Inferring User Demographics and Social Strategies in Mobile Social Networks

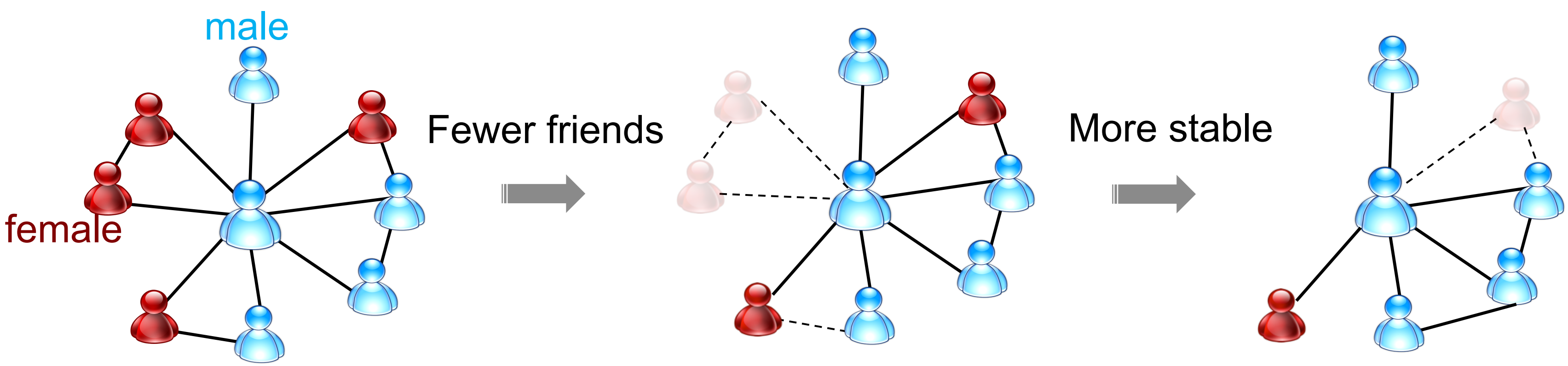
Yuxiao Dong[#], Yang Yang[+], Jie Tang[+], Yang Yang[#], Nitesh V. Chawla[#]

[#]University of Notre Dame [+]Tsinghua University

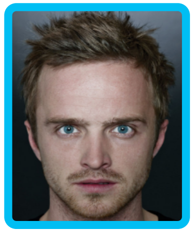
Did you know:

As of 2014, there are **7.3** billion mobile phones, larger than the global population.

Users average **22** calls, **23** messages, and **110** status checks **per day**.



Younger → Older



have **2x** more social friends / **4x** more opposite-gender circles than

Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla. Inferring User Demographics and Social Strategies in Mobile Social Networks. KDD 2014.

Big Mobile Data

- Real-world large-scale mobile data
  - An anonymous country.
  - No communication content.
  - Aug. 2008 – Sep. 2008.
  - > 7 million mobile users + demographic information.
    - Gender: Male (55%) / Female (45%)
    - Age: Young (18-24) / Young-Adult (25-34) / Middle-Age (35-49) / Senior (>49)
  - > 1 billion communication records (call and message).

- Two networks:

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

What We Do

- How do people communicate / interact with each other with mobile phones?
  - Infer human social strategies on demographics.

- To what extent can user demographic profiles be inferred from their mobile communication interactions?
  - Infer user demographics based on social strategies.

- Applications:
  - Viral marketing
  - Personalized services
  - User modeling
  - Customer churn warning
  - …

Infer human social strategies on demographics

user demographics + mobile social network
→ social strategies

Social Strategy

- Human needs are defined according to the existential categories of being, having, doing, and interacting[1]. Two basic human needs[2] are to
  - Meet new people → Social needs.
  - Strengthen existing relationships → Social needs.

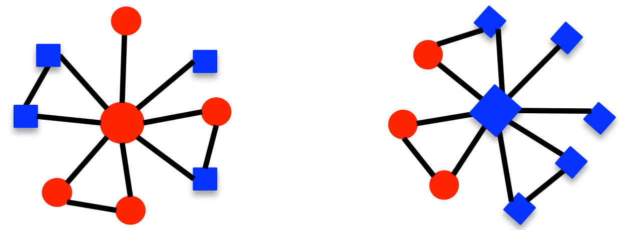
- Social strategies are used by people to meet social needs.
  - Human needs are constant across historical time periods.
  - However, the strategies by which these needs are satisfied change over time[1,3] .

- Barabasi and Dunbar[3]:
  - "Women are more focused on opposite-sex relationships than men during the reproductively active period of their lives." … "As women age, their attention ships from their spouse to younger females---their daughters."
  - "Human social strategies have more complex dynamics than previously assumed."

1. http://en.wikipedia.org/wiki/Fundamental_human_needs
2. M.J. Piskorski. Social strategies that work. Harvard Business Review. Nov. 2011.
3. V. Palchykov, K. Kaski, J. Kertesz, A.-L. Barabasi, R. I. M. Dunbar. Sex differences in intimate relationships. Scientific Reports 2012.

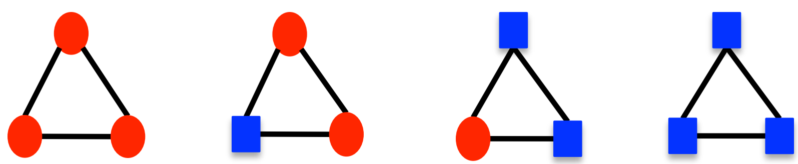
Social Strategy

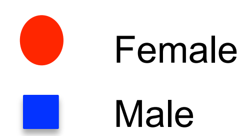
- We study demographic-based social strategy with respect to the micro-level network structures.
  - Ego network



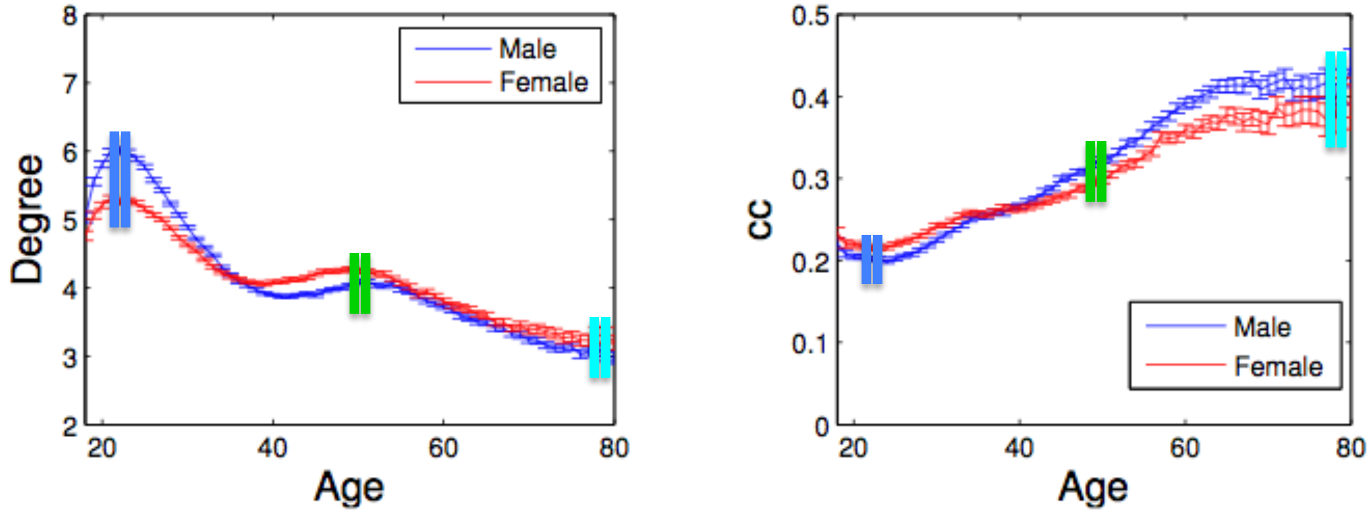
  - Social tie

  - Social triad





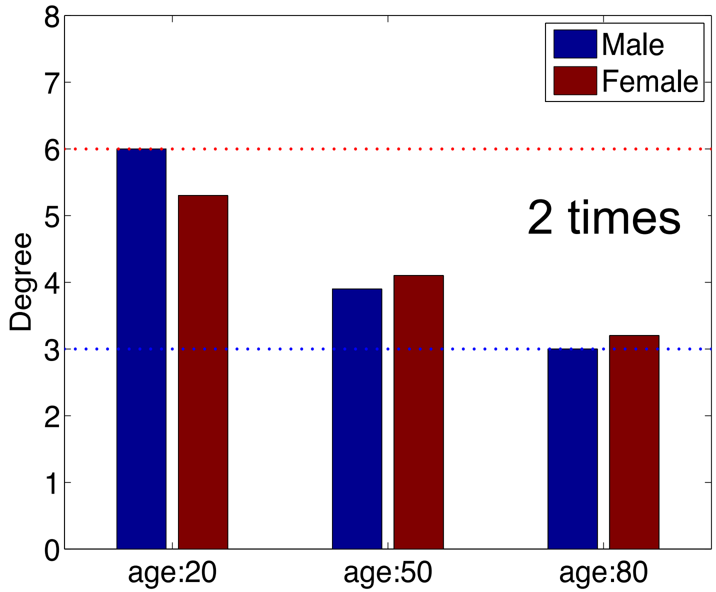
Social Strategy: Ego Network

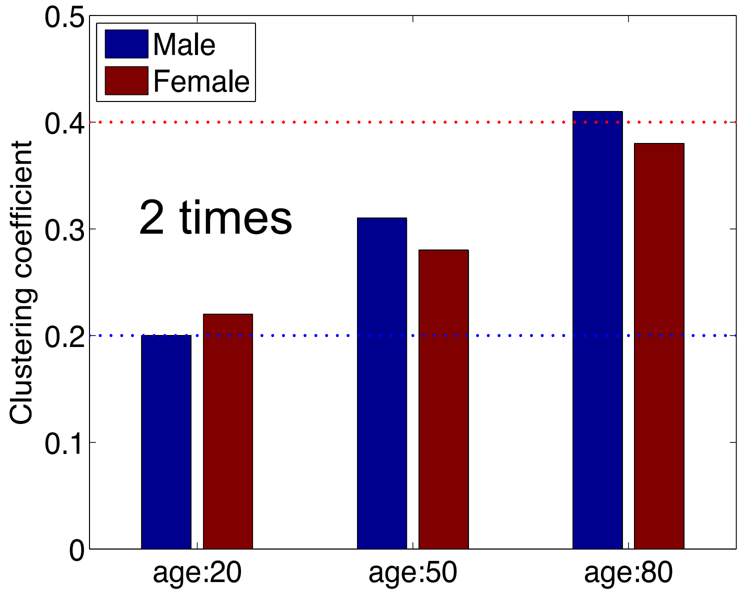


(a) Degree Centrality (b) Triadic Closure

Correlations between user demographics and network properties.

Social Strategy: Ego Network





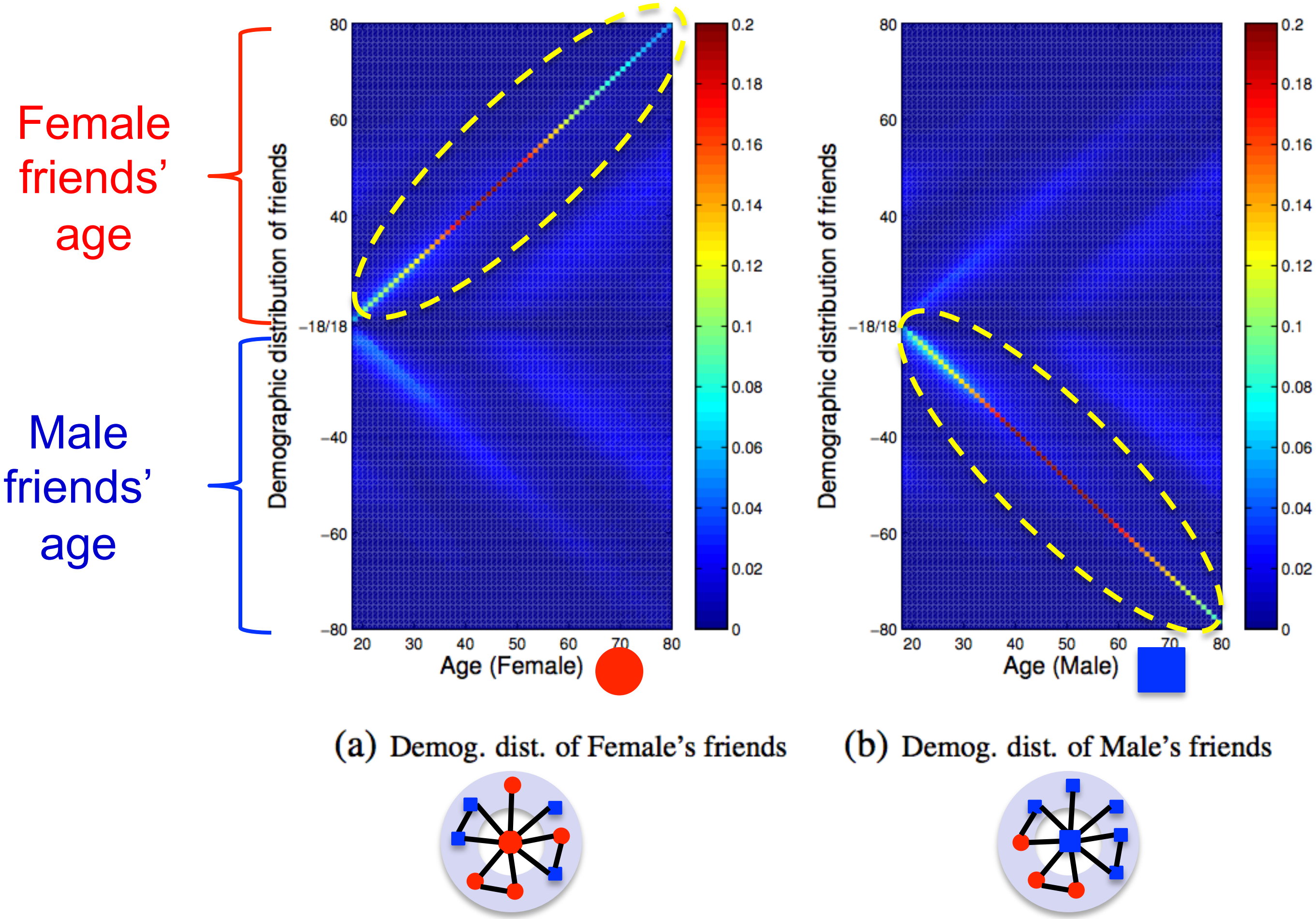
Correlations between user demographics and network properties.

Social Strategies: Young people are active in broadening their social circles, while seniors have the tendency to maintain small but close connections.

Social Strategy: Ego Network

In your mobile phone contact list,
do you have more **female** or **male** friends?

Social Strategy: Ego Network



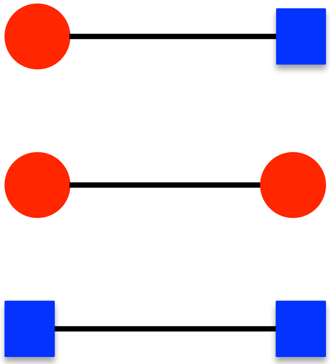
(a) Demog. dist. of Female's friends (b) Demog. dist. of Male's friends

X: age of central user.
Y: age of friends.
Positive Y: female friends;
Negative Y: male friends;
Spectrum: distribution

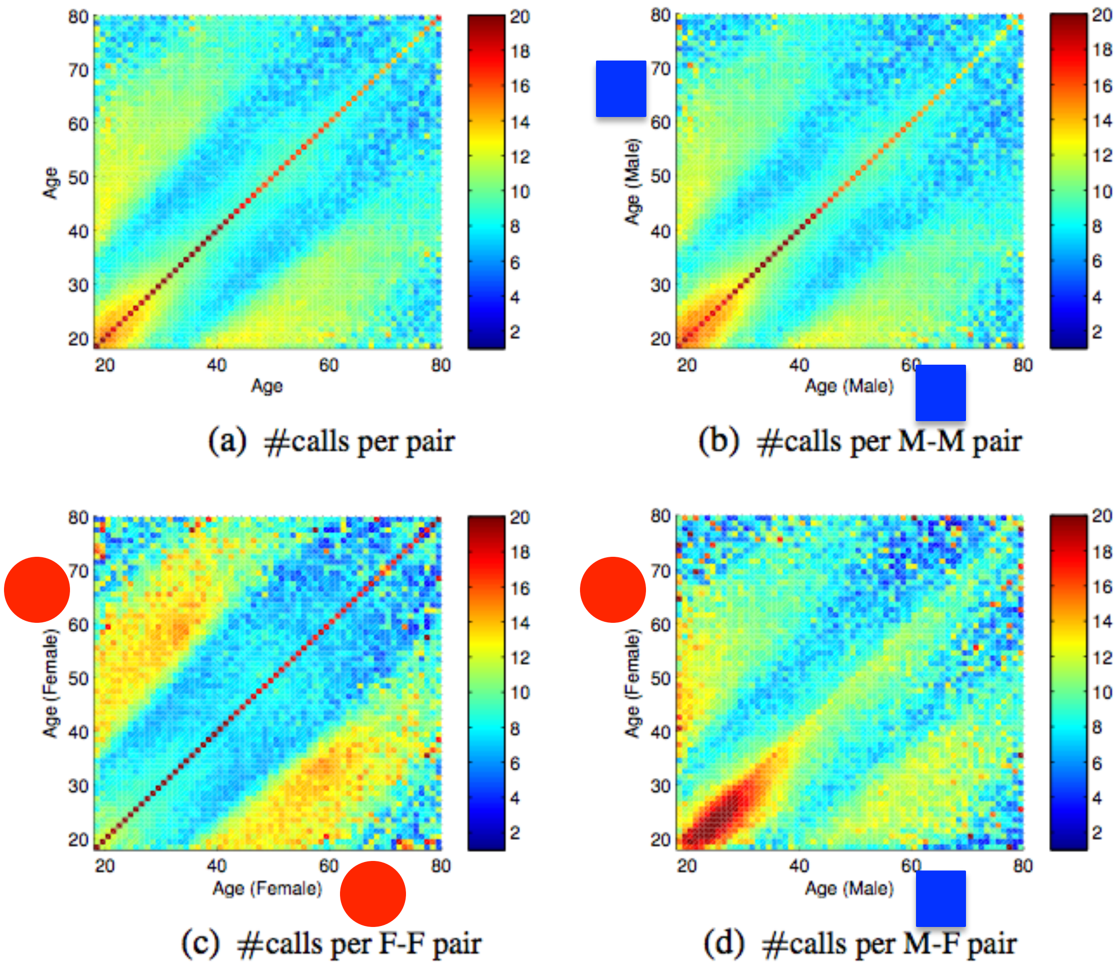
Social Strategies: People tend to communicate with others of both similar gender and age, i.e., demographic homophily.

Social Strategy: Social Tie

How frequently do you call your mother vs. your significant other?



Social Strategy: Social Tie

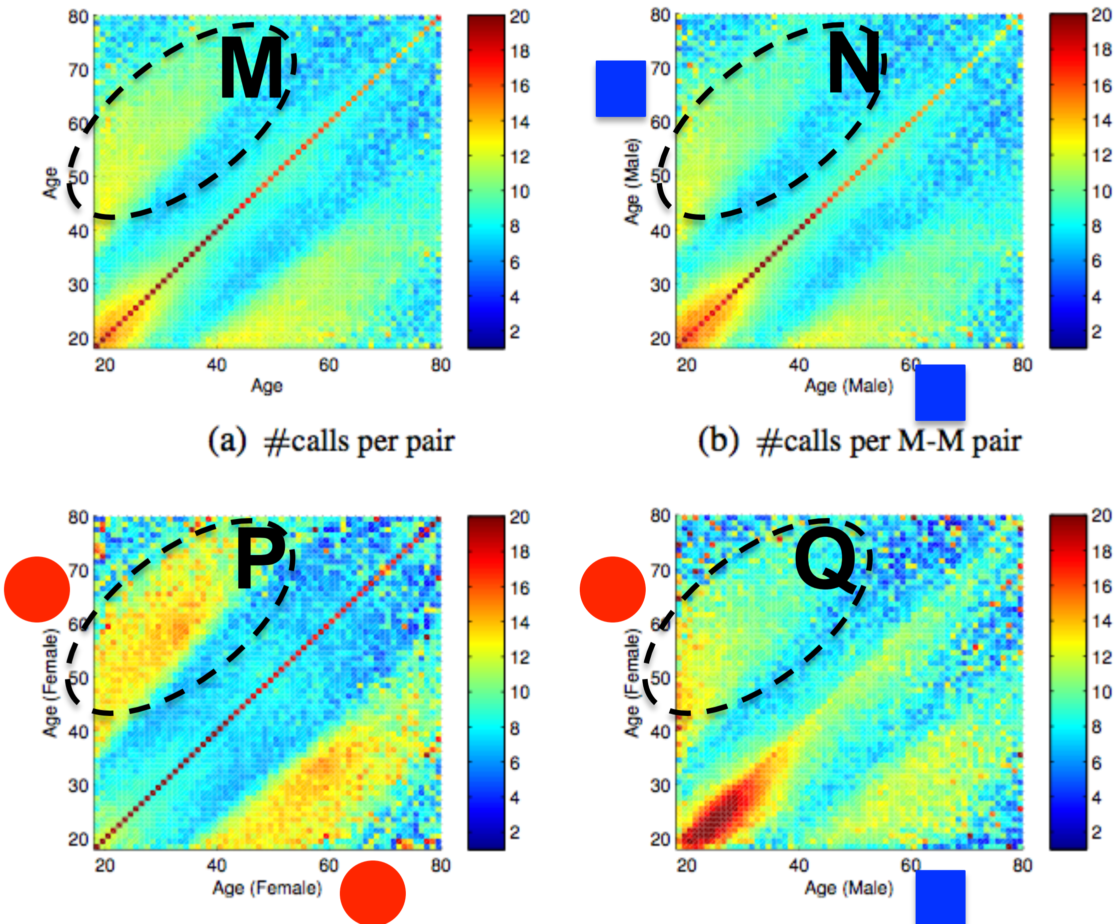


X: age of one user.
Y: age of the other user.

Spectrum: #calls per month

(a), (b), (c) are symmetric.

Social Strategy: Social Tie



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

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

X: age of one user.
Y: age of the other user.

Spectrum: #calls per month

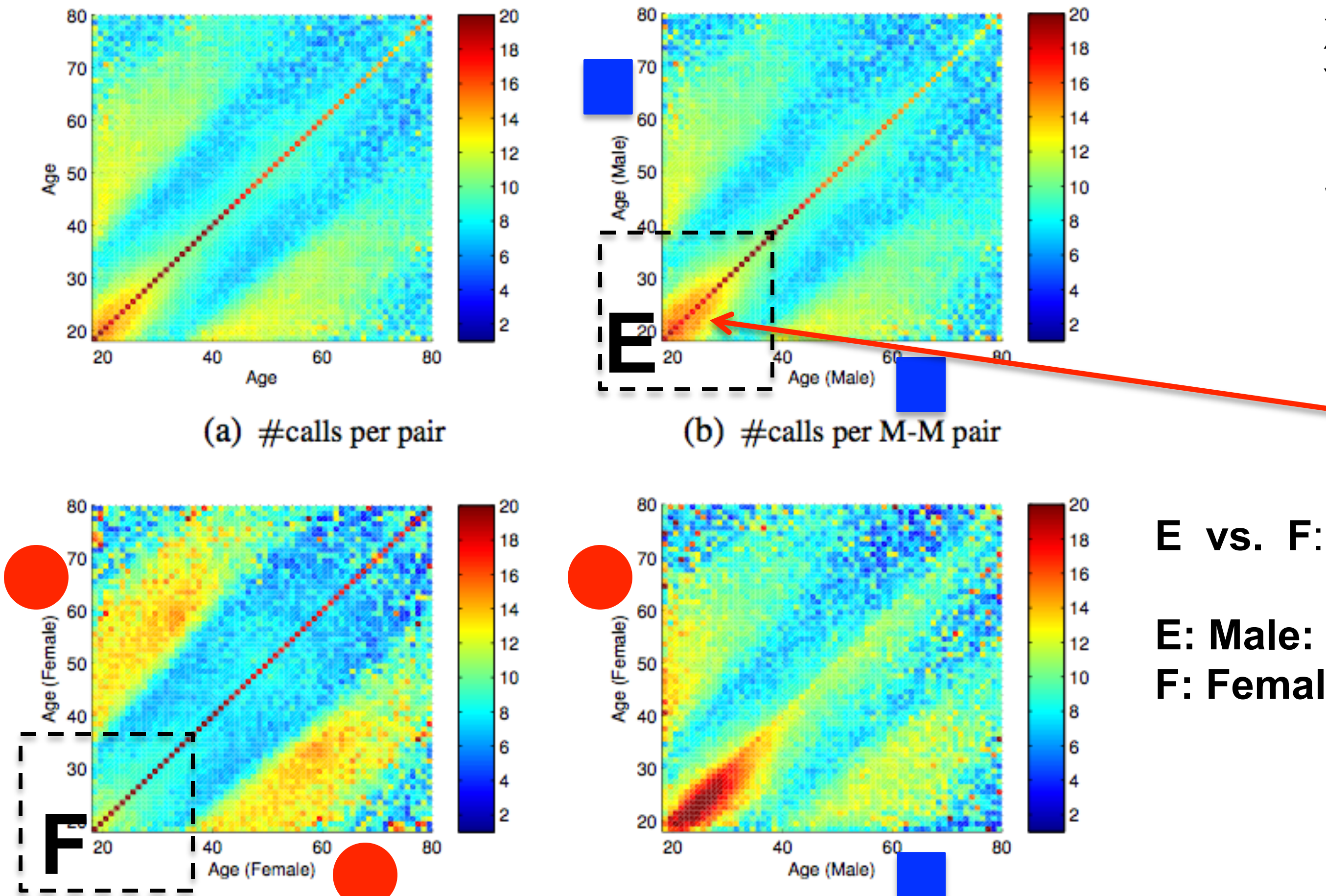
(a), (b), (c) are symmetric.

M,N,P,Q:

10~15 calls per month are made between parents and children.

Social Strategies: Frequent cross-generation interactions are maintained to bridge age gaps.

Social Strategy: Social Tie



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

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

X: age of one user.
Y: age of the other user.

Spectrum: #calls per month

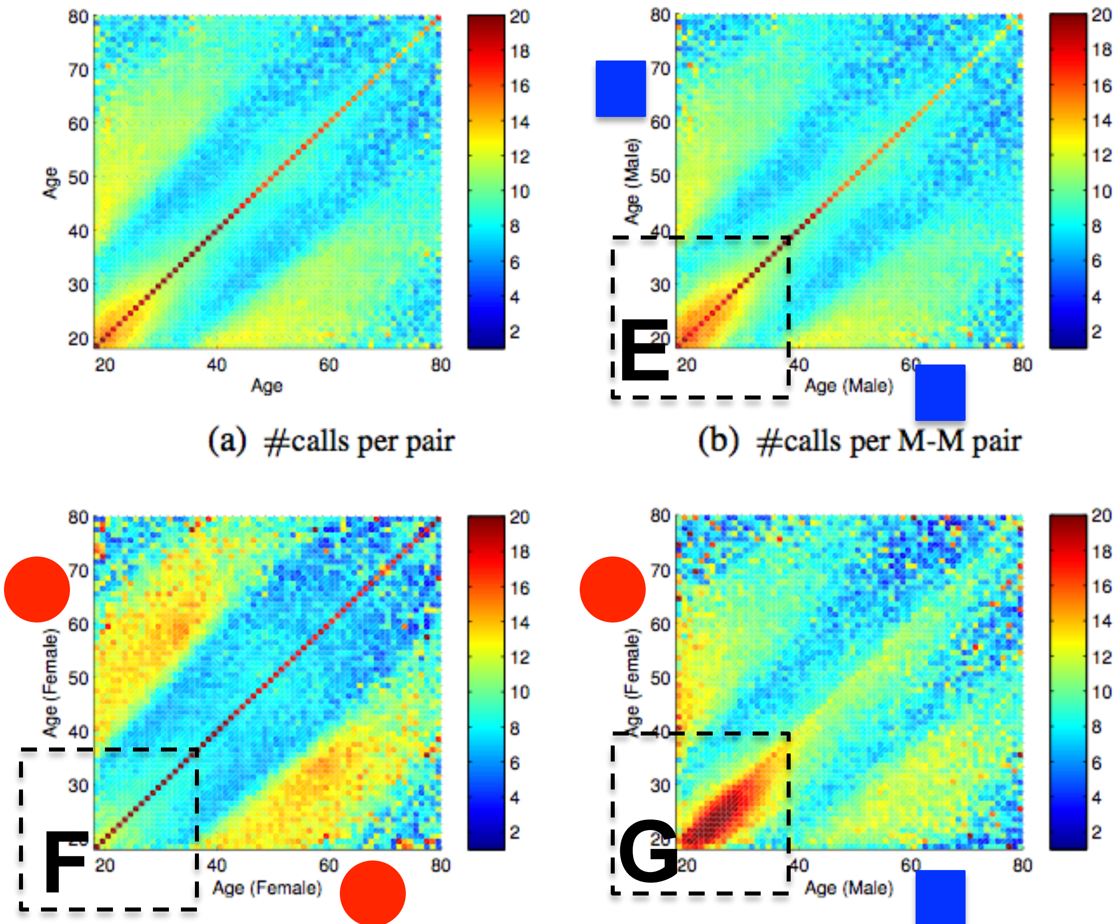
(a), (b), (c) are symmetric.

E vs. F:

E: Male: ±5 years old interactions
F: Female: only same-age interaction

Social Strategies: Young male maintain more frequent and broader social connections than young females.

Social Strategy: Social Tie



(a) #calls per pair

(b) #calls per M-M pair

(c) #calls per F-F pair

(d) #calls per M-F pair

X: age of one user.
Y: age of the other user.

