A DISSERTATION

Computational Lens on Big Social and Information Networks

Yuxiao Dong

Committee: Nitesh V. Chawla (Chair), David Chiang, Omar Lizardo, Zoltán Toroczkai

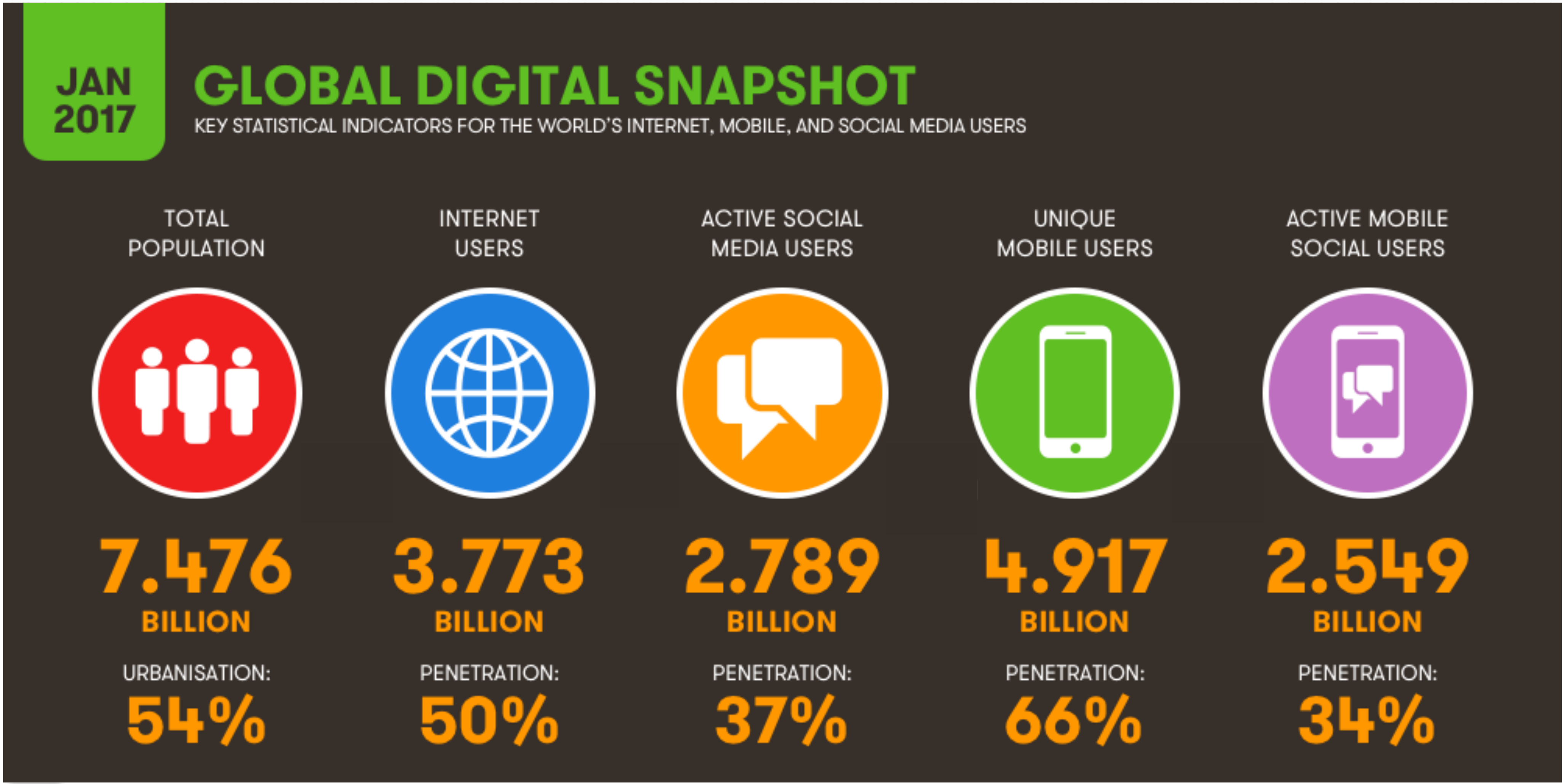
Department of Computer Science and Engineering
Interdisciplinary Center for Network Science and Applications (*iCeNSA*)
University of Notre Dame

UNIVERSITY OF NOTRE DAME

iCeNSA
Interdisciplinary Center for
Network Science & Applications

9:30AM, Thursday, Feb 09, 2017

The Era of Digitally Networked World



JAN 2017

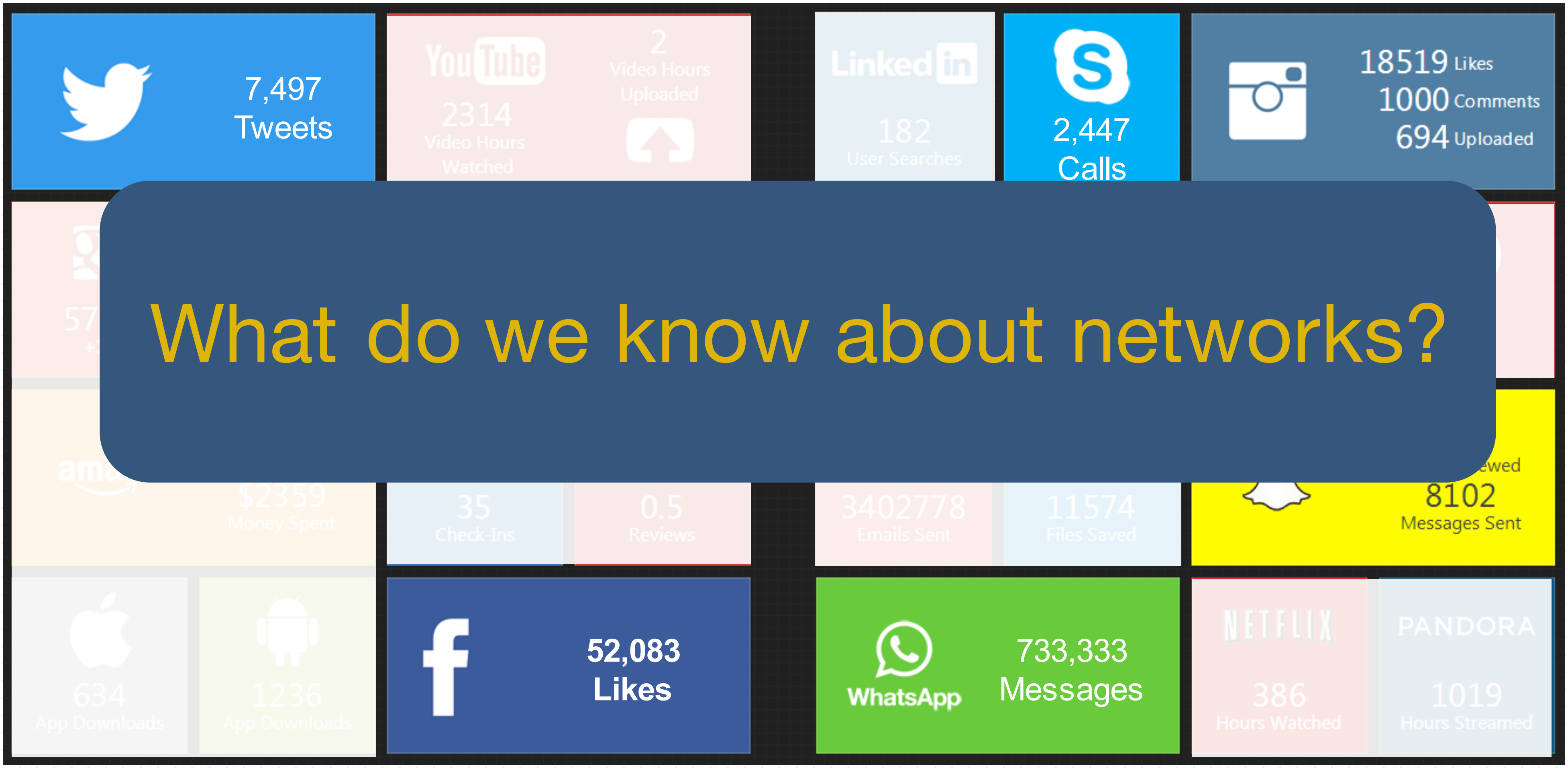
GLOBAL DIGITAL SNAPSHOT

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
7.476 BILLION	3.773 BILLION	2.789 BILLION	4.917 BILLION	2.549 BILLION
URBANISATION: 54%	PENETRATION: 50%	PENETRATION: 37%	PENETRATION: 66%	PENETRATION: 34%

http://wearesocial.com/uk/blog/2017/01/digital-in-2017-global-overview

The Era of Digitally Networked World



What do we know about networks?

As of Feb. 01, 2017. http://www.internetlivestats.com/one-second/

Network Science

- **Social Sciences:** Two-step Flow [Lazarsfeld, 1944], Homophily [Lazarsfeld & Merton, 1954], Balance Theory [Helder et al. 1958], Small World [Migram, 1960], Weak Tie [Granovetter, 1973], Dunbar's Numbers [Dunbar, 1992], Structural Hole [Burt, 1992], Cultural Network [Lizardo, 2006], Three Degree of Influence [Christakis & Fowler, 2007]

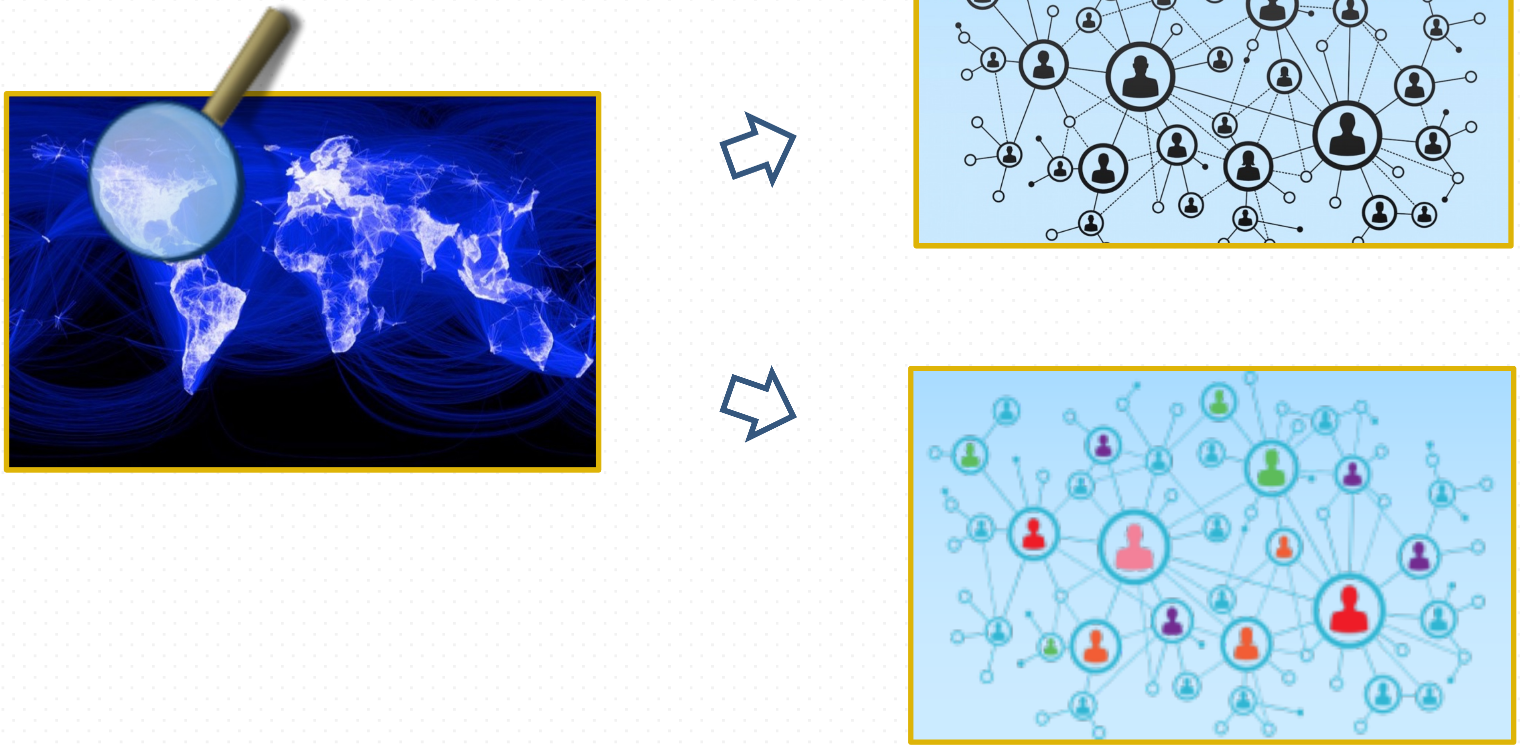
What to study about Networks?

- e n],

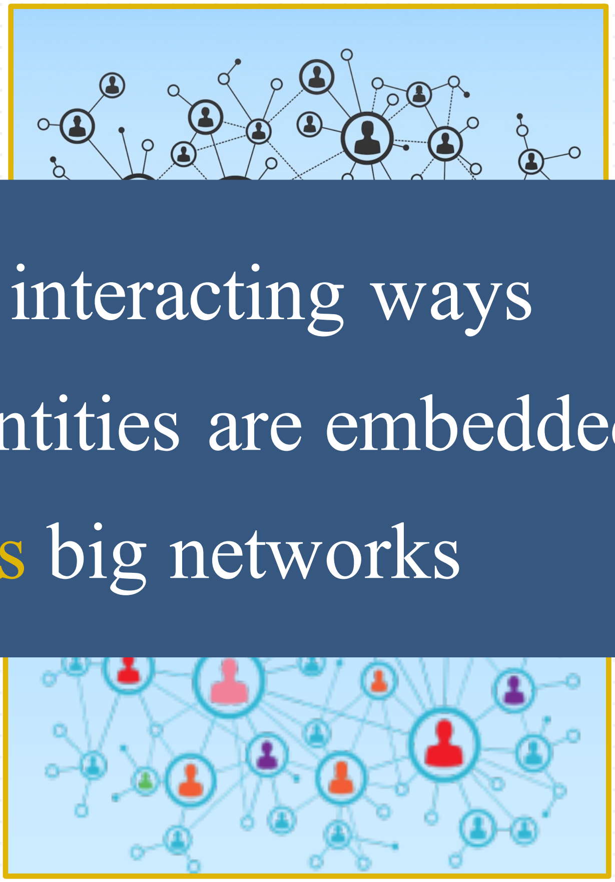
- W [Domingos & Richardson 2001 & Kempe, Kleinberg, Tardos, 2003], Link Prediction [Liben-Nowell & Kleinberg, 2003], Graph Evolution [Leskovec et. al, 2005], Network Heterogeneity [Sun et al., 2009], Four Degrees of Separation [Backstrom et al. 2012]

- **Computational Social Science** [Lazer et al. 2009, Watts 2013]

This Thesis Studies



This Thesis Studies



the diverse interacting ways

that different entities are embedded

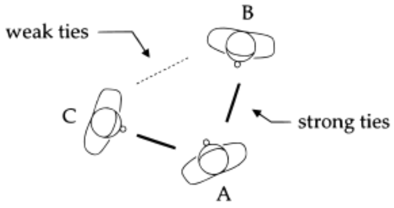
in various big networks

Computational Lens on Networks

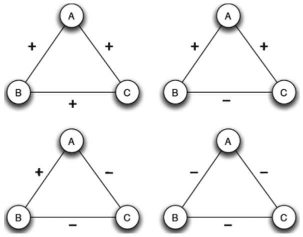
Demographics

Diversity

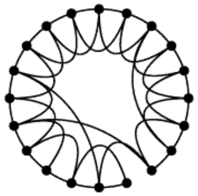
Weak/Strong Ties
[Granovetter,1973]

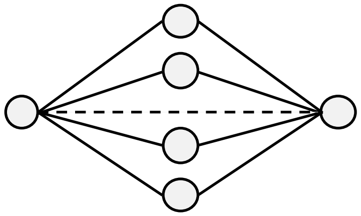


Social Balance
[Heider et al., 1958]

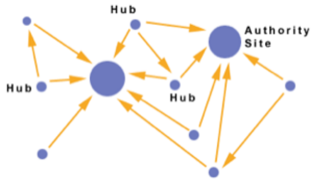


Small World
[Milgram, 1967]
[Watts, Strogatz, 1998]

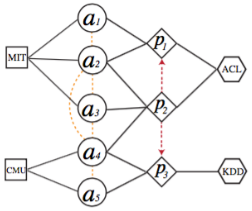




Homophily
[Lazarsfeld & Merton, 1954]

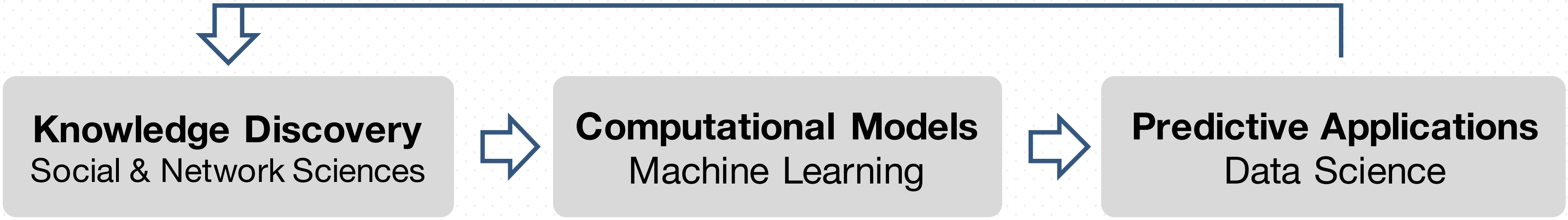


Authorities & Hubs
[Kleinberg, 1997]



Network Heterogeneity
[Sun & Han, 2012]

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

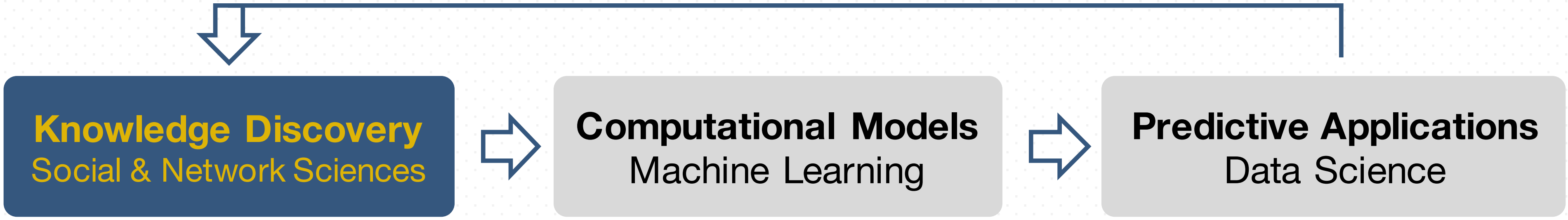
Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

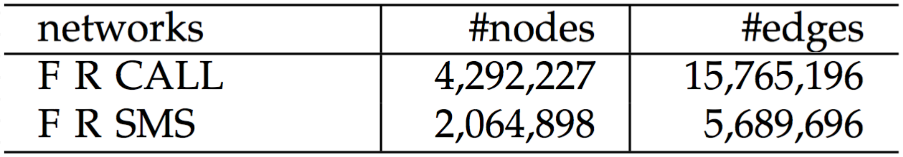
- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

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

Dong, Yang, Tang, Yang, Chawla, Inferring User Demographics and Social Strategies in Mobile Social Networks. In ***ACM KDD 2014***
Featured on United Nations Global Pulse, ND News, ACM TechNews, etc.

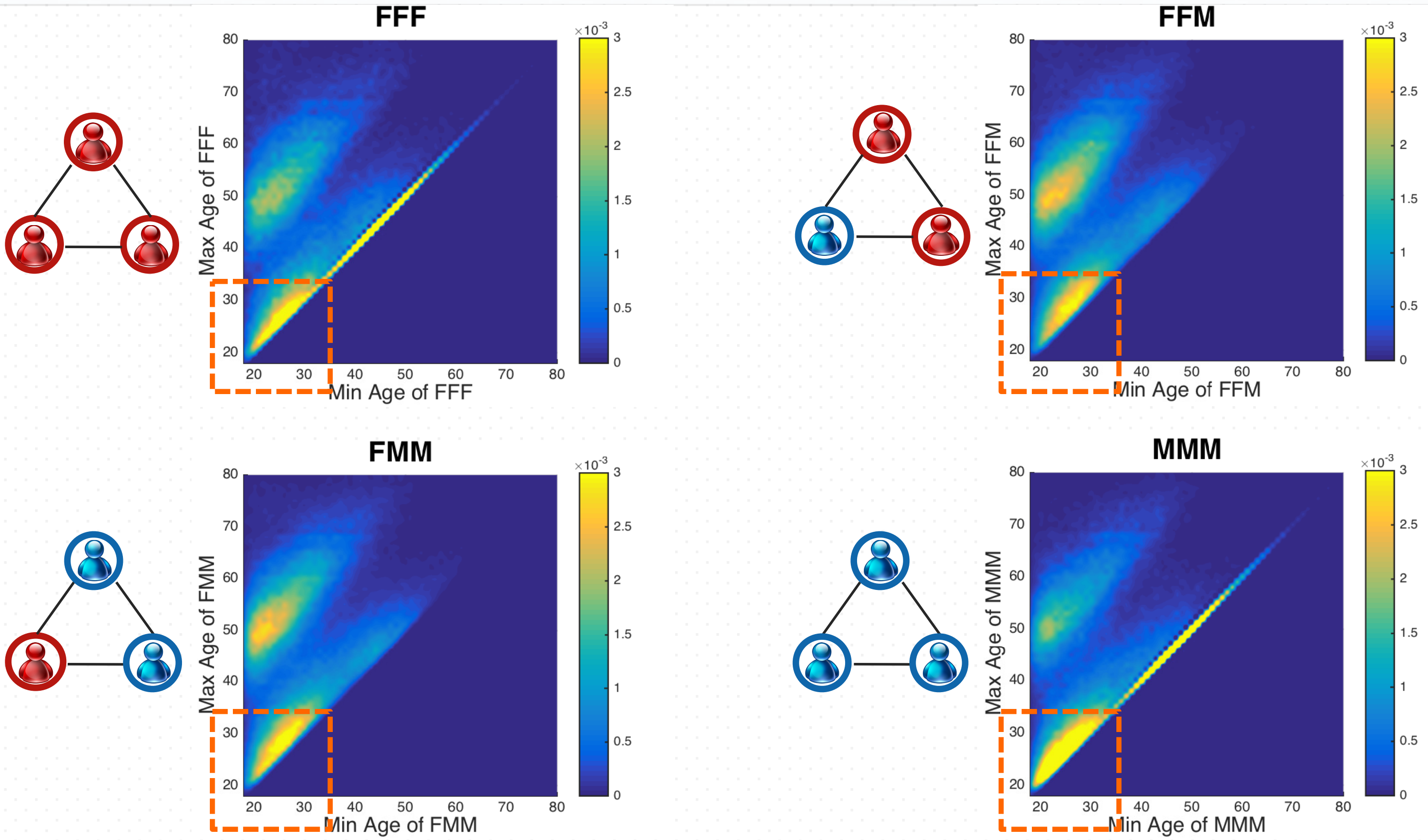
Big Mobile Network Data

- A **nation-wide** large mobile communication data
  - Over 1 billion call & message records between Aug. and Sep. 2008
  - Reciprocal, undirected, and weighted networks: CALL & SMS



networks	#nodes	#edges
F R CALL	4,292,227	15,765,196
F R SMS	2,064,898	5,689,696

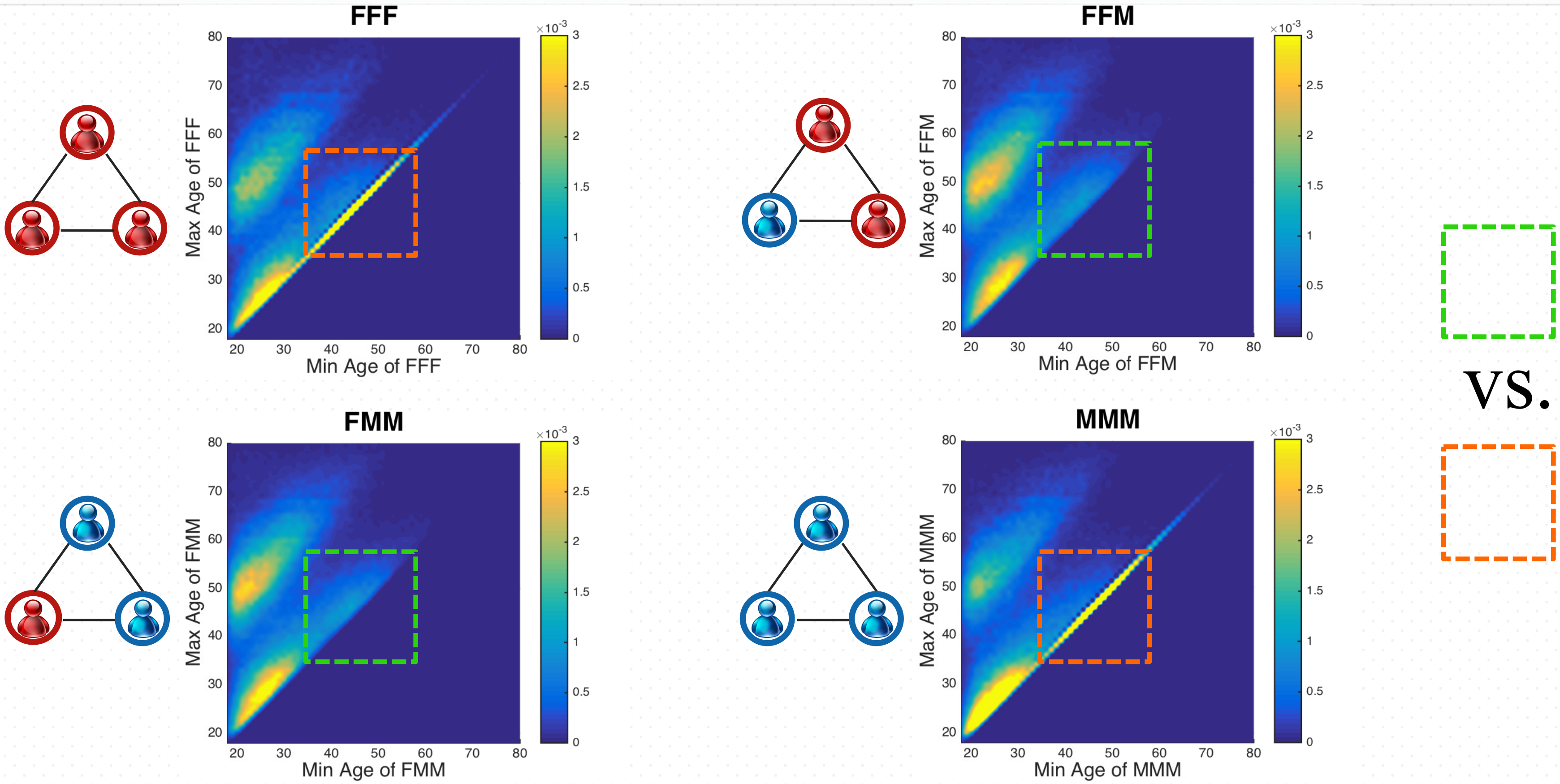
How many different triadic social circles do we have?



