Challenges in Mining Large-Scale Network Data

Jure Leskovec (@jure)
Stanford University

Including joint work with L. Backstrom, D. Huttenlocher, M. Gomez-Rodriguez, J. Kleinberg, J. McAuley, S. Myers

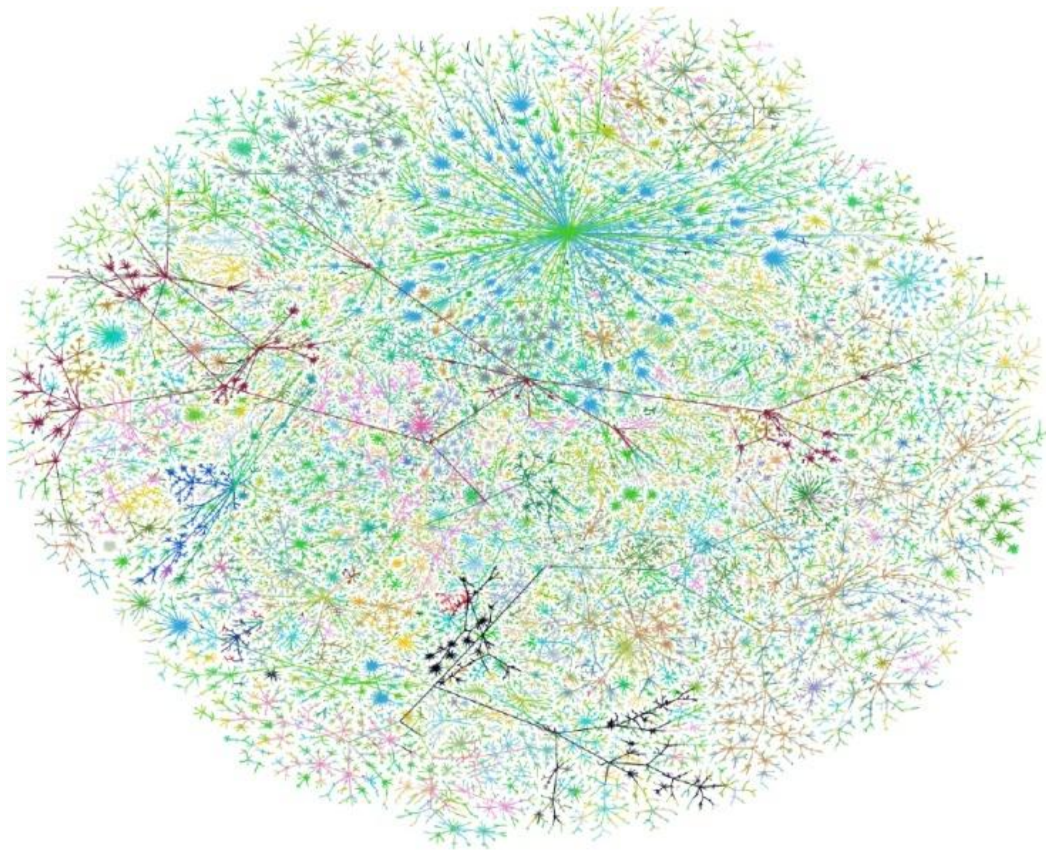
Data Mining & Networks

- Data mining has rich history and methods for analyzing …
  - … tabular data
  - … textual data
  - … time series & streams
  - … market baskets

  Bag of features

- What about relations and dependencies?

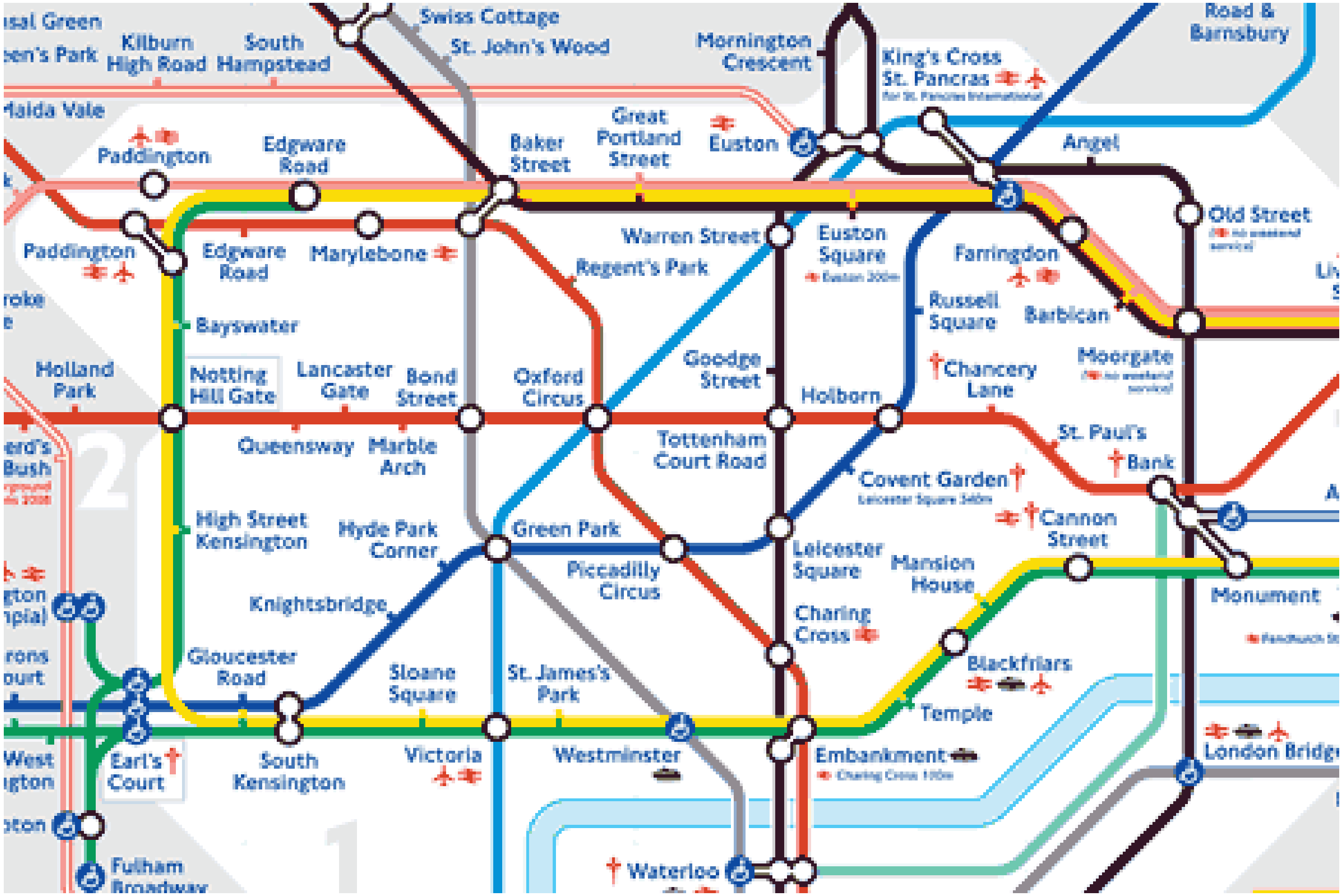
Network: A First Class Citizen



Networks allow for modeling dependencies!

Networks

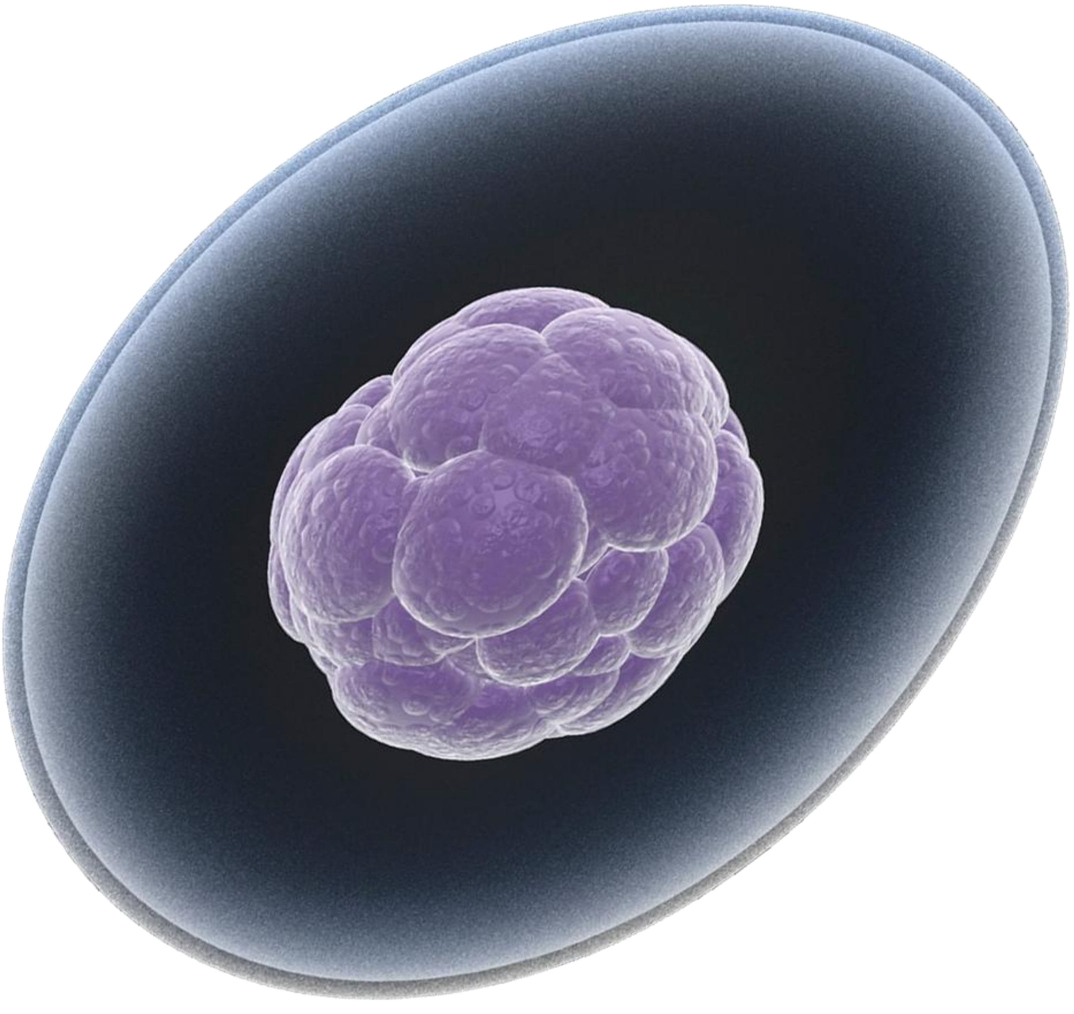
are a general language for describing real-world systems



Infrastructure


NASDAQ
KB Home
-2.2%
-1.16
50.96
Microsoft
+1.5%
+0.44
30.89
+2.0%
+0.35
17.84
BUY ORDERS
19.12
+1.10
+6.1%
19.09
OPLK
+1.2%
+0.27
23.46
+1.0%
+0.17
17.15
NDTV PROFIT

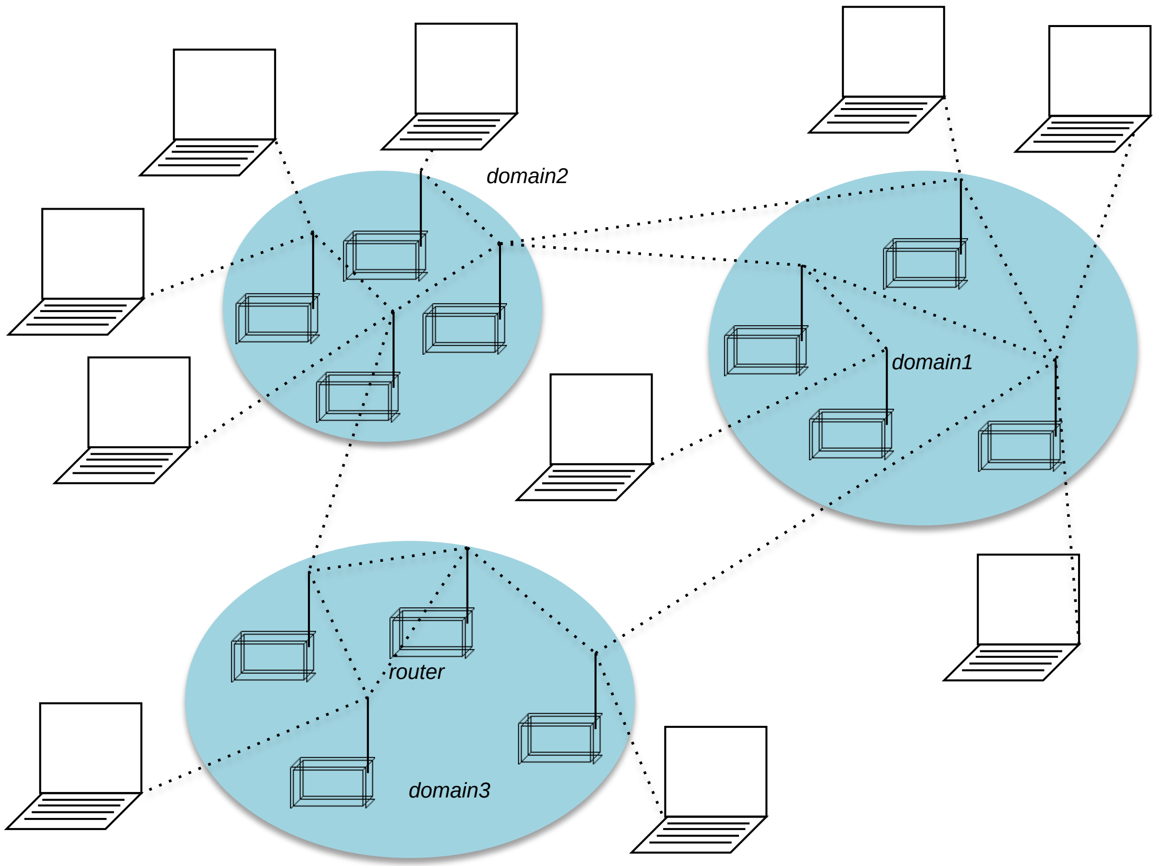
Economy



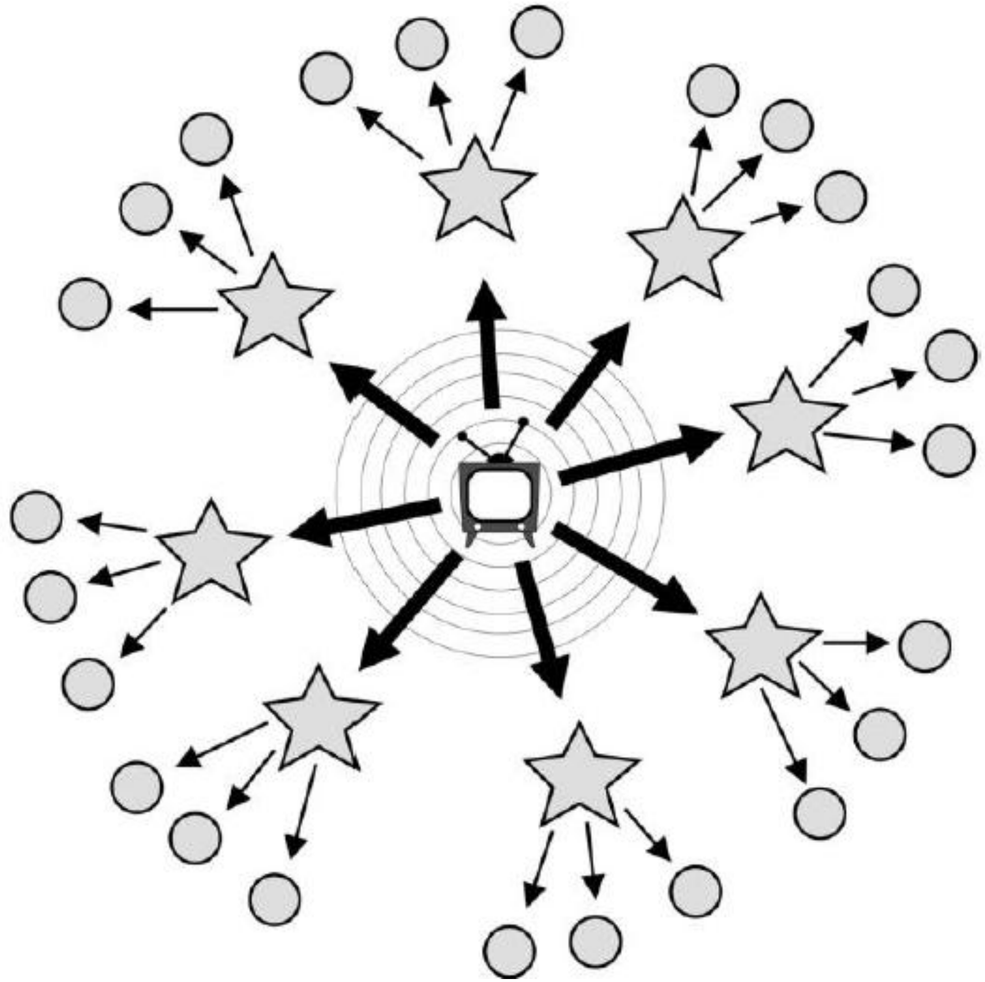
Human cell

Brain

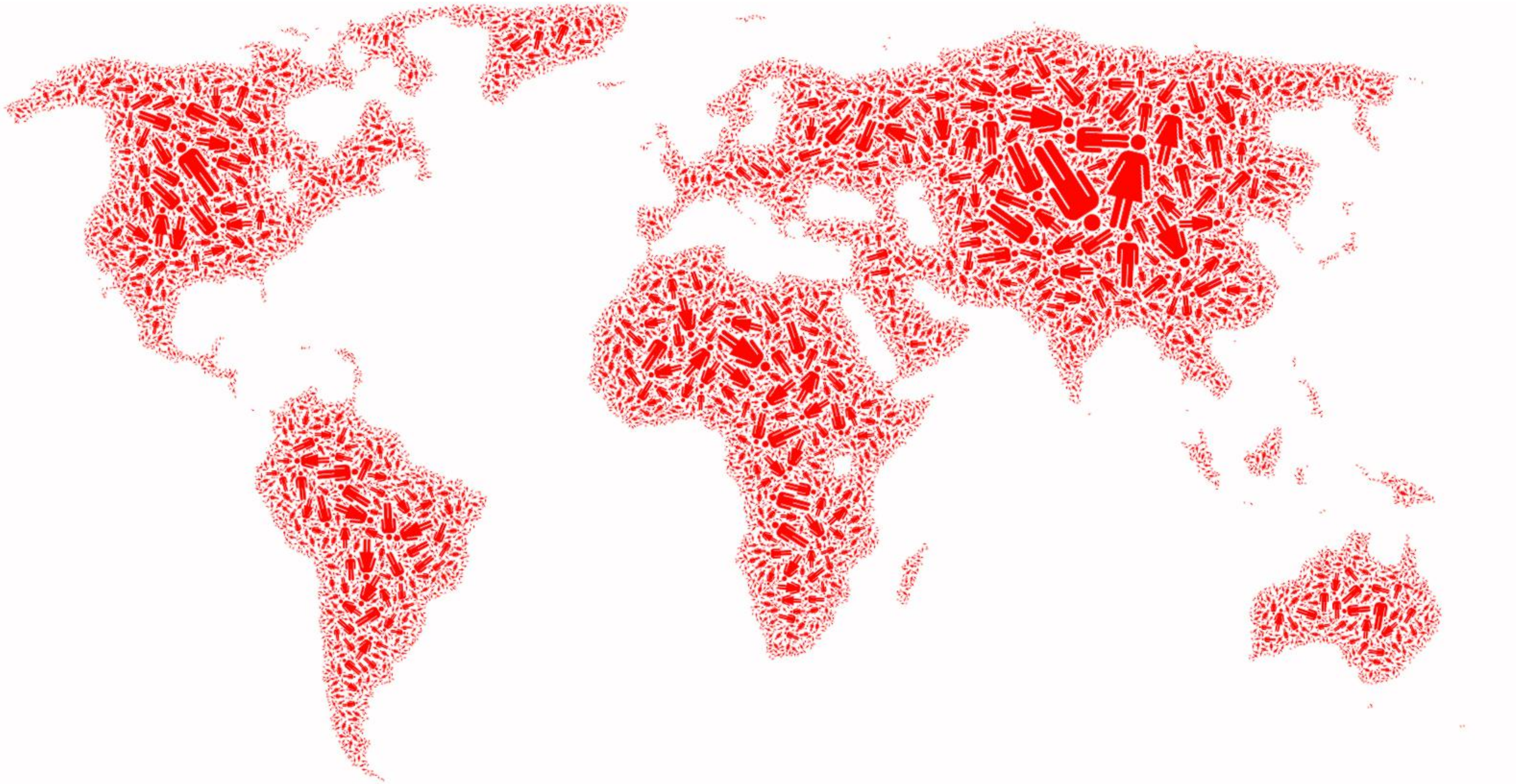
Friends & Family



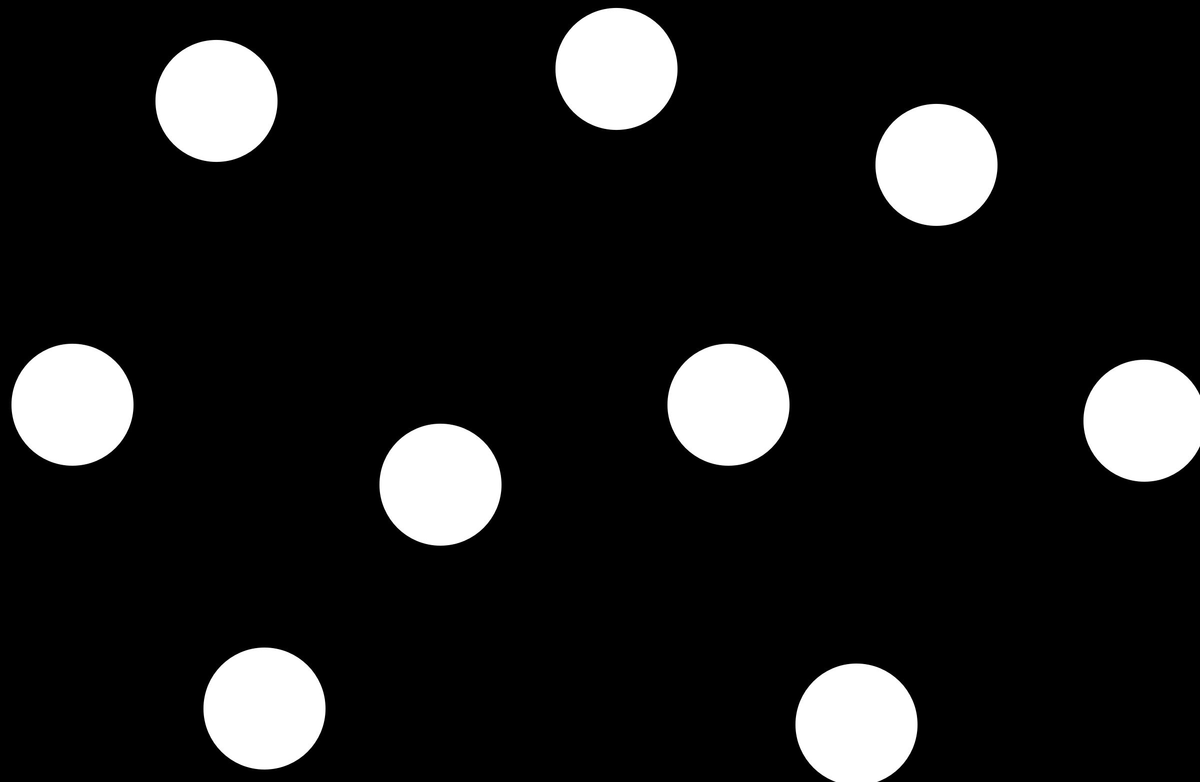
Internet



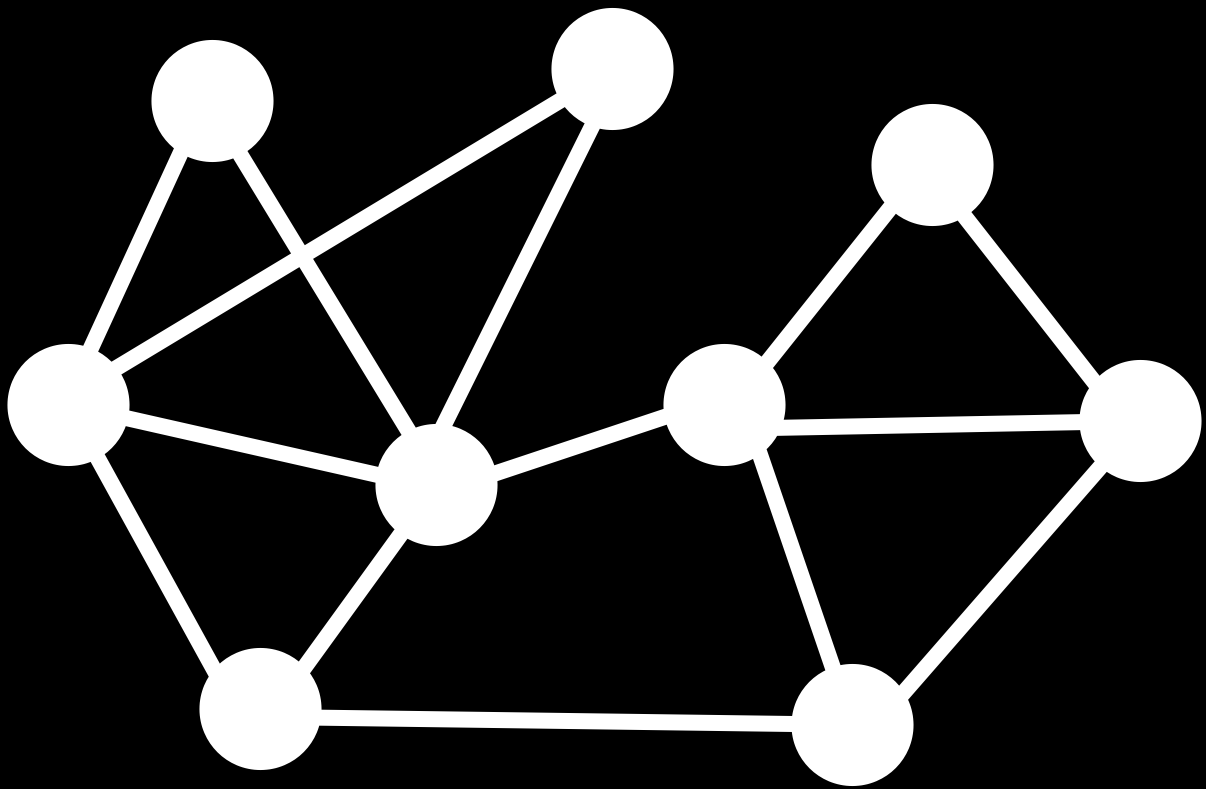
Media & Information



Society

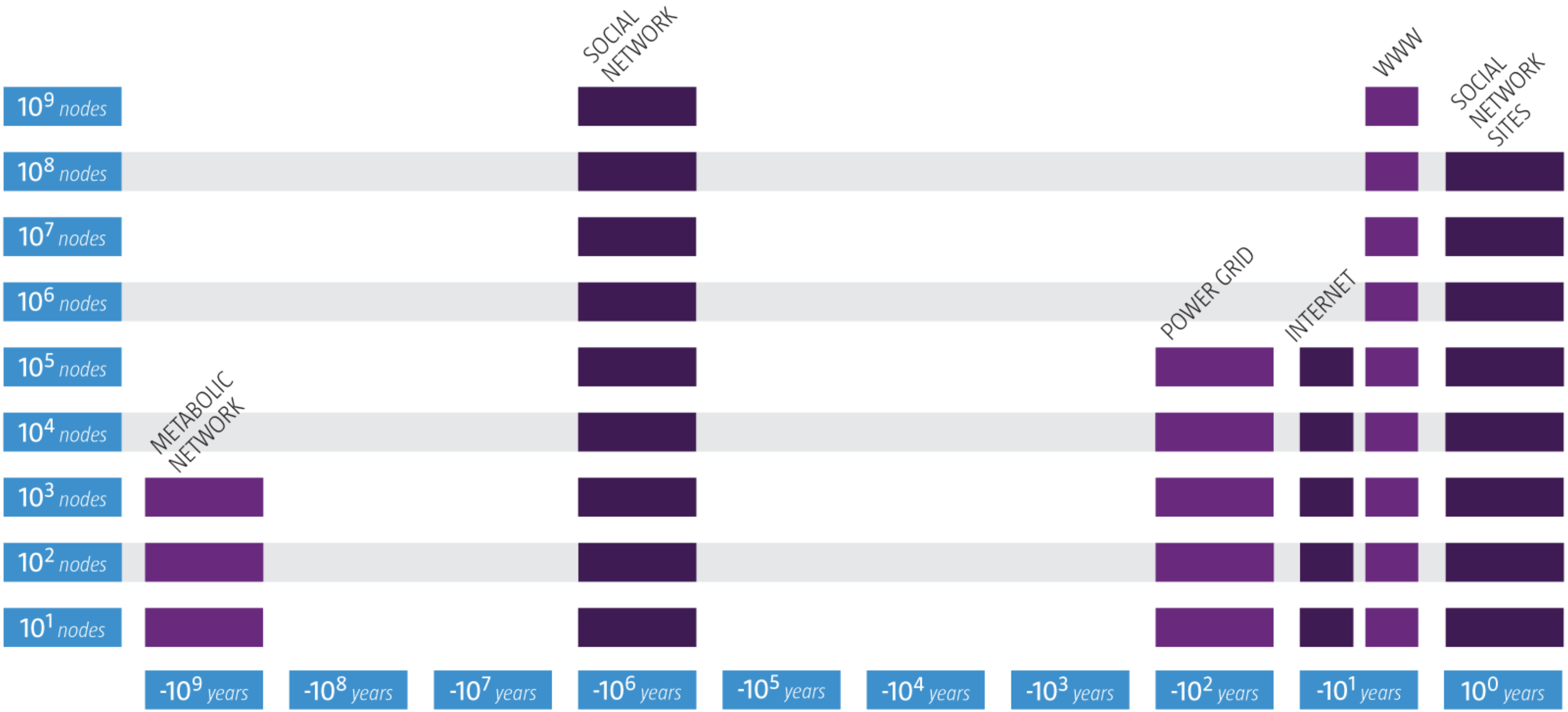


Network!



Network!

The Life of Networks



Networks, why now?

Transformation of Computing

Online friendships
[Ugander-Karrer-Backstrom-Marlow, '11]

Corporate e-mail communication
[Adamic-Adar, '05]

- **Web: a Social and a Technological network**
- **Profound transformation in:**
  - How knowledge is produced and shared
  - How people interact and communicate
  - **The scope of CS as a discipline**

Mining Network Data

- **Network data brings several questions:**
  - **Working with network data is messy**
    - Not just "wiring diagrams" but also dