@SIGIR 2018

User Modeling on Demographic Attributes in Big Mobile Social Networks

Yang Yang
Northwestern University

User Modeling on Demographic Attributes in Big Mobile Social Networks.
Yuxiao Dong, Nitesh V. Chawla, Jie Tang, Yang Yang, Yang Yang.
ACM TOIS 2017

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

The Era of Digitally Networked World

JAN 2018

DIGITAL AROUND THE WORLD IN 2018

KEY STATISTICAL INDICATORS FOR THE WORLD'S INTERNET, MOBILE, AND SOCIAL MEDIA USERS

TOTAL POPULATION	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS	UNIQUE MOBILE USERS	ACTIVE MOBILE SOCIAL USERS
7.593 BILLION	4.021 BILLION	3.196 BILLION	5.135 BILLION	2.958 BILLION
URBANISATION: 55%	PENETRATION: 53%	PENETRATION: 42%	PENETRATION: 68%	PENETRATION: 39%

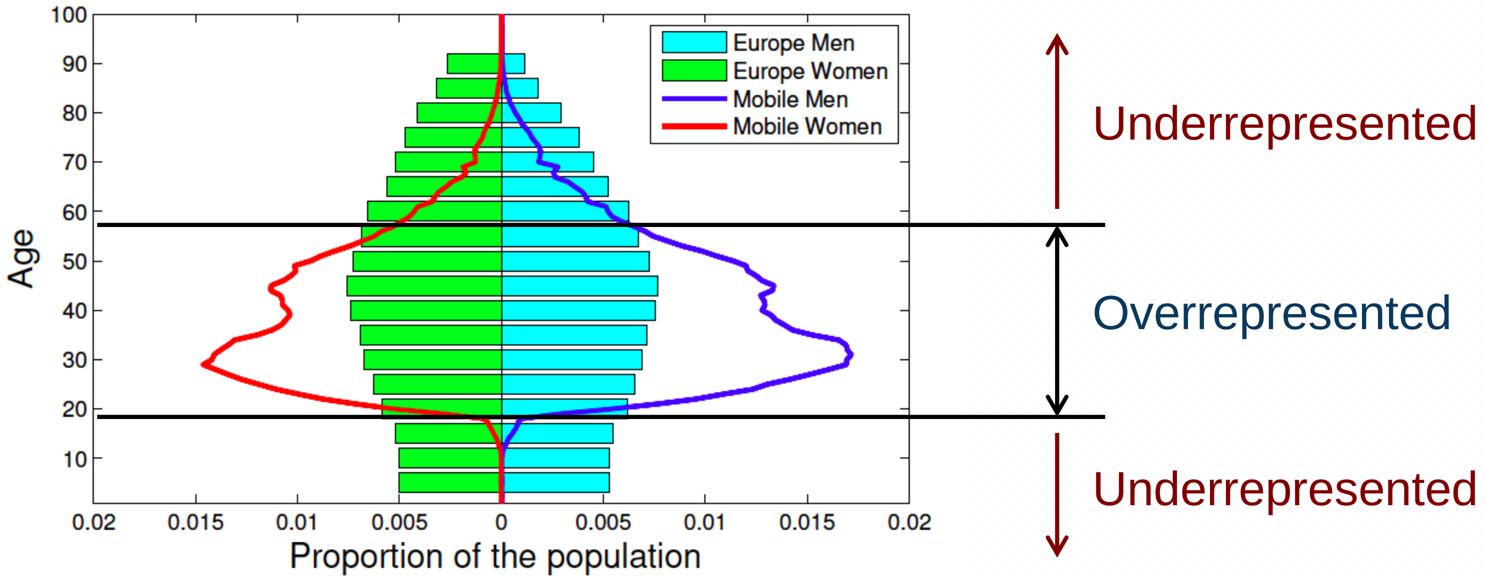
http://wearesocial.com/uk/blog/2018/01/digital-in-2018-global-overview

As of 2018, there were **5.135** billion mobile subscriptions, large global penetration.
Users average **22** calls, **23** messages, and **110** status checks per day[2].

1. http://www.dailymail.co.uk/sciencetech/article-2449632/How-check-phone-The-average-person-does-110-times-DAY-6-seconds-evening.html
2. https://www.enisa.europa.eu/media/press-releases/using-national-roaming-to-mitigate-mobile-network-outages201d-new-report-by-eu-cyber-security-agency-enisa

Big Mobile Network Data

- A **nation-wide** large mobile communication data
  - Over 7 million users: male 55% / Female 45%
  - Over 1 billion call & message records between Aug. and Sep. 2008
  - Reciprocal, undirected, and weighted networks: CALL & SMS



Europe and Mobile (CALL) population pyramids.

User Profiling on Demographics

Human Social Needs & Social Strategies

- Human needs are defined according to the existential categories of
  - being, having, doing, and **interacting**[1].

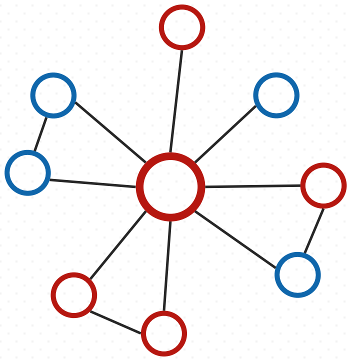
- Two basic social needs are to[2]
  - Meet new people
  - Strengthen existing relationships

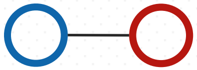
- Social strategies are used by people to meet social needs[1,2,3].
  - **What are the social strategies of people with different demographics?**
  - Demographics: **gender, age**, social status, etc.

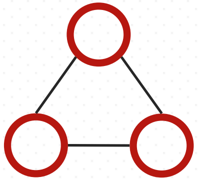
1. http://en.wikipedia.org/wiki/Fundamental_human_needs
2. M.J. Piskorski. Social strategies that work. Harvard Business Review. Nov. 2011.
3. V. Palchykov, K. Kaski, J. Kertesz, A.-L. Barabasi, R. I. M. Dunbar. Sex differences in intimate relationships. Scientific Reports 2012.

How do people of different gender and age connect & interact with each other?

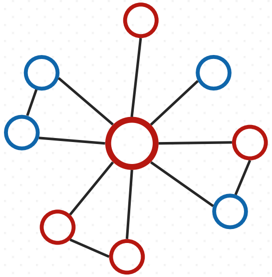
Micro: Ego, Social Tie, & Triad

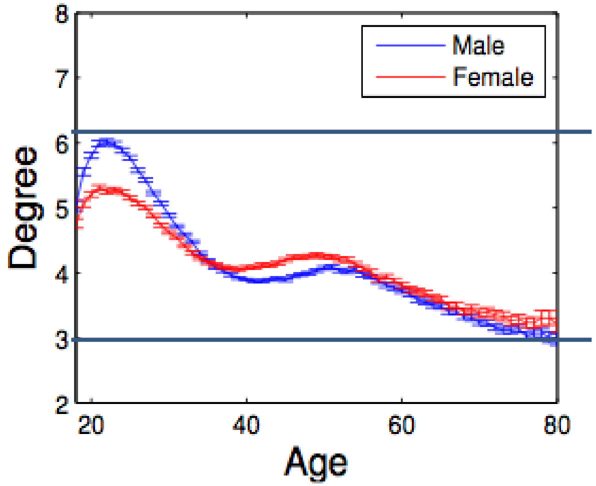




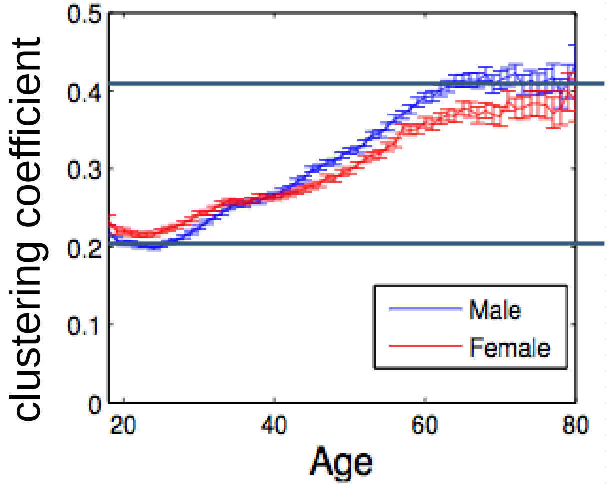


Ego Networks





(a) Degree Centrality

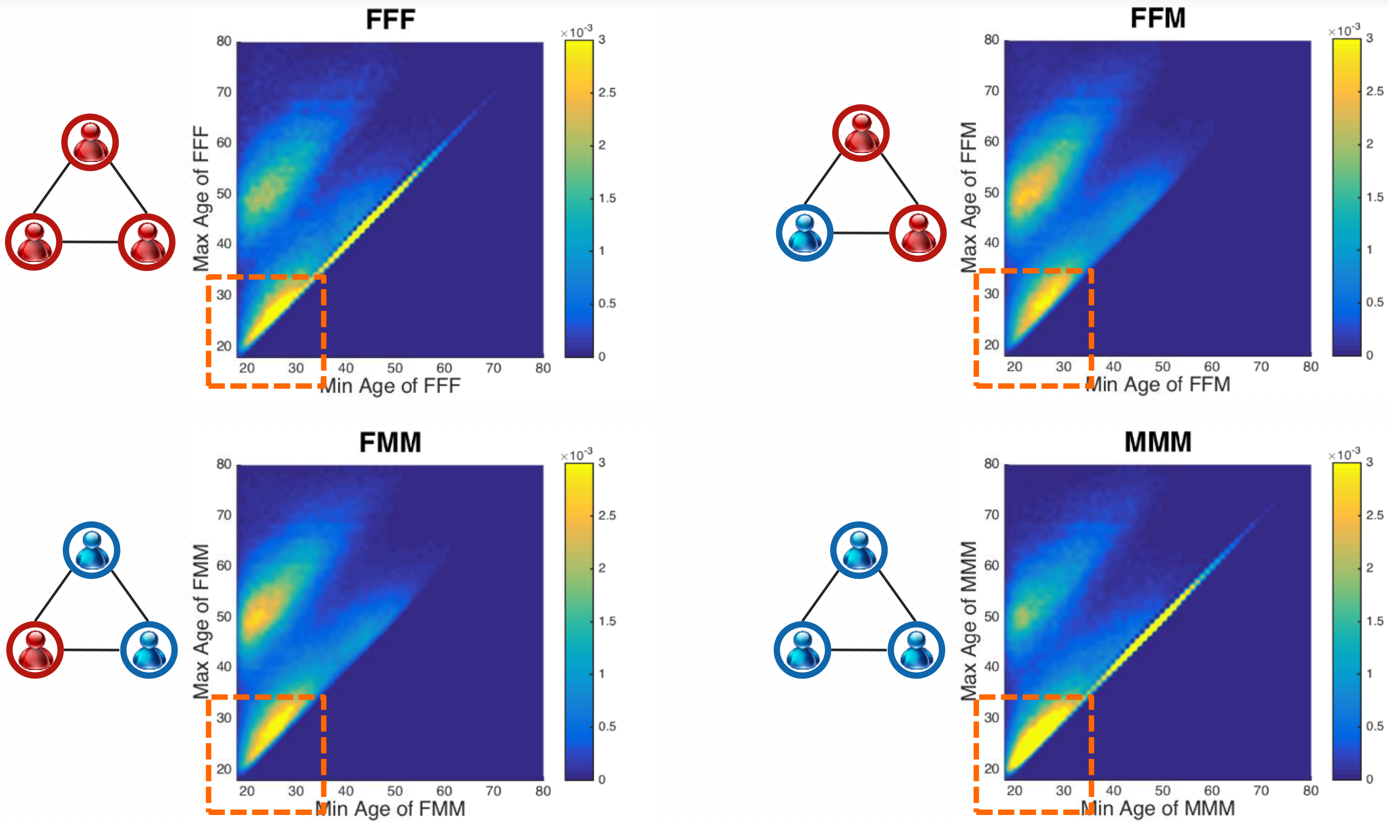


(b) Triadic Closure

- Younger people are active in broadening their social circles, while older people tend to maintain smaller but more closed connections.

Results in the CALL network, and similar observations are also found from SMS.

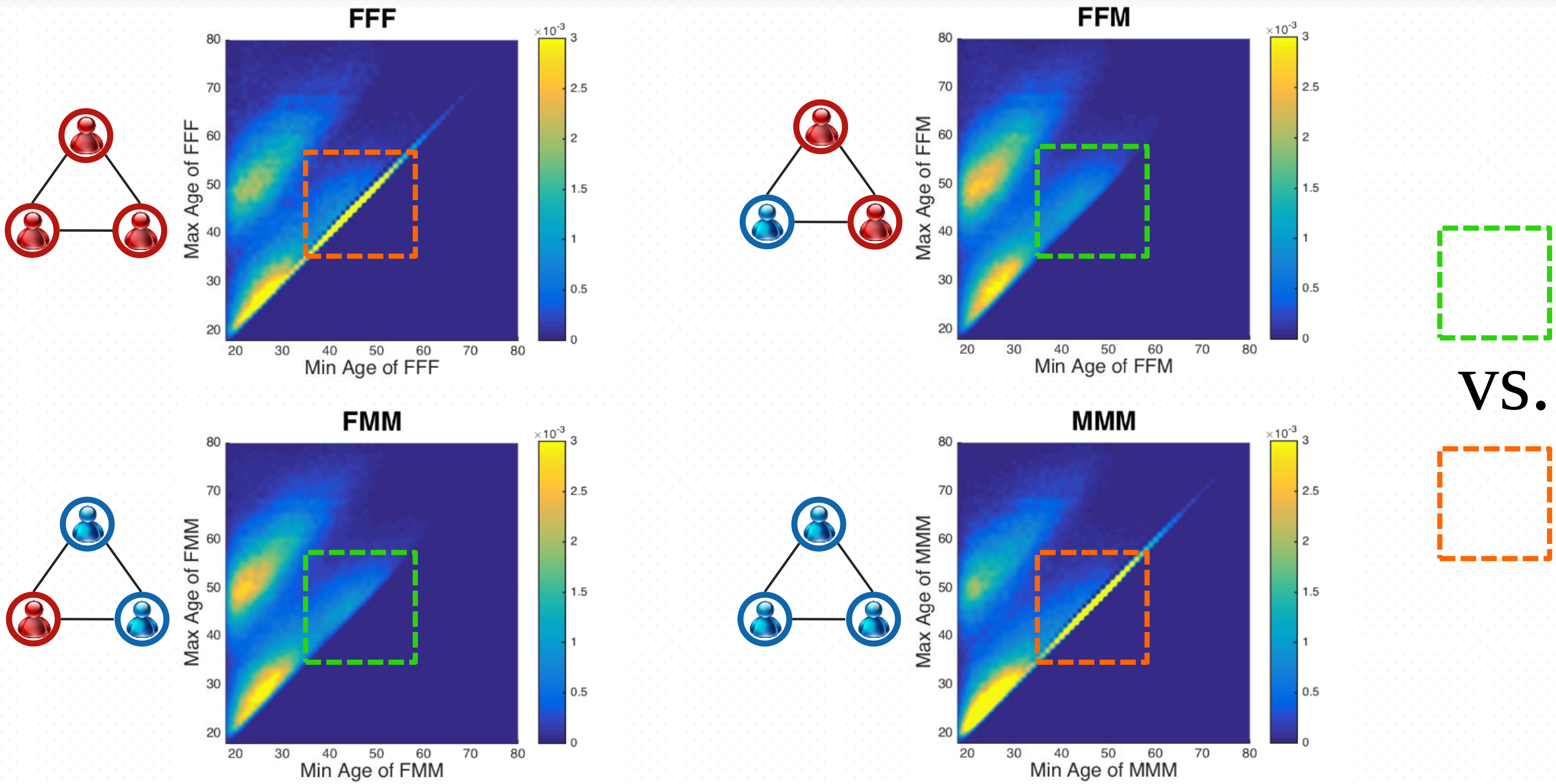
How many different triadic social circles do we have?



- People expand both same-gender and opposite-gender social groups.

Results in the CALL network, and similar observations are also found from SMS.

Demographic Triad Distribution



- The opposite-gender social groups disappear.
- The same-gender social groups last for a lifetime.

Results in the CALL network, and similar observations are also found from SMS.

Null Model

- Users' gender and age are randomly shuffled
- Randomly shuffle 10,000 times

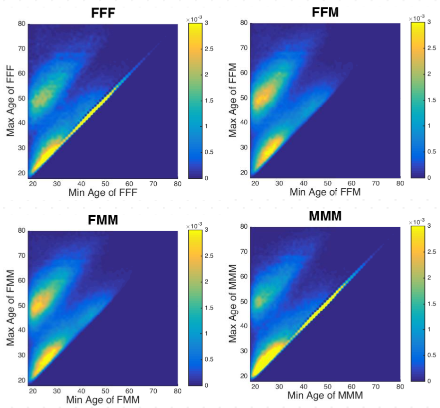
- x: empirical result from real data
- $\tilde{x}$: shuffled results
- $\mu(\tilde{x})$: the average of shuffled data
- $\sigma(\tilde{x})$: the standard deviation of shuffled data

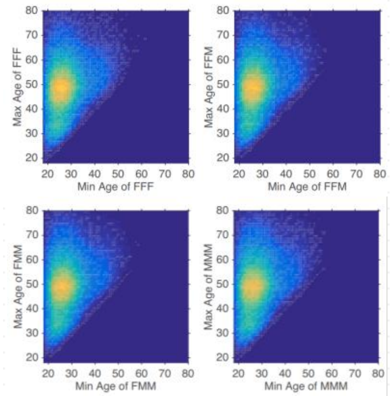
- $z(x)$: *z-score* $$z(x) = \frac{x - \mu(\tilde{x})}{\sigma(\tilde{x})}$$