- ♣ 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

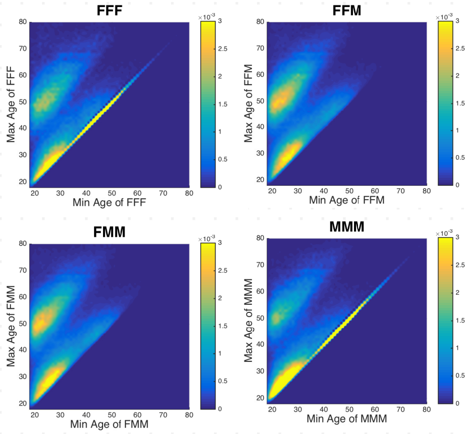
- x: empirical result from real data
- $\tilde{x}$: shuffled results
- $\mu(\tilde{x})$: the average of shuffled data
- $\sigma(\tilde{x})$: the standard deviation of shuffled data

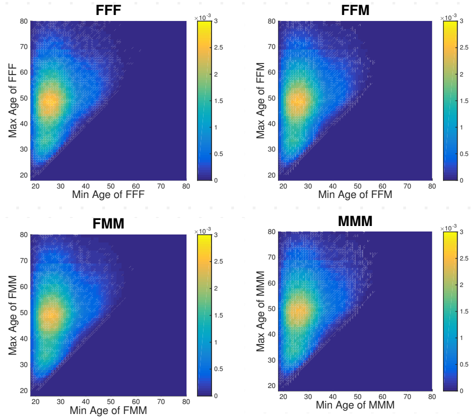
- $z(x)$: *z-score* $$z(x) = \frac{x - \mu(\tilde{x})}{\sigma(\tilde{x})}$$

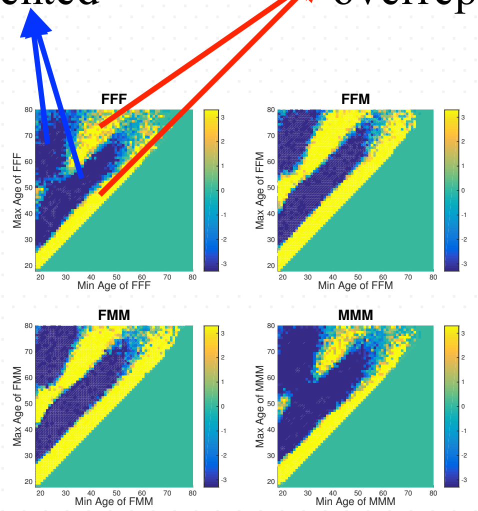
Demographic Triad Distribution

$z <$ -3.3
underrepresented

$z >$ 3.3
overrepresented





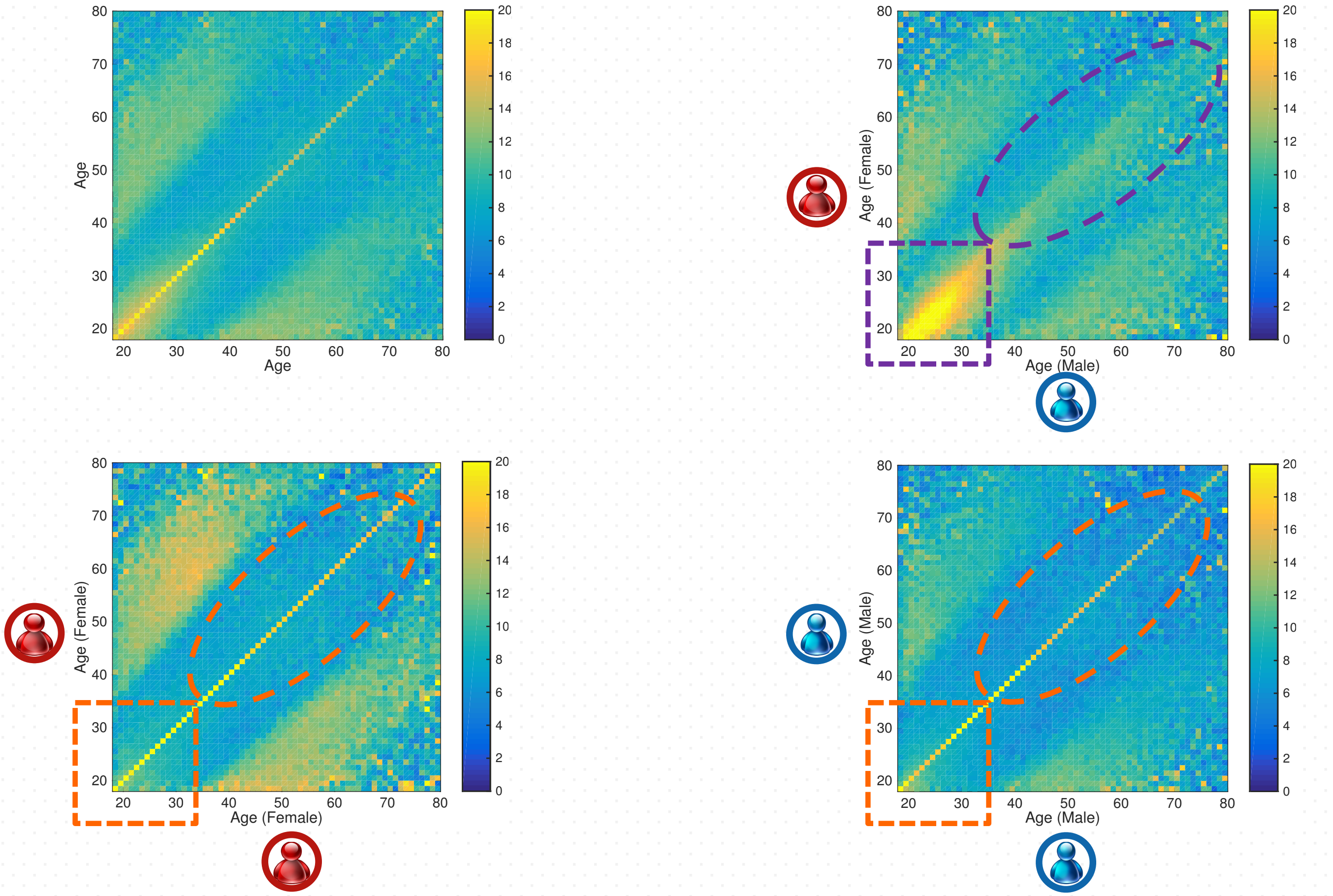


- ♣ x: empirical result from **real** data
- ♣ $\mu(\tilde{x})$: the average of **shuffled** data
- ♣ $z(x)$: *z-score*

- ♣ 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?



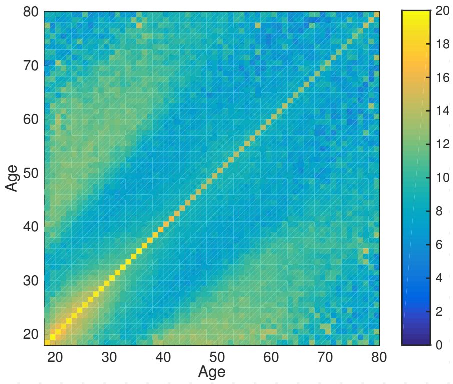
VS.

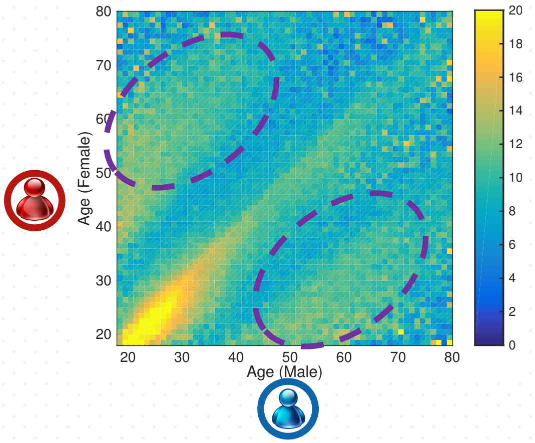
Color:
#calls/per month

- 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

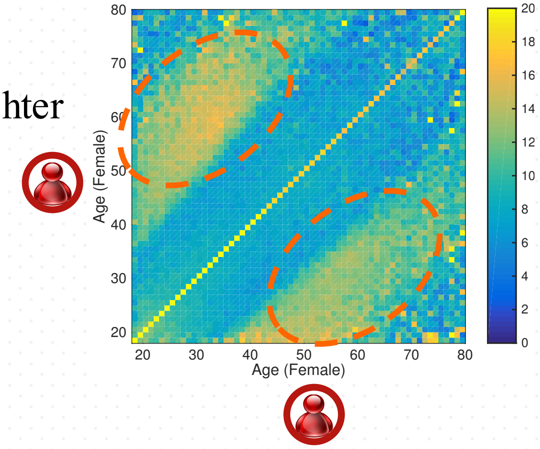
Social Tie Strength

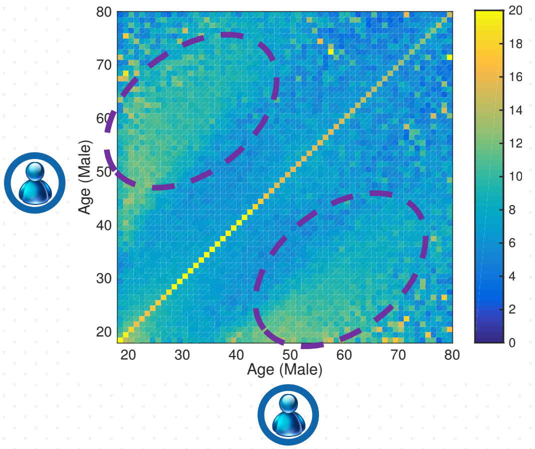




e.g., mom--son
dad--daughter

e.g., mom--daughter



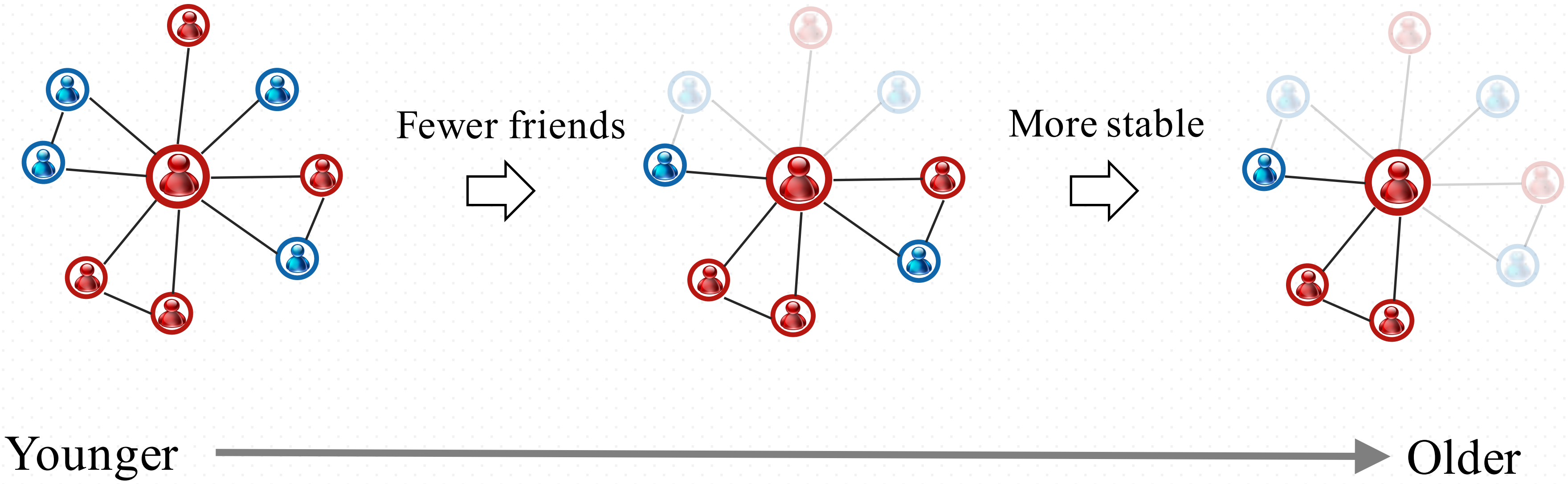


e.g., dad--son

- ♣ 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



Fewer friends

More stable

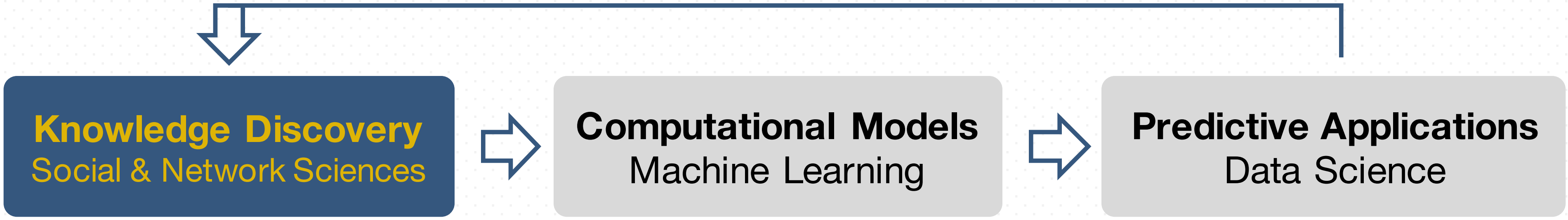
Younger

Older

more friends
- same-gender
- opposite-gender

fewer friends
- only same-gender
- closed circles

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

Small Worlds

- “Given two individuals selected randomly from the population, what is the probability that the minimum number of intermediaries required to link them is 0, 1, 2, …, k?”
- Mail ~300 letters from Boston to r uals in Texas

  --- Travers and Milgram, **1960s**
- Send 60,000 Emails ple at different countries

  --- Dodds, Muhamad, &Watts, **2003**

Algorithmic Search (people)

- MSN network of 80 million nodes & 1,3 bi s: 6.6

  kovec & Horvitz, **2008**
- Facebook graph of des & 69 billion edges: 4.74

  --- Backstrom et al., **2012**

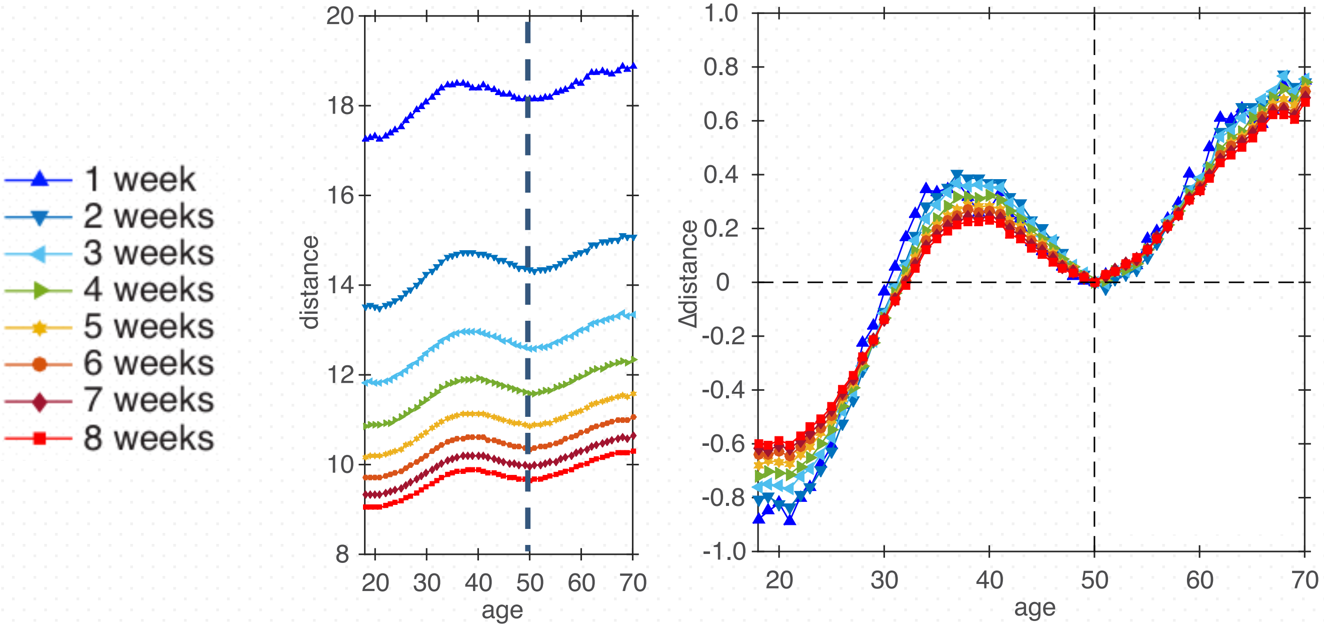
Topological Search (BFS)

1. J. Travers, S. Milgram. An experimental study of the small world problem. **Sociometry** 32, 1969.
2. P. S. Dodds, R. Muhamad, D. J. Watts. An experimental study of search in global social networks. **Science** 301, 2003.
3. J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In ACM **WWW’08**,
4. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna. Four degress of separation. In ACM **WebSci’12**.

- ♣ How do “small worlds” relate to individual demographics?
- ♣ What are the distances between the young and the old, and males and females?

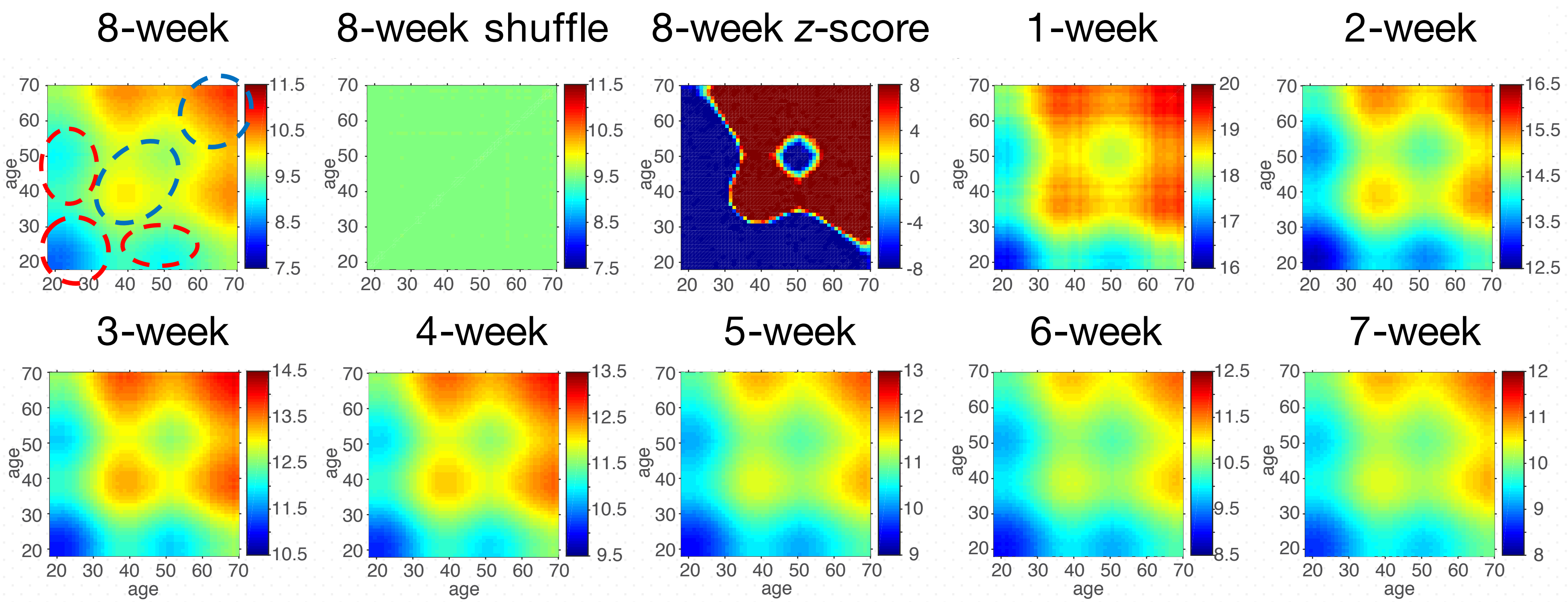
Dong*, Lizardo*, Chawla. Do the young live in a “smaller world” than the old? Age-specific degrees of separation in human communication. *arXiv:1606.07556*

Age-Specific Small Worlds



- ♣ The young live in the smallest world
- ♣ The old live in the least small world

Age-Specific Small Worlds



8-week

8-week shuffle

8-week *z*-score

1-week

2-week

3-week

4-week

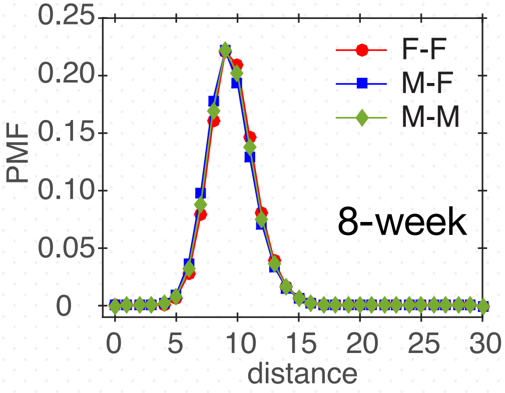
5-week

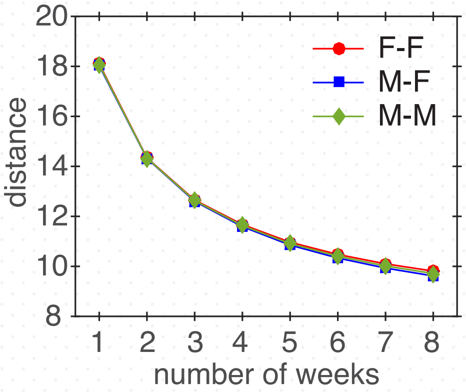
6-week

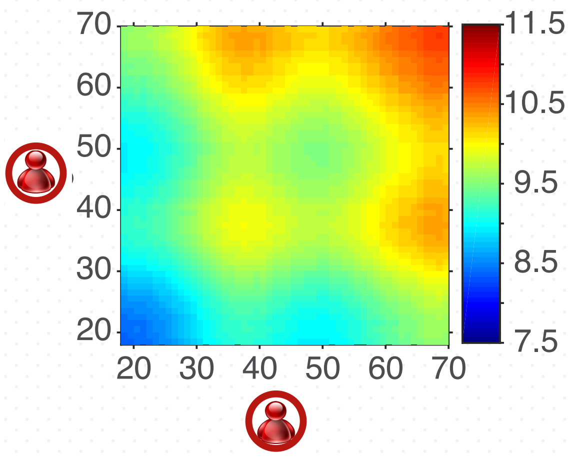
7-week

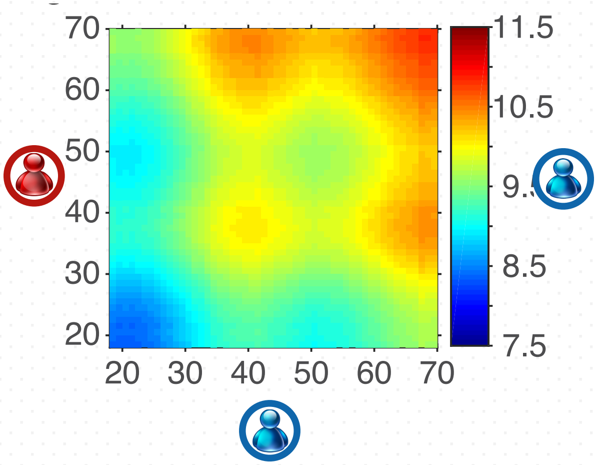
- ♣ The young are close to the young
- ♣ The old are far from the old

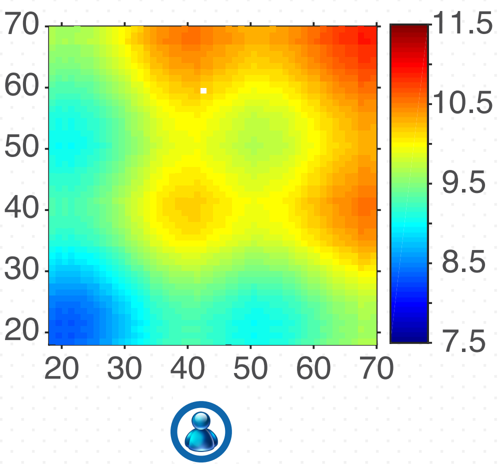
Non Gender Differences in Small Worlds



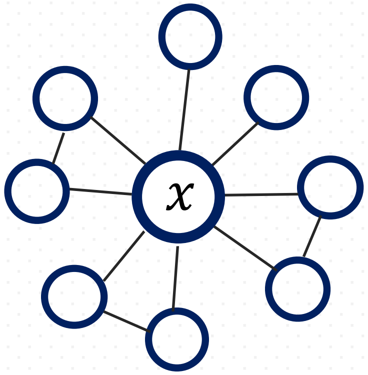


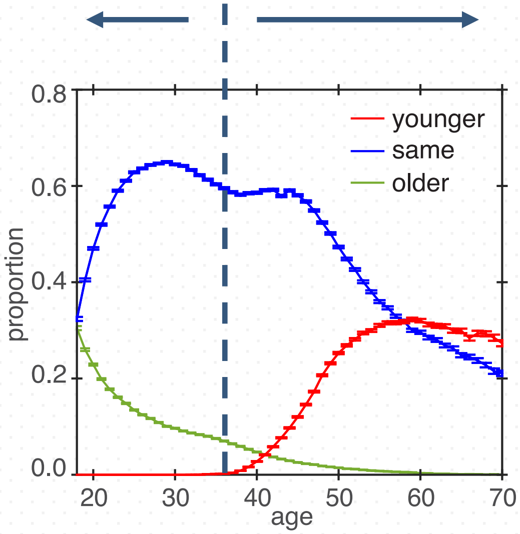


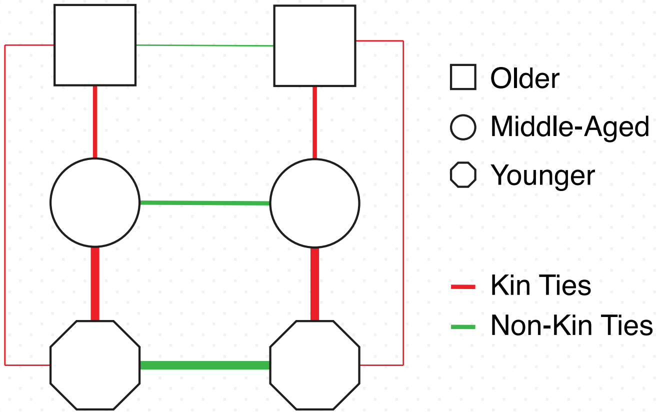




Model of Kin & Non-Kin Ties across Ages







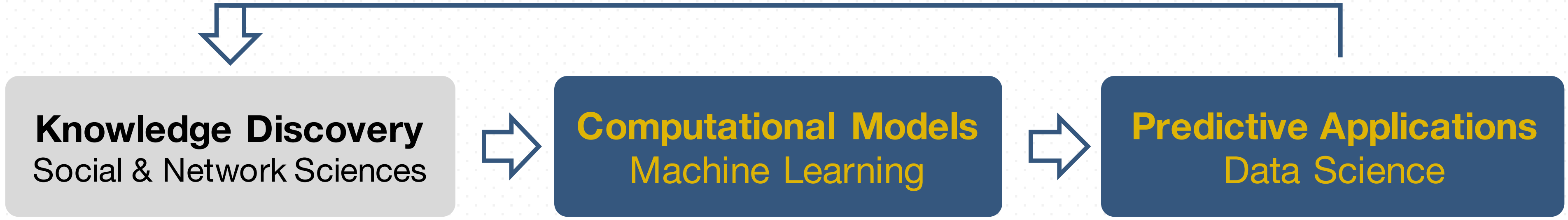
The younger generation
$(x - 30, x - 20)$

The same generation
$x \mp 5$

The older generation
$(x + 20, x + 30)$

- Most informal socializing outside of the family occurs among people of similar age.
- Kin Ties are the primary link connecting individuals across generations.

