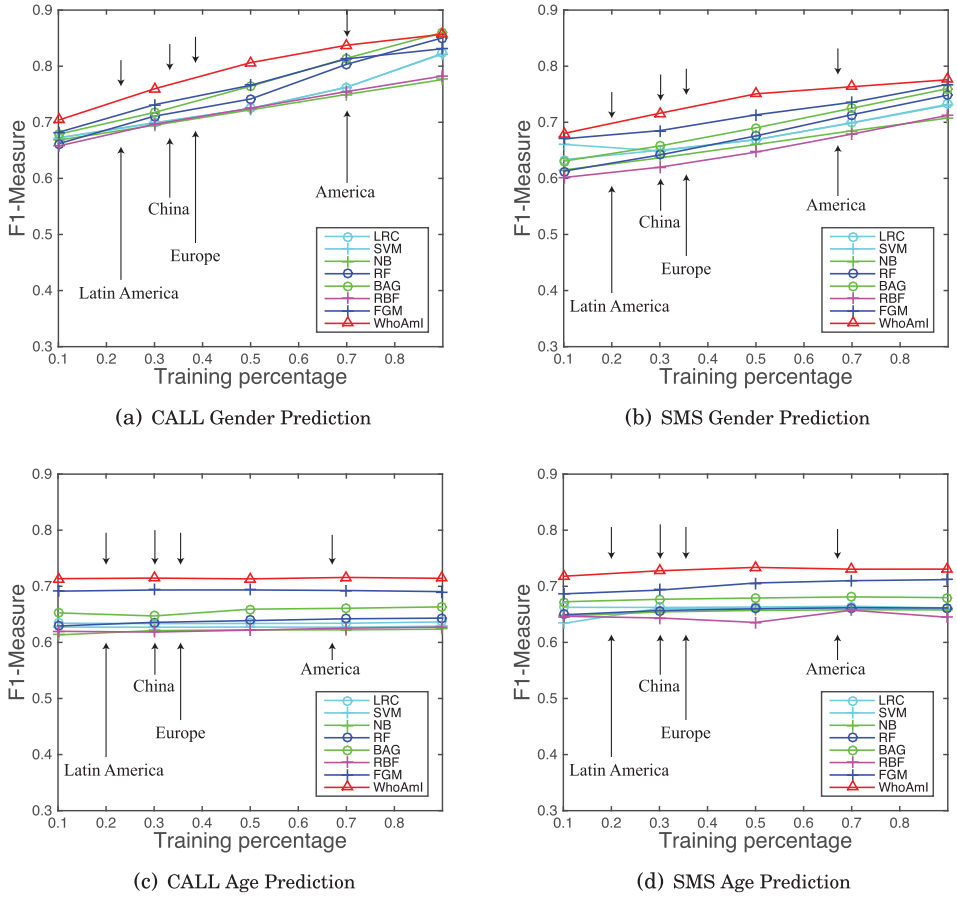
Fig. 18. Feature Contribution Analysis. WhoAmI is the proposed model. WhoAmI-d is the basic version of WhoAmI without modeling the correlation between gender and age. WhoAmI-df stands for further ignoring friend features. WhoAmI-dc stands for further ignoring circle features. WhoAmI-dcf stands for ignoring both friend and circle features.

Users' areas are determined by their postal codes during subscription registration. Each sub-network 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 less than 2% drop in performance. Basically, our learning algorithm can converge in 100 iterations, and each iteration costs about 2 (SMS) or 5min (CALL) for one single processor. By leveraging a distributed learning algorithm, our WhoAmI model is efficient even for large-scale networks with millions of nodes.

Application—Predicting Prepaid Users. As introduced before, mobile operators may not have the demographic information of prepaid users, and the percentages of prepaid users in mobile operators of different countries are different, such as 95% in India, 80% in Latin America, 70% in China, 65% in Europe, and 33% in America. We use different ratios of users as training data and the remaining as testing data. In this way, we can simulate the effects of different percentages of prepaid users on predictive performance. Figure 19 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 Figures 19(a) and 19(b). This indicates the positive effects of training data size on predicting the gender of mobile users. Specifically, we can see that in this simulation, the performance for predicting the gender of prepaid users can reach ∼70% in India (5% users as training) in terms of weighted F1-Measure, ∼75% in China (30% users as training), and ∼83% in America (67% users as training). The smooth lines in Figures 19(c) and 19(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 WhoAmI method with different sizes of training data.

