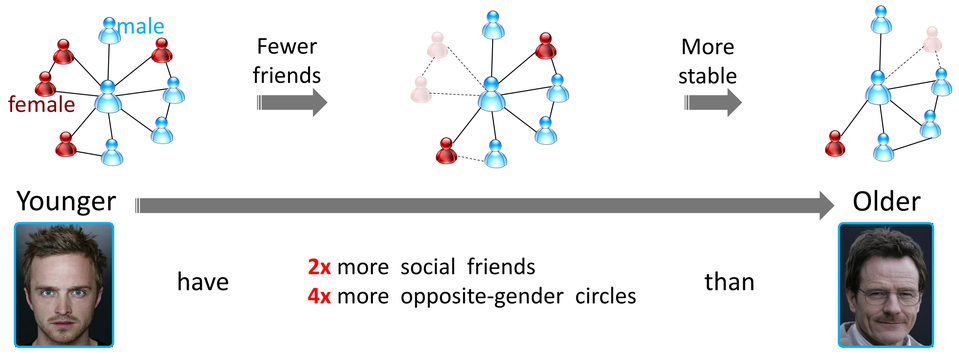
Inferring User Demographics and Social Strategies in Mobile Social Networks

Yuxiao Dong — University of Notre Dame — ydong1@nd.edu
Yang Yang — Tsinghua University — yyang.thu@gmail.com
Jie Tang — Tsinghua University — jietang@tsinghua.edu.cn
Yang Yang — University of Notre Dame — yyang1@nd.edu
Nitesh V. Chawla — University of Notre Dame — nchawla@nd.edu

UNIVERSITY OF NOTRE DAME
Tsinghua University

Did You Know?

As of 2014, there are **7.3 billion** mobile users, larger than the global population.
In the U.S., users average **22** calls, **23** messages, and **110** status checks per day.
Overall, they made **3 billion** mobile phone calls and **6 billion** messages each day.



Summary:

1. Social strategies are used by people in social activities to meet their social needs, i.e., to connect with new people and to strengthen existing relationships.

2. Different people with different demographic profiles (gender/age) make use of different social strategies to fulfill social needs for belonging, love, and affection.

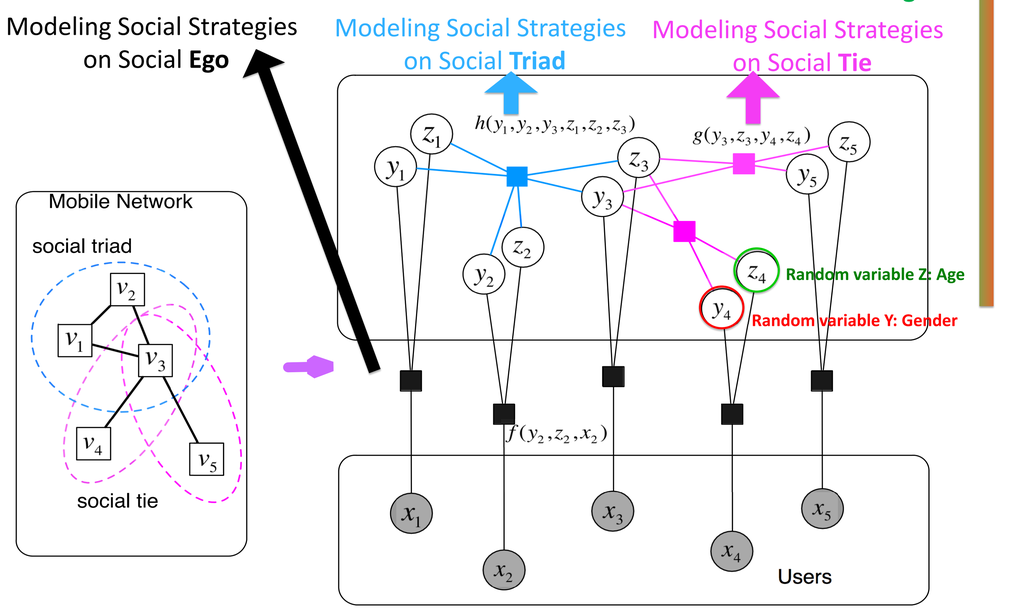
3. The proposed ***WhoAmI*** method can infer 80% of the users' genders from their mobile phone call behaviors, and 73% of the users' ages from text message behaviors.

Data:

A real-world large mobile network;
An anonymous country, two months;
1,000,000,000 communication records.