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

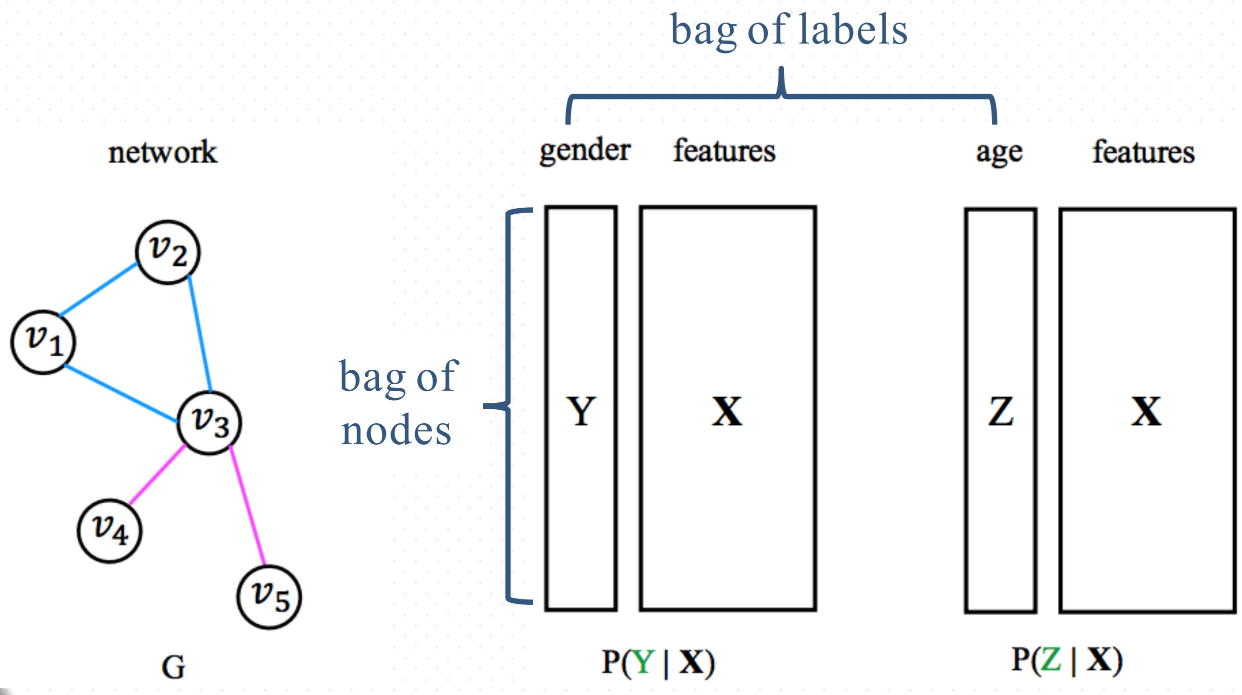
- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

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

- Dong, Zhang, Tang, Chawla, Wang. CoupledLP: Link Prediction in Coupled Networks. In ***ACM KDD*** *2015*.
- Dong, Chawla, Tang, Yang, Yang. User Modeling on Demographic Attributes in Big Mobile Social Networks. In ***ACM TOIS*** *2017*.

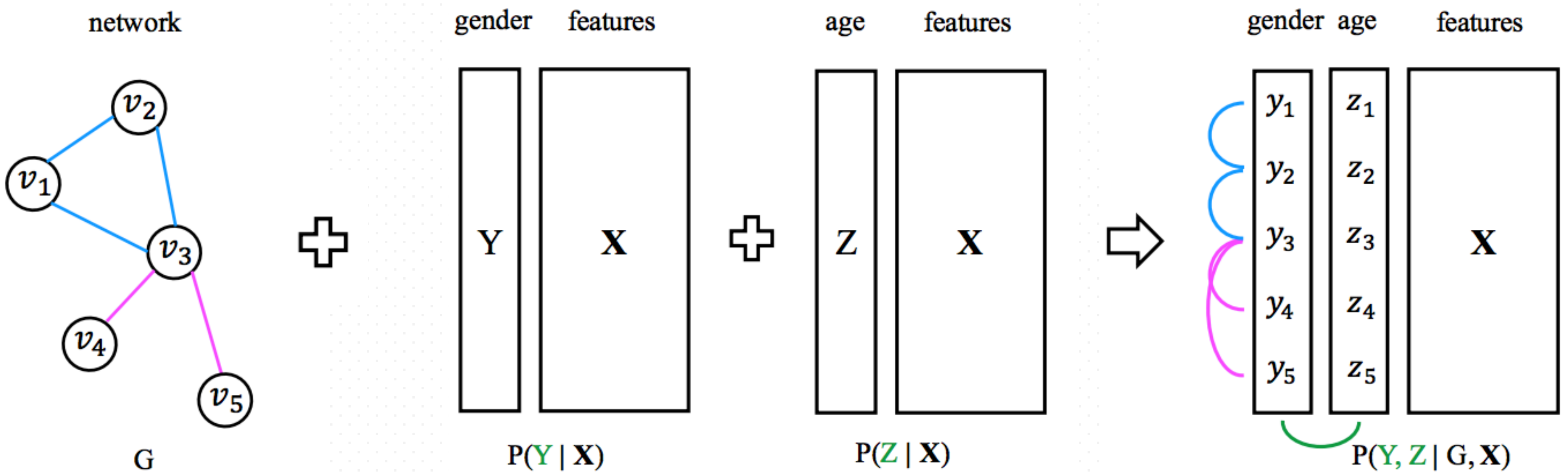
Demographic Prediction

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

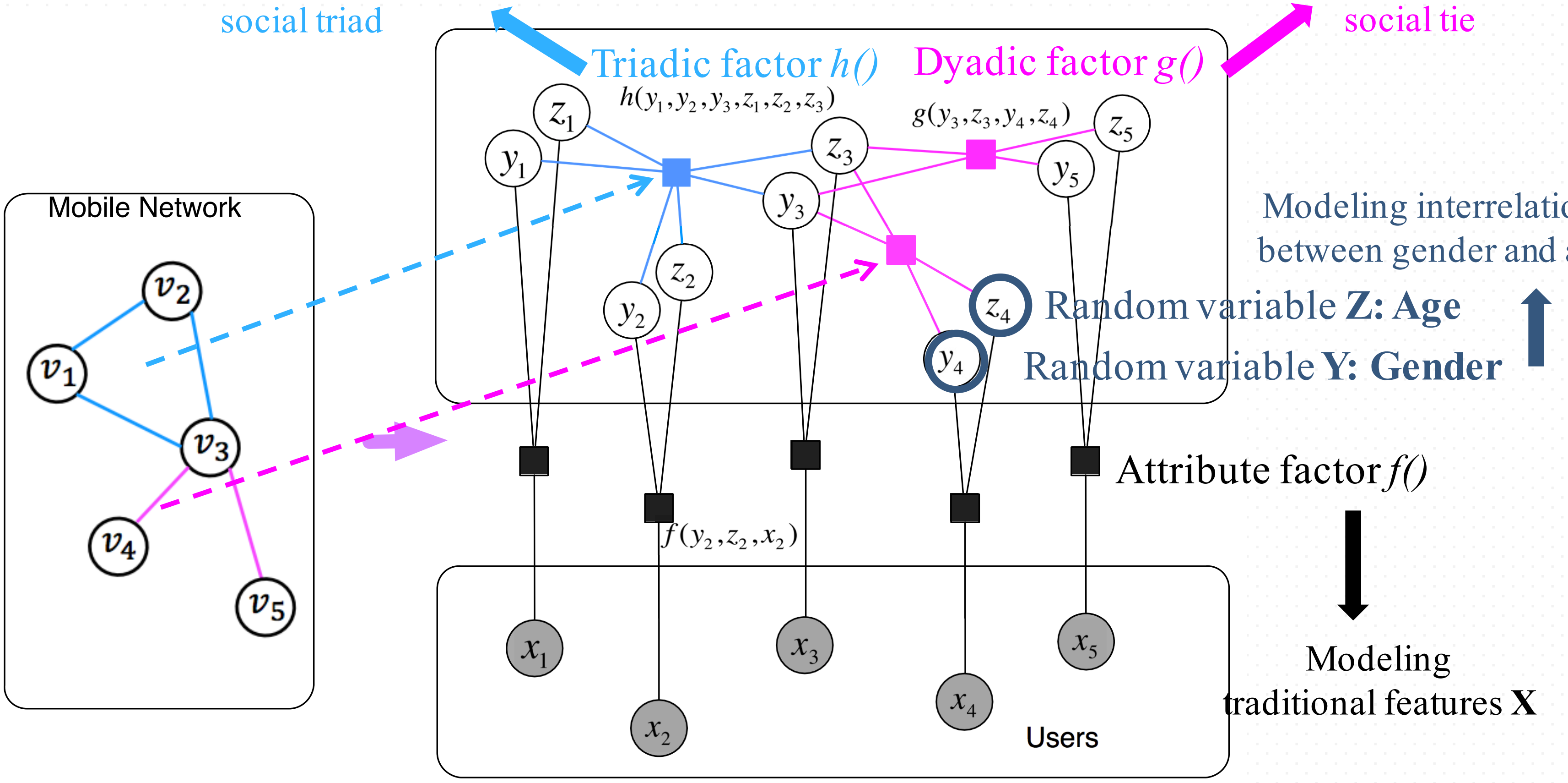


Demographic Prediction

- ♣ 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



Modeling social strategies on social triad

Triadic factor *h()*

Dyadic factor *g()*

Modeling social strategies on social tie

Modeling interrelations between gender and age

Random variable **Z: Age**

Random variable **Y: Gender**

Attribute factor *f()*

Modeling traditional features **X**

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

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

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

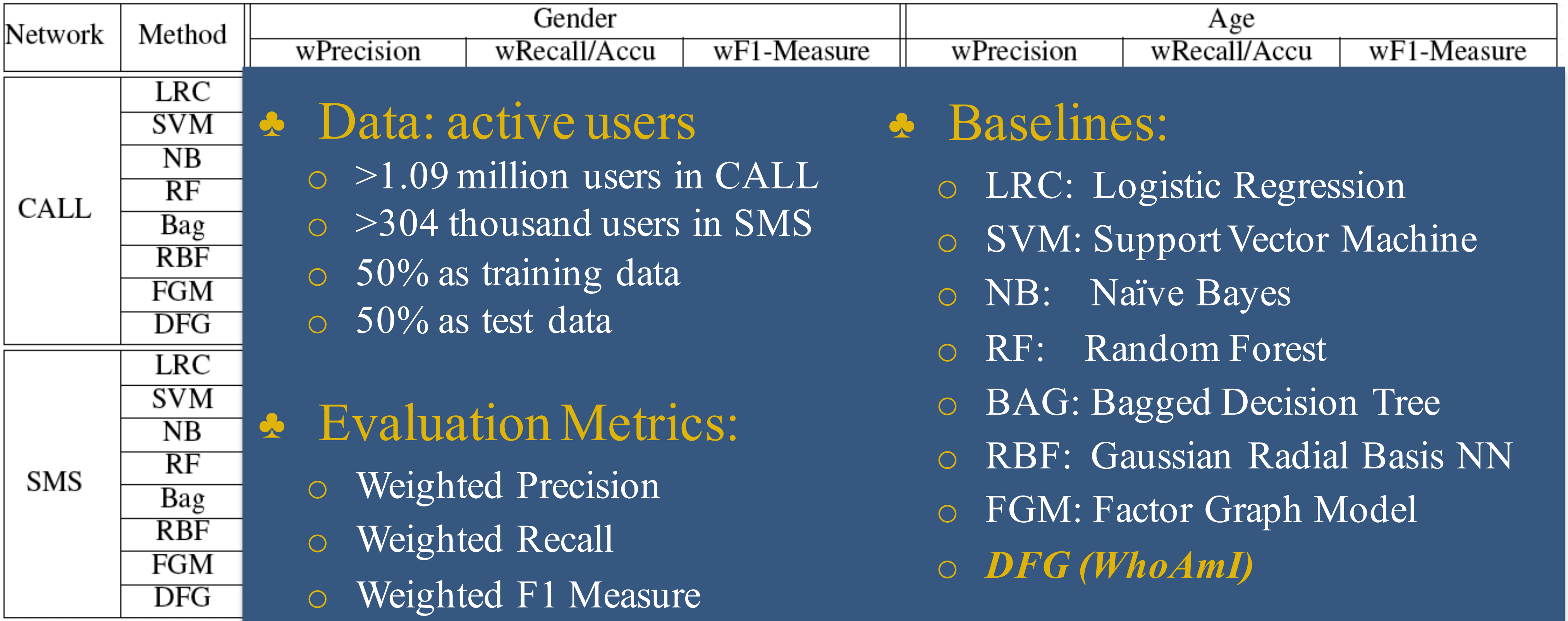
$$\frac{\partial\mathcal{O}(\theta)}{\partial\gamma}=\boldsymbol{E}[\sum_{c_{ijk}\in G}h'(\cdot)]-\boldsymbol{E}_{P_\gamma(Y,Z|\mathbf{X},G)}[\sum_{c_{ijk}\in G}h'(\cdot)]$$

⇨ 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

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: active users

- >1.09 million users in CALL
- >304 thousand users in SMS
- 50% as training data
- 50% as test data

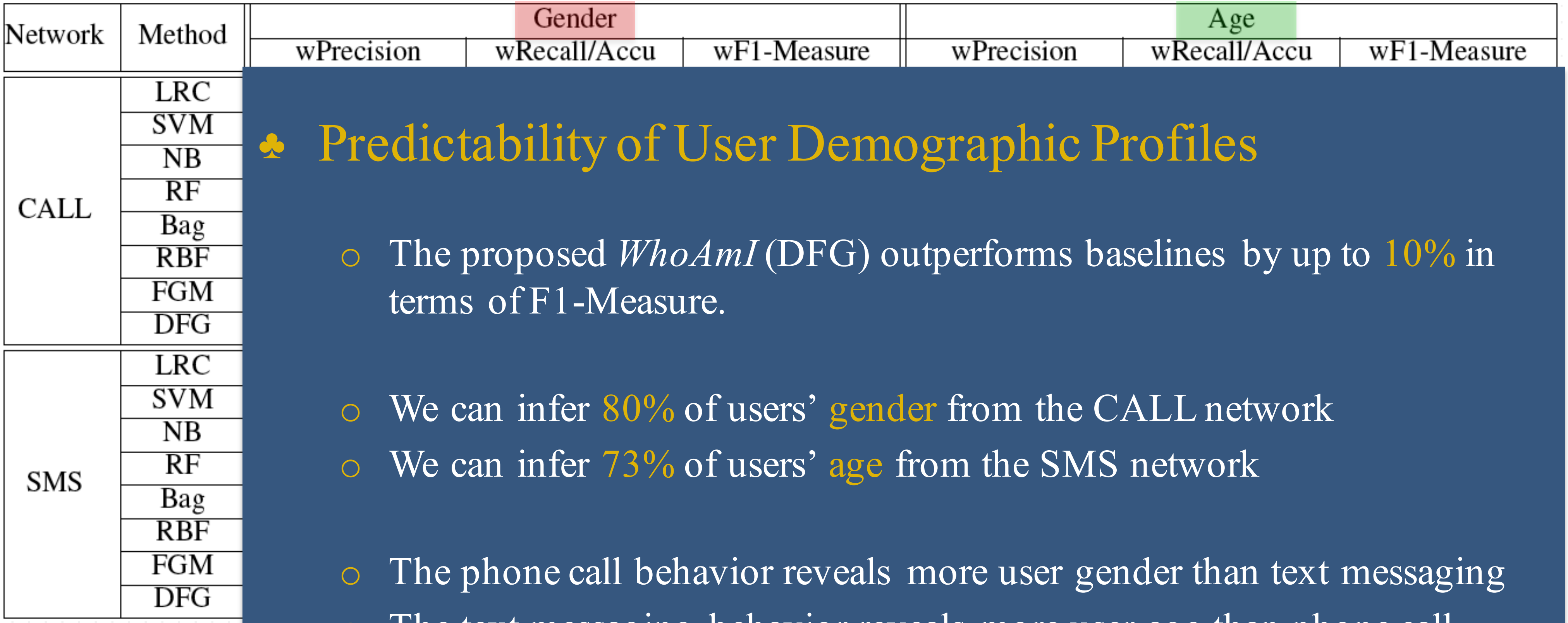
♣ Evaluation Metrics:

- 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

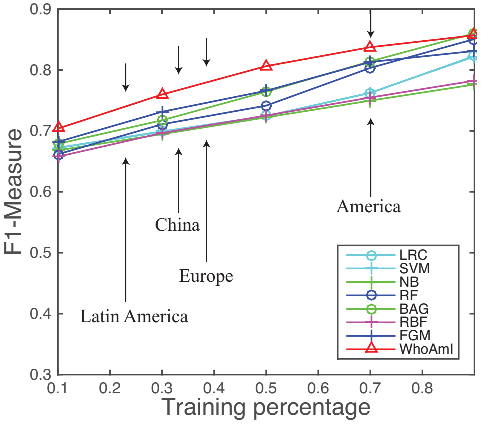
Application 1: Postpaid → Prepaid

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

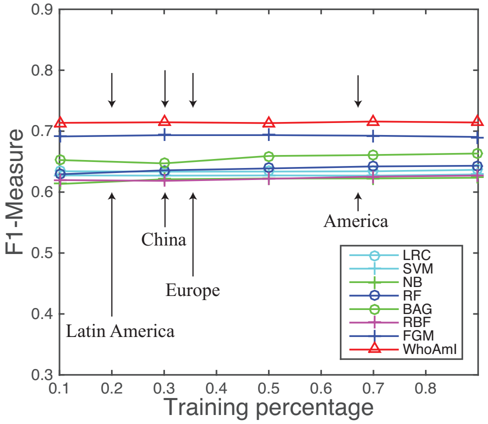
- *Prepaid* services (pay-as-you-go) allow users to be anonymous --- no need to provide any user-specific information.
  - 95% of mobile users in India
  - 80% of mobile users in Latin America
  - 70% of mobile users in China
  - 65% of mobile users in Europe
  - 33% of mobile users in the United States

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

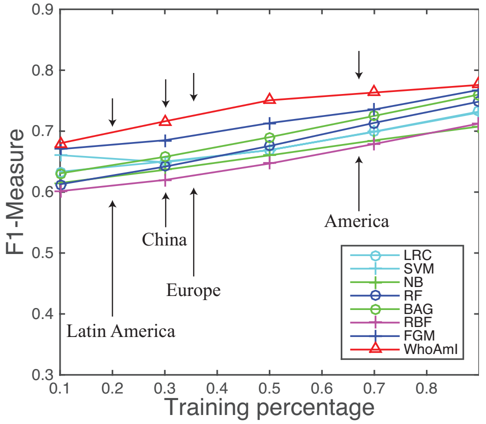
Application 1: Postpaid → Prepaid



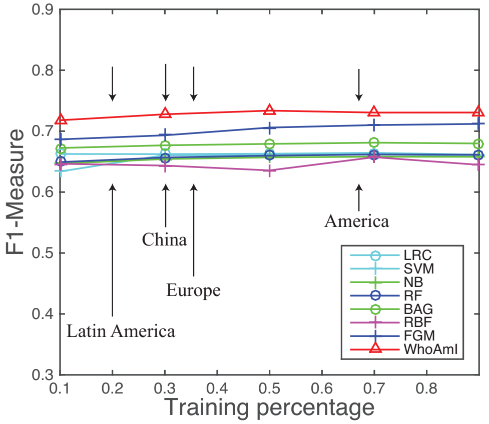
CALL Gender



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

Sprint

2015.08.08 10:30

2015.08.08 10:48

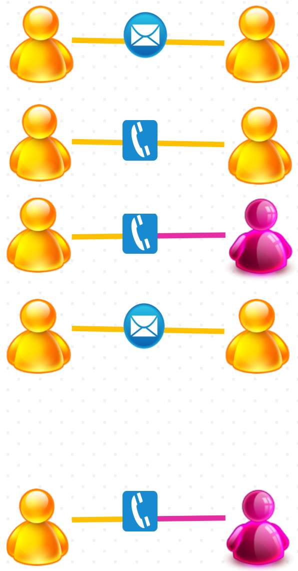
2015.08.08 11:01

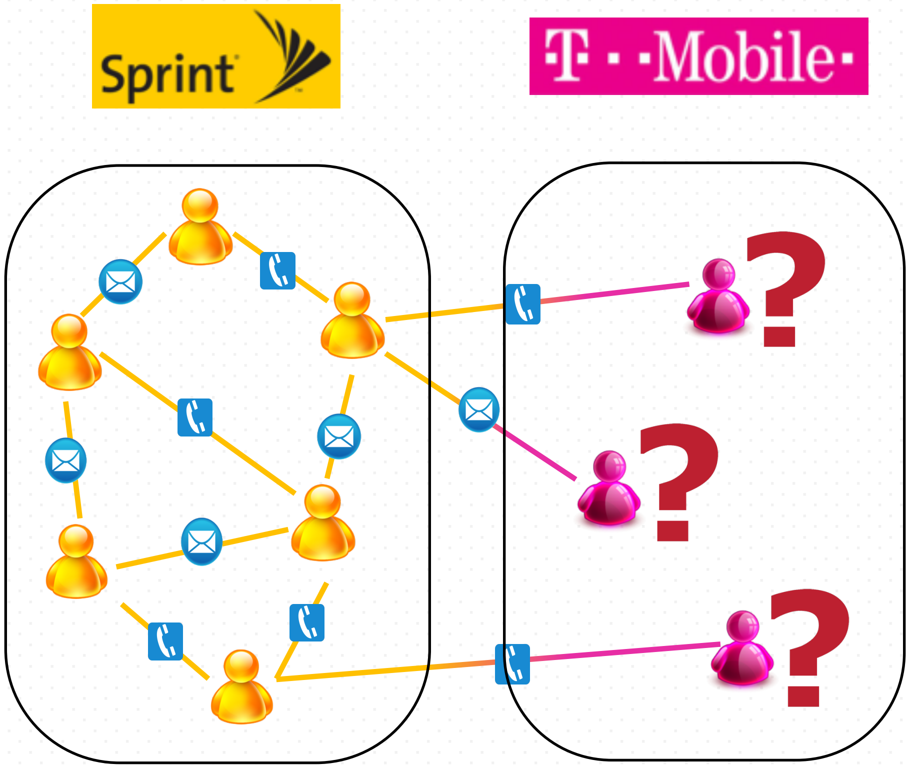
2015.08.08 11:29

……

2016.01.01 00:00

……

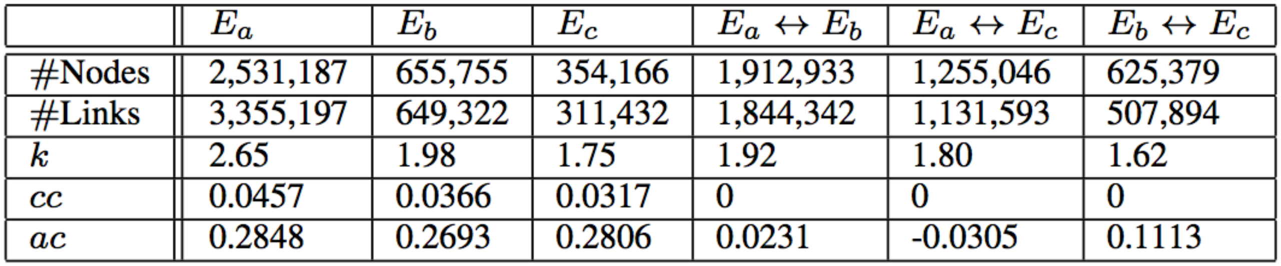




Coupled Demographic Prediction

Coupled Network Data

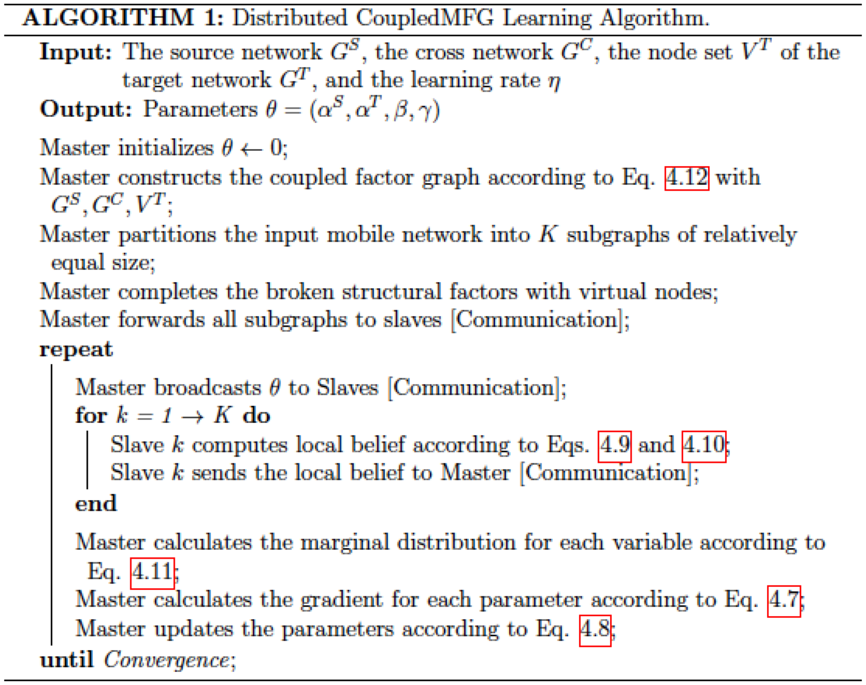
- ♣ Real-world large mobile communication data
  - Over 1 billion call & message records between Aug. to Sep. 2008
  - Undirected and weighted networks
  - Three major mobile operators E_a, E_b, E_c



	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

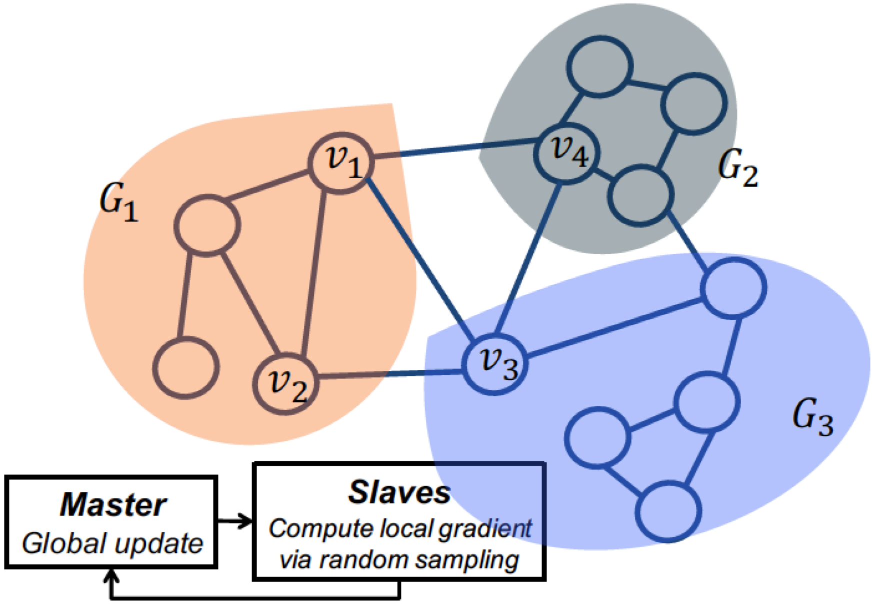
k: average degree
cc: clustering coefficient
ac: associative coefficient