Spectrum: #calls per month

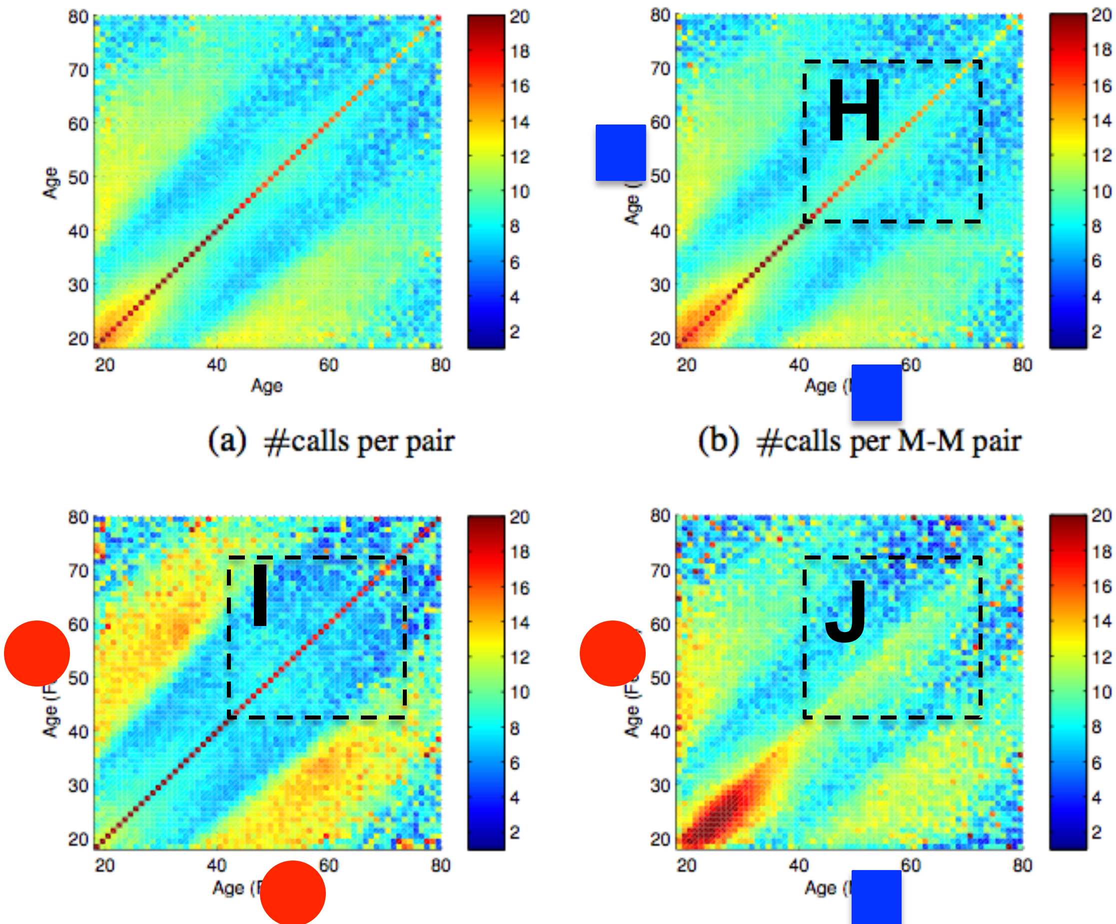
(a), (b), (c) are symmetric.

E,F vs. G:

G: f-m: >30 calls per months
E/F: m-m or f-f: 10~15 calls

Social Strategies: Opposite-gender interactions are much more frequent than those between young same-gender users.

Social Strategy: Social Tie



(a) #calls per pair

(b) #calls per M-M pair

(c) #calls per F-F pair

(d) #calls per M-F pair

X: age of one user.
Y: age of the other user.

Spectrum: #calls per month

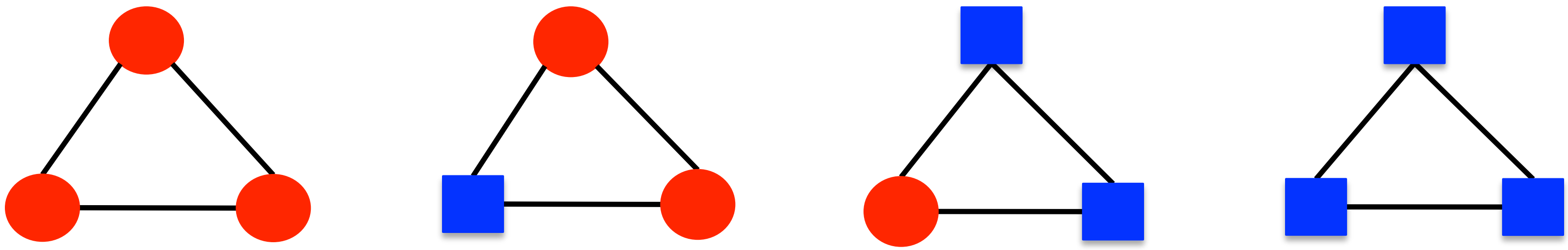
(a), (b), (c) are symmetric.

H,I vs. J:

Social Strategies: When people become mature, reversely, same-gender interactions are more frequent than those between opposite-gender users.

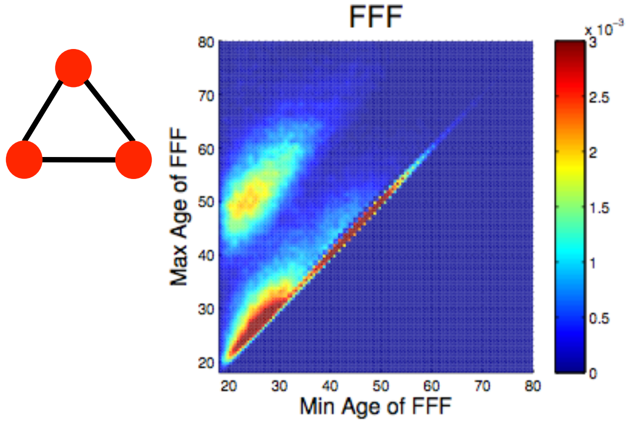
Social Strategy: Social Triad

How do people maintain their social triadic relationships across their lifetime?

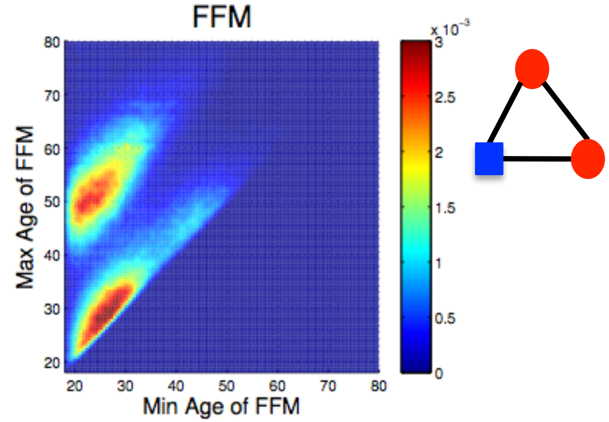


1. D. Easley, J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge U. Press. 2010

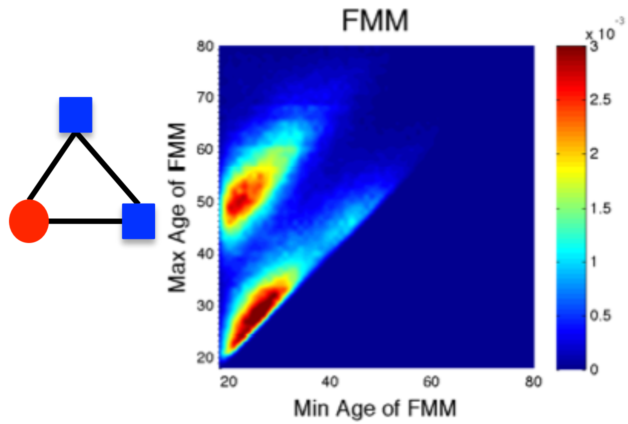
Social Strategy: Social Triad



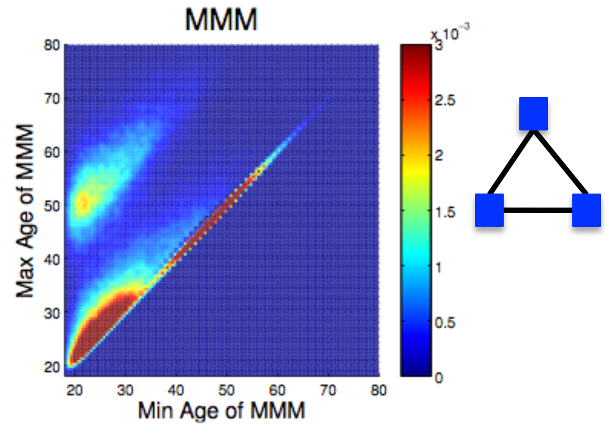
(a) Triad FFF demog. dist.



(b) Triad FFM demog. dist.



(c) Triad FMM demog. dist.

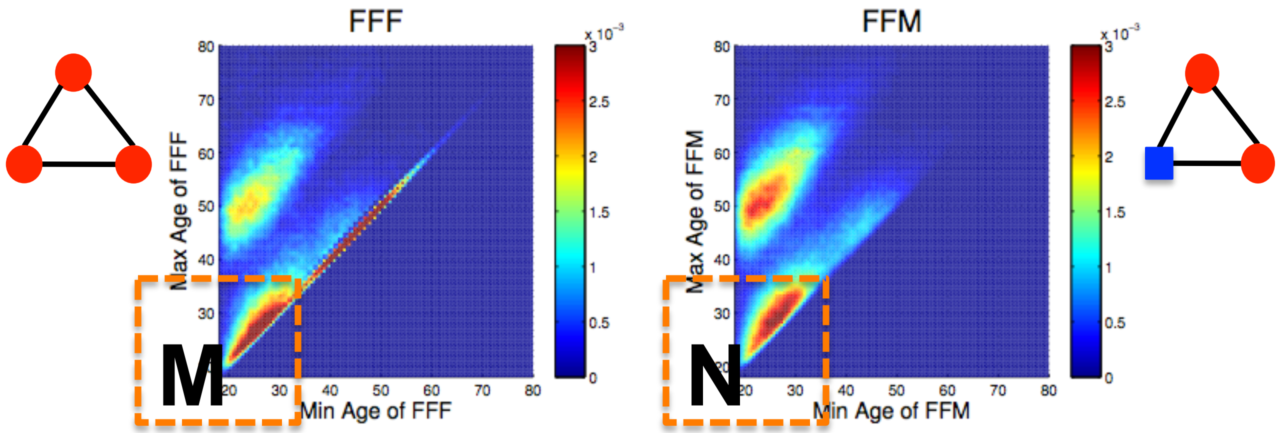


(d) Triad MMM demog. dist.