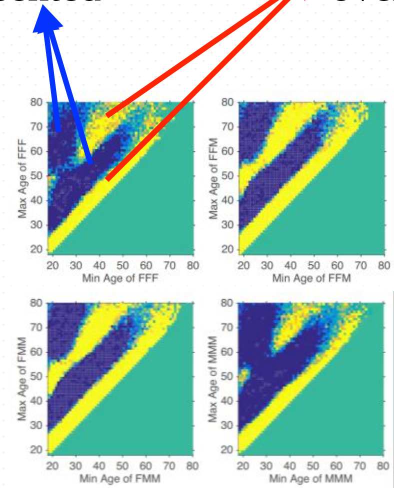
Demographic Triad Distribution

z < -3.3
underrepresented

z > 3.3
overrepresented





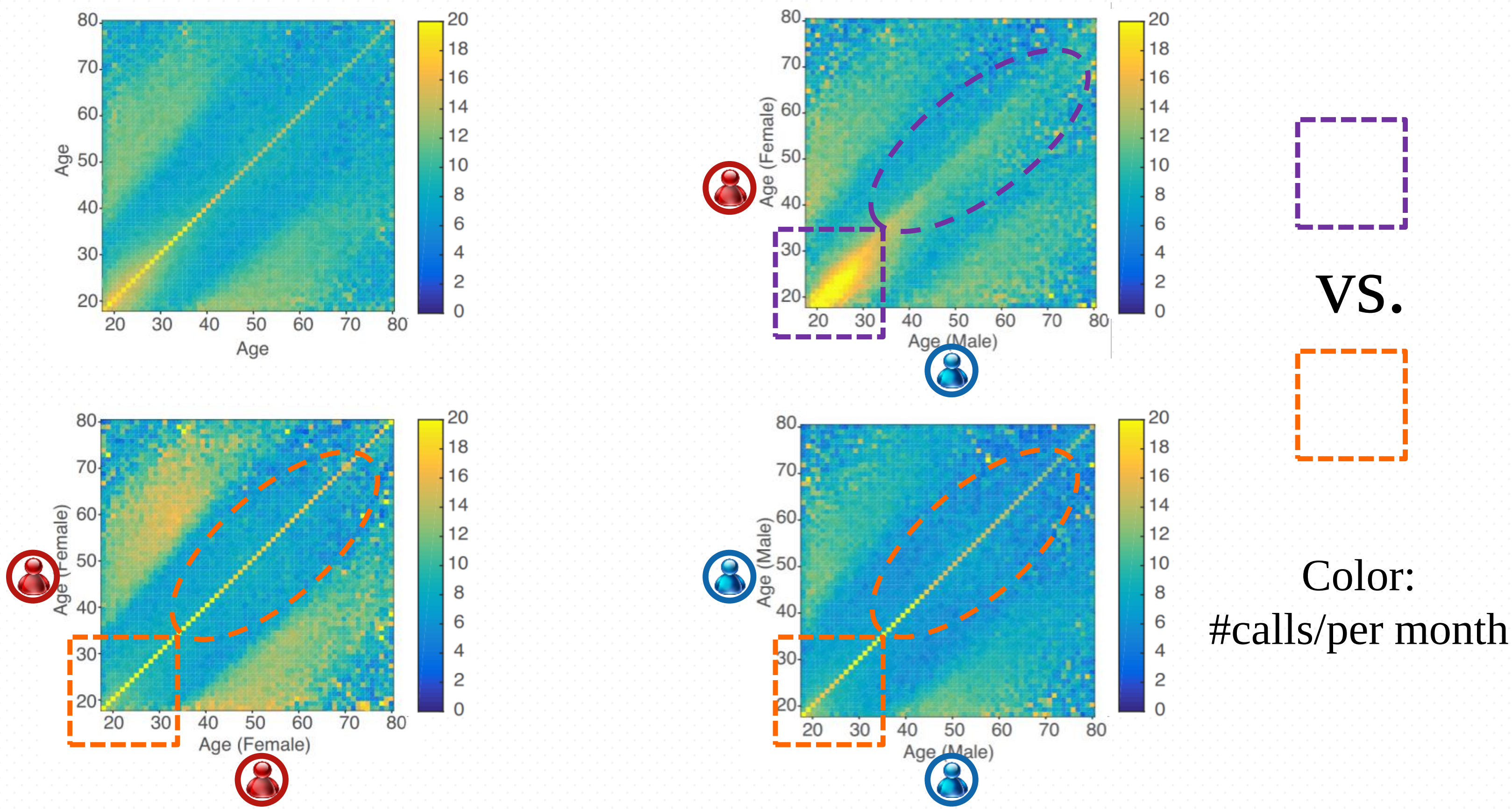


- x: empirical result from **real** data
- $\mu(\tilde{x})$: the average of **shuffled** data
- $z(x)$: *z-score*

- The results are statistically significant

Results in the CALL network, and similar observations are also found from SMS.

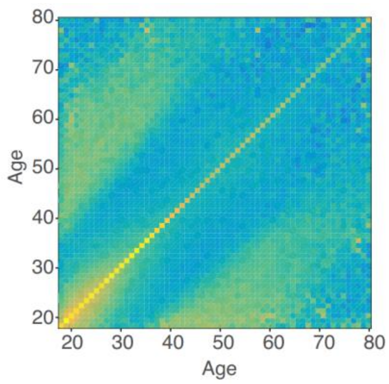
How frequently do you call your mom vs. your significant other?

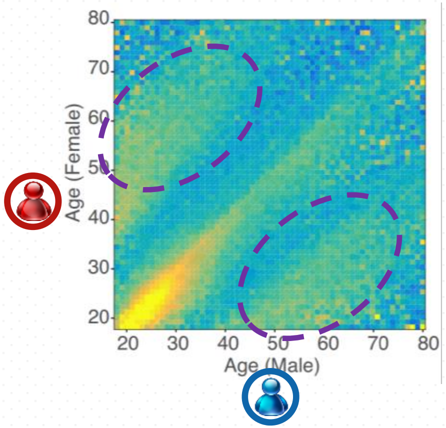


- Interactions between young girls and boys are much more frequent than those between two girls or two boys.

Results in the CALL network, and similar observations are also found from SMS.

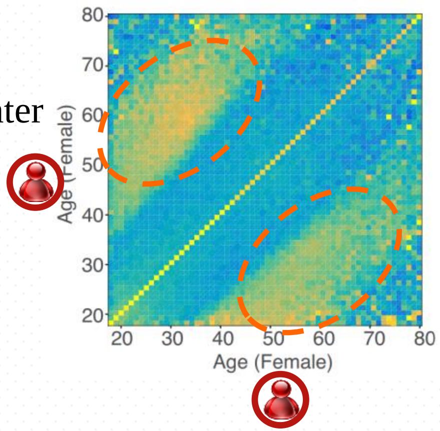
Social Tie Strength

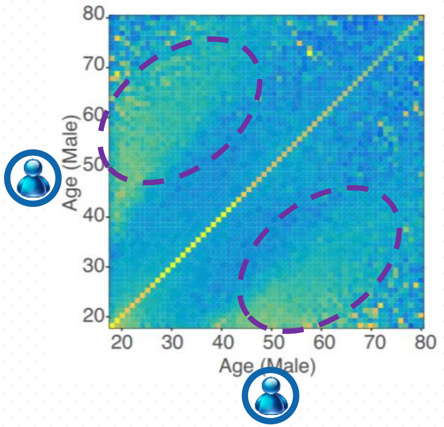




e.g., mom--son
dad--daughter

e.g., mom--daughter



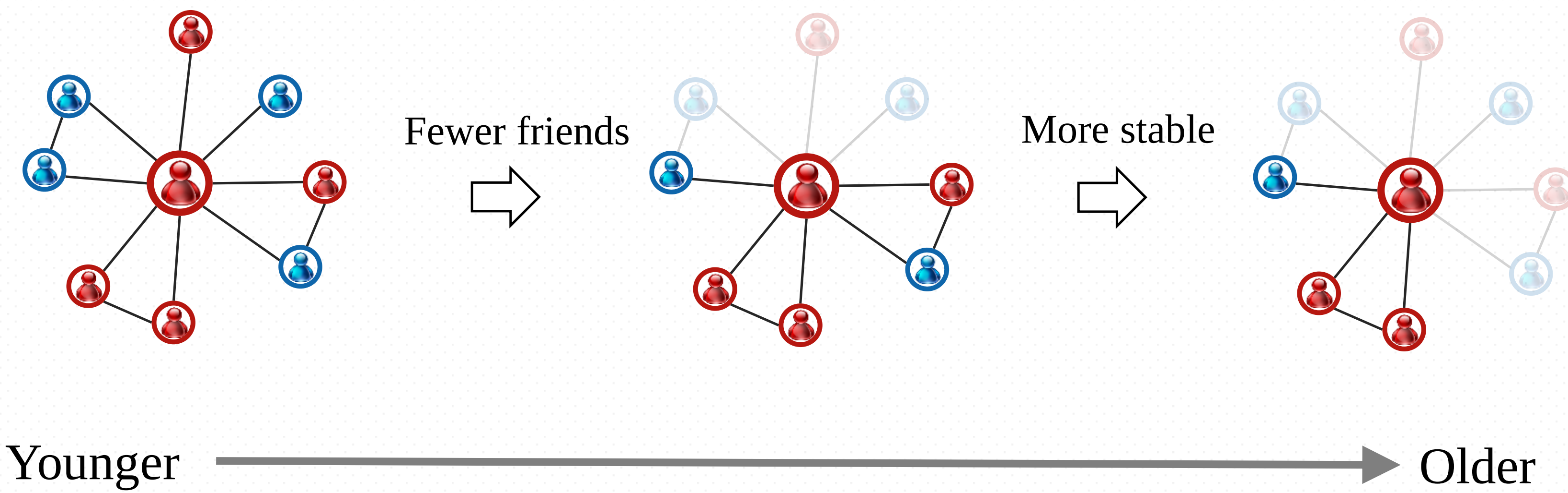


e.g., dad--son

- Cross-generation interactions between two females are more frequent than those between two males or one male and one female.

Results in the CALL network, and similar observations are also found from SMS.

Social Strategies across the Lifespan



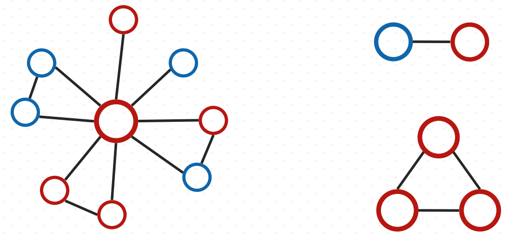
more friends: same-gender, opposite-gender

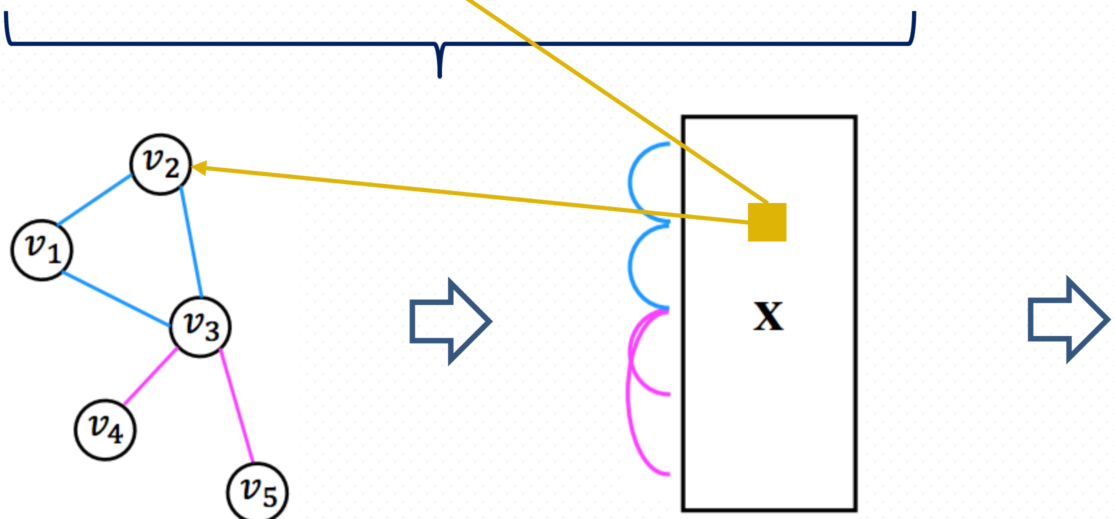
fewer friends: only same-gender, closed circles

Can we know who we are based on our social networks?

Network Mining and Learning Paradigm

Node Centralities:
- degree
- betweenness
- clustering coefficient
- PageRank
- Eigenvector
- …