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 $\eta$
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 $\theta$ 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;
```

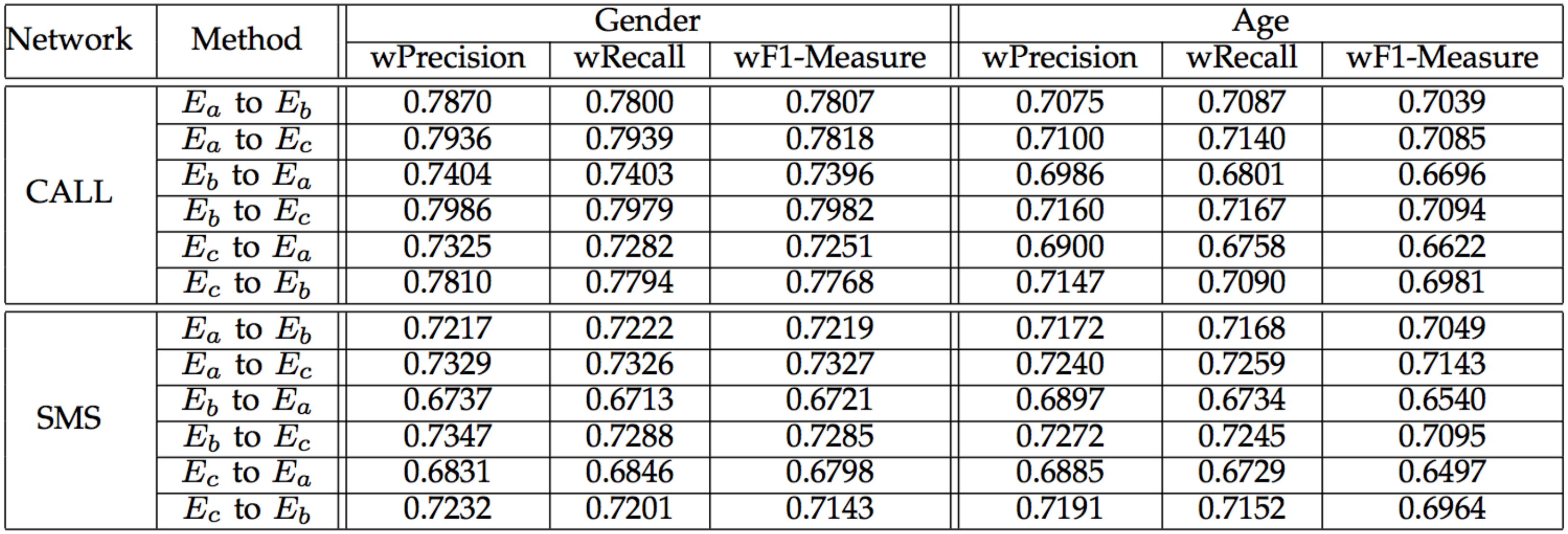


Master
Global update

Slaves
Compute local gradient via random sampling

MPI based

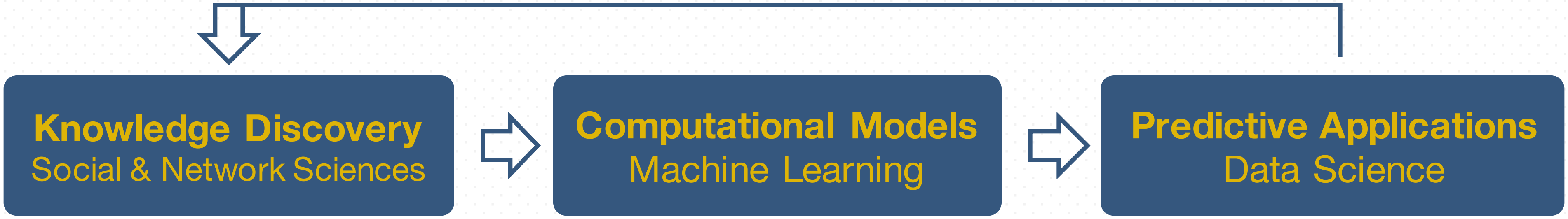
Coupled Demographic Prediction



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.

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

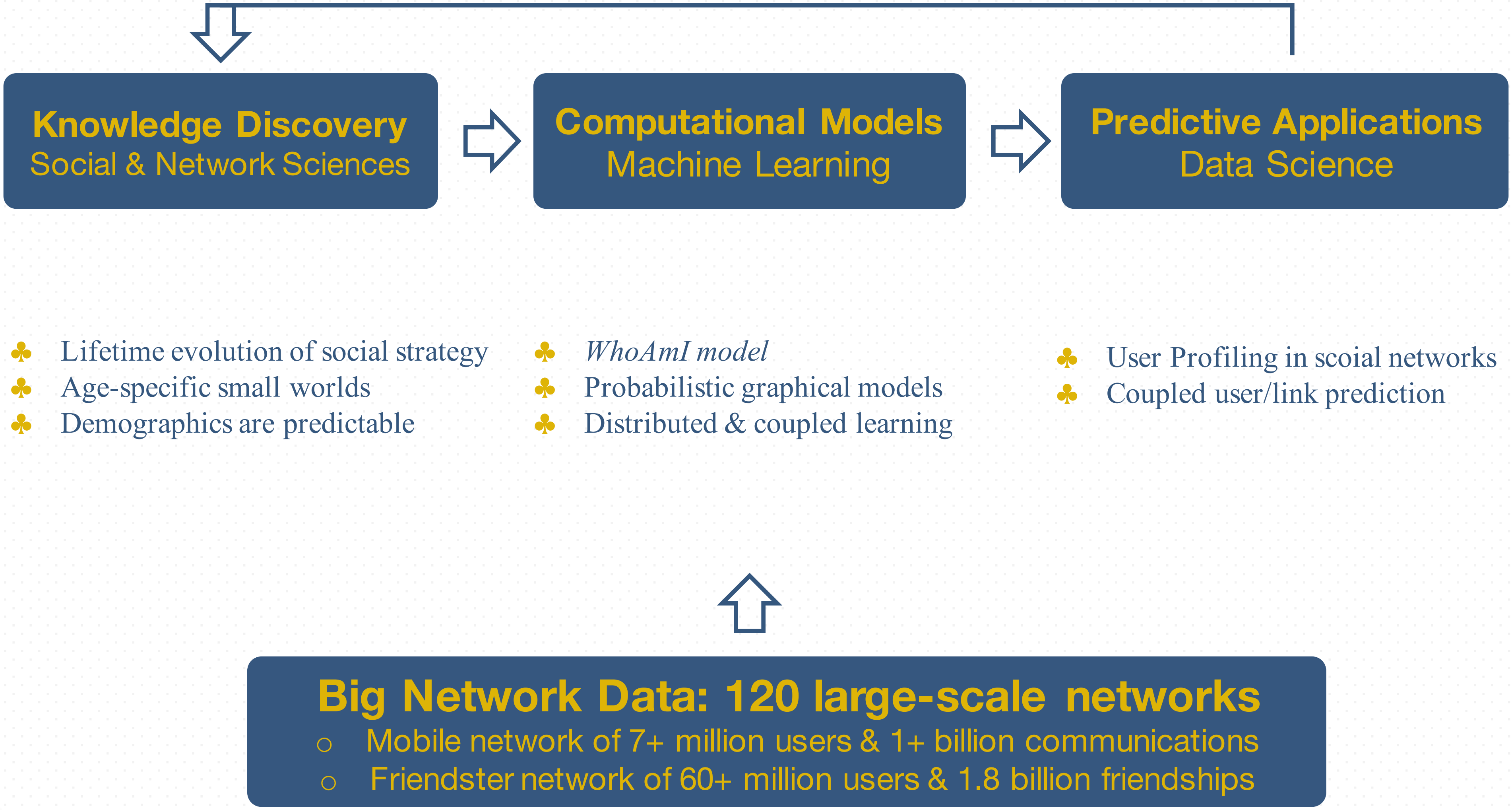
Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

- ♣ Lifetime evolution of social strategy
- ♣ Age-specific small worlds
- ♣ Demographics are predictable

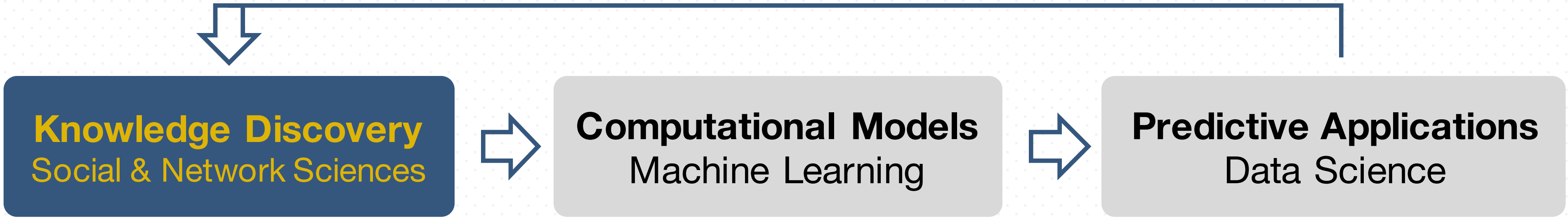
- ♣ *WhoAmI model*
- ♣ Probabilistic graphical models
- ♣ Distributed & coupled learning

- ♣ User Profiling in scoial networks
- ♣ Coupled user/link prediction

Big Network Data: 120 large-scale networks

- ○ Mobile network of 7+ million users & 1+ billion communications
- ○ Friendster network of 60+ million users & 1.8 billion friendships

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

How does the structural diversity of common neighborhoods influence link existence & network organization ?

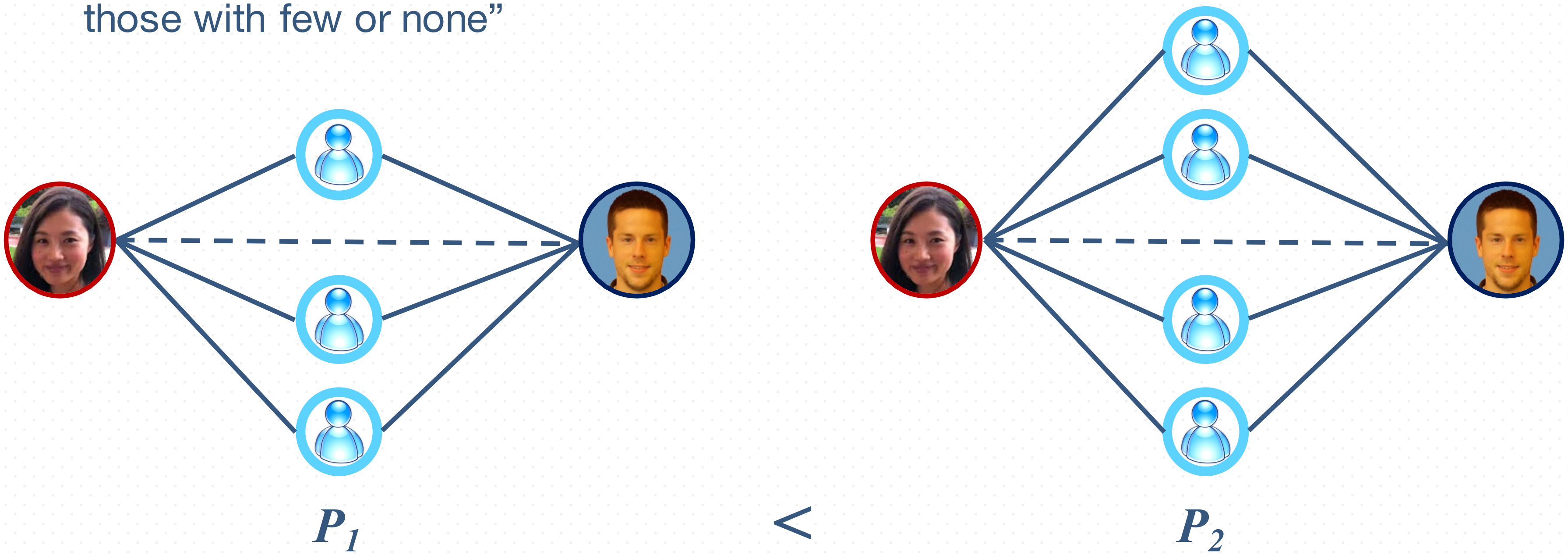
Dong, Johnson, Xu, Chawla. Structural Diversity and Homophily: A Study Across One Hundred Big Networks. In **ACM KDD 2017**

Updated on May, 2017

Structural Homophily

“Love those who are like themselves” ---*Aristotle*

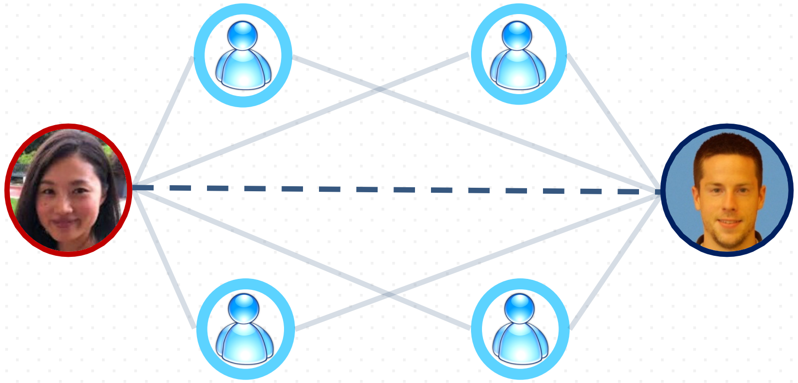
“People with many common friends are more likely to become acquainted than those with few or none”

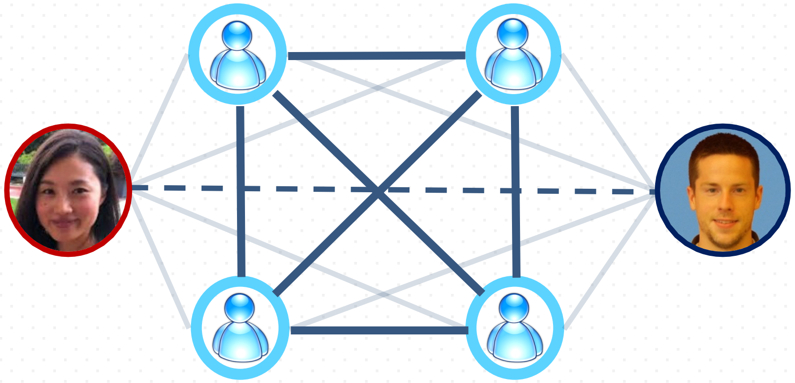


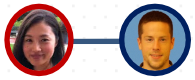
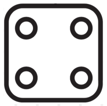
- M. E. J. Newman. Clustering and preferential attachment in growing networks. **Phys. Rev. E**. 2001.
- M. McPherson, L. Smith-Lovin, J. M. Cook. Birds of a feature: homophily in social networks. **Annual Review of Sociology**. 2001.

Common Neighbor (CN) Subgraph

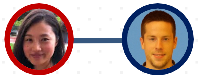
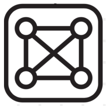
P(connect | common-neighbor-subgraph)





P_1 ( | )

?

P_2 ( | )

more diverse

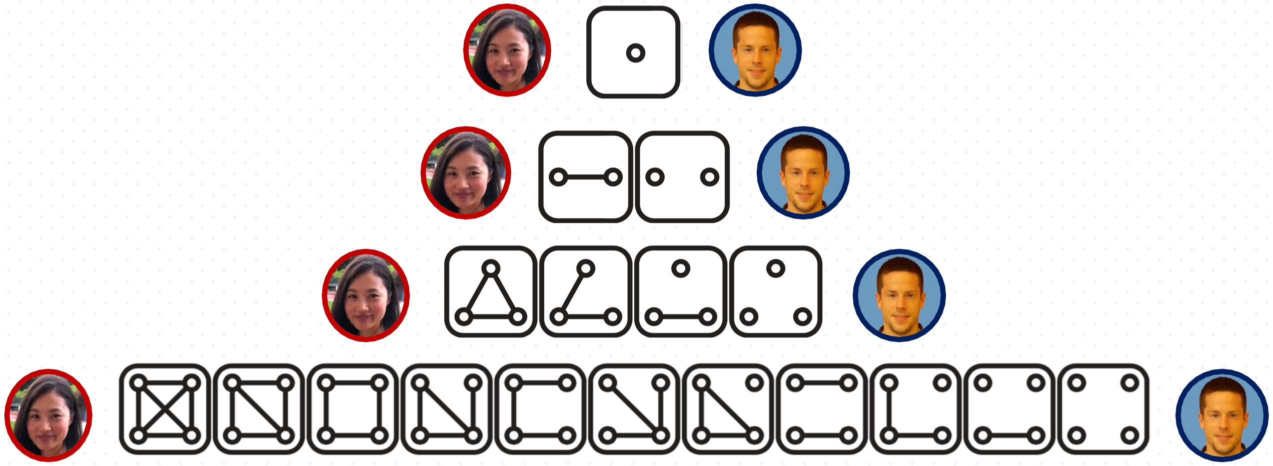
less diverse

Structural Diversity: #components of a common neighbor subgraph

- M. Granovetter. Problems of explanation in economic sociology. Networks and organizations: Structure, form, and action, 25:56, 1992.
- B. Uzzi. Social structure and competition in interfirm networks: the paradox of embeddedness. **Administrative science quarterly**. 1997.
- J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. PNAS, 109(16):5962–5966, 2012.

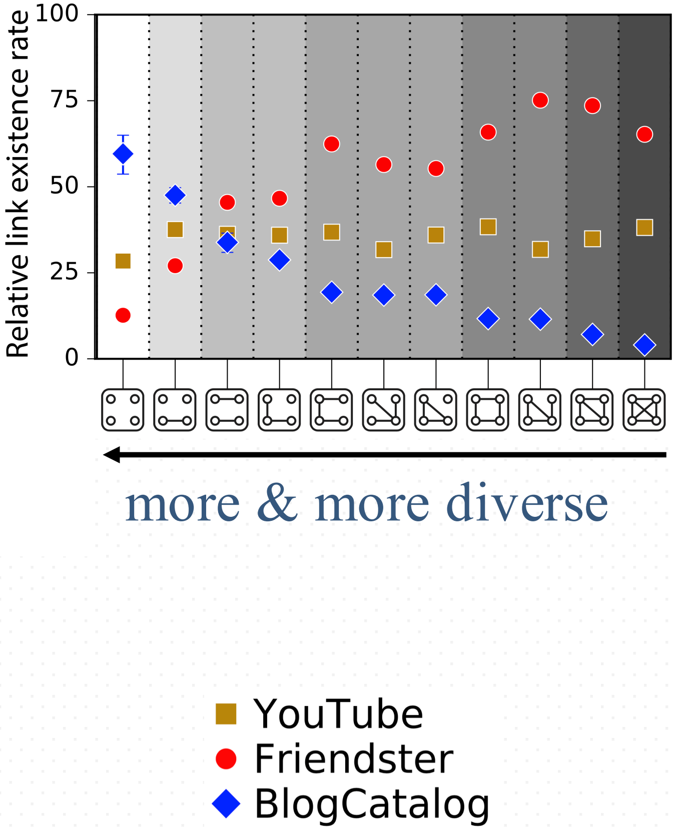
Common Neighbor (CN) Subgraph

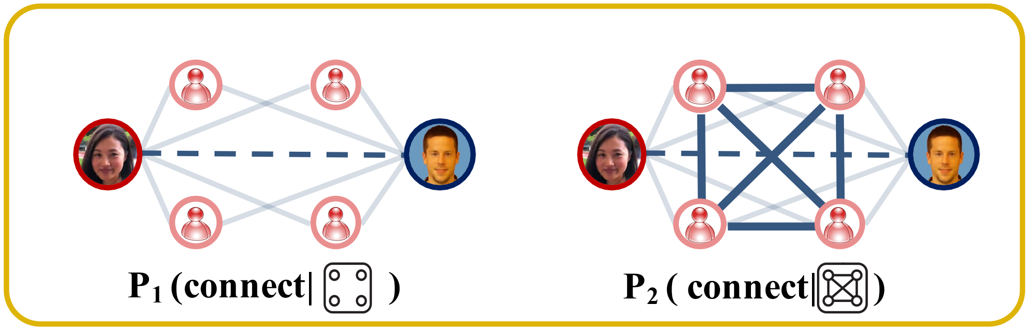
P(connect | common-neighbor-subgraph)



Network	# nodes	# edges	# pairs with ≥1 CN	Data source
Friendster	65,608,366	1,806,067,135	546 billion	SNAP
BlogCatalog	88,784	2,093,195	612 million	ASU
YouTube	1,134,890	2,987,624	1 billion	MPI-SWS

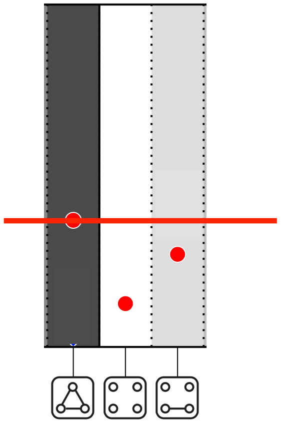
Structural Diversity of CN Subgraph Affects Link Existence

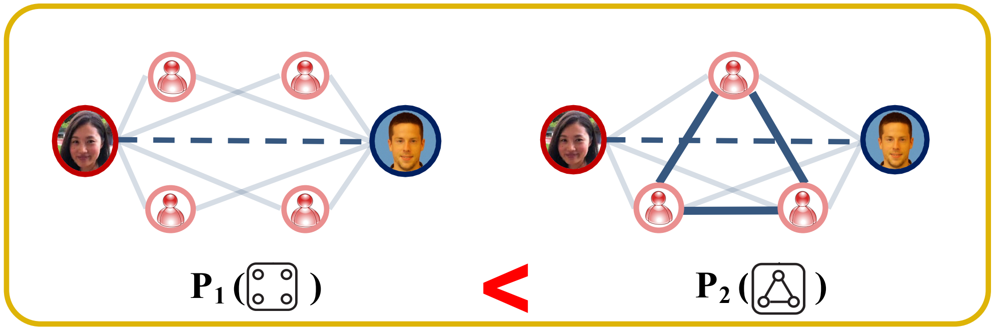




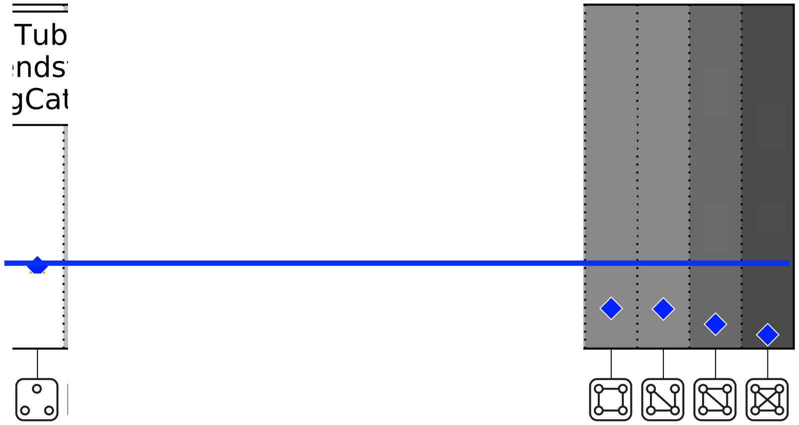
$$\frac{P_1(\text{connect} \mid ⁞⁞)}{P_2(\text{connect} \mid \boxtimes)} \approx \frac{1}{5}$$

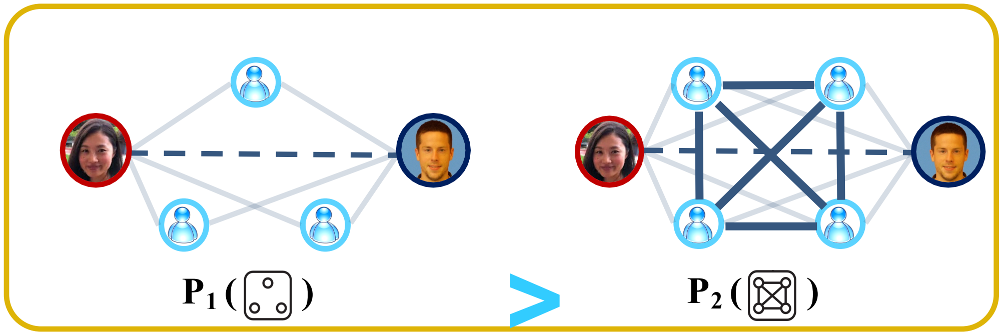
The Violation of Structural Homophily





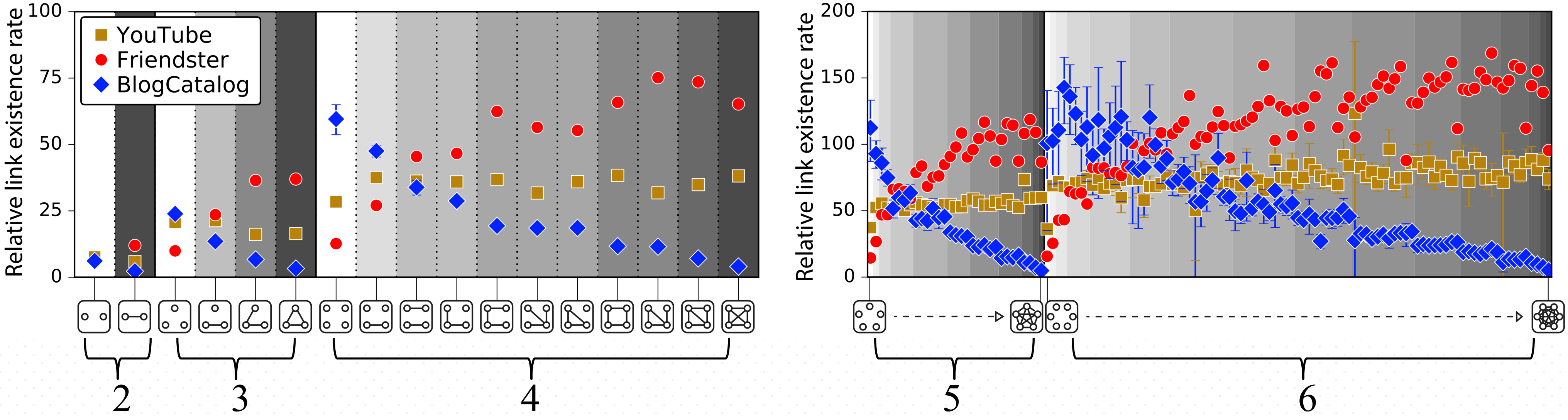
friendster





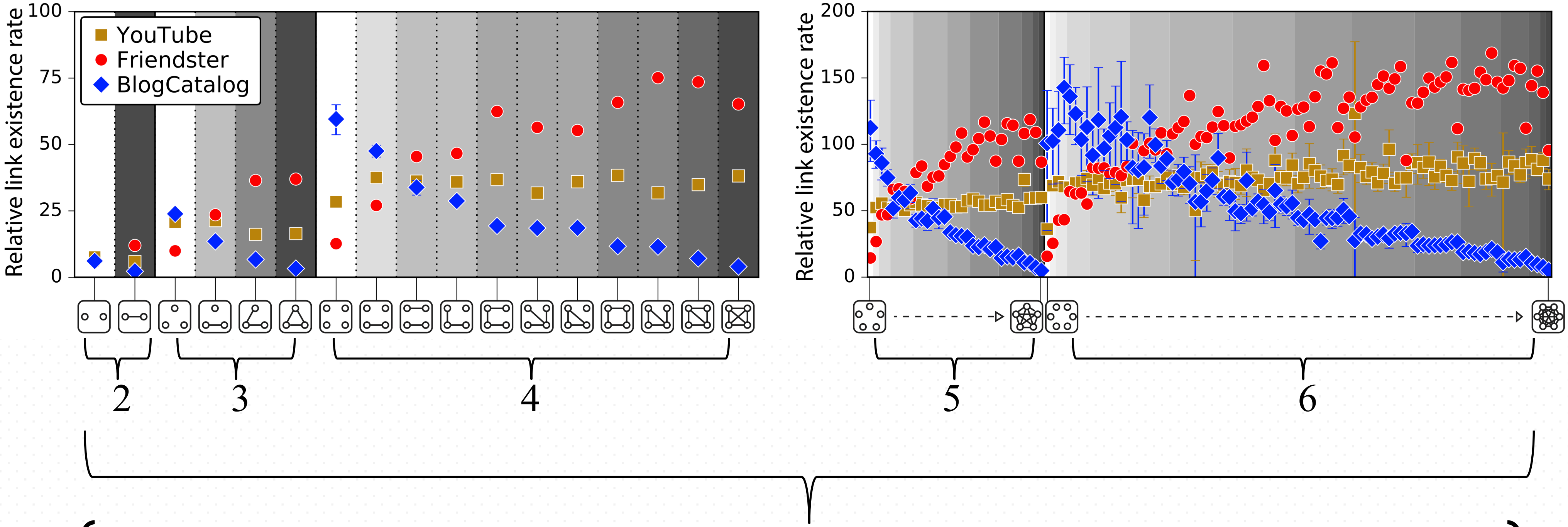
blogcatalog

Structural Diversity of Common Neighborhood



♣ The diversity of common neighborhood affects link formation and also violates the principle of homophily.

Common Neighborhood Signature



$\left(y_2^1, y_2^2, \; y_3^1, y_3^2, y_3^3, y_3^4, y_4^1, y_4^2 \; \dots \; y_4^{11}, y_5^1 \; \dots \dots \right)$

Massive Social & Information Networks

AMiner
ASU
KONECT
MPI-SWS
Notre Dame
Net Repo
Newman
SNAP

80 real networks from

+

ER
BA
WS
Kronecker

40 random networks by

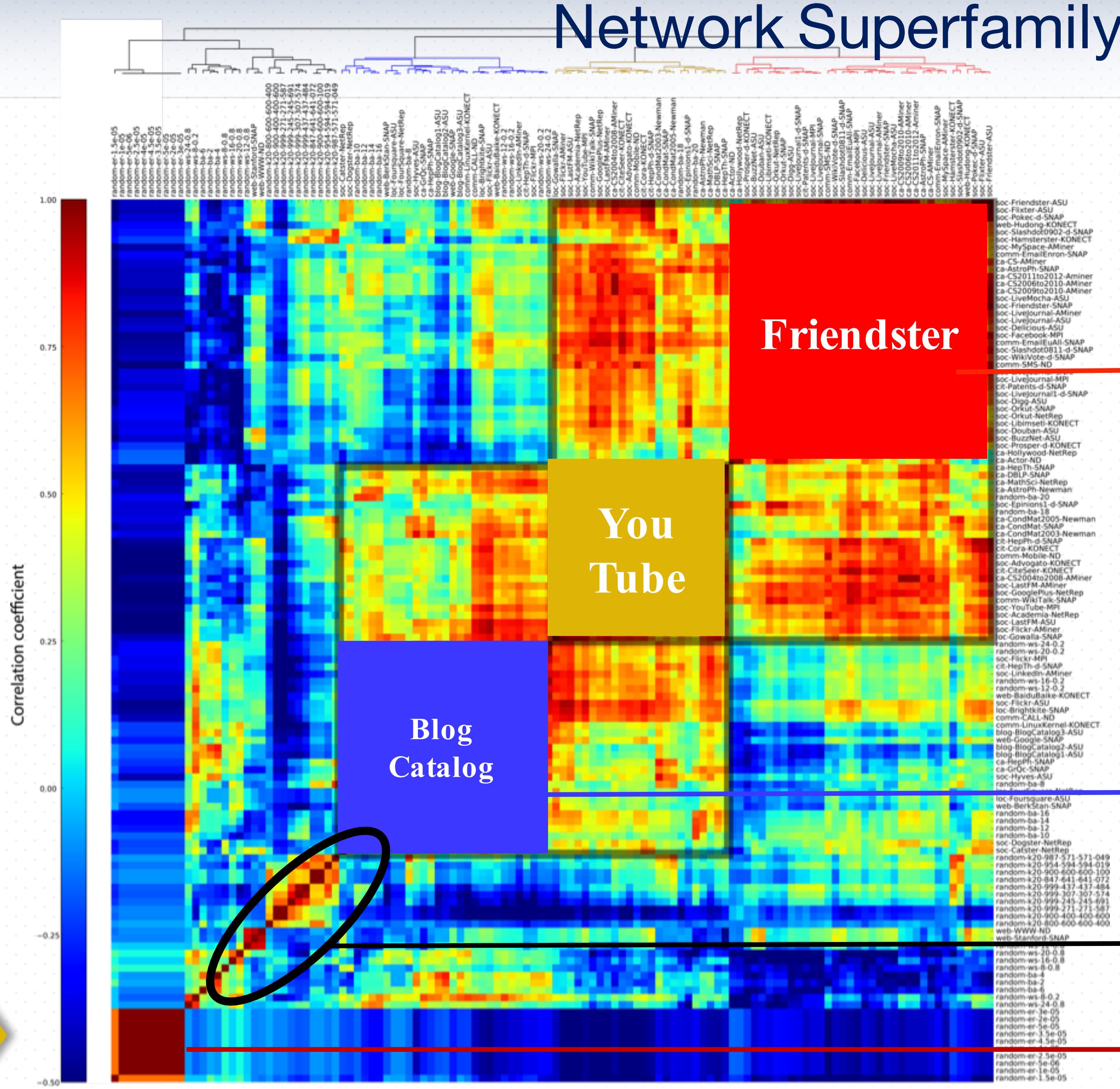
♣ For each network

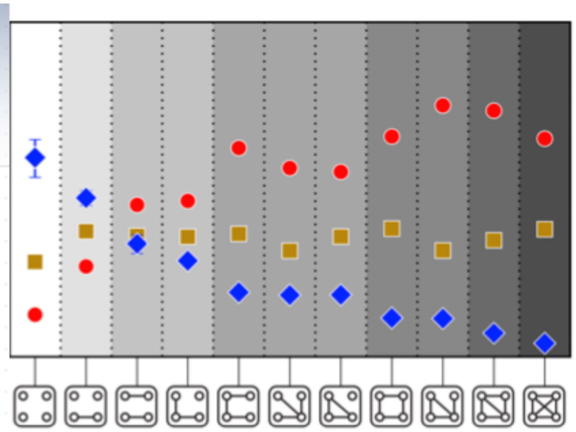
- ○ Get its common neighborhood signature $\boldsymbol{v}$

♣ For each pair of two networks

- ○ Get the correlation coefficient $\rho(\boldsymbol{v}_i, \boldsymbol{v}_j)$ between their common neighborhood signatures $\boldsymbol{v}_i, \boldsymbol{v}_j$

Network Superfamily





Facebook

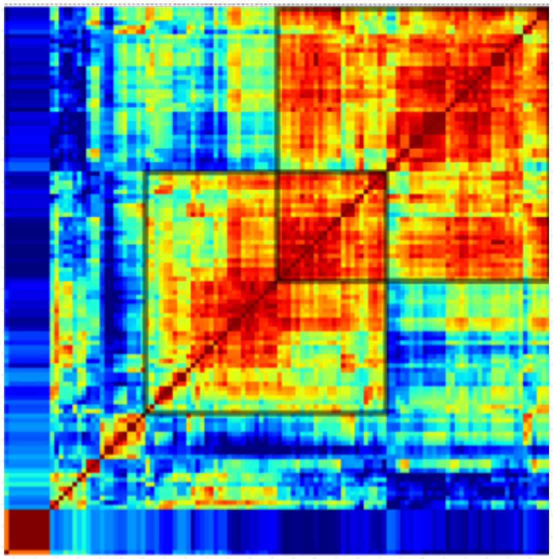
Color: correlation coefficient

LinkedIn
Twitter
WS($\beta \to 0$)

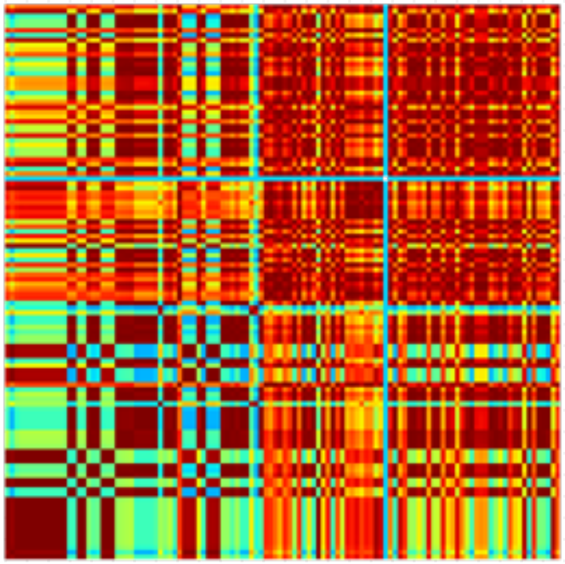
BA, WS($\beta \to 1$)

ER, WS($\beta \to 1$)

Network Superfamily



Common Neighborhood Signature

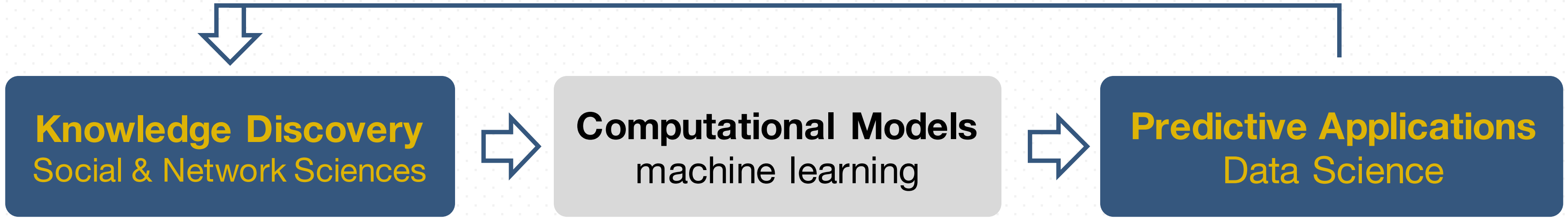


Subgraph Significance Profile
[Milo et al. 2004]

- ♣ Common neighborhood signature serves as a fundamental property of a network, and unveils unique network superfamilies.

R. Milo, et al. Superfamilies of evolved and designed networks. *Science* 2004.

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
machine learning

Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

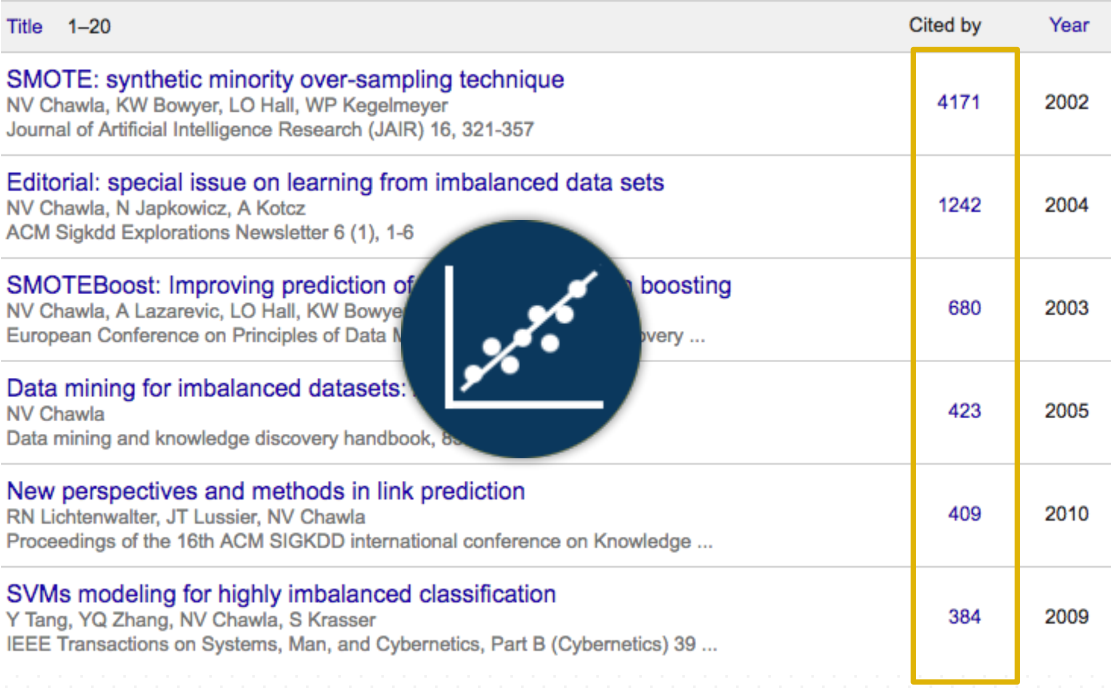
- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

How can we increase our social impact ?

- Dong, Johnson, Chawla. Will This Paper Increase Your h-index? Scientific Impact Prediction. In ***ACM WSDM** 2015.*
  Best Paper Award Nomination
- Dong*, Johnson*, Chawla. Can Scientific Impact Be Predicted? **IEEE Trans. on Big Data** 2016.

Science of Science