Network	#nodes	#edges
CALL	7,440,123	32,445,941
SMS	4,505,958	10,913,601

WhoAmI -- Demographic Prediction



Double Dependent-Variable Factor Graph (DFG)
Inferring gender and age simultaneously

Joint Distribution

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

Factor Initialization

Attribute Factor

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

Dyadic Factor

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

Triadic Factor

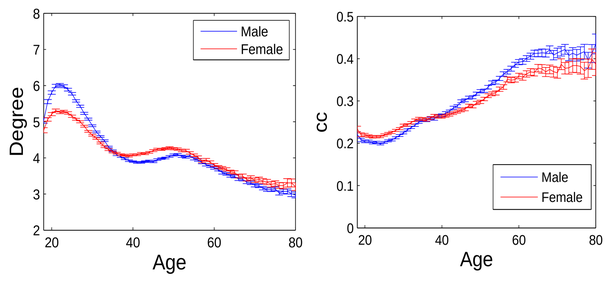
$$h(\mathbf{y}_c, \mathbf{z}_c) = \begin{cases} \frac{1}{W_{c_1}} \exp\{\gamma_1 \cdot h'_1(y_i, y_j, y_k)\} \\ \frac{1}{W_{c_2}} \exp\{\gamma_2 \cdot h'_2(y_i, y_j, z_i)\} \\ \cdots \\ \frac{1}{W_{c_{20}}} \exp\{\gamma_{20} \cdot h'_{20}(z_i, z_j, z_k)\} \end{cases}$$

Objective Function

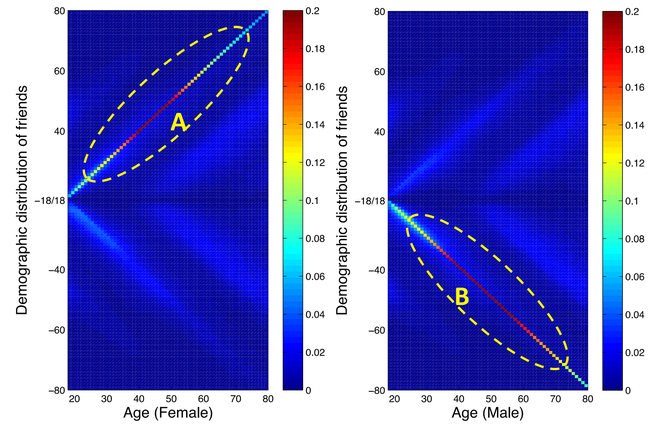
$$\mathcal{O}(\alpha, \beta, \gamma) = \sum_{v_i \in V} \alpha_{y_i z_i} \mathbf{x}_i + \sum_{e_{ij} \in E} \sum_{p=1}^{6} \beta_p g'_p(\cdot) + \sum_{c_{ijk} \in G} \sum_{q=1}^{20} \gamma_q h'_q(\cdot) - \log W$$

Social Strategy Analysis

Social Strategies on Social Ego

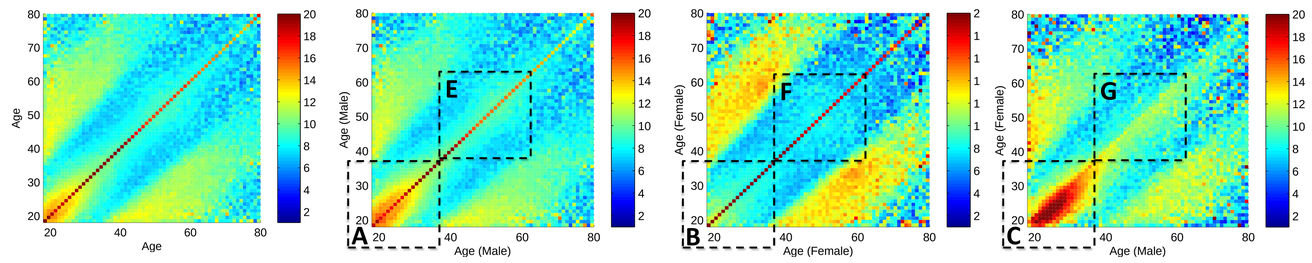


1. Young people are active in broadening their social circles (larger degree), while seniors have the tendency to maintain small but close connections (smaller degree & higher *cc*).