7.3. Coupled Network Demographic Prediction Across Multiple Mobile Operators

We further study how the coupled variant of the WhoAmI method can be used by a mobile operator to infer the demographic profiles of its competitors' users. As the example illustrated in Figure 3, a mobile operator O_1 could have the communication records of its users and also the communication logs between its users and users of



(a) CALL Gender Prediction (b) SMS Gender Prediction

(c) CALL Age Prediction (d) SMS Age Prediction

Fig. 19. Application. Performance of demographic prediction with different percentages of postpaid users.

Table V. The Number of Active CALL Users Across Different Operators. #Edges in $O_i \rightarrow O_i$ Represents the Number of Edges Between Two O_i Users, and #Edges in $O_i \rightarrow O_j$ Represents the Number of Edges Having One O_i Endpoint and the Other O_j Endpoint

	$O_0 \rightarrow O_0$	$O_0 \rightarrow O_1$	$O_0 \rightarrow O_2$	$O_1 \rightarrow O_1$	$O_1 \rightarrow O_0$	$O_1 \rightarrow O_2$	$O_2 \rightarrow O_2$	$O_2 \rightarrow O_0$	$O_2 \rightarrow O_1$
#users	608,589	608,589	608,589	292,848	292,848	292,848	183,893	183,893	183,893
#edges	1,291,086	534,064	342,845	424,394	534,064	205,487	208,452	342,845	205,487
degree	2.12	0.88	0.56	1.45	1.82	0.70	1.13	1.86	1.12

another operator O_2 [Dong et al. 2015]. It would be very useful for the operator O_1 to have the demographic profiles of users of the competitor O_2 for business intelligence.

In this mobile dataset, there are three major mobile operators. We denote each of the three operators as O_0, O_1, and O_2, respectively. Tables V and VI list the numbers of active users in the CALL and SMS networks of each operator, and the numbers of edges within and across different operators. We train the coupled WhoAmI model by taking one operator's network as the source network and another one's as the target network. In total, we construct six pairs of prediction cases in the CALL and SMS networks, respectively, that is, O_0 to O_1, O_0 to O_2, O_1 to O_0, O_1 to O_2, O_2 to O_0, and O_2 to O_1.

Table VI. The Number of Active SMS Users Across Different Operators. #Edges in $O_i \rightarrow O_i$ Represents The Number of Edges Between Two O_i Users, and #Edges in $O_i \rightarrow O_j$ Represents the Number of Edges having One O_i Endpoint and the Other O_j Endpoint

	$O_0 \rightarrow O_0$	$O_0 \rightarrow O_1$	$O_0 \rightarrow O_2$	$O_1 \rightarrow O_1$	$O_1 \rightarrow O_0$	$O_1 \rightarrow O_2$	$O_2 \rightarrow O_2$	$O_2 \rightarrow O_0$	$O_2 \rightarrow O_1$
#users	161,547	161,547	161,547	87,556	87,556	87,556	56,634	56,634	56,634
#edges	257,154	123,192	72,313	93,342	123,192	46,807	37,660	72,313	46,807
degree	1.59	0.76	0.45	1.06	1.41	0.53	0.66	1.28	0.83

Table VII. Performance of Coupled Network Demographic Prediction Across Multiple Mobile Operators

Network	Method	Gender			Age		
		wPrecision	wRecall	wF1-Measure	wPrecision	wRecall	wF1-Measure
CALL	O_0 to O_1	0.7870	0.7800	0.7807	0.7075	0.7087	0.7039
	O_0 to O_2	0.7936	0.7939	0.7818	0.7100	0.7140	0.7085
	O_1 to O_0	0.7404	0.7403	0.7396	0.6986	0.6801	0.6696
	O_1 to O_2	0.7986	0.7979	0.7982	0.7160	0.7167	0.7094
	O_2 to O_0	0.7325	0.7282	0.7251	0.6900	0.6758	0.6622
	O_2 to O_1	0.7810	0.7794	0.7768	0.7147	0.7090	0.6981
SMS	O_0 to O_1	0.7217	0.7222	0.7219	0.7172	0.7168	0.7049
	O_0 to O_2	0.7329	0.7326	0.7327	0.7240	0.7259	0.7143
	O_1 to O_0	0.6737	0.6713	0.6721	0.6897	0.6734	0.6540
	O_1 to O_2	0.7347	0.7288	0.7285	0.7272	0.7245	0.7095
	O_2 to O_0	0.6831	0.6846	0.6798	0.6885	0.6729	0.6497
	O_2 to O_1	0.7232	0.7201	0.7143	0.7191	0.7152	0.6964

Table VII shows the strong predictability of users' demographic attributes across each pair of mobile operators. In general, we can see that the predictive performance is very promising compared to the results in Table IV. Specifically, the results demonstrate that the coupled WhoAmI method offers a 67%∼80% predictability for inferring competitor users' gender and a greater than 65% potential for the inference of their age. In other words, a mobile operator would know the demographic profiles of as many as more than half of its competitors' users, enabling the real-world application of business intelligence in telecommunication, such as acquiring new users from competitors through precision marketing.

We also notice that the prediction cases with a larger mobile operator (more users) as the training data and a smaller operator as the targeting data perform better than those with them exchanged, that is, the cases O_0 to O_1, O_0 to O_2, and O_1 to O_2 outperform the cases O_1 to O_0, O_2 to O_0, and O_2 to O_1, where the size $|O_0| > |O_1| > |O_2|$. Recall that the coupled prediction task is set in real-world scenarios (cf. Figure 3), that is, the source operator can only observe partial information about the target network, making it infeasible to compute the user distribution distances between its users and target operator users. However, to reason about the outperformance when predicting from O_{large} to O_{small}, we report the average number of connections of users from each operator in Tables V and VI. In a composite network of two operators, such as O_0 (large) and O_1 (small), O_1 users on average have more O_0 connections than O_1 connections (1.82 versus 1.45 in CALL and 1.41 versus 1.06 in SMS). In other words, users in a small operator associate more with users of a large operator than users of the same operator. Not surprisingly, users in the large operator O_0 have higher rates of same-operator contacts than of O_1 connections (2.12 versus 0.88 in CALL and 1.59 versus 0.76 in SMS). Consequently, the large operator O_{large} is able to collect rich structural information about target users from its competitors O_{small} who have smaller user base, due to those targets communicate more intensively with O_{large} users than

themselves—O_{small}. This enables its advantage of more accurately inferring its competitors' users, facilitating its marketing strategies and outcomes.

8. 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 behavior [Dong et al. 2014; Saramaki and Moro 2015; Blondel et al. 2015].