“An emerging area of interest in research on the ‘science of science’ is the ***prediction*** *of future* ***impact****.”*



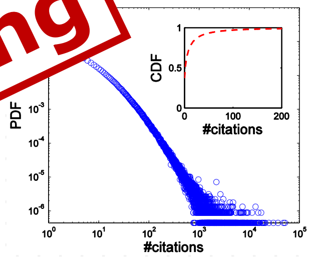
Title 1–20	Cited by	Year
SMOTE: synthetic minority over-sampling technique NV Chawla, KW Bowyer, LO Hall, WP Kegelmeyer Journal of Artificial Intelligence Research (JAIR) 16, 321-357	4171	2002
Editorial: special issue on learning from imbalanced data sets NV Chawla, N Japkowicz, A Kotcz ACM Sigkdd Explorations Newsletter 6 (1), 1-6	1242	2004
SMOTEBoost: Improving prediction of … boosting NV Chawla, A Lazarevic, LO Hall, KW Bowye European Conference on Principles of Data M … very ...	680	2003
Data mining for imbalanced datasets: NV Chawla Data mining and knowledge discovery handbook, 8	423	2005
New perspectives and methods in link prediction RN Lichtenwalter, JT Lussier, NV Chawla Proceedings of the 16th ACM SIGKDD international conference on Knowledge ...	409	2010
SVMs modeling for highly imbalanced classification Y Tang, YQ Zhang, NV Chawla, S Krasser IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 ...	384	2009

James A. Evans. Future Science, **Science** 342 (44), 2013
R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In ACM **JCDL’12**.
D. Wang, C. Song, A.-L. Barabasi. Quantifying long-term scientific impact. **Science**, 342 (6154), 2013.

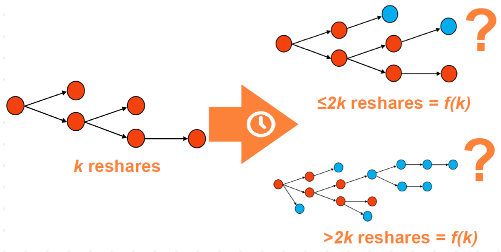
Scientific Impact Prediction

- ♣ Predicting the #citations of each paper

Challenging



- ♣ Predicting whether a cascade will double in size (*k* reshares → *2k*)



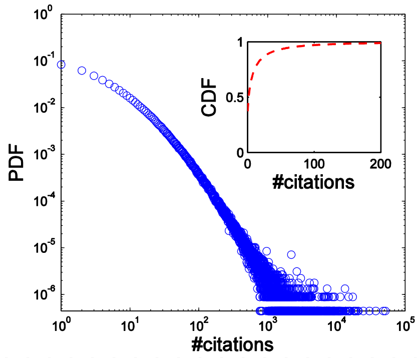
[Cheng et al. 2014]

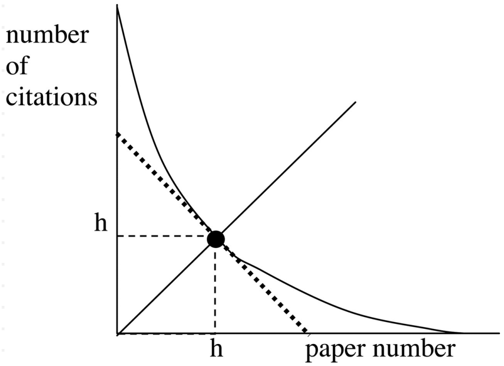
Will This Paper Increase Your *h*-index?

J. Cheng, L. Adamic, A. Dow, J. Kleinberg, J. Leskovec. Can cascades be predicted? In ACM **WWW'14**.

Scientific Impact Prediction

- Given a paper and its author information at *t*:
  - What is its author's future *h*-index, ***h'*** , within a timeframe Δt ?
  - Will this paper published at *t* will contribute to his future *h*-index, *h'*, within a timeframe Δt ?





Data & Factors

- **A real-world academic dataset**
  - ○ Arnetminer
  - ○ 1,712,433 authors
  - ○ 2,092,356 papers
  - ○ 4,258,615 collaborations
  - ○ 8,024,869 citations
  - ○ *arnetminer.org/AMinerNetwork*

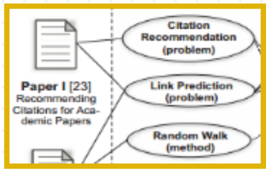
Content

Venue

Author

Will This Paper Increase Your h-index?
Scientific Impact Prediction

Paper


Citation Recommendation (problem)
Paper I [23] Recommending Citations for Academic Papers
Link Prediction (problem)
Random Walk (method)

Reference

Social

Temporal

24 Factors from 6 groups

J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su. ArnetMiner: Extraction and mining of academic social networks. In ACM **KDD'08**.

Factors Driving Impact Growth

♣ Publishing in academically ***diverse topics*** *is difficult* to further one's scientific impact, at least as measured by an increase in one's *h*-index.

♣ A scientific ***researcher's authority*** on the topic of a paper is the most decisive factor in determining whether the paper contributes to his or her *h*-index.

♣ The ***level of the venue*** in which a given paper is published is another crucial factor in determining the probability that it will contribute to its authors' *h*-indices.

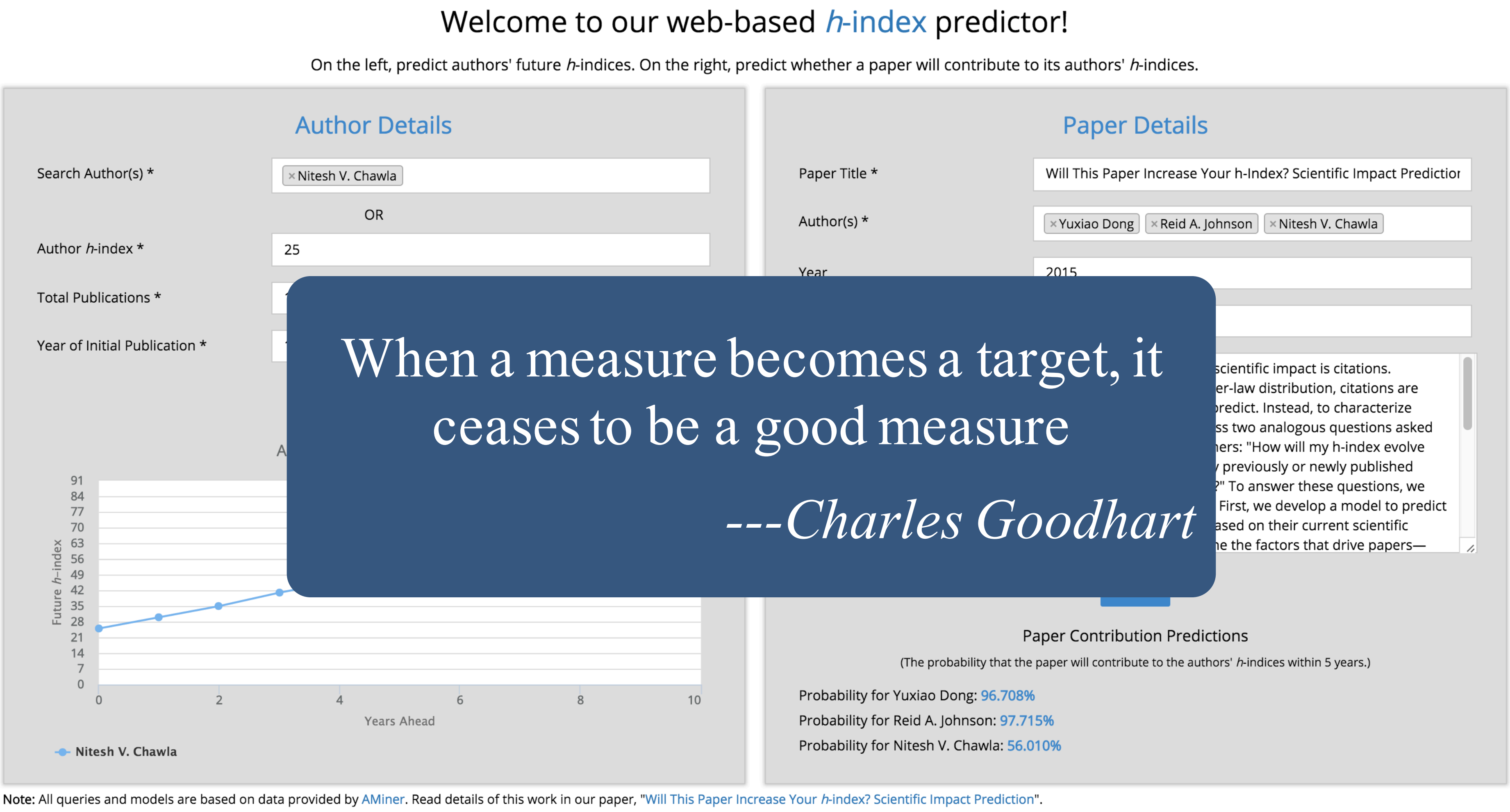
Predictability of Scientific Impact

- ♣ **Task 2.1 (*t* = 2007, Δ*t* = 5):** predict whether the number of citations for each paper published **in** 2007 will be larger than or equal to the **max-*h*-index** author's future *h*-index in 2012.
  - ○ Features: 24 factors
  - ○ Half training, half test

♣ Future scientific impact can be predicted from the past.

Method	Precision	Recall	F1	AUC	Acc.	Pre@3	MAP
Random Guess	**0.210**	0.500	0.296	0.500	0.500	0.589	0.413
Logistic Regression	0.823	0.592	0.689	0.929	0.887	0.892	0.944

Online Demo



Welcome to our web-based *h*-index predictor!

On the left, predict authors' future *h*-indices. On the right, predict whether a paper will contribute to its authors' *h*-indices.

Author Details

Search Author(s) * ×Nitesh V. Chawla

OR

Author *h*-index * 25

Total Publications *

Year of Initial Publication *

Paper Details

Paper Title * Will This Paper Increase Your h-Index? Scientific Impact Predictior

Author(s) * ×Yuxiao Dong ×Reid A. Johnson ×Nitesh V. Chawla

Year 2015

scientific impact is citations.
er-law distribution, citations are
predict. Instead, to characterize
ss two analogous questions asked
hers: "How will my h-index evolve
previously or newly published
?" To answer these questions, we
First, we develop a model to predict
ased on their current scientific
he the factors that drive papers—

Paper Contribution Predictions

(The probability that the paper will contribute to the authors' *h*-indices within 5 years.)

Probability for Yuxiao Dong: 96.708%

Probability for Reid A. Johnson: 97.715%

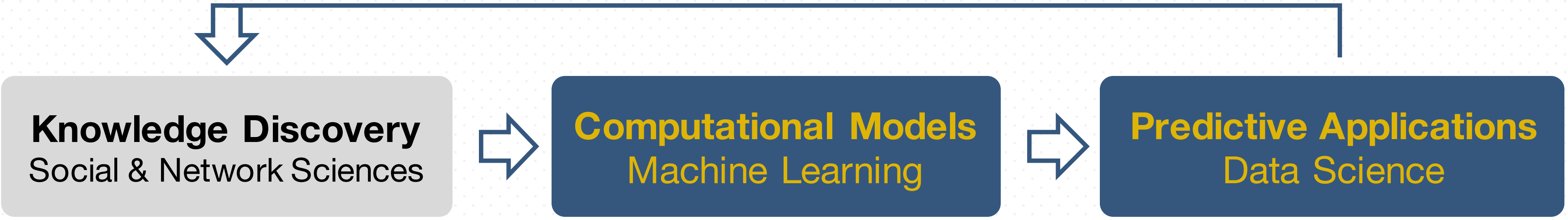
Probability for Nitesh V. Chawla: 56.010%

Note: All queries and models are based on data provided by AMiner. Read details of this work in our paper, "Will This Paper Increase Your *h*-index? Scientific Impact Prediction".

> When a measure becomes a target, it ceases to be a good measure
>
> *---Charles Goodhart*

Web: Reid A. Johnson.

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

How to represent the diverse types of nodes in heterogeneous networks ?

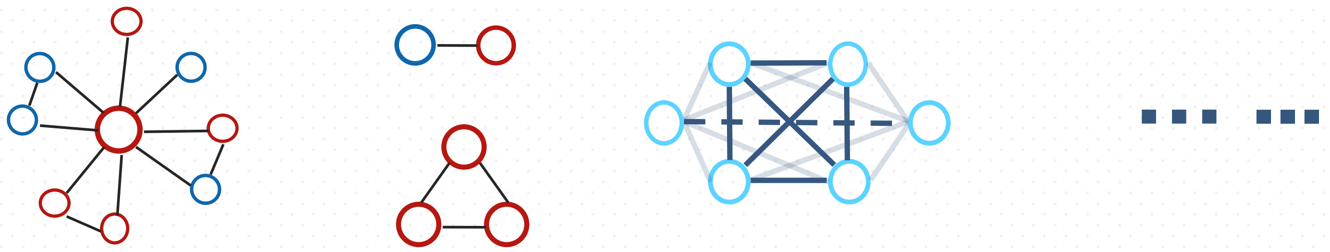