hand-crafted feature matrix

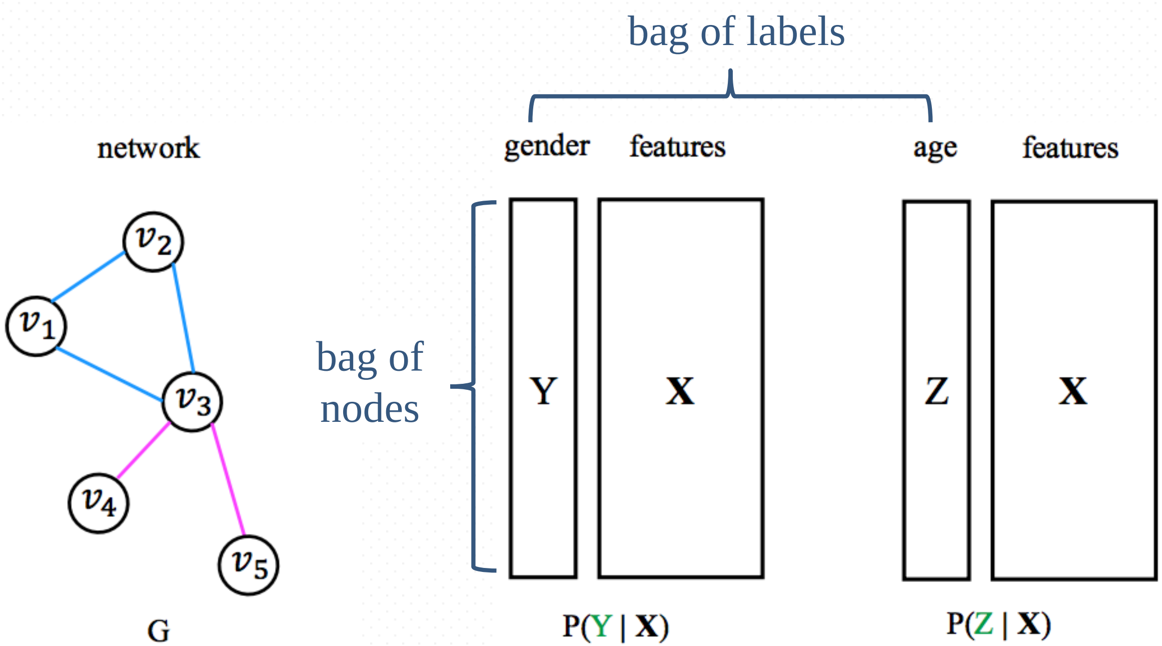
Network Mining Tasks
- **node attribute inference**
- community detection
- similarity search
- link prediction
- social recommendation
- …

feature engineering

machine learning models

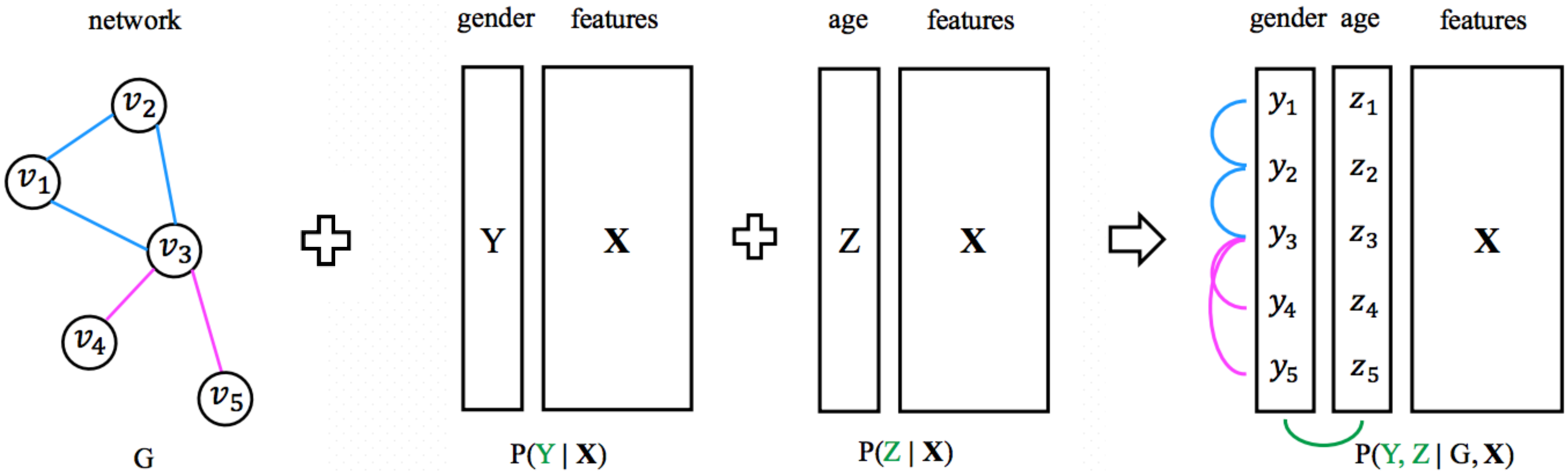
Predicting User Demographic Attributes

- Infer Users' Gender *Y* and Age *Z* Separately.
  - Model correlations between gender *Y* and attributes **X**;
  - Model correlations between age *Z* and attributes **X**;