To better model the macro properties of mobile communication networks, Onnela et al. [2007] examine the local and global structure of a society-wide mobile communication network. Hidalgo and Rodriguez-Sickert [2008] investigate the communication persistence in mobile phone networks. Seshadri et al. [2008] 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 clique structures follow the power-law distribution in mobile networks [Du et al. 2009]. Recently, the emergence of work on human mobility [Gonzalez et al. 2008; Wang et al. 2011; Dong et al. 2015a; Zheng 2015] and mobile communication networks [Aledavood et al. 2015; Stopczynski et al. 2014; Gao et al. 2013], where human activities are tracked by mobile phones, provides us a means of understanding and predicting mobile social behavior. Eagle et al. [2009] try to infer the friendship network in mobile phone data. Shie et al. [2013] aim to discover the valuable user behavior patterns by mining in mobile commerce environments. Miritello et al. [2013] discover that people follow underpinning strategies to interact with each other due to limited communication capacity. Meng et al. [2016] study the correlations and differences between mobile and online networking behavior. Calabrese et al. [2014] and Blondel et al. [2015] survey the problems, techniques, and results by using mobile phones network data. 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 behavior 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 [Hu et al. 2007], gaming [Szell and Thurner 2013], and search [Bi et al. 2013] behaviors. Herring surveys how online communications facilitate gender equality, in particular, empowering women to achieve social identity that are difficult in offline environment [Herring 2003]. Leskovec and Horvitz [2008] examine the interplay of the MSN network and user demographic attributes. Mislove et al. [2011] study the demographics of Twitter users. Tang et al. [2008] extract and model the researcher profiles in large-scale collaboration networks. Michelson and Macskassy [2011] analyze both the text and the network connectivity of the blogs to infer the demographics of bloggers. Dong et al. [2013] investigate the mobile call duration behavior in mobile social networks and find that young females tend to make long phone calls [Smoreda and Licoppe 2000], in particular in the evening. Llimona et al. [2015] study the impact of gender and call duration on self-reported customer satisfaction. Chakrabarti et al. [2014] also learn a label propagation model to infer users' public profiles in Facebook social network. Additionally, researchers have used network information to identify user status differences in email [Dong et al. 2015b; Hu and Liu 2012] and LinkedIn networks [Zhao et al. 2013]. Nokia research organized the 2012 Mobile Data Challenge to infer mobile user demographics by using 200 individual communication records without network information [Mo et al. 2012; Ying et al. 2012]. Kovanen et al. [2013] 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 interrelation among different demographic attributes.

9. CONCLUSION

In this article, we model users' social decisions on connecting and maintaining relationships conditioned on their demographic profiles in large-scale mobile communication networks. Significant social strategies are stemmed from the big mobile data. We find young people put more focus on enlarging social circles; as they age, they have the tendency to maintain small but closed social relationships. We also observe striking gender differences in social triadic relationships across individuals' lifespans. Specifically, the relationships among three same-gender individuals are persistently maintained over a lifetime, while the opposite-gender triadic relationships disappear when they enter into their middle-age. Our null model demonstrates the statistical significance of the evolution of social strategies in human communication. We further engage in answering the question of to what extent user demographics can be revealed from mobile communication interactions. We formalize a demographic prediction problem to simultaneously infer users' gender and age, and further propose the WhoAmI method to solve it. Experimental results in phone call and text messaging networks demonstrate both the effectiveness and efficiency of our proposed model. Meanwhile, we identify a new problem—coupled network demographic prediction across multiple mobile operators. To address the unique challenges in this task, we present a coupled variant of the WhoAmI method. Our results unveil the predictability of user demographics across competitor networks, enabling the real-world application scenario of business intelligence in telecommunication.

Despite the promising discoveries and predictive performance of the present work, there is still large room left for future work. First, although we examine the social strategies in two large-scale mobile networks with millions of users, the results are limited to the data we used, that is, the mobile communications from one specific country. On one hand, there may exist variances on social strategies used by people across different cultural backgrounds, political systems, and geographical boundaries. Therefore, it is natural to examine the observed results in other countries upon the available data. On the other hand, although previous studies have demonstrated that mobile communications can be used as a proxy to represent human communications, it would generalize our findings beyond mobile channels if online social networks with demographic information could be investigated. Second, mobile communications are associated with dynamic information, making it necessary to further couple our studies between network structures and user demographics with social dynamics. Third, in addition to studying phone calls and text messages separately, it would be interesting to investigate social strategies and predict user demographics from the mobile network as a whole by combining the phone call and text messaging networks into one network. Finally, some other social strategies and theories can be explored and validated for modeling user social networking behavior. In addition, examining how the inferred demographics can help other topics in social network analysis, such as influence propagation, community detection, and network evolution, would also be very meaningful.

REFERENCES

Talayeh Aledavood, Eduardo López, Sam G. B. Roberts, Felix Reed-Tsochas, Esteban Moro Egido, Robin I. M. Dunbar, and Jari Saramäki. 2015. Channel-specific daily patterns in mobile phone communication. *CoRR*, abs/1507.04596 (2015).

Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *WWW'13*. 131–140.

Vincent D. Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *arXiv:1502.03406* (2015).

F. Calabrese, L. Ferrari, and V. Blondel. 2014. Urban sensing using mobile phones network data: A survey of research. *ACM Comput. Surv.* (2014).