X: minimum age of 3 users.
Y: maximum age of 3 users.

Spectrum: distribution

Social Strategy: Social Triad

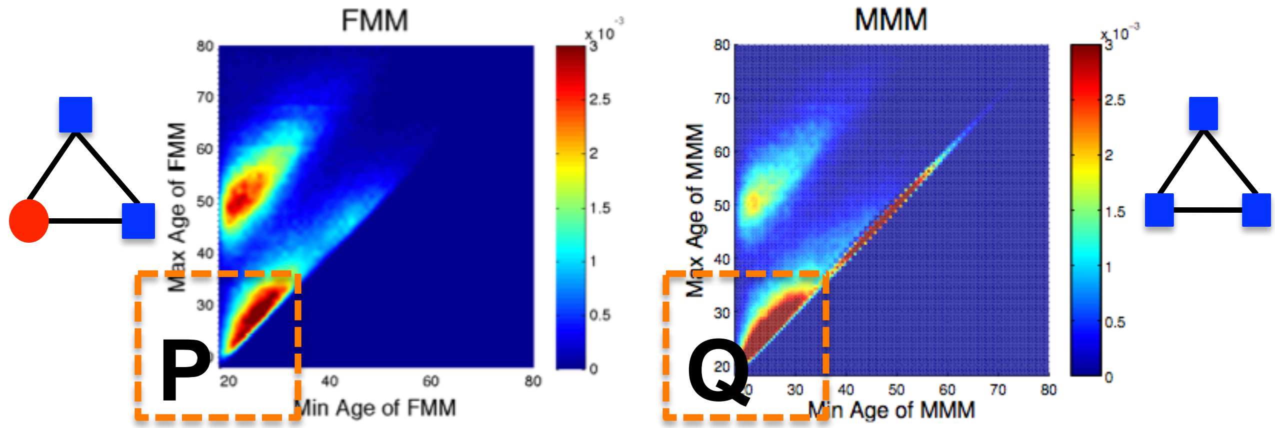


(a) Triad FFF demog. dist.

(b) Triad FFM demog. dist.

X: minimum age of 3 users.
Y: maximum age of 3 users.

Spectrum: distribution



(c) Triad FMM demog. dist.

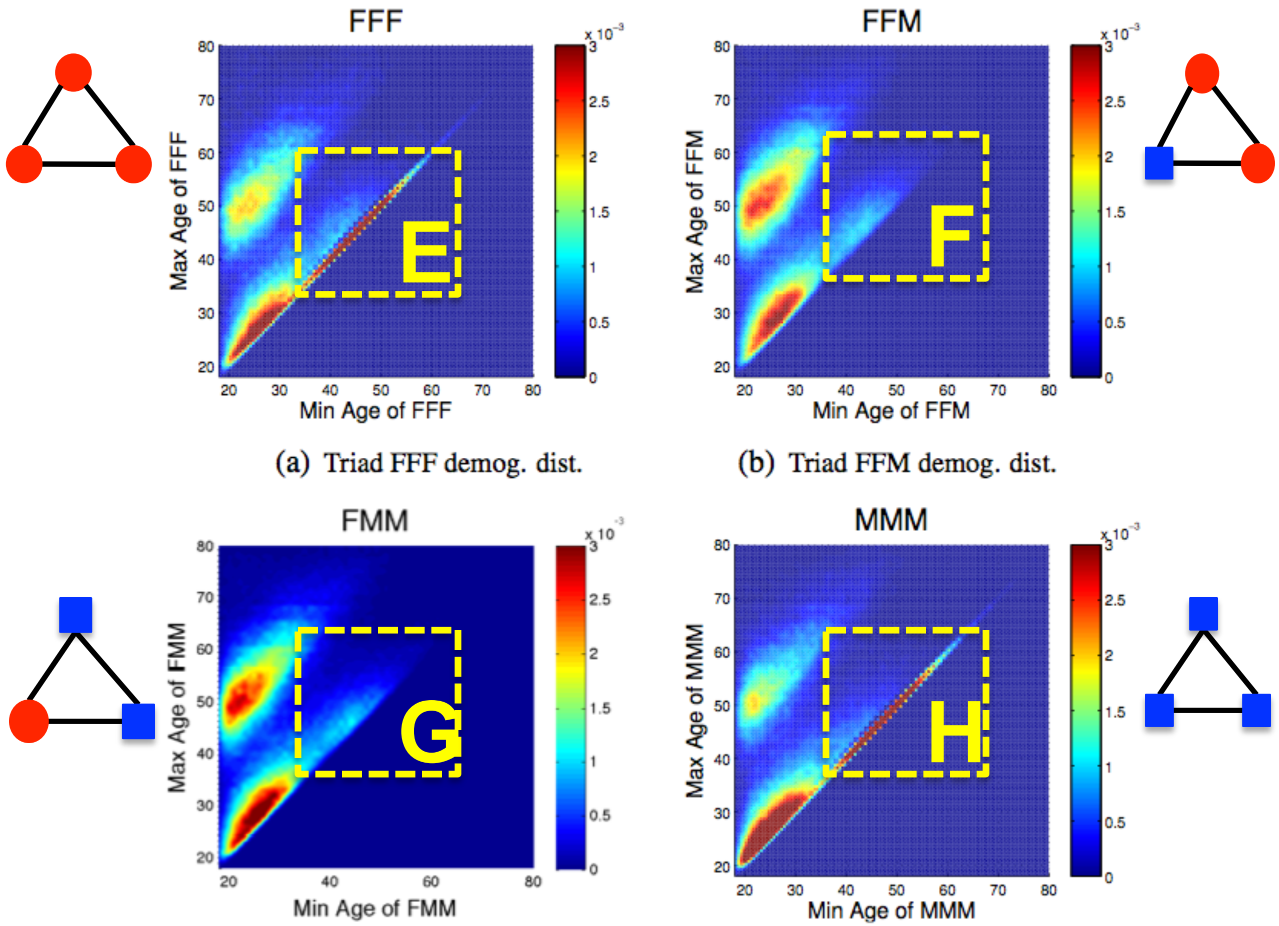
(d) Triad MMM demog. dist.

M,N,P,Q:

Intense red areas

Social Strategies: People expand both same-gender and opposite-gender social groups during the dating and reproductively active period.

Social Strategy: Social Triad



(a) Triad FFF demog. dist.

(b) Triad FFM demog. dist.

(c) Triad FMM demog. dist.

(d) Triad MMM demog. dist.

X: minimum age of 3 users.
Y: maximum age of 3 users.

Spectrum: distribution

E,H vs. F,G:

#same-gender triads are ~6 times more than #opposite-gender triads.

Social Strategies: People's attention to opposite-gender groups quickly disappears, and the insistence and social investment on same-gender social groups lasts for a lifetime.

Infer user demographics based on social strategies

social strategies + mobile social network
→ user demographics

Problem: Demographic Prediction

- Gender or Age Classification
  - Infer users' gender *Y* and age *Z* separately.
  - Model correlations between gender *Y* and attributes **X**;
  - Model correlations between age *Z* and attributes **X**;

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

Output:
$, \mathbf{X}) \rightarrow (Y^U)$

Miss the interrelation between Y and Z !

In
$G = (V^L, V^U, E, Z^L), \mathbf{X}$

Output:
$f(G, \mathbf{X}) \rightarrow (Z^U)$

Problem: Demographic Prediction

- Double Dependent-Variable Classification
  - Infer users' gender Y and age Z simultaneously.
  - Model correlations between gender Y and attributes $\mathbf{X}$;
  - Model correlations between age Z and attributes $\mathbf{X}$;
  - Model interrelations between Y and Z;

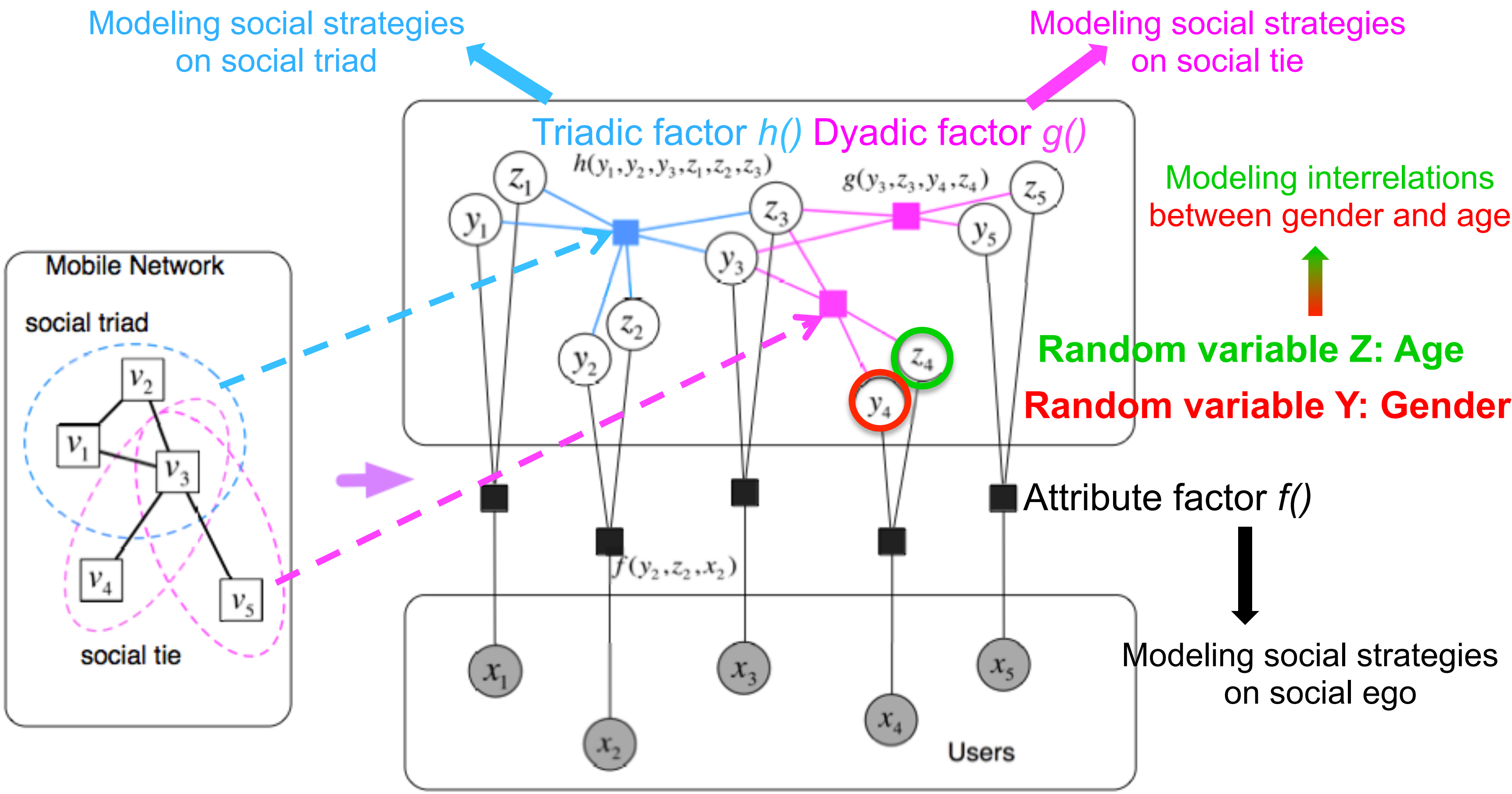
Input:
$G = (V^L, V^U, E, Y^L, Z^L), \mathbf{X}$