Dong, Chawla, Swami. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In ***ACM KDD 2017*.**

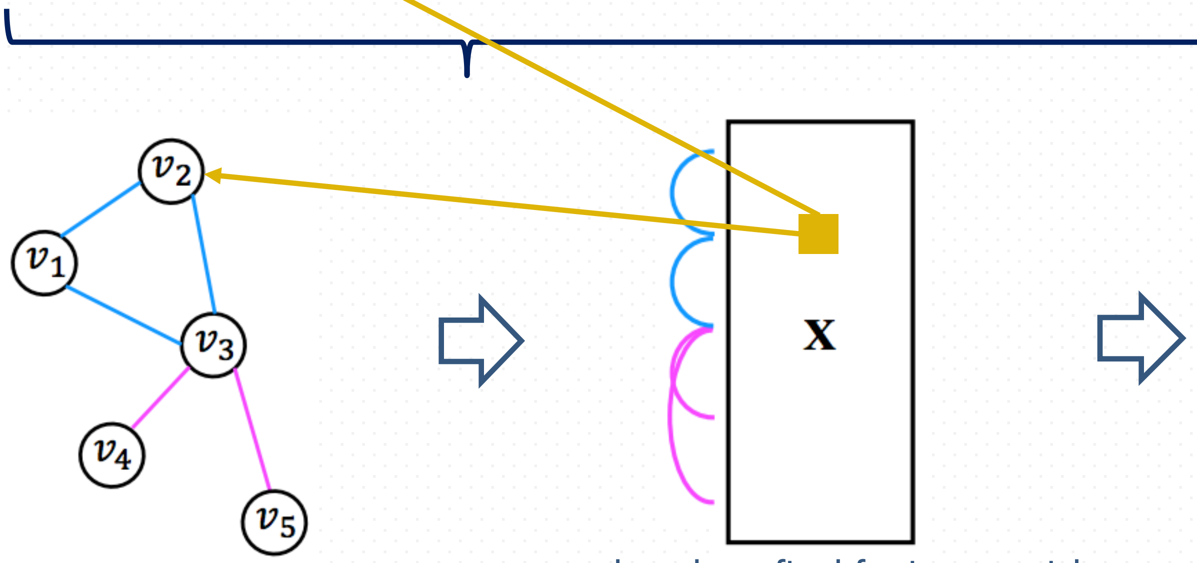
Updated on May, 2017

Network Mining and Learning Paradigm

Node Centralities:

- degree
- betweenness
- clustering coefficient
- PageRank
- Gigenvector
- …





hand-crafted feature matrix

Network Mining Tasks

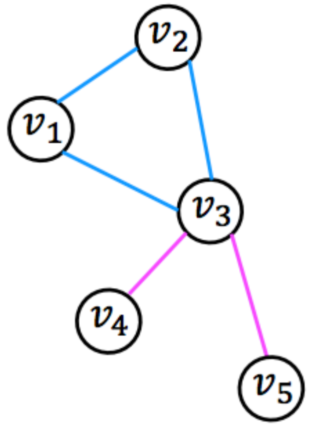
- ♣ node attribute inference
- ♣ community detection
- ♣ similarity search
- ♣ link prediction
- ♣ social recommendation
- ♣ …

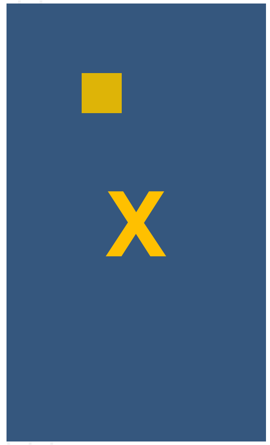
feature engineering

machine learning models

Network Mining and Learning Paradigm

? (deep) neural network based feature representation learning





latent representation matrix

Network Mining Tasks

- node attribute inference
- community detection
- similarity search
- link prediction
- social recommendation
- …

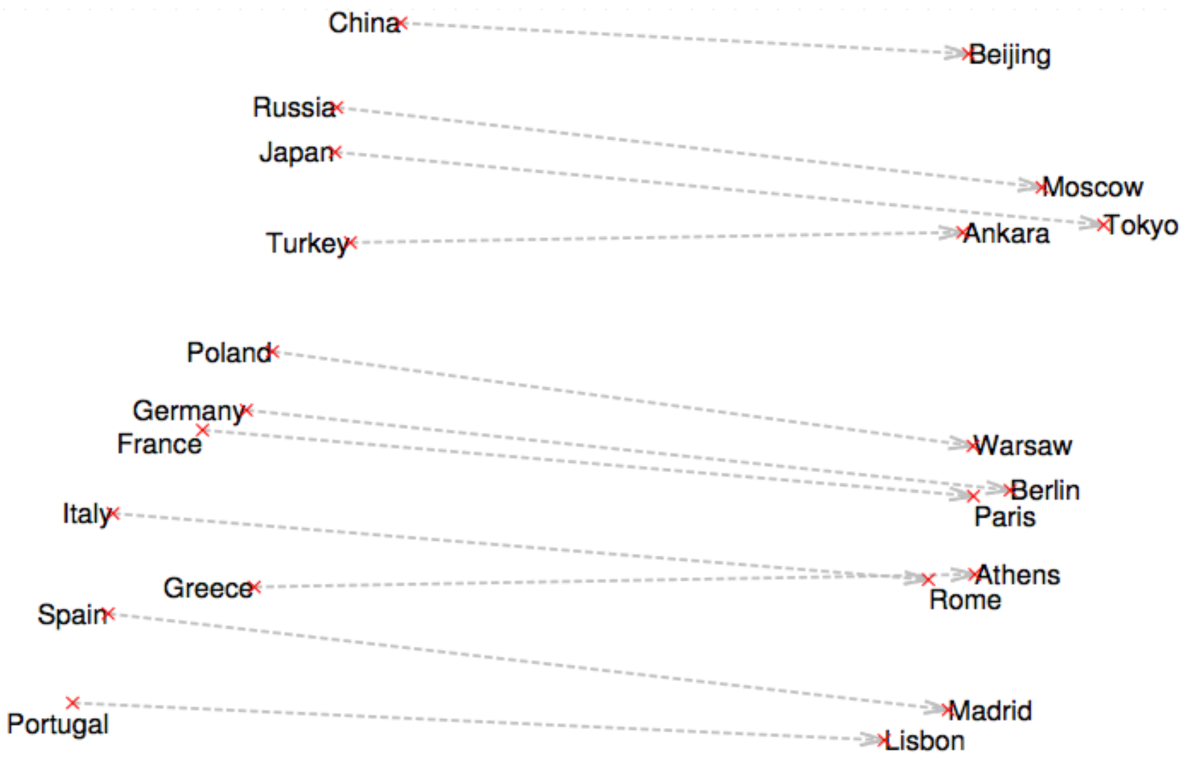
feature learning

machine learning models

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. **IEEE TPAMI**, 35(8):1798–1828, 2013.
Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. **Nature**, 521(7553):436–444, 2015.

Word Representation Learning in NLP

- ♣ Input: a text corpus $D = \{W\}$
- ♣ Output: $\boldsymbol{X} \in R^{|W| \times d}, d \ll |W|$, d-dim vector $\boldsymbol{X}_w$ for each word *w*.

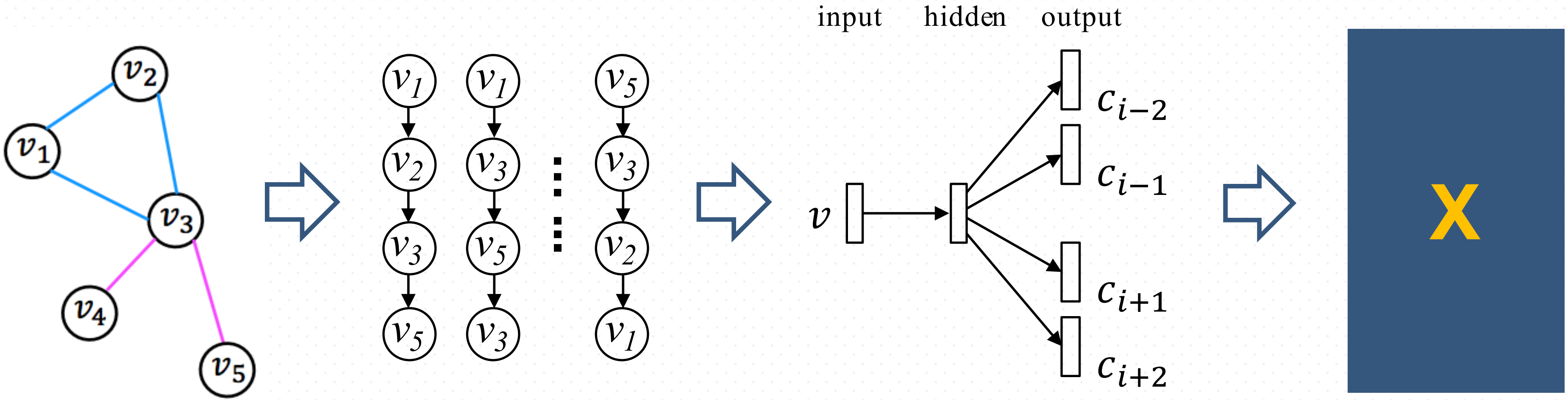


[1] T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean. Distributed representations of words and phrases and their compositionality. In ***NIPS*** *'13*, pp. 3111-31119.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv:1301.3781*, 2013.

Network Representation Learning

- ♣ Input: a network $G = (V, E)$
- ♣ Output: $\boldsymbol{X} \in R^{|V|\times d}, d \ll |V|$, d-dim vector $\boldsymbol{X}_v$ for each node v.



random walk paths (sentences)

word2vec in NLP

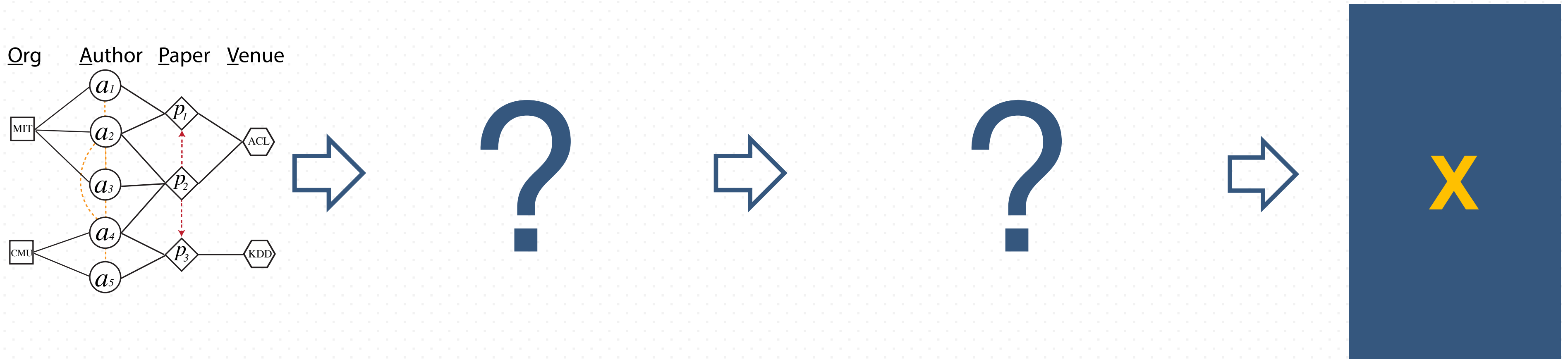
latent representation vector

node2vec [KDD16], DeepWalk [KDD14]

1. B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in ***KDD*** *' 14*, pp. 701–710.
2. A. Grover, J. Leskovec. node2vec: Scalable Feature Learning for Networks. in ***KDD*** *'16*, pp. 855—864.
3. T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean. Distributed representations of words and phrases and their compositionality. In ***NIPS*** *'13*, pp. 3111-31119.
4. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781*, 2013.

Heterogeneous Network Representation Learning

- ♣ Input: **a heterogeneous information network** $G = (V, E, T)$
- ♣ Output: $\boldsymbol{X} \in R^{|V|\times d}, d \ll |V|$, d-dim vector $\boldsymbol{X}_v$ for each node v.

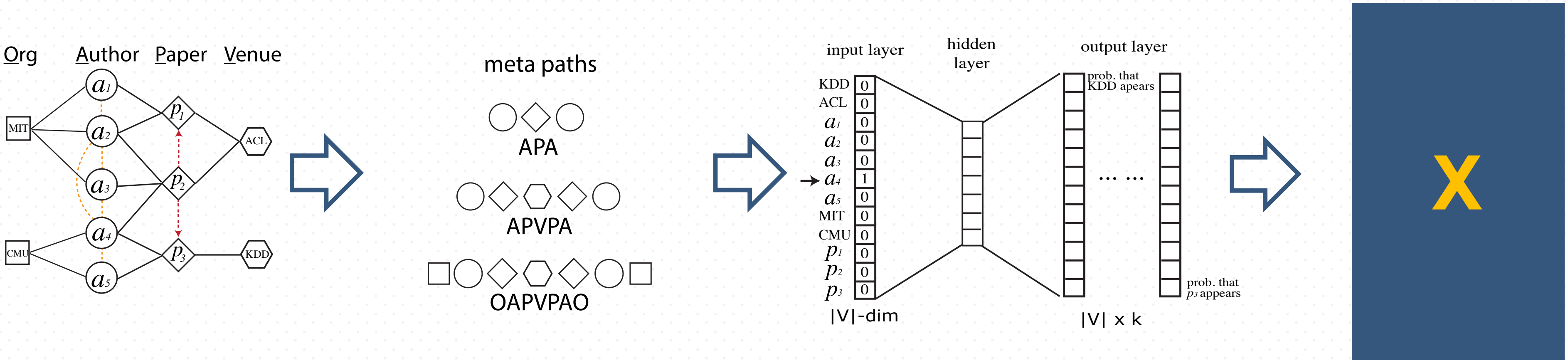


latent representation vector

metapath2vec

- ♣ Input: **a heterogeneous information network** $G = (V, E, T)$
- ♣ Output: $\boldsymbol{X} \in R^{|V| \times d}, d \ll |V|$, d-dim vector $\boldsymbol{X}_v$ for each node v.

probabilistic meta paths heterogeneous skip-gram



latent representation vector

- ○ transition probability

$$p(v^{i+1}|v_t^i, \mathcal{P}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t^i) \in E, \ \phi(v^{i+1}) = t+1 \\ 0 & (v^{i+1}, v_t^i) \in E, \ \phi(v^{i+1}) \neq t+1 \\ 0 & (v^{i+1}, v_t^i) \notin E \end{cases}$$

To predict the context node c_t (type t) given a node v, metapath2vec encourages all types of nodes to appear in this context position

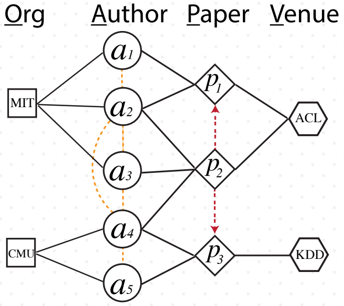
1. Y. Sun, J. Han. Mining heterogeneous information networks: Principles and Method
2. T. Mikolov, et al. Distributed representations of words and phrases and their compos

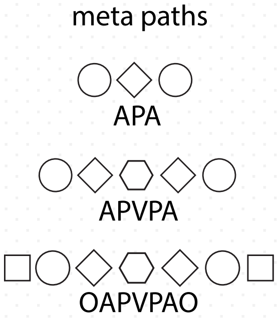
metapath2vec++

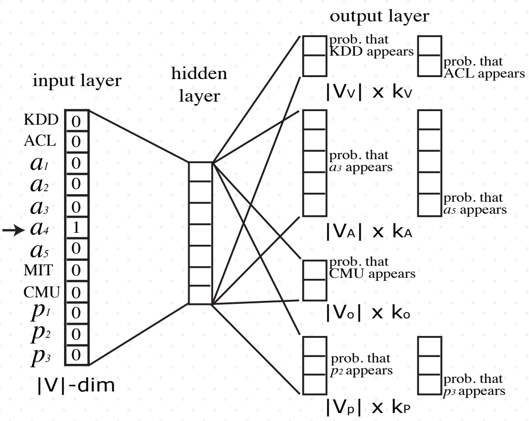
- Input: **a heterogeneous information network** $G = (V, E, T)$
- Output: $\boldsymbol{X} \in R^{|V| \times d}, d \ll |V|$, d-dim vector $\boldsymbol{X}_v$ for each node *v.*

probabilistic meta paths

heterogeneous skip-gram
heterogeneous negative sampling







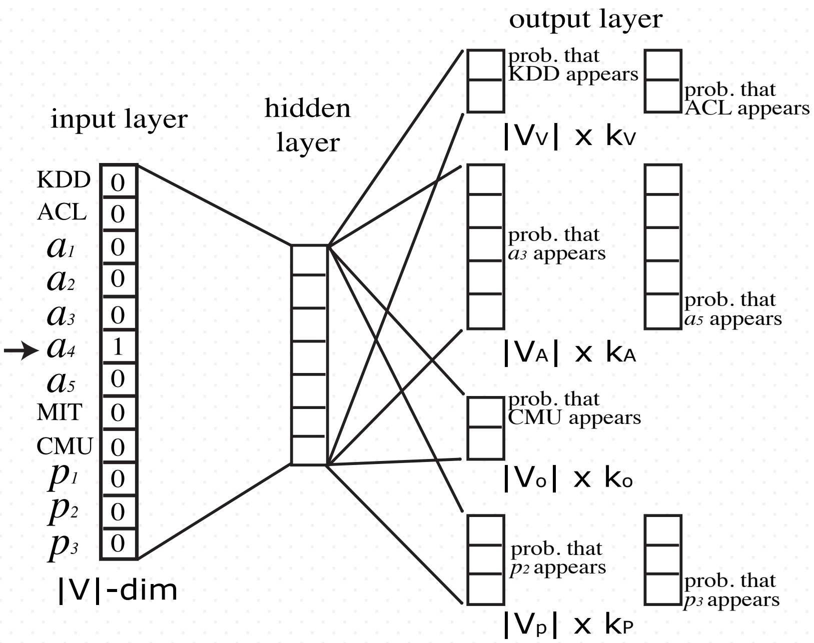
latent representation vector

- transition probability

$$p(v^{i+1}|v_t^i, \mathcal{P}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t^i) \in E, \ \phi(v^{i+1}) = t+1 \\ 0 & (v^{i+1}, v_t^i) \in E, \ \phi(v^{i+1}) \neq t+1 \\ 0 & (v^{i+1}, v_t^i) \notin E \end{cases}$$

1. T. Mikolov, et al. Distributed representations of words and phrases and their compositionality. In ***NIPS** '13*.

metapath2vec++



- ♣ network maximization

$$\arg\max_{\theta} \prod_{t \in T_V} \prod_{(v, c_t) \in G} p(c_t | v; \theta)$$

- ♣ softmax in *metapath2vec*

$$p(c_t | v; \theta) = \frac{e^{X_{c_t}} \cdot e^{X_v}}{\sum_{u \in V} e^{X_u} \cdot e^{X_v}}$$

- ♣ softmax in *metapath2vec++*

$$p(c_t | v; \theta) = \frac{e^{X_{c_t}} \cdot e^{X_v}}{\sum_{u_t \in V_t} e^{X_{u_t}} \cdot e^{X_v}}$$

- ♣ objective function (negative sampling)

$$\mathcal{O}(\mathbf{X}) = \log \sigma(X_{c_t} \cdot X_v) + \sum_{k=1}^{K} \mathbb{E}_{u_t^k \sim P_t(u_t)} [\log \sigma(-X_{u_t^k} \cdot X_v)]$$

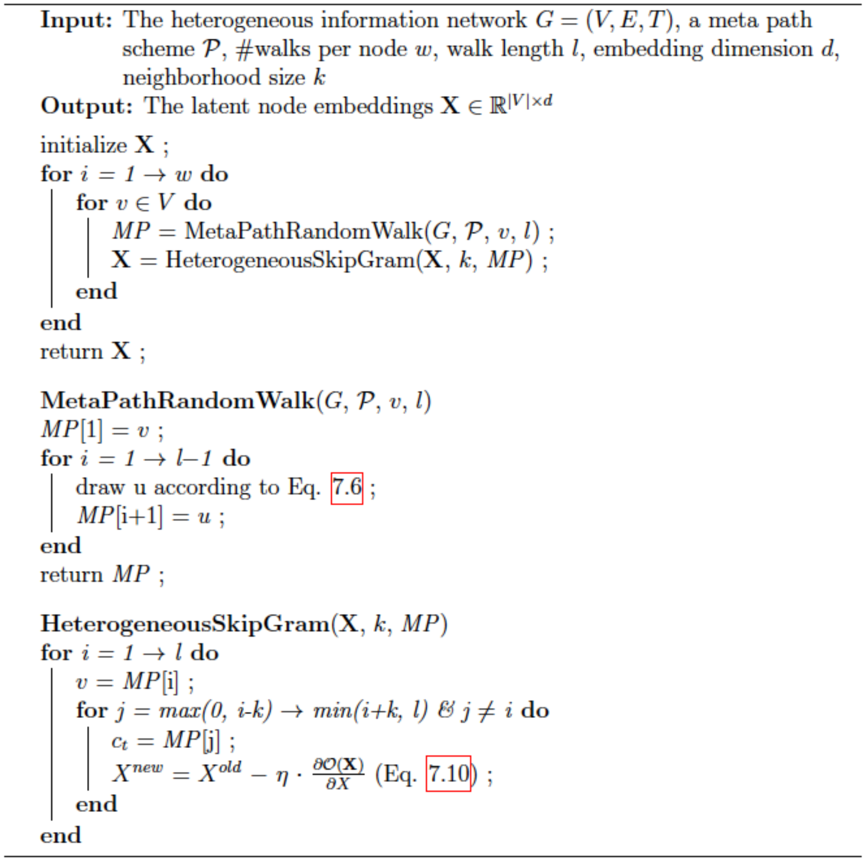
- ♣ stochastic gradient descent

$$\frac{\partial \mathcal{O}(\mathbf{X})}{\partial X_{u_t^k}} = (\sigma(X_{u_t^k} \cdot X_v - \mathbb{I}_{c_t}[u_t^k]))X_v$$

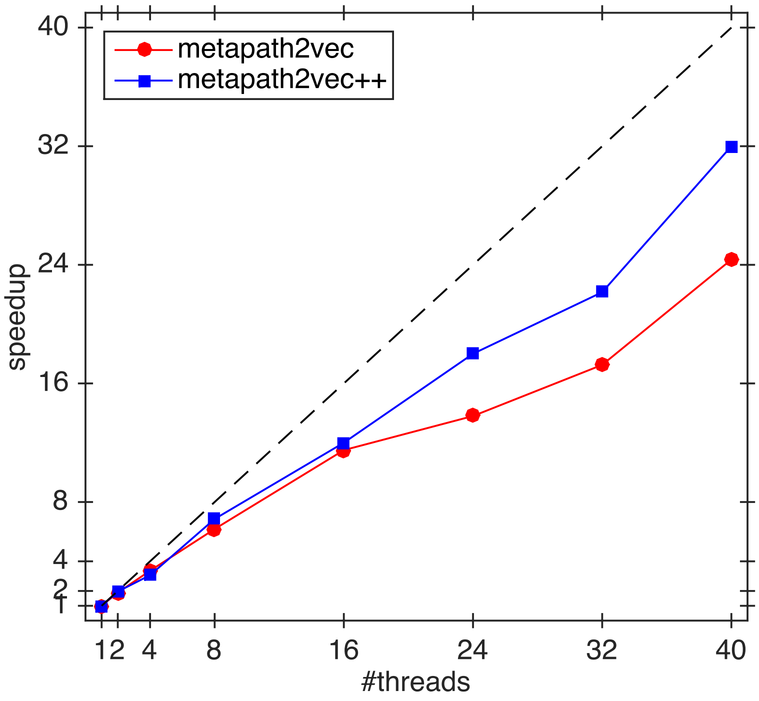
$$\frac{\partial \mathcal{O}(\mathbf{X})}{\partial X_v} = \sum_{k=0}^{K} (\sigma(X_{u_t^k} \cdot X_v - \mathbb{I}_{c_t}[u_t^k]))X_{u_t^k}$$

[1] T. Mikolov, et al. Distributed representations of words and phrases and their compositionality. In ***NIPS*** *'13.*

metapath2vec++

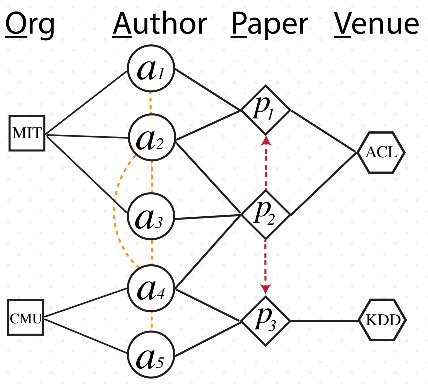