Deepayan Chakrabarti, Stanislav Funiak, Jonathan Chang, and Sofus A. Macskassy. 2014. Joint inference of multiple label types in large networks. In *ICML'14*. 874–882.

Benjamin Cornwell. 2011. Age trends in daily social contact patterns. *Res. Aging* 33, 5 (2011), 598–631.

Yuxiao Dong, Fabio Pinelli, Yiannis Gkoufas, Zubair Nabi, Francesco Calabrese, and Nitesh V. Chawla. 2015a. Inferring unusual crowd events from mobile phone call detail records. In *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 474–492.

Yuxiao Dong, Jie Tang, Nitesh V. Chawla, Tiancheng Lou, Yang Yang, and Bai Wang. 2015b. Inferring social status and rich club effects in enterprise communication networks. *PLoS ONE* 10 (03 2015), e0119446.

Yuxiao Dong, Jie Tang, Tiancheng Lou, Bin Wu, and Nitesh V. Chawla. 2013. How long will she call me? Distribution, social theory and duration prediction. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 16–31.

Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. 2014. Inferring user demographics and social strategies in mobile social networks. In *KDD'14*. ACM, 15–24.

Yuxiao Dong, Jing Zhang, Jie Tang, Nitesh V. Chawla, and Bai Wang. 2015. CoupledLP: Link prediction in coupled networks. In *KDD'15*. ACM, 199–208.

Nan Du, Christos Faloutsos, Bai Wang, and Leman Akoglu. 2009. Large human communication networks: Patterns and a utility-driven generator. In *KDD'09*. ACM, 269–278.

Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. 2009. Inferring social network structure using mobile phone data. *Proc. Natl. Acad. Sci. U.S.A.* 106, 36 (2009).

David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Mária Ercsey-Ravasz, Ryan N. Lichtenwalter, Nitesh V. Chawla, and Zoltán Toroczkai. 2012. Range-limited centrality measures in complex networks. *Phys. Rev. E* 85, 6 (Jun 2012), 066103.

Linton C. Freeman. 1982. Centered graphs and the structure of ego networks. *Math. Soc. Sci.* 3, 3 (1982), 291–304.

Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. 2013. Modeling temporal effects of human mobile behavior on location-based social networks. In *CIKM'13*. 1673–1678.

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

Mark Granovetter. 1973. The strength of weak ties. *Am. J. Sociology* 78, 6 (1973), 1360–1380.

Mark Granovetter. 1985. Economic action and social structure: The problem of embeddedness. *Amer. J. Sociol.* (1985).

Susan C. Herring. 2003. Gender and power in on-line communication. In *Handbook of Language and Gender*. Wiley-Blackwell, 202.

Cesar A. Hidalgo and C. Rodriguez-Sickert. 2008. The dynamics of a mobile phone network. *Physica A: Stat. Mech. Appl.* 387, 12 (2008), 3017–3024.

Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. In *WWW'07*. 151–160.

Xia Hu and Huan Liu. 2012. Social status and role analysis of Palin's email network. In *WWW'12 Companion*. ACM, 531–532.

Lauri Kovanen, Kimmo Kaski, János Kertész, and Jari Saramäki. 2013. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *PNAS* 110, 45 (2013), 18070–18075.

David Krackhardt. 1992. *The Strength of Strong Ties*. Cambridge, Harvard Business School Press, Hershey, USA.

Frank R. Kschischang, Brendan J. Frey, and Hans A. Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Trans. Internet Technol.* 47 (2001), 498–519.

P. F. Lazarsfeld and R. K. Merton. 1954. Friendship as a social process: A substantive and methodological analysis. *Freedom and Control in Modern Society,*Van Nostrand, New York (1954), 8–66.

Jure Leskovec and Eric Horvitz. 2008. Planetary-scale views on a large instant-messaging network. In *WWW'08*. ACM, 915–924.

Quim Llimona, Jordi Luque, Xavier Anguera, Zoraida Hidalgo, Souneil Park, and Nuria Oliver. 2015. Effect of gender and call duration on customer satisfaction in call center big data. In *INTERSPEECH'15*.

Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, and Xiaowen Ding. 2013. Learning to predict reciprocity and triadic closure in social networks. *ACM Trans. Knowl. Discov. Data* 7, 2 (2013), 5:1–5:25.

Peter V. Marsden. 1987. Core discussion networks of americans. *Amer. Sociol. Rev.* (1987), 122–131.

Manfred Max-Neef, Antonio Elizalde, and Martin Hopenhayn. 1992. Development and human needs. *Real-life Economics: Understanding Wealth Creation* (1992), 197–213.

M. Mead. 1970. *Culture and Commitment: A Study of the Generation Gap*. Natural History Press.

L. Meng, Y. Hulovatyy, A. Striegel, and T. Milenković. 2016. On the interplay between individuals' evolving interaction patterns and traits in dynamic multiplex social networks. *IEEE Trans. Netw. Sci. Eng.* 3, 1 (2016), 32–43.

Matthew Michelson and Sofus A. Macskassy. 2011. What blogs tell us about websites: A demographics study. In *WSDM'11*. ACM, 365–374.

Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. 2013. Limited communication capacity unveils strategies for human interaction. *Sci. Rep.* 3 (2013).

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *ICWSM'11*.

Kaixiang Mo, Ben Tan, Erheng Zhong, and Qiang Yang. 2012. Your phone understands you. In *Nokia MDC'12*.

Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. 1999. Loopy belief propagation for approximate inference: An empirical study. In *UAI'99*. 467–475.

J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. 2007. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.* (2007).

Vasyl Palchykov, Kimmo Kaski, János Kertész, Albert-László Barabási, and Robin I. M. Dunbar. 2012. Sex differences in intimate relationships. *Sci. Rep.* 2:370 (2012).

Stephen W. Raudenbush and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Vol. 1. Sage.

Jari Saramaki and Esteban Moro. 2015. From seconds to months: multi-scale dynamics of mobile telephone calls. *arXiv:1504.01479* (2015).

Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskovec. 2008. Mobile call graphs: Beyond power-law and lognormal distributions. In *KDD'08*. ACM, 596–604.

Xiaolin Shi, Lada A. Adamic, and Martin J. Strauss. 2007. Networks of strong ties. *Physica A: Stat. Mech. Appl.* 378, 1 (2007), 33–47.

Bai-En Shie, S. Yu Philip, and Vincent S. Tseng. 2013. Mining interesting user behavior patterns in mobile commerce environments. *Appl. Intell.* 38, 3 (2013), 418–435.

Zbigniew Smoreda and Christian Licoppe. 2000. Gender-specific use of the domestic telephone. *Soc. Psych. Quart.* 63, 3 (2000), 238–252.

Richard C. Sprinthall. 2011. *Basic Statistical Analysis*. Pearson.

Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Jakob Eg Larsen, and Sune Lehmann. 2014. Measuring large-scale social networks with high resolution. *PLOS One* 9, 4 (2014), e95978.

Michael Szell and Stefan Thurner. 2013. How women organize social networks different from men. *Sci. Rep.* 3 (July 2013).

Jie Tang, Tiancheng Lou, Jon Kleinberg, and Sen Wu. 2016. Transfer learning to infer social ties across heterogeneous networks. *ACM Trans. Inf. Syst.* 34, 2, Article 7 (April 2016).

Jie Tang, Sen Wu, and Jimeng Sun. 2013. Confluence: Conformity influence in large social networks. In *KDD'13*. ACM, 347–355.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and mining of academic social networks. In *KDD'08*. 990–998.

Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *KDD'11*. ACM, 1100–1108.

Josh Ying, Yao-Jen Chang, Chi-Min Huang, and Vincent S. Tseng. 2012. Demographic prediction based on user's mobile behaviors. In *Nokia MDC'12*.

Yuchen Zhao, Guan Wang, Philip S. Yu, Shaobo Liu, and Simon Zhang. 2013. Inferring social roles and statuses in social networks. In *KDD'13*. 695–703.

Yu Zheng. 2015. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol. (TIST)* 6, 3 (2015), 29.

Received June 2016; revised October 2016; accepted December 2016