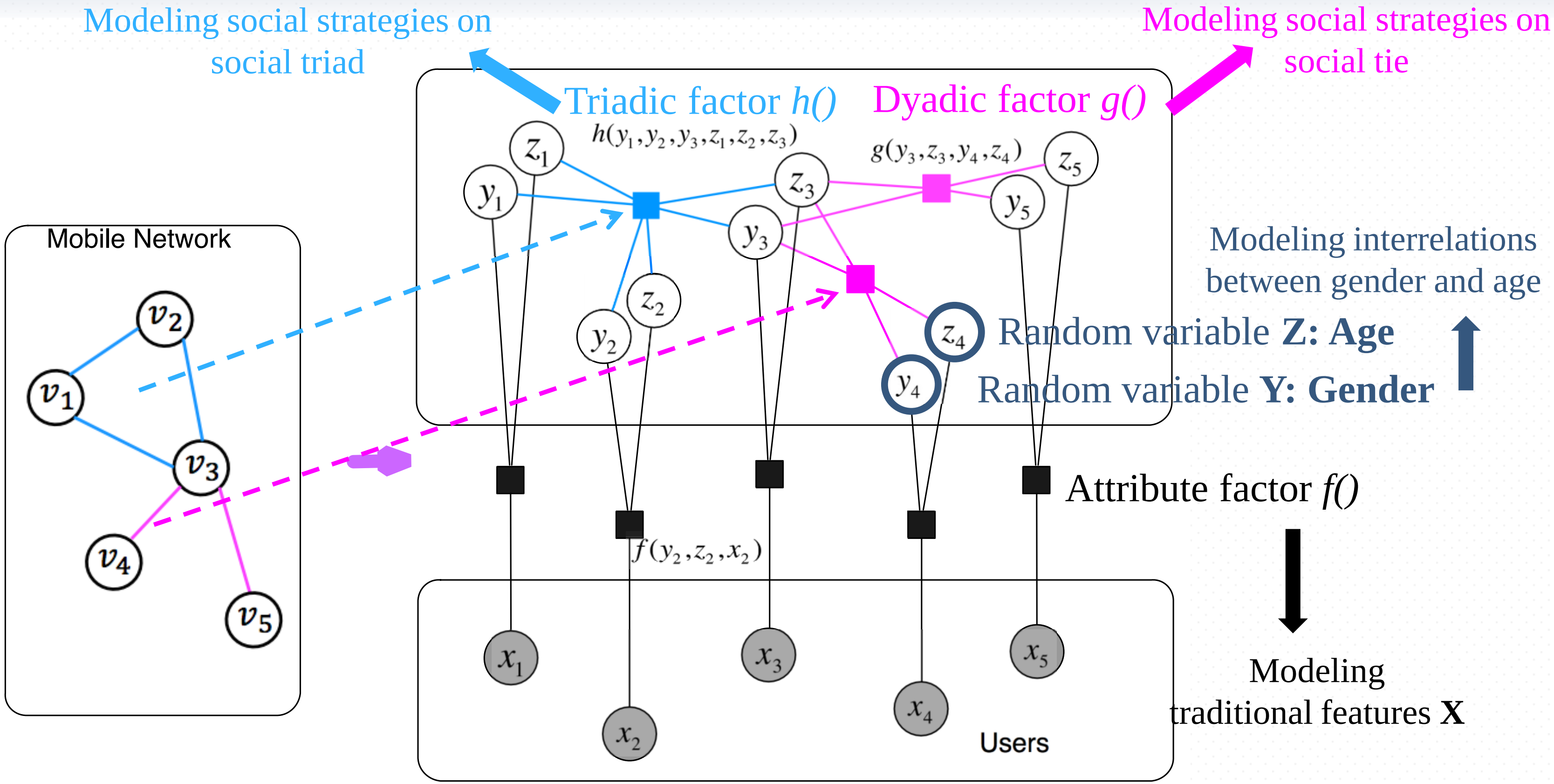


Demographic Prediction

- Infer Users' Gender Y and Age Z Simultaneously.
  - Model correlations between gender Y and attributes $\mathbf{X}$, Network G and Y;
  - Model correlations between age Z and attributes $\mathbf{X}$, Network G and Z;
  - Model interrelations between Y and Z;



WhoAmI Method



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

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

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

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

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

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

Circles?
Loopy Belief Propagation

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

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

Experiments: Feature Definition

- 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

WhoAmI: Experiments

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: mobile phone users
  - >1.09 million users in CALL
  - >304 thousand users in SMS
  - 50% as training data
  - 50% as test data
- Evaluation Metrics:
  - Weighted Precision
  - Weighted Recall
  - Weighted F1 Measure
  - Accuracy
- Baselines:
  - LRC: Logistic Regression
  - SVM: Support Vector Machine
  - NB: Naïve Bayes
  - RF: Random Forest
  - BAG: Bagged Decision Tree
  - RBF: Gaussian Radial Basis NN
  - FGM: Factor Graph Model
  - ***DFG (WhoAmI)***

Demographic Predictability

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						

- Predictability of User Demographic Profiles
  - The proposed *WhoAmI* (DFG) outperforms baselines by up to 10% in terms of F1-Measure.
  - We can infer 80% of users' gender from the CALL network
  - We can infer 73% of users' age from the SMS network
  - The phone call behavior reveals more user gender than text messaging
  - The text messaging behavior reveals more user age than phone call

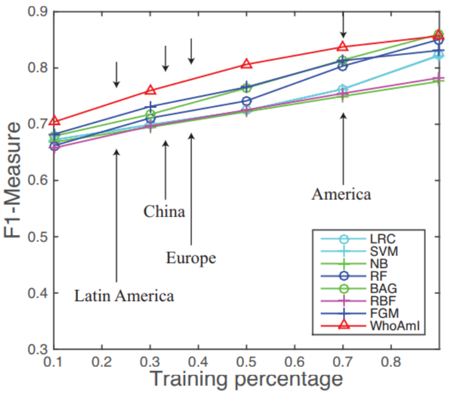
Application 1: Postpaid → Prepaid

- ***Postpaid*** mobile users are required to create an account by providing detailed demographic information (e.g., name, age, gender, etc.).

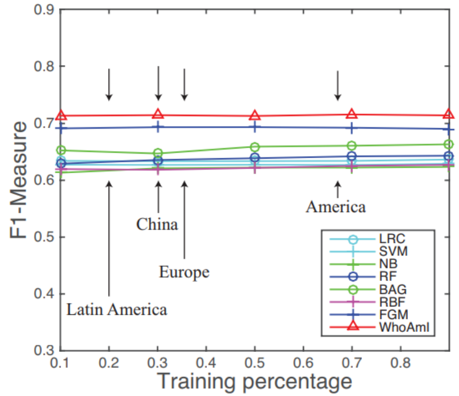
- ***Prepaid*** services (pay-as-you-go) allow users to be anonymous --- no need to provide any user-specific information.
  - 95% of mobile users in India
  - 80% of mobile users in Latin America
  - 70% of mobile users in China
  - 65% of mobile users in Europe
  - 33% of mobile users in the United States

- Train the model on postpaid users and infer prepaid users' demographics

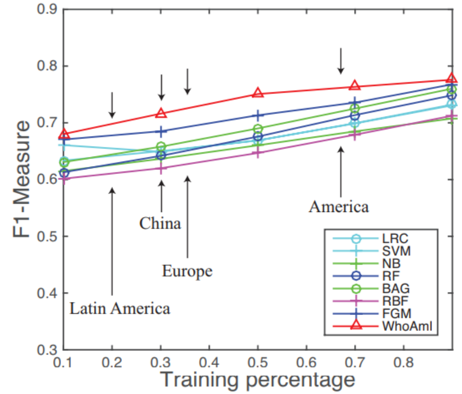
Application 1: Postpaid → Prepaid



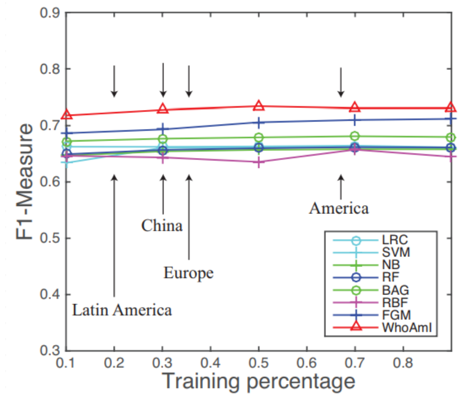
CALL Gender



CALL Age



SMS Gender



SMS Age

- Slide the training ratio to match proportion of postpaid users per country
- Train the model on postpaid users and infer prepaid users' demographics