```
Input: The heterogeneous information network $G = (V, E, T)$, a meta path
       scheme $\mathcal{P}$, #walks per node $w$, walk length $l$, embedding dimension $d$,
       neighborhood size $k$
Output: The latent node embeddings $\mathbf{X} \in \mathbb{R}^{|V| \times d}$
initialize $\mathbf{X}$ ;
for $i = 1 \to w$ do
    for $v \in V$ do
        $MP$ = MetaPathRandomWalk($G$, $\mathcal{P}$, $v$, $l$) ;
        $\mathbf{X}$ = HeterogeneousSkipGram($\mathbf{X}$, $k$, $MP$) ;
    end
end
return $\mathbf{X}$ ;

MetaPathRandomWalk($G$, $\mathcal{P}$, $v$, $l$)
$MP[1] = v$ ;
for $i = 1 \to l{-}1$ do
    draw u according to Eq. 7.6 ;
    $MP$[i+1] = $u$ ;
end
return $MP$ ;

HeterogeneousSkipGram($\mathbf{X}$, $k$, $MP$)
for $i = 1 \to l$ do
    $v = MP[i]$ ;
    for $j = max(0, i{-}k) \to min(i{+}k, l)$ & $j \neq i$ do
        $c_t = MP[j]$ ;
        $X^{new} = X^{old} - \eta \cdot \frac{\partial \mathcal{O}(\mathbf{X})}{\partial X}$ (Eq. 7.10) ;
    end
end
```



- ♣ every sub-procedure is easy to parallelize
- ♣ 24-32X speedup by using 40 cores

Network Mining and Learning Paradigm



metapath2vec
metapath2vec++

X

latent representation vector

Network Applications

- ♣ **node attribute inference**
- ♣ **community detection**
- ♣ **similarity search**
- ♣ link prediction
- ♣ social recommendation
- ♣ …

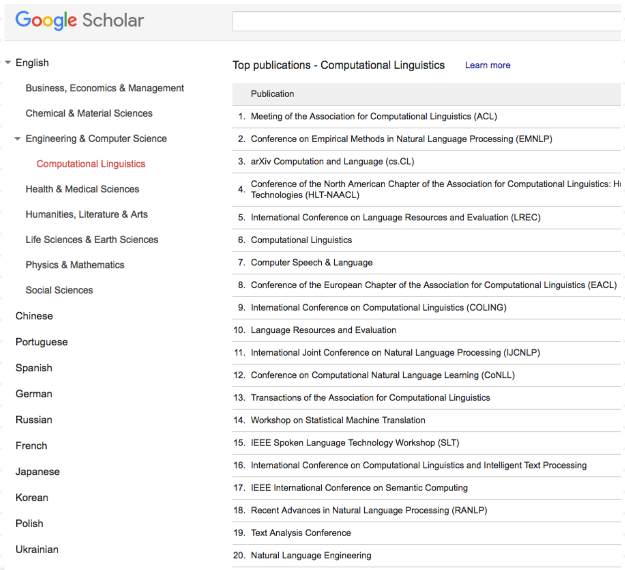
feature learning

machine learning models

Experiments

Heterogeneous Data

- AMiner CS publications
  - 8 categories of research areas


Google Scholar
English
Business, Economics & Management
Chemical & Material Sciences
Engineering & Computer Science
Computational Linguistics
Health & Medical Sciences
Humanities, Literature & Arts
Life Sciences & Earth Sciences
Physics & Mathematics
Social Sciences
Chinese
Portuguese
Spanish
German
Russian
French
Japanese
Korean
Polish
Ukrainian
Learn more

Top publications - Computational Linguistics

	Publication
1.	Meeting of the Association for Computational Linguistics (ACL)
2.	Conference on Empirical Methods in Natural Language Processing (EMNLP)
3.	arXiv Computation and Language (cs.CL)
4.	Conference of the North American Chapter of the Association for Computational Linguistics: H Technologies (HLT-NAACL)
5.	International Conference on Language Resources and Evaluation (LREC)
6.	Computational Linguistics
7.	Computer Speech & Language
8.	Conference of the European Chapter of the Association for Computational Linguistics (EACL)
9.	International Conference on Computational Linguistics (COLING)
10.	Language Resources and Evaluation
11.	International Joint Conference on Natural Language Processing (IJCNLP)
12.	Conference on Computational Natural Language Learning (CoNLL)
13.	Transactions of the Association for Computational Linguistics
14.	Workshop on Statistical Machine Translation
15.	IEEE Spoken Language Technology Workshop (SLT)
16.	International Conference on Computational Linguistics and Intelligent Text Processing
17.	IEEE International Conference on Semantic Computing
18.	Recent Advances in Natural Language Processing (RANLP)
19.	Text Analysis Conference
20.	Natural Language Engineering

Baselines

- DeepWalk [KDD ’14]
- node2vec [KDD ’16]
- LINE [WWW ’15]
- PTE [KDD ’15]

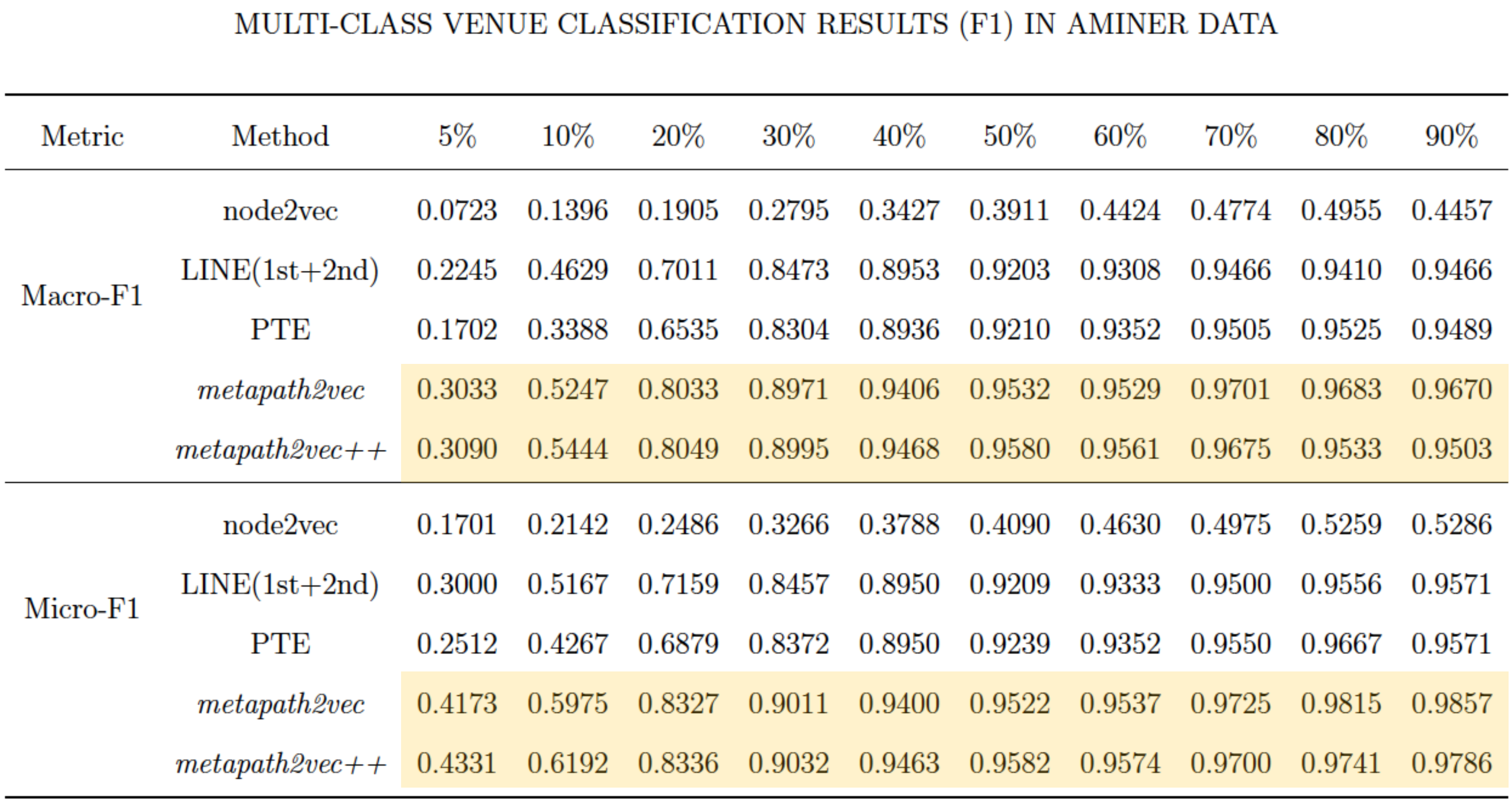
Parameters

- #walks: 1000
- walk-length: 100
- #dimensions: 128
- neighborhood size: 7

Mining Tasks

- Multi-class node classification
  - logistic regression
- node clustering
  - k-means
- similarity search
  - cosine similarity

Application 1: Multi-Class Node Classification



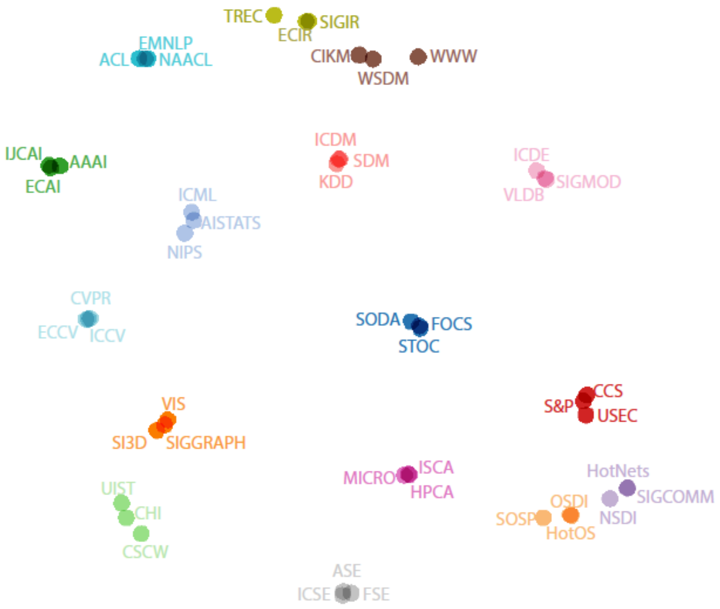
MULTI-CLASS VENUE CLASSIFICATION RESULTS (F1) IN AMINER DATA

Metric	Method	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Macro-F1	node2vec	0.0723	0.1396	0.1905	0.2795	0.3427	0.3911	0.4424	0.4774	0.4955	0.4457
	LINE(1st+2nd)	0.2245	0.4629	0.7011	0.8473	0.8953	0.9203	0.9308	0.9466	0.9410	0.9466
	PTE	0.1702	0.3388	0.6535	0.8304	0.8936	0.9210	0.9352	0.9505	0.9525	0.9489
	metapath2vec	0.3033	0.5247	0.8033	0.8971	0.9406	0.9532	0.9529	0.9701	0.9683	0.9670
	metapath2vec++	0.3090	0.5444	0.8049	0.8995	0.9468	0.9580	0.9561	0.9675	0.9533	0.9503
Micro-F1	node2vec	0.1701	0.2142	0.2486	0.3266	0.3788	0.4090	0.4630	0.4975	0.5259	0.5286
	LINE(1st+2nd)	0.3000	0.5167	0.7159	0.8457	0.8950	0.9209	0.9333	0.9500	0.9556	0.9571
	PTE	0.2512	0.4267	0.6879	0.8372	0.8950	0.9239	0.9352	0.9550	0.9667	0.9571
	metapath2vec	0.4173	0.5975	0.8327	0.9011	0.9400	0.9522	0.9537	0.9725	0.9815	0.9857
	metapath2vec++	0.4331	0.6192	0.8336	0.9032	0.9463	0.9582	0.9574	0.9700	0.9741	0.9786

Application 2: Node Clustering

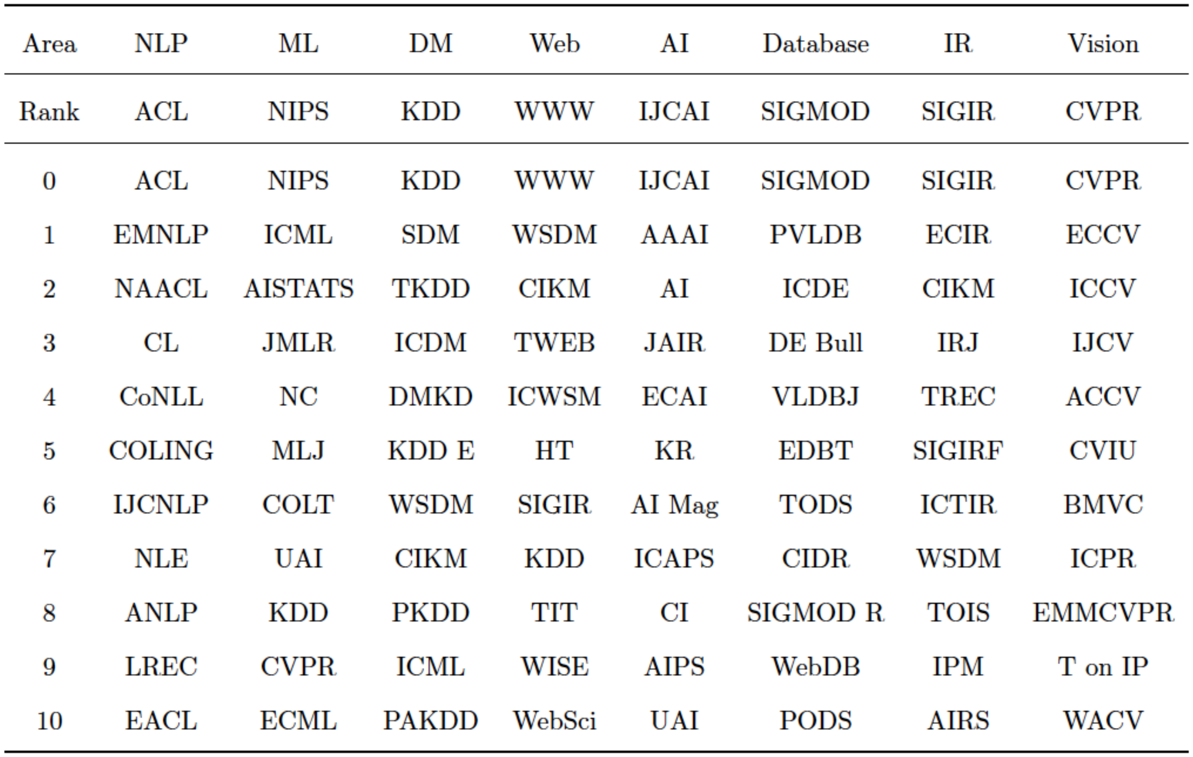
NODE CLUSTERING RESULTS (NMI) IN AMINER DATA

methods	venue	author
node2vec	0.1952	0.2941
LINE (1st+2nd)	0.8967	0.6423
PTE	0.9060	0.6483
metapath2vec	0.9274	0.7470
metapath2vec++	0.9261	0.7354



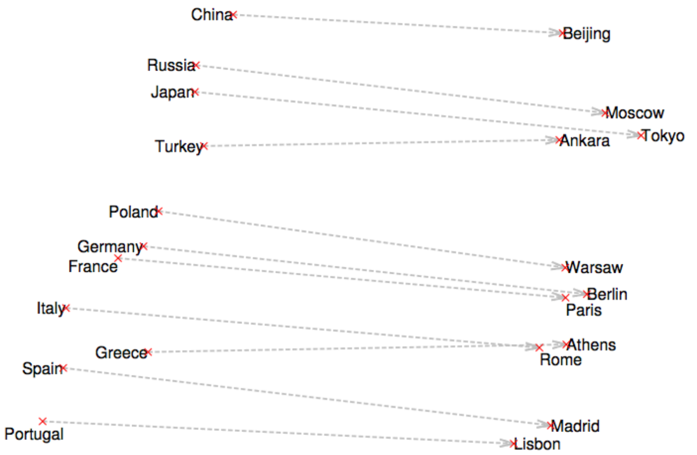
http://projector.tensorflow.org/

Application 3: Similarity Search

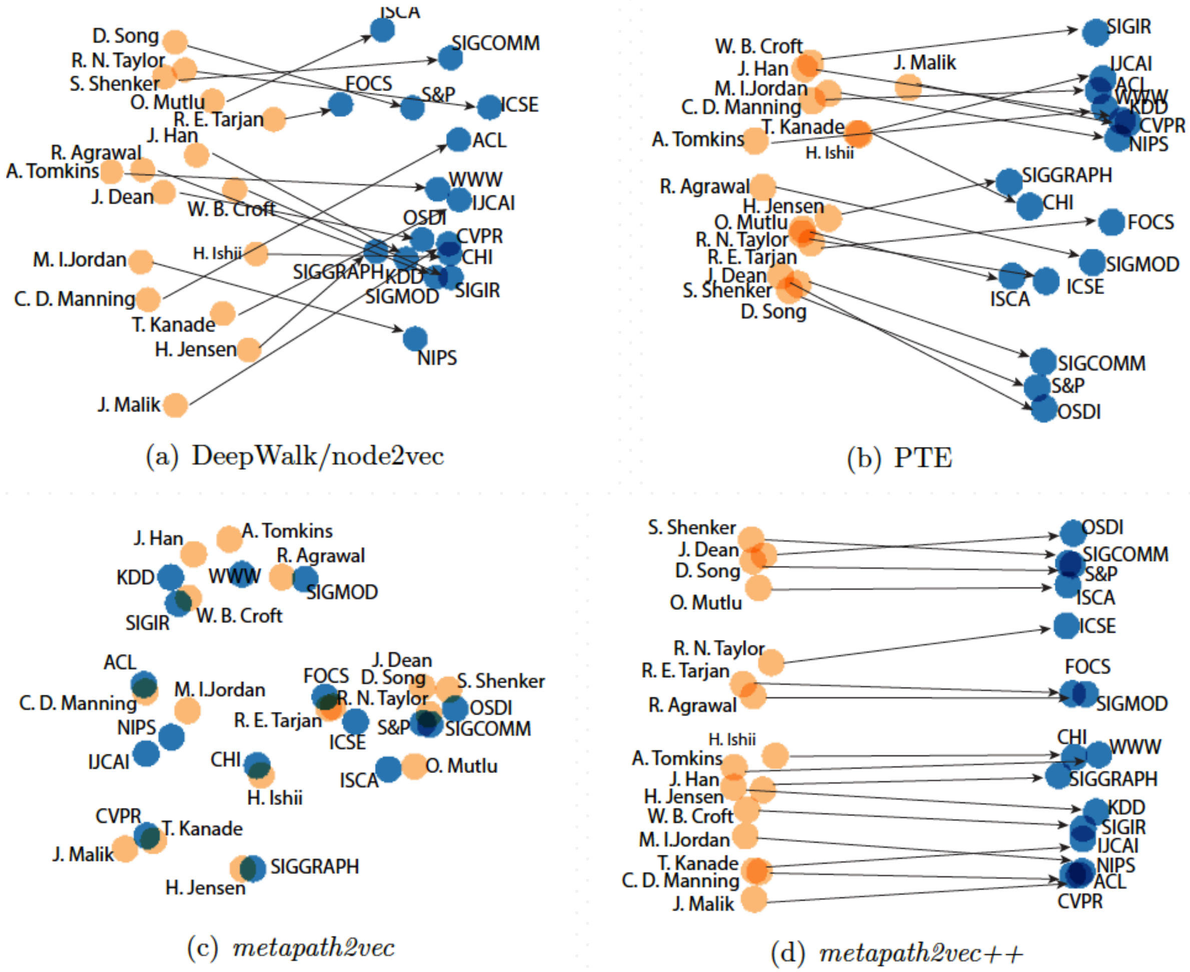


Area	NLP	ML	DM	Web	AI	Database	IR	Vision
Rank	ACL	NIPS	KDD	WWW	IJCAI	SIGMOD	SIGIR	CVPR
0	ACL	NIPS	KDD	WWW	IJCAI	SIGMOD	SIGIR	CVPR
1	EMNLP	ICML	SDM	WSDM	AAAI	PVLDB	ECIR	ECCV
2	NAACL	AISTATS	TKDD	CIKM	AI	ICDE	CIKM	ICCV
3	CL	JMLR	ICDM	TWEB	JAIR	DE Bull	IRJ	IJCV
4	CoNLL	NC	DMKD	ICWSM	ECAI	VLDBJ	TREC	ACCV
5	COLING	MLJ	KDD E	HT	KR	EDBT	SIGIRF	CVIU
6	IJCNLP	COLT	WSDM	SIGIR	AI Mag	TODS	ICTIR	BMVC
7	NLE	UAI	CIKM	KDD	ICAPS	CIDR	WSDM	ICPR
8	ANLP	KDD	PKDD	TIT	CI	SIGMOD R	TOIS	EMMCVPR
9	LREC	CVPR	ICML	WISE	AIPS	WebDB	IPM	T on IP
10	EACL	ECML	PAKDD	WebSci	UAI	PODS	AIRS	WACV

Visualization



word2vec [Mikolov, 2013]



(a) DeepWalk/node2vec

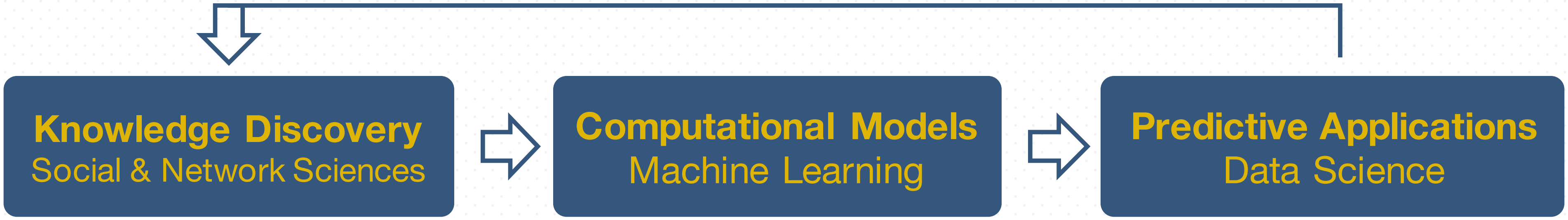
(b) PTE

(c) *metapath2vec*

(d) *metapath2vec++*

http://projector.tensorflow.org/

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

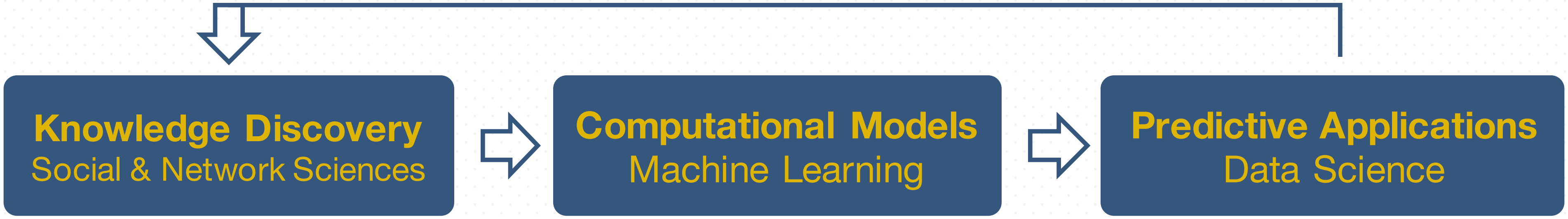
Predictive Applications
Data Science

Demographics	Local: ***Social Ties, Triads***	Global: ***Small Worlds***	Graphical Models: ***Demographic Prediction***
Diversity	Local: ***Common Neighborhood*** Global: ***Network Superfamily***	Topic: ***Social Impact***	Neural Networks: ***Heterogeneity Embedding***

Big Network Data: 120 large-scale networks

- Mobile network of 7+ million users & 1+ billion communications
- Friendster network of 60+ million users & 1.8 billion friendships

Computational Lens on Networks



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

- Common neighborhood signature
- Structural diversity violates homophily
- Authority facilitates influence growth

- *metapath2vec(++) model*
- Heterogeneous network embedding

- Future social impact prediction
- HIN mining and analysis tasks

- Lifetime evolution of social strategy
- Age-specific small worlds
- Demographics are predictable

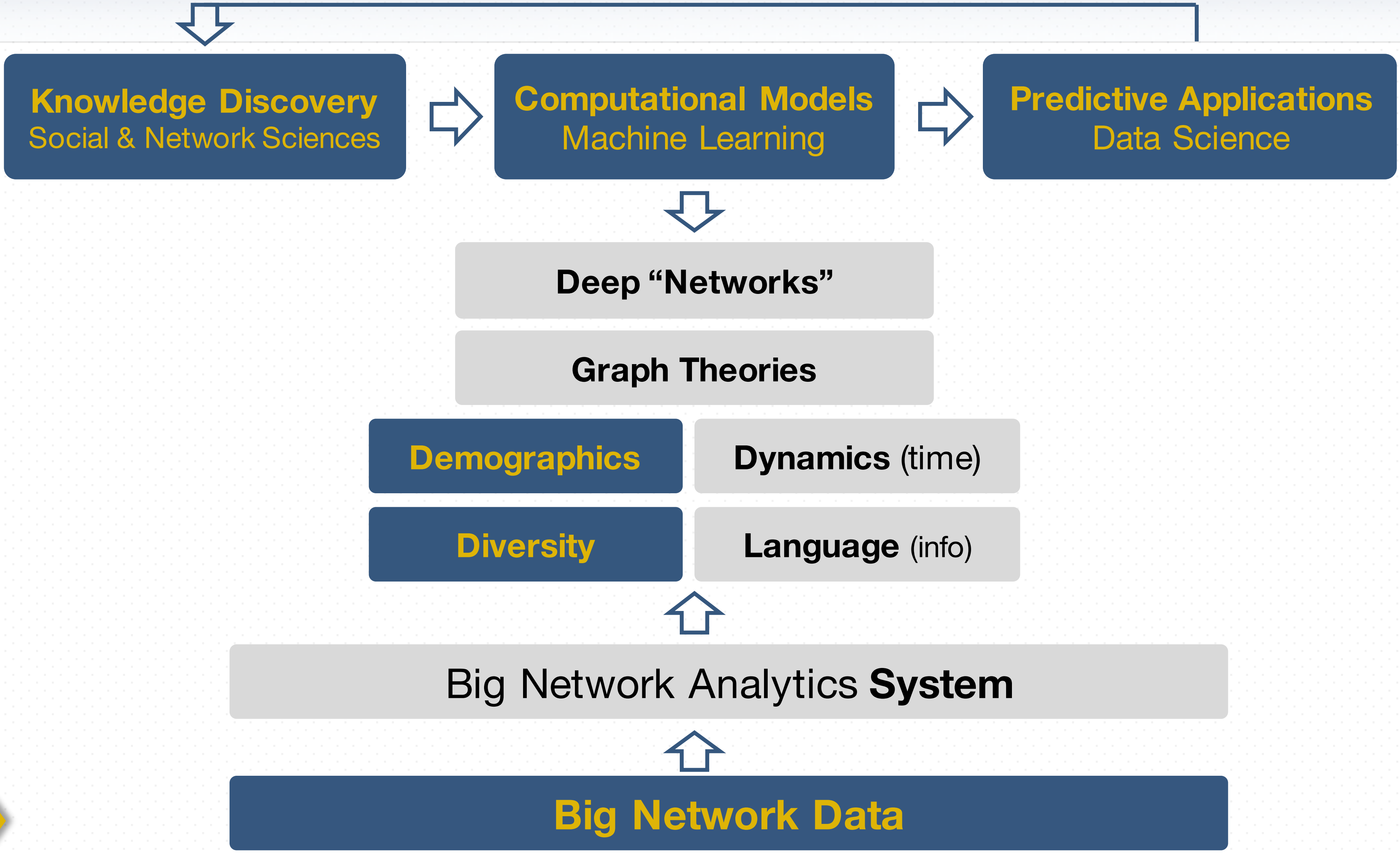
- *WhoAmI model*
- Probabilistic graphical models
- Distributed & coupled learning

- User Profiling in social networks
- Coupled user/link prediction

Big Network Data: 120 large-scale networks

- ○ Mobile network of 7+ million users & 1+ billion communications
- ○ Friendster network of 60+ million users & 1.8 billion friendships

Future Directions



Knowledge Discovery
Social & Network Sciences

Computational Models
Machine Learning

Predictive Applications
Data Science

Deep "Networks"

Graph Theories

Demographics

Dynamics (time)

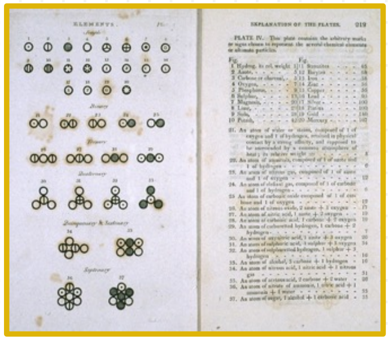
Diversity

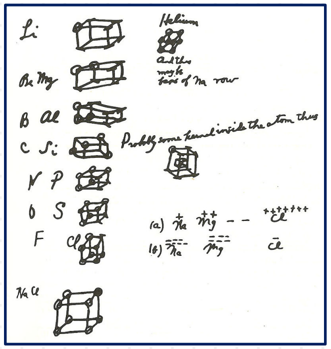
Language (info)

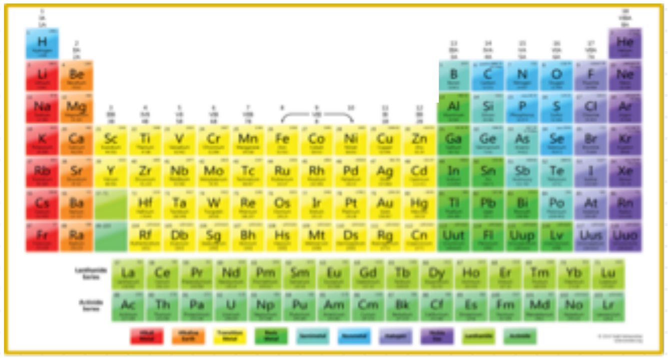
Big Network Analytics **System**

Big Network Data

Future^2: Back to Physical World


ELEMENTS
EXPLANATION OF THE PLATES.



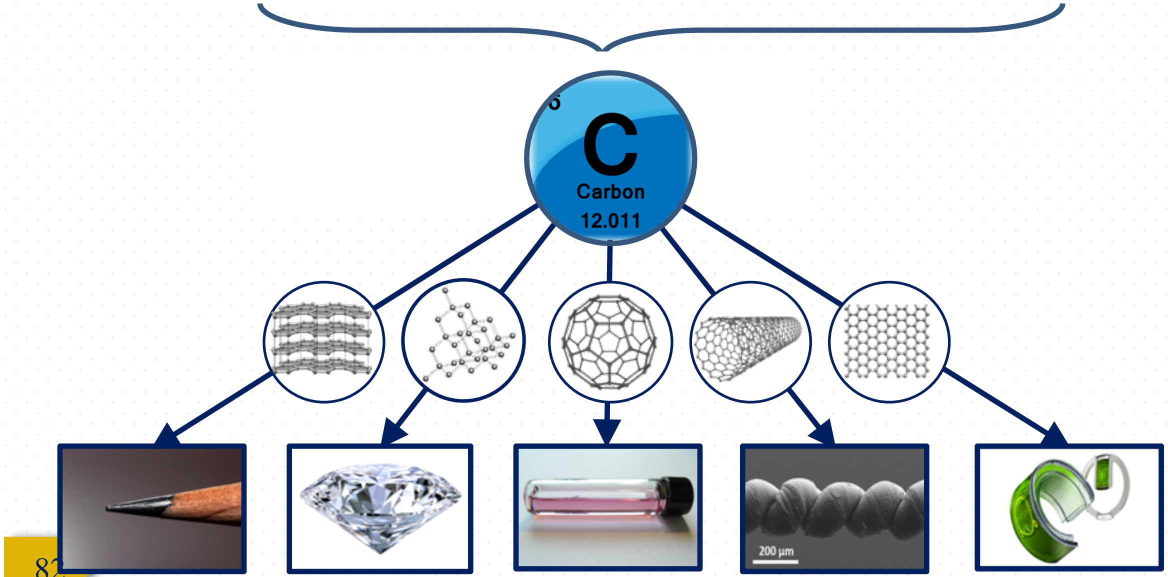


H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba		Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra		Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Uut	Fl	Uup	Lv	Uus	Uuo
			La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
			Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Atom
[Dalton, 1808]

Covalent Bond (Interactions)
[Gilbert Lewis, 1902 & 1916]

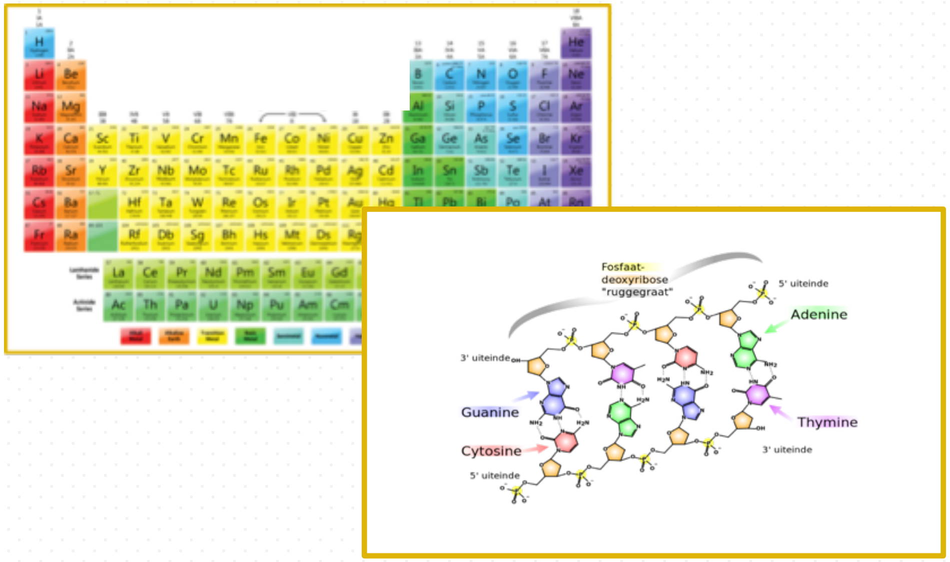
Periodic Table
[Mendeleev, 1869]



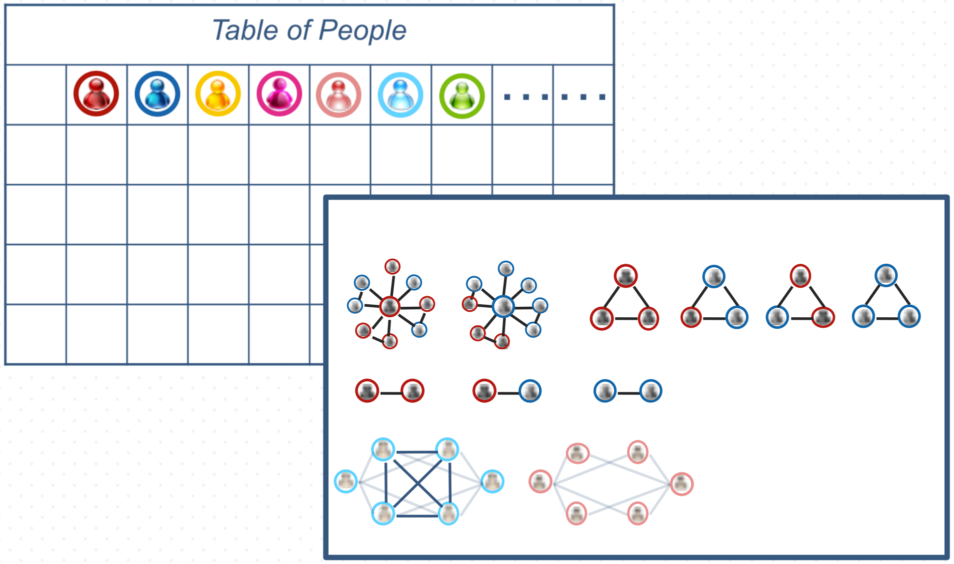
graphite diamond fullerene nanotube graphene

Future^2: Fundamental Elements & Principles in Social Networks

Different Atoms
[Mendeleev, **1869**]



Different Interactions
[Gilbert Lewis, **1902**]

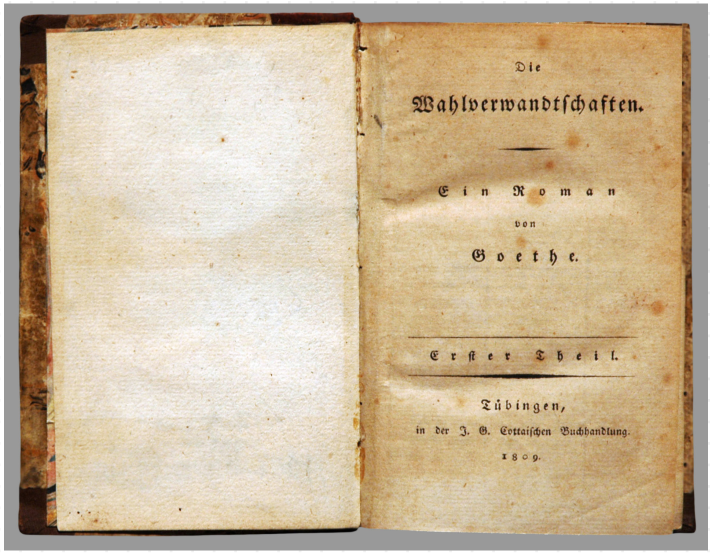


Physical World: Networks of Atoms

Social Space: Networks of People

Future^2: Fundamental Elements & Principles in Social Networks

"***Elective Affinities***[1] by Johann Goethe in 1809 is supposed to be the first work to model human relationships as chemical reactions or chemical processes … "[2]



Die

Wahlverwandtſchaften.

Ein Roman

von

Goethe.

Erſter Theil.

Tübingen,