Output:
$f(G, \mathbf{X}) \rightarrow (Y^U, Z^U)$

- Gender:
  - Male (55%) / Female (45%)
- Age:
  - Young (18-24) / Young-Adult (25-34) / Middle-Age (35-49) / Senior (>49)

WhoAmI Method

---A double dependent-variable factor graph



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

Code is available at: http://arnetminer.org/demographic

Arnetminer

WhoAmI: Model Initialization

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

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:

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

Interrelations *between gender Y & age Z*

Triadic factor:

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

Code is available at: http://arnetminer.org/demographic

WhoAmI: Objective Function

Objective function:

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

Model learning: gradient descent

$$\frac{\partial \mathcal{O}(\theta)}{\partial \alpha} = \boldsymbol{E}[\sum_{v_i \in V} f(y_i, z_i, \mathbf{x}_i)] - \boldsymbol{E}_{P_\alpha(Y,Z|X)}[\sum_{v_i \in V} f(y_i, z_i, \mathbf{x}_i)]$$

$$\frac{\partial \mathcal{O}(\theta)}{\partial \beta} = \boldsymbol{E}[\sum_{e_{ij} \in E} g(\mathbf{y}_e, \mathbf{z}_e)] - \boldsymbol{E}_{P_\beta(Y,Z|\mathbf{X},G)}[\sum_{e_{ij} \in E} g(\mathbf{y}_e, \mathbf{z}_e)]$$

$$\frac{\partial \mathcal{O}(\theta)}{\partial \gamma} = \boldsymbol{E}[\sum_{c_{ijk} \in G} h(\mathbf{y}_c, \mathbf{z}_c)] - \boldsymbol{E}_{P_\gamma(Y,Z|\mathbf{X},G)}[\sum_{c_{ijk} \in G} h(\mathbf{y}_c, \mathbf{z}_c)]$$

Circles?→LBP[1]

1. K. P. Murphy, Y. Weiss, M. I. Jordan. Loopy Belief Propagation for Approximate Inference: An Empirical Study. UAI'99.

Code is available at: http://arnetminer.org/demographic

Experiment

Network	Method	Gender			Age		
		wPrecision	wRecall/Accu	wF1-Measure	wPrecision	wRecall/Accu	wF1-Measure
CALL	LRC						
	SVM						
	NB						
	RF						
	Bag						
	RBF						
	FGM						
	DFG						
SMS	LRC						
	SVM						
	NB						
	RF						
	Bag						
	RBF						
	FGM						
	DFG						

Data: active users (#contacts >=5 in two months)

>1.09 million users in CALL
>304 thousand users in SMS

50% as training data
50% as test data

Experiment

Network	Method	Gender			Age		
		wPrecision	wRecall/Accu	wF1-Measure	wPrecision	wRecall/Accu	wF1-Measure
CALL	LRC						
	SVM						
	NB						
	RF						
	Bag						
	RBF						
	FGM						
	DFG						
SMS	LRC						
	SVM						
	NB						
	RF						
	Bag						
	RBF						
	FGM						
	DFG						

Baselines:

LRC: Logistic Regression
SVM: Support Vector Machine
NB: Naïve Bayes
RF: Random Forest
BAG: Bagged Decision Tree
RBF: Gaussian Radial Basis Function Neural Network
FGM: Factor Graph Model

DFG: WhoAmI: Double Dependent-Variable Factor Graph

Experiment

Network	Method	Gender			Age		
		wPrecision	wRecall/Accu	wF1-Measure	wPrecision	wRecall/Accu	wF1-Measure
CALL	LRC						
	SVM						
	NB						
	RF						
	Bag						
	RBF						
	FGM						
	DFG						
SMS	LRC						
	SVM						
	NB						
	RF						
	Bag						
	RBF						
	FGM						
	DFG						

Evaluation Metrics:

- Weighted Precision
- Weighted Recall
- Weighted F1 Measure
- Accuracy

Experiment

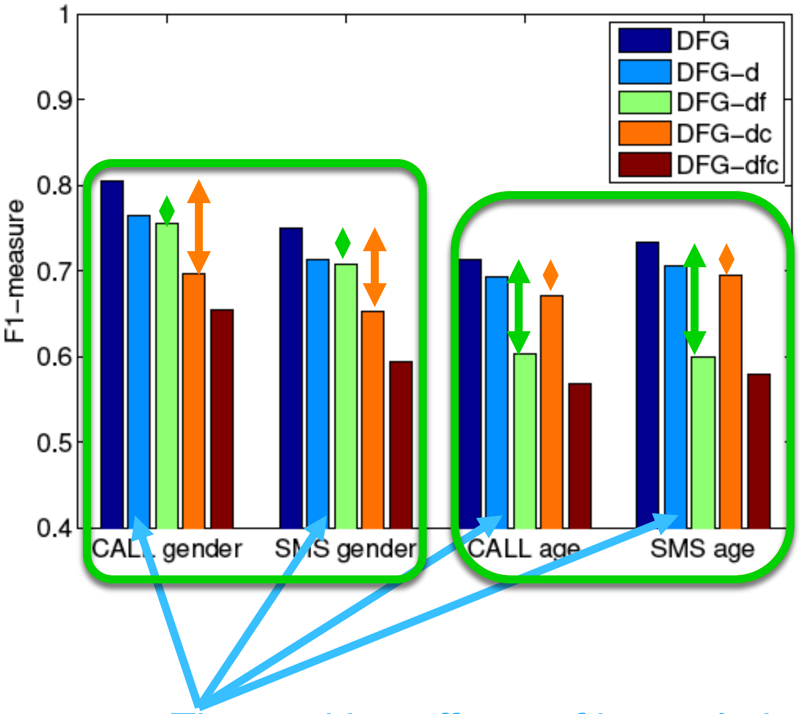
Network	Method	Gender			Age		
		wPrecision	wRecall/Accu	wF1-Measure	wPrecision	wRecall/Accu	wF1-Measure
CALL	LRC	0.7327 (0.0003)	0.7289 (0.0003)	0.7245 (0.0005)	0.6350 (0.0005)	0.6466 (0.0003)	0.6337 (0.0005)
	SVM	0.7327 (0.0004)	0.7287 (0.0003)	0.7242 (0.0003)	0.6369 (0.0004)	0.6463 (0.0005)	0.6273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

The proposed *WhoAmI* (DFG) outperforms baselines by up to 10% in terms of F1.

We can infer 80% of the users' GENDER in the CALL network correctly.
The CALL behaviors reveal more users' GENDER information than SMS.

We can infer 73% of the users' AGE in the SMS network correctly.
The SMS behaviors reveal more users' AGE information than CALL.

Experiment: Results



DFG-d: stands for ignoring the interrelations between gender and age.

DFG-df: stands for further ignoring **tie** features.

DFG-dc: stands for further ignoring **triad** features.

DFG-dcf: stands for further ignoring **tie** and **triad** features.

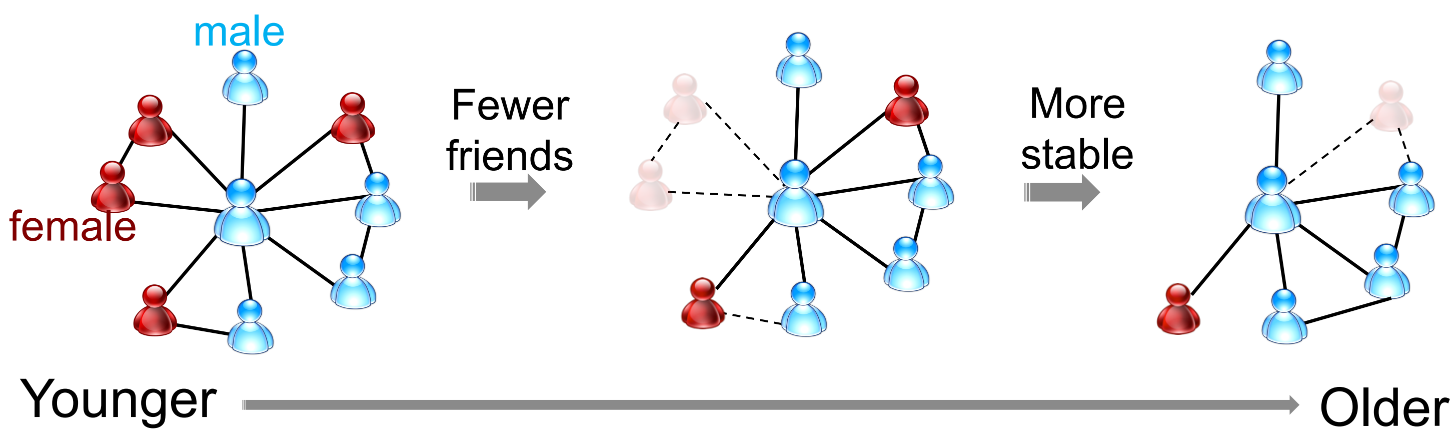
The positive effects of interrelations between gender and age.