Application 2: Coupled Networks

2015.08.08 10:30

2015.08.08 10:48

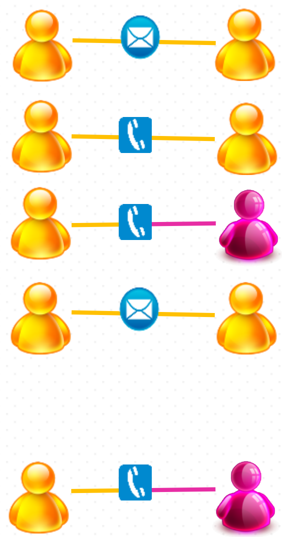
2015.08.08 11:01

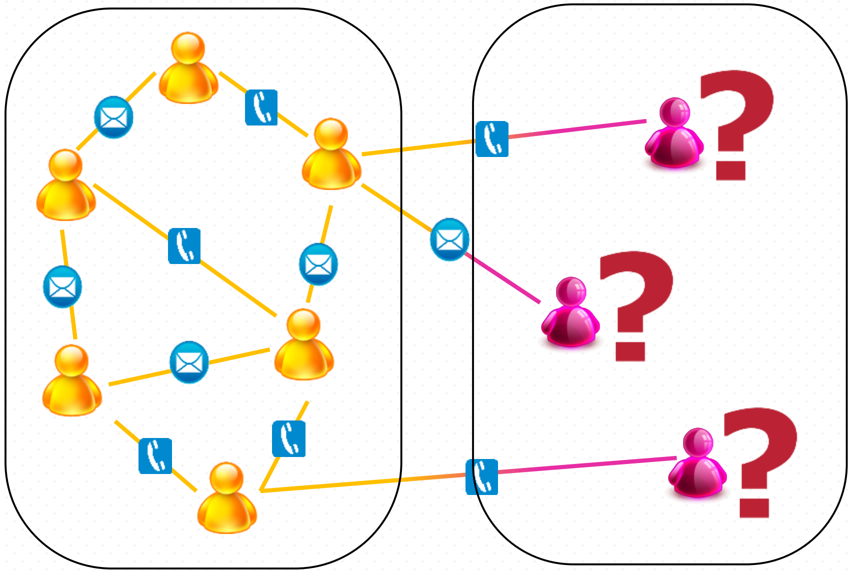
2015.08.08 11:29

……

2016.01.01 00:00

……





Coupled Demographic Prediction

Coupled Network Data

- Real-world large mobile communication data
  - Over 1 billion call & message records between Aug. to Sep. 2008
  - Undirected and weighted networks
  - Three major mobile operators E_a, E_b, E_c

	E_a	E_b	E_c	$E_a \leftrightarrow E_b$	$E_a \leftrightarrow E_c$	$E_b \leftrightarrow E_c$
#Nodes	2,531,187	655,755	354,166	1,912,933	1,255,046	625,379
#Links	3,355,197	649,322	311,432	1,844,342	1,131,593	507,894
k	2.65	1.98	1.75	1.92	1.80	1.62
cc	0.0457	0.0366	0.0317	0	0	0
ac	0.2848	0.2693	0.2806	0.0231	-0.0305	0.1113

k: average degree
cc: clustering coefficient
ac: associative coefficient

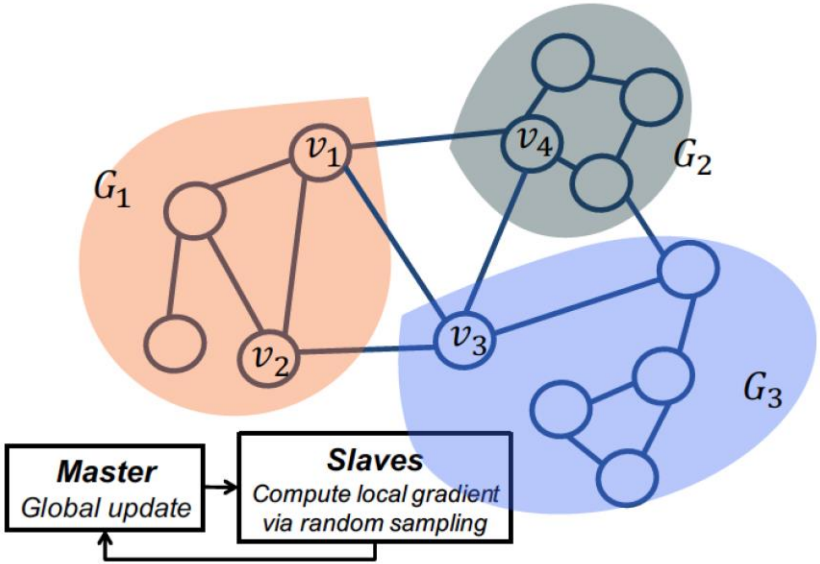
WhoAmI: Distributed Coupled Learning

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 Eq. 4.12 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 Eqs. 4.9 and 4.10;
        Slave k sends the local belief to Master [Communication];
    end
    Master calculates the marginal distribution for each variable according to Eq. 4.11;
    Master calculates the gradient for each parameter according to Eq. 4.7;
    Master updates the parameters according to Eq. 4.8;
until *Convergence*;



MPI based

Coupled Demographic Prediction

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

- Train the model on my own users and infer the demographics of my competitor' users.
- Infer 73~79% of gender information and 66~70% of age of a competitor's users.

- Discover the evolution of social strategies across lifespan
- Propose Probabilistic Graphical Model---Multi-Label Factor Graph (***WhoAmI***)---for node attribute prediction in networks
- Demonstrate the predictability of users' gender and age from mobile communication networks & two applications in telecommunications.

Thank you!

User Modeling on Demographic Attributes in Big Mobile Social Networks.
Yuxiao Dong, Nitesh V. Chawla, Jie Tang, Yang Yang, Yang Yang.
In ACM Transactions on Information Systems, 2017 (**TOIS 2017**).

Inferring User Demographics and Social Strategies in Mobile Social Networks.
Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla.
In ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2014 (**KDD'14**).

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