in der J. G. Cottaiſchen Buchhandlung.
1809.

[1] Johann W. Goethe. Elective Affinities. Cottaische Publisher. 1809.

[2] Jeremy Adler. Goethe's Use of Chemical Theory in his Elective Affinities. Cambridge University Press. 1990.

References

1. G. Alexanderson, Euler and Konigsberg's bridges: a historical view. **Bulletin of the American Mathematical Society** 43 (4): 567. 2006.
2. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna. Four degress of separation. In ACM **WebSci'12**.
3. S. Milgram. The Small-World Problem. **Psychology Today**. 1967.
4. D. Watts, S. Strogatz. Collective Dynamics of Small-World Networks. **Nature 393**, 440-442.
5. A.-L Barabasi, R. Albert. Emergence of scaling in random networks. **Science 286** (5439): 590-512. 1999.
6. R. Dunbar. Neocortex size as a constraint on group size in primates. **Human Evolution**, 1992, 20: 469–493.
7. R. S. Burt. Structural holes: The social structure of competition. Harvard university press. 2009.
8. M. McPherson, L. Smith-Lovin, J. M. Cook. Birds of a feature: homophily in social networks. **Annual Review of Sociology**. 2001.
9. J. Ugander, L. Backstrom, C. Markow, J. Kleinberg. Structural Diversity in Social Contagion. **PNAS** 109(16) 5962-5966, 2012
10. M. Ercsey-Ravasz, Z. Toroczkai. Centrality scaling in large networks. **Physical Review Letters 105** (3), 038701, 2010.
11. J. Kleinberg. Authoritative sources in a hyperlinked environment. IBM **TR 10076, 1997** and In ACM-SIAM **SODA'98**.
12. M. Faloutsos, P. Faloutsos, C. Faloutsos. On power-law relationships of the internet topology. In ACM **SIGCOMM'99**.
13. J. Leskovec, J. Kleinberg, C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In ACM **KDD'05**.
14. J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási. Structure and tie strengths in mobile communication networks. **PNAS 104**, 2007.
15. V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, R. I. M. Dunbar. Sex differences in intimate relationships. In Nature **Scientific Reports 2**, 370, 2012.
16. M. E. J. Newman. Clustering and preferential attachment in growing networks. **Phys. Rev. E**. 2001.
17. D. Liben-Norwell, J. M. Kleinberg. The Link Prediction Problem for Social Networks. In ACM **CIKM'03**.
18. Y. LeCun, Y. Bengio, G. Hinton. Deep Learning. **Nature 521**, 436-444, 2015.
19. F. R. Kschischang, B. J. Frey, H. A. Loeliger. Factor graphs and the sum-product algorithm. **IEEE TOIT**, 47:498–519, 2001.
20. H.-A. Loeliger. An introduction to factor graphs. **IEEE Signal Processing Magazine**, 21(1):28–41, 2004.

References (cont.)

21. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su. ArnetMiner: Extraction and mining of academic social networks. In ACM **KDD'08**.
22. J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogeneous networks. In ACM **WSDM'12**.
23. T. Lou, J. Tang, J. Hopcroft, Z. Fang, X. Ding. Learning to predict reciprocity and triadic closure in social networks. ACM **TKDD**, 7(2):5:1–5:25, 2013.
24. J. Zhang, B. Liu, J. Tang, T. Chen, J. Li. Social influence locality for modeling retweeting behaviors. In **IJCAI'13**.
25. Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In ACM **WSDM'12**.
26. J. Leskovec, E. Horvitz. Planetary-scale views on a large instant-messaging network. In ACM **WWW'08**.
27. R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In ACM **JCDL'12**.
28. R. Lichtenwalter, J. T. Lussier, N. V. Chawla. New Perspectives and Methods in Link Prediction. In ACM **KDD'10**.
29. J. E. Hirsch. An index to quantify an individuals' scientific research output. **PNAS 102** (45). 2005.
30. D. Wang, C. Song, A.-L. Barabasi. Quantifying long-term scientific impact. **Science 342** (6154), 2013.
31. B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. **Science 342** (6157):468–472, 2013.
32. James A. Evans. Future Science, **Science 342**, 44, 2013
33. J. Cheng, L. Adamic, A. Dow, J. Kleinberg, J. Leskovec. Can cascades be predicted? In ACM **WWW'14**.
34. K. P. Murphy, Y. Weiss, M. I. Jordan. Loopy Belief Propagation for Approximate Inference: Am Empirical Study. In **UAI'99**.
35. M. Cha, H. Haddadi, F. Benevenuto, P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In AAAI **ICWSM'10**.
36. J. Dalton. A new system of chemical philosophy. 1808.
37. Picture --- Physical world global: http://diva-diary.com/can-you-name-the-50-most-important-capital-cities-in-the-world
38. Picture --- Network background: http://pacificaweb.com/social-media-marketing.html
39. Picture --- DNA structure: https://nl.wikipedia.org/wiki/Desoxyribonucle%C3%AFnezuur
40. Photos: Personal academic website or department roster.

Publications (covered)

1. Yuxiao Dong, Reid A. Johnson, Jian Xu, Nitesh V. Chawla. Structural Diversity and Homophily: A Study Across One Hundred Big Networks. In ***ACM KDD'17***. Updated on May, 2017
2. Yuxiao Dong, Nitesh V. Chawla, Ananthram Swami, Ram Ramanathan. metapath2vec: Scalable Representation Learning for Heterogeneous Information Networks. In ***ACM KDD'17*** Updated on May, 2017
3. Yuxiao Dong, Nitesh V. Chawla, Jie Tang, Yang Yang, Yang Yang. User Modeling on Demographic Attributes in Big Mobile Social Networks. In ACM Transactions on Information Systems (***ACM TOIS 2017***), accepted.
4. Yuxiao Dong*, Reid A. Johnson*, Nitesh V. Chawla. Can Scientific Impact Be Predicted?. In IEEE Transactions on Big Data (***IEEE TBD 2016***), 2016. ***Equal Contributions**.
5. Yuxiao Dong, Reid A. Johnson, Nitesh V. Chawla. Will This Paper Increase Your h-index? Scientific Impact Prediction. In ***ACM WSDM'15***. **Best Paper Award Nomination.**
6. Yuxiao Dong, Jing Zhang, Jie Tang, Nitesh V. Chawla, Bai Wang. CoupledLP: Link Prediction in Coupled Networks. In ***ACM KDD'15***.
7. Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla. Inferring User Demographics and Social Strategies in Mobile Social Networks. In ***ACM KDD'14***.

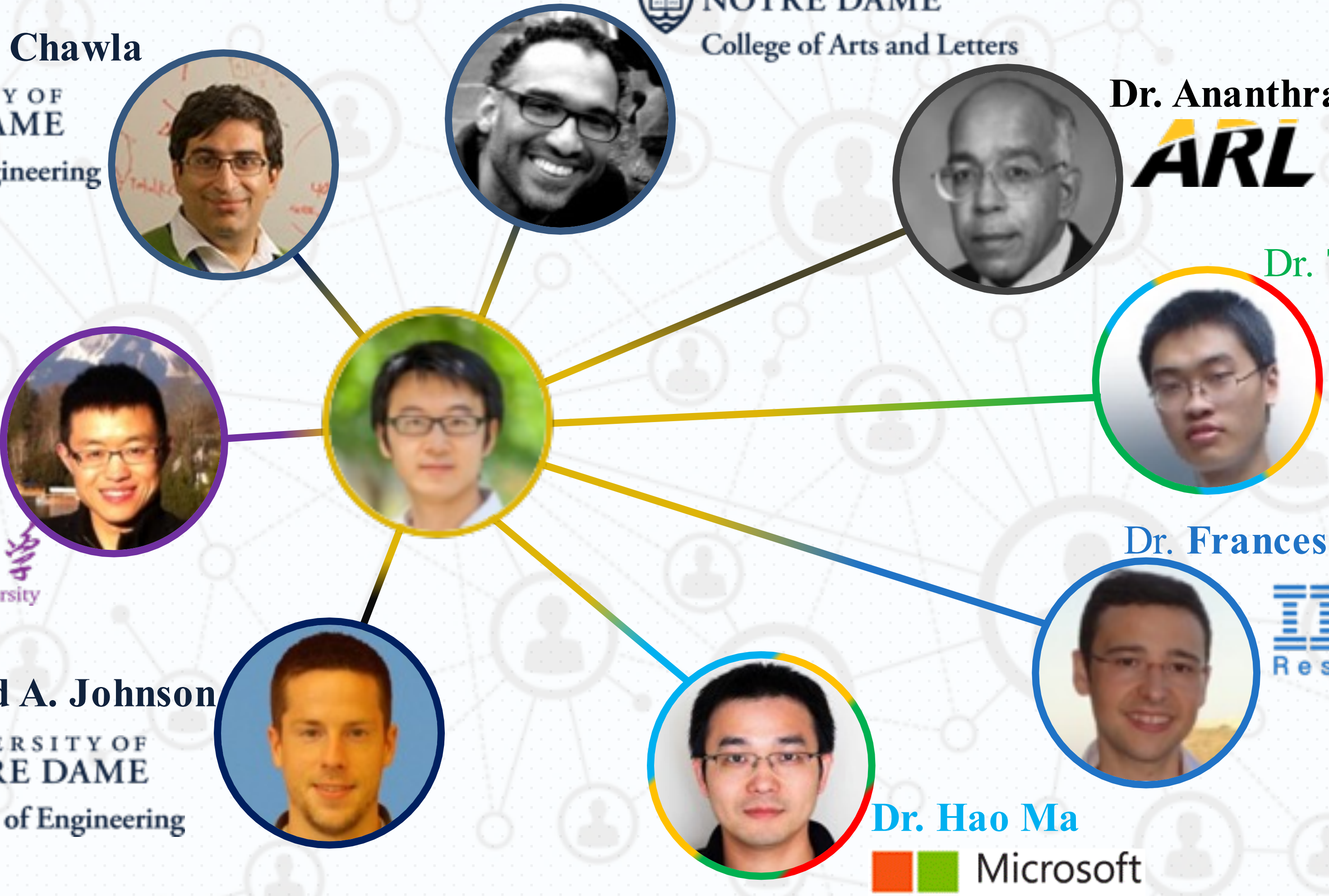
Pre-Prints (covered)

1. Yuxiao Dong*, Omar Lizardo*, Nitesh V. Chawla. Do the Young Live in a "Smaller World" than The Old? Age-Specific Degrees of Separation in Mobile Communication. http://arxiv.org/abs/1606.07556. ***Equal Contributions**.

Publications (others)

1. Siddharth Pal, Yuxiao Dong, Bishal Thapa, Nitesh V Chawla, Ananthram Swami, Ram Ramanathan. Deep Learning for Network Analysis: Problems, Approaches and Challenges. In ***MILCOM'16***.
2. Yuxiao Dong. User Modeling in Large Social Networks. In ***ACM WSDM'16 DC***. Doctoral Consortium paper, 1 page.
3. Ashwin Bahulkar, Boleslaw K. Szymanski, Omar Lizardo, Yuxiao Dong, Yang Yang, Nitesh V. Chawla. Analysis of Link Formation, Persistence and Dissolution in NetSense Data. In SNAA'16. **Best Paper Award Nomination.**
4. Yuxiao Dong, Jie Tang, Nitesh V. Chawla, Tiancheng Lou, Yang Yang, Bai Wang. Inferring Social Status and Rich Club Effects in Enterprise Communication Networks. In ***PLOS ONE 2015***.
5. Yuxiao Dong, Reid A. Johnson, Yang Yang, Nitesh V. Chawla. Collaboration Signatures Reveal Scientific Impact. In ***ACM/IEEE ASONAM'15***.
6. Yuxiao Dong, Fabio Pinelli, Yiannis Gkoufas, Zubair Nabi, Francesco Calabrese, Nitesh V. Chawla. Inferring Unusual Crowd Events From Mobile Phone Call Detail Records. In ***ECML/PKDD'15***.
7. Yang Yang, Yuxiao Dong, Nitesh V. Chawla. Predicting Node Degree Centrality with the Node Prominence Profile. ***Scientific Reports 2014***.
8. Chuan Shi, Yanan Cai, Di Fu, Yuxiao Dong. A Link Clustering Based Overlapping Community Detection Algorithm. Data and Knowledge Engineering 2013. **Highly Cited Research Award in DKE 2017.**
9. Yuxiao Dong, Jie Tang, Tiancheng Lou, Bin Wu, Nitesh V. Chawla. How Long will She Call Me? Distribution, Social Theory and Duration Prediction. In ***ECML/PKDD'13***.
10. Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nitesh V. Chawla, Jinghai Rao, Huanhuan Cao. Link Prediction and Recommendation across Heterogeneous Social Networks. In ***IEEE ICDM'12***. **Top 3 Most Cited Papers Among 151 ICDM'12 Papers.**

Thanks All Collaborators



Prof. **Omar Lizardo**

UNIVERSITY OF NOTRE DAME
College of Arts and Letters

Prof. **Nitesh V. Chawla**

UNIVERSITY OF NOTRE DAME
College of Engineering

Dr. Ananthram Swami

ARL

Dr. **Tiancheng Lou**

Google
Baidu Research

Prof. **Jie Tang**

清華大學
Tsinghua University

Dr. **Francesco Calabrese**

IBM Research
vodafone

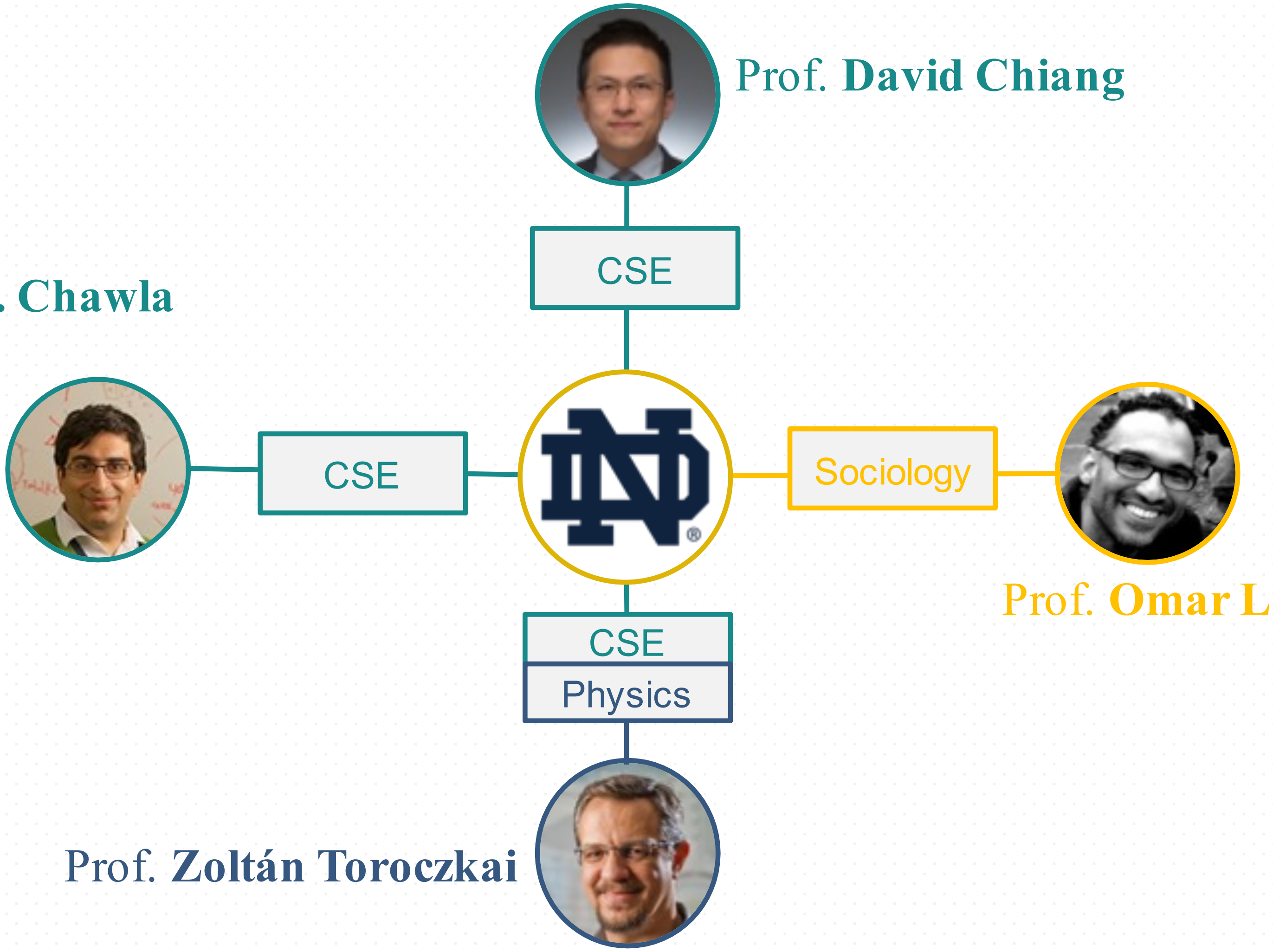
Prof. **Reid A. Johnson**

UNIVERSITY OF NOTRE DAME
College of Engineering

Dr. Hao Ma

Microsoft Research

Examination Committee



Prof. **David Chiang**

CSE

Prof. **Nitesh V. Chawla**
Chair

CSE

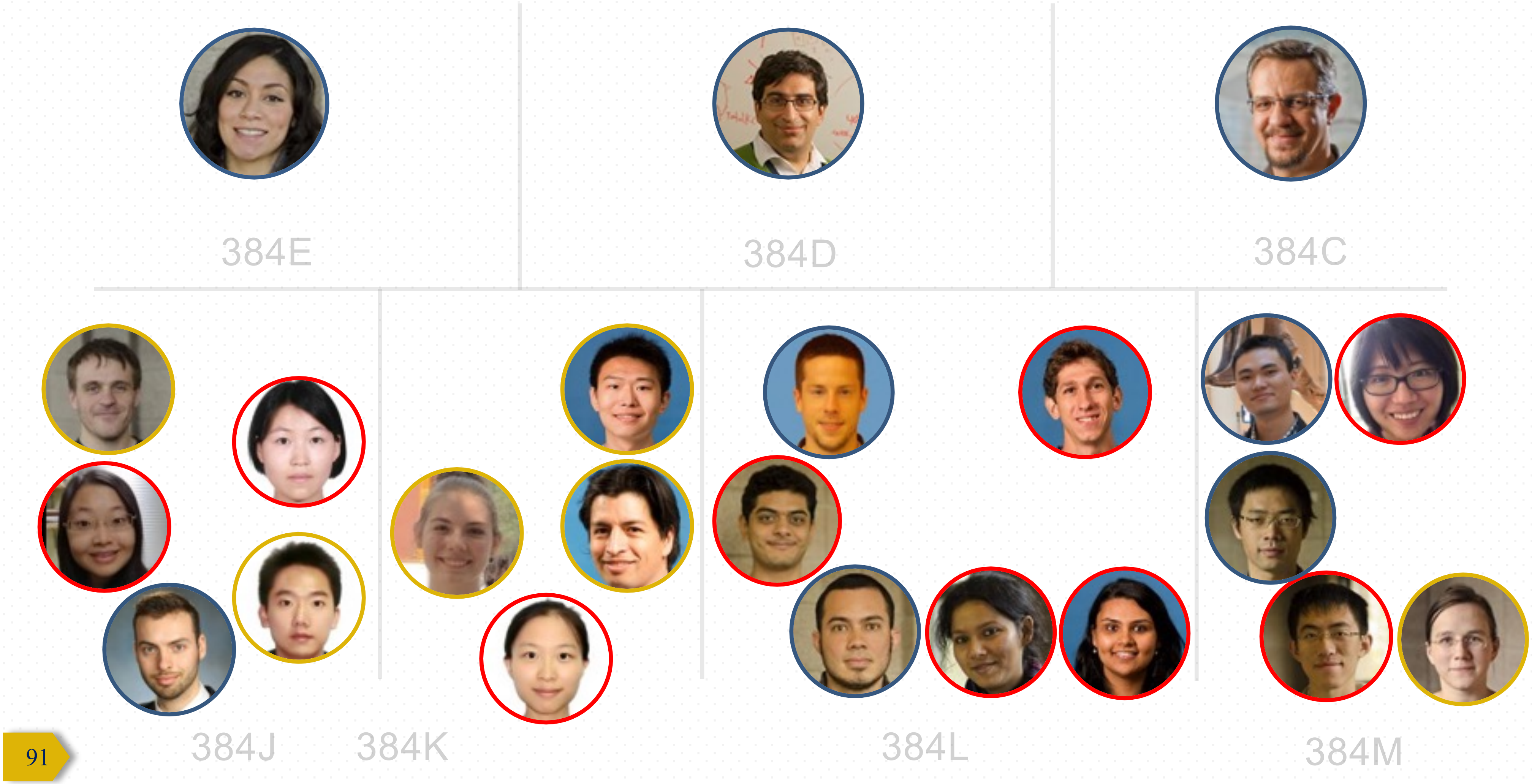
Sociology

Prof. **Omar Lizardo**

CSE
Physics

Prof. **Zoltán Toroczkai**

384 Nieuwland Hall & CSE Department



Thanks Jasmine & Joyce

Thank you!