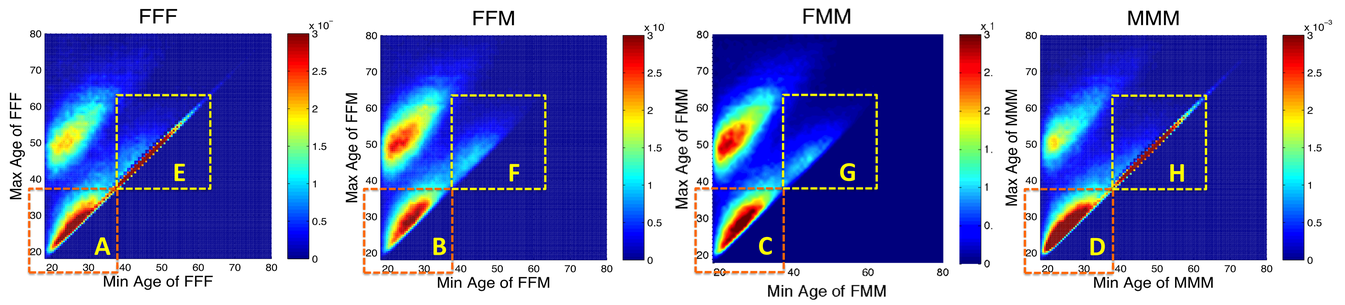
1. A, B: People tend to communicate with others of both similar gender and age, i.e., homophily.

Social Strategies on Social Tie



1. A vs. B: Young males maintain more frequent and broader social connections than young females. .
2. A/B vs. C: Opposite-gender interactions are much more frequent than those between young same-gender users.
   E/F vs. G: When becoming mature, same-gender interactions are more frequent than those of opposite-gender users.

Social Strategies on Social Triad

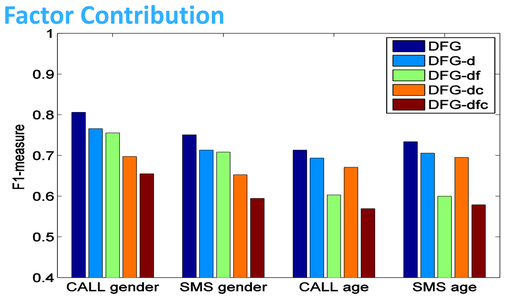


1. A,B,C,D: People expand both same-gender and opposite-gender social groups during the dating period.
2. F/G vs. E/H: People's attention to opposite-gender groups quickly disappears, and the insistence and social investment on same-gender social groups lasts for a lifetime.

Performance

Network		Gender wPre.	Gender wRec.	Gender wF1	Age wPre.	Age wRec.	Age wF1
CALL	LRC	0.732	0.728	0.724	0.635	0.646	0.633
	SVM	0.732	0.728	0.724	0.636	0.646	0.627
	NB	0.722	0.722	0.722	0.624	0.622	0.622
	RF	0.743	0.731	0.741	0.638	0.648	0.638
	BAG	0.764	0.764	0.764	0.660	0.668	0.659
	FGM	0.765	0.766	0.765	0.699	0.698	0.693
	DFG	0.808	0.807	0.806	0.726	0.714	0.713
SMS	LRC	0.676	0.765	0.668	0.670	0.689	0.663
	SVM	0.674	0.675	0.669	0.665	0.688	0.660
	NB	0.623	0.665	0.660	0.656	0.658	0.657
	RF	0.639	0.674	0.675	0.662	0.677	0.659
	BAG	0.690	0.691	0.690	0.690	0.698	0.679
	FGM	0.713	0.713	0.713	0.715	0.715	0.705
	DFG	0.758	0.754	0.750	0.740	0.730	0.733

Factor Contribution



DFG: the proposed model
DFG-d: no interrelations between gender / age
DFG-df: further ignoring friend features
DFG-dc: further ignoring circle features
DFG-dcf: ignoring both friend&circle features

Distributed Learning

Message Passing Interface (MPI)
9-10× speedup with 16 cores
<2% drop in performance
Converge in 100 iterations, each costs 2-5 minutes

Paper ID

851

iCeNSA
Interdisciplinary Center for Network Science & Applications

Arnetminer

Code: http://arnetminer.org/demographic

20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 24-27, 2014 New York City

References

1. N. Eagle, A. S. Pentland, D. Lazer. Inferring social network structure using mobile phone data. PNAS, 2009.
2. V. Palchykov, K. Kaski, J. Kertesz, A. –L. Barabasi, R. I. M. Dunbar. Sex differences in intimate relationships. SciRep, 2012.
3. D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabasi. Human mobility, social ties, and link prediction. In KDD'11.
4. J. Leskovec, E. Horvitz. Planetary-scale views on a large instant-messaging network. In WWW'08.
5. Y. Dong, J. Tang, N. V. Chawla. How long will she call me? Distribution, social theory and duration prediction. In ECML/PKDD'13.