Social Triad features are more powerful for inferring users' gender.

Social Tie features are more powerful for inferring users' age.

Conclusion

- Unveil the demographic-based social strategies used by people to meet their social needs:



- Propose ***WhoAmI***, a Double Dependent-Variable Factor Graph, for inferring users' genders and ages simultaneously.

- Demonstrate the proposed ***WhoAmI*** method in a large-scale mobile social network.

Acknowledgements

- Army Research Laboratory
- U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA)
- National High-Tech R&D Program
- Natural Science Foundation of China
- National Basic Research Program of China

Thank You!

Inferring User Demographics and Social Strategies in Mobile Social Networks

Yuxiao Dong[#], Yang Yang[+], Jie Tang[+], Yang Yang[#], Nitesh V. Chawla[#]

[#]University of Notre Dame [+]Tsinghua University

Code is available at: http://arnetminer.org/demographic

Big Network Data

- **1.26 billion** users
- **700 billion** minutes/month

- **280 million** users
- **80% of users** are 80-90's

- **555 million** users
- **.5 billion** tweets/day

- **560 million** users
- **influencing** our daily life

- **79 million** users per month
- **9.65 billion** items/year

- **500 million** users
- **35 billion** on 11/11

- **800 million** users
- **~50% revenue** from network life

Big Network Data

facebook

- **1.26 billion** users
- **700 billion** minutes/month

人人网

- **280 million** users
- **80% of users** are 80-90's

twitter

- **555 million** users
- **.5 billion** twee

$19 billion acquisition
Feb. 2014

新浪微博 weibo.com

- **560 million** users
- **influencing** our daily life

amazon.com

- **79 million** users per month
- **9.65 billion** items/year

- **500 mi** users
- **35 billion** on 11/11

WhatsApp

- **illion** users
- **revenue** from network life

Big Mobile Network Data

- **7.3 billion** mobile devices in 2014[1]
- **>100%** of global population

1. http://www.itu.int/ International Telecommunications Union (ITU) at 2013 Mobile World Congress.

Big Mobile Network Data

- In 2013, 97% of adults have a mobile phone in the US[1]
  - made 3 billion phone calls per day
  - sent 6 billion text messages per day
- This talk (15 mins):
  - 21 million calls & 42 million messages

- On average, in one day each mobile user in the US[2]
  - makes, receives or avoids 22 phone calls
  - sends or receives text messages 23 times
  - checks her/his phone 110 times.

1. http://www.accuconference.com/blog/Cell-Phone-Statistics.aspx
2. http://www.dailymail.co.uk/news/article-2276752/Mobile-users-leave-phone-minutes-check-150-times-day.html

Related work

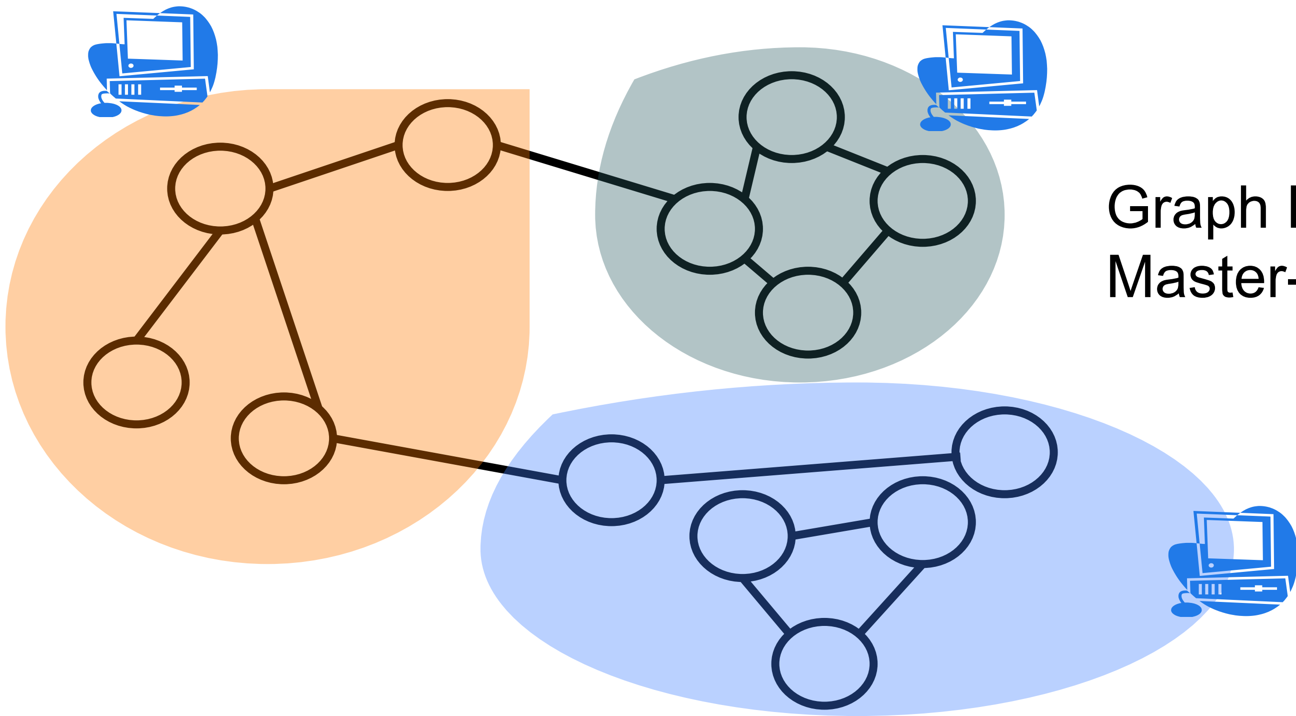
- Previous work on mobile social networks mainly focuses on macro-level models[1,2].
  - No Demographics.
- Reality Mining[3]
  - The friendship network of 100 specific users (student of faculty in MIT).
  - Demographics + Human interactions.
- The 2012 Nokia Mobile Data Challenge[4]
  - Infer user demographics by using communication records of 200 users.

1. J.P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabasi. Structure and tie strengths in mobile communication networks. PNAS 2007.
2. M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, J. Leskovec. Mobile call graphs: Beyond power-law and lognormal distributions. KDD'08.
3. http://realitycommons.media.mit.edu/
4. https://research.nokia.com/page/12000

WhoAmI: Distributed Learning

Master
Global update

Slave
Compute local gradient via random sampling



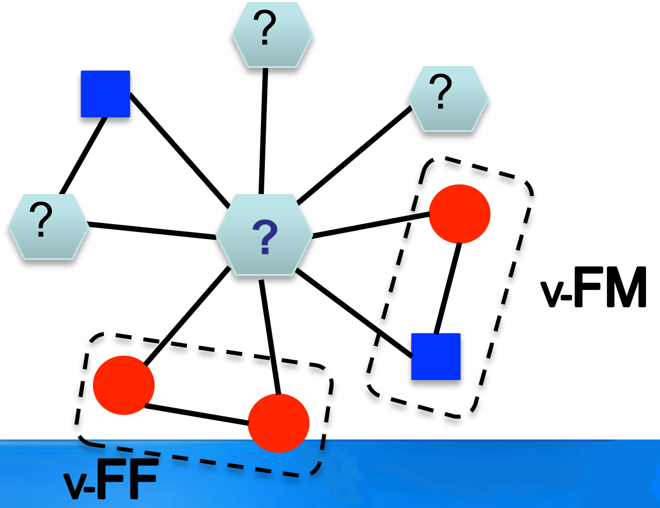
Graph Partition by Locations
Master-Slave Computing

Inevitable loss of correlation factors!

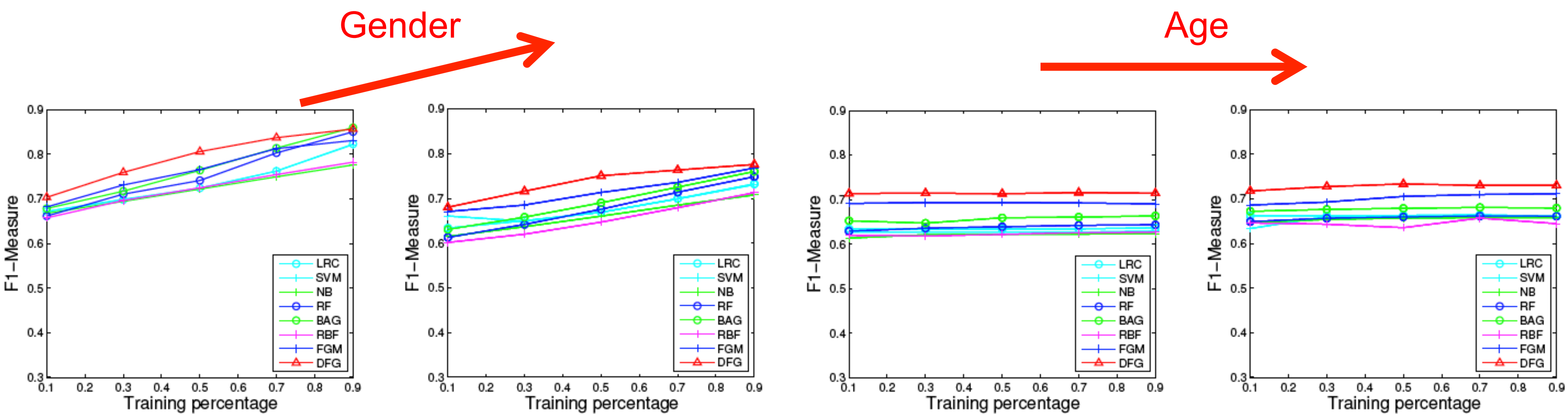
1. Jie Tang, Sen Wu, Jimeng Sun. Confluence: Conformity influence in large social networks. KDD'13.

Experiment: Features

- Given one node v and its ego network:
  - Individual feature:
    - Individual attribute: degree, neighbor connectivity, clustering coefficient, embeddedness and weighted degree.
  - Friend feature:
    - Friend attribute: # of connections to female/male, young/young-adult/middle-age/senior friends (from labeled friends).
    - Dyadic factor: both labeled and unlabeled friends for social tie structures in v's ego network.
  - Circle feature:
    - Circle attribute: # of demographic triads, i.e., v-FF, v-FM, v-MM; v-AA, v-AB, v-AC, v-AD, v-BB, v-BC, v-BD, v-CC, v-CD, v-DD. (A/B/C/C denote the young/young-adult/middle-age/senior)
    - Triadic factor: both labeled and unlabeled friends for social triad structures in v's ego network.
- LCR/SVM/NB/RF/Bag/RBF:
  - Individual/Friend/Circle Attributes
- FGM/DFG
  - Individual/Friend/Circle Attributes
  - Structure feature: Dyadic factors
  - Structure feature: Triadic factors



Experiment: Results

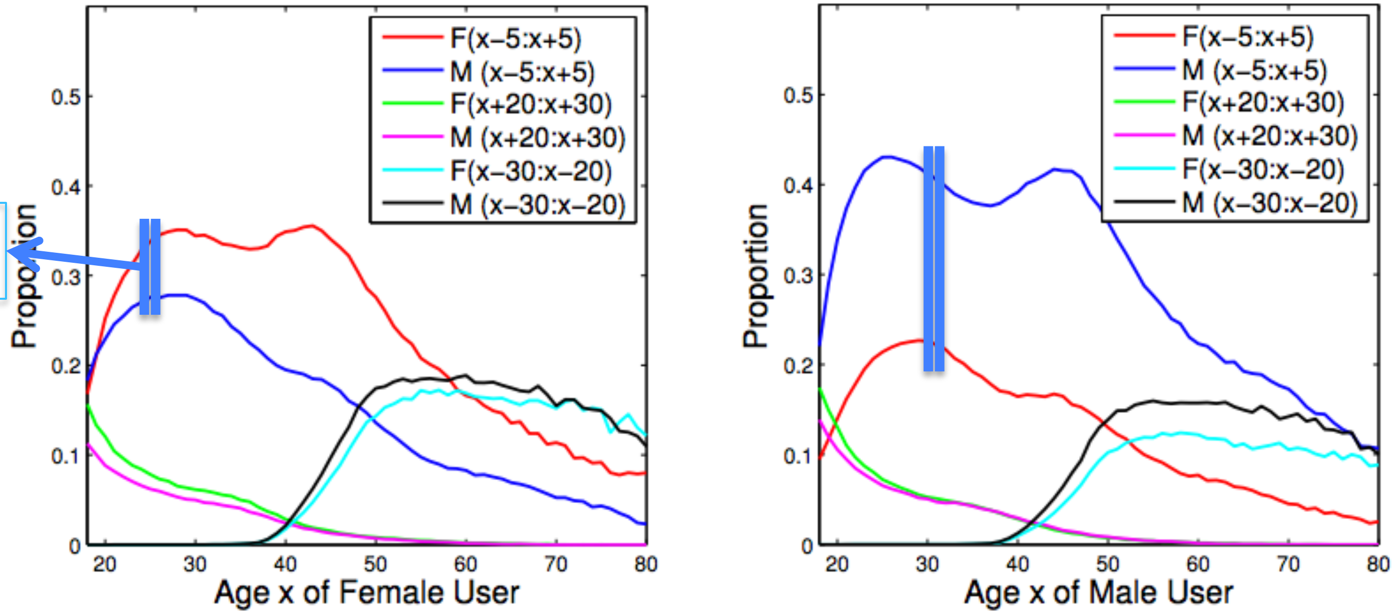


(a) CALL Gender Prediction (b) SMS Gender Prediction (c) CALL Age Prediction (d) SMS Age Prediction

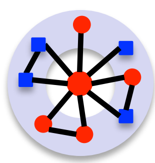
Performance of demographic prediction with different percentage of labeled data

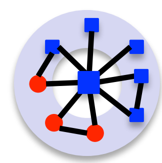
Social Strategy: Ego Network

same-generation friends



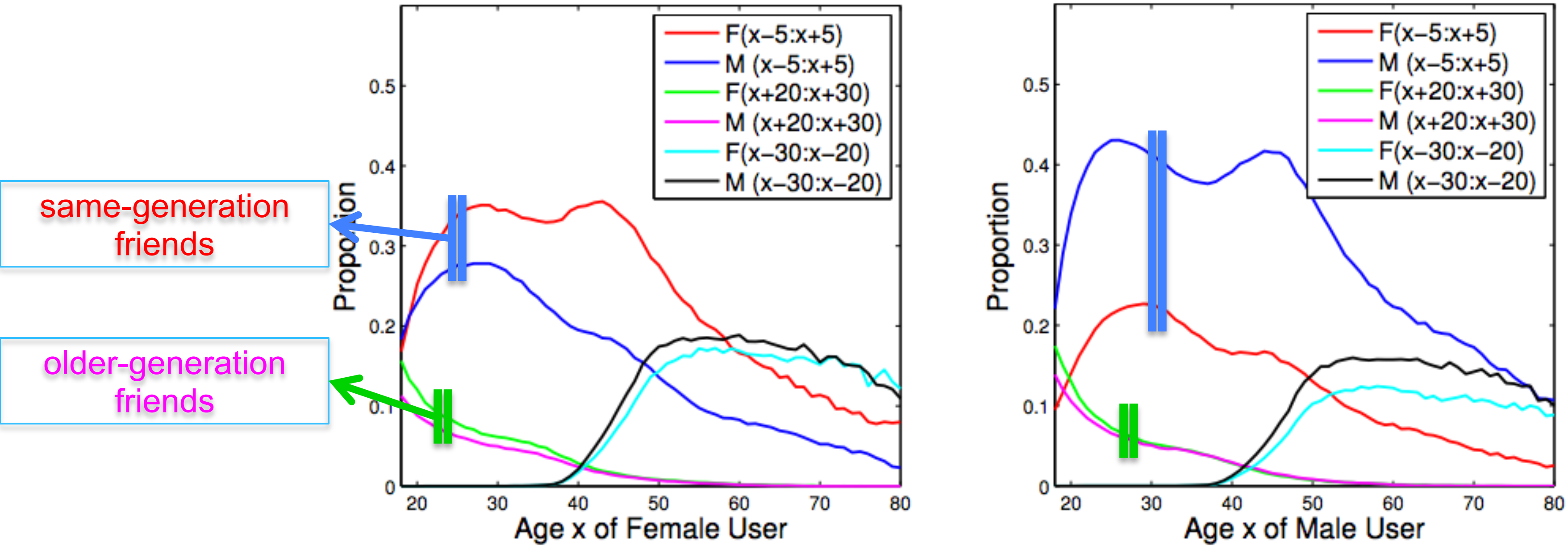
(a) Proportion of Female's friends' age (b) Proportion of Male's friends' age



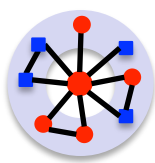


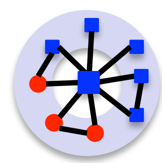
Social Strategies: The young put increasing focus on the same generation, but decrease it after entering middle-age.

Social Strategy: Ego Network



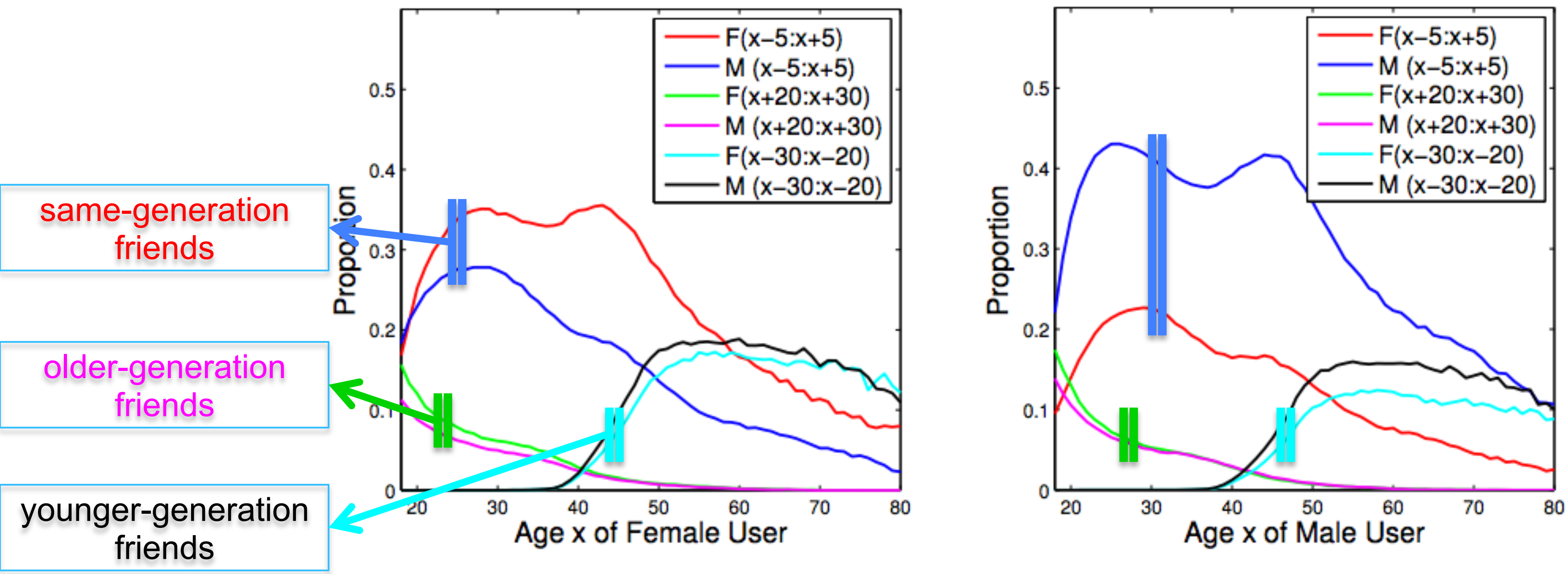
(a) Proportion of Female's friends' age (b) Proportion of Male's friends' age



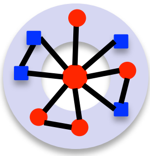


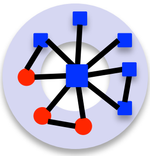
Social Strategies: The young put decreasing focus on the older generation across their lifespans.

Social Strategy: Ego Network



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





Social Strategies: The middle-age people devote more attention on the younger generation even along with the sacrifice of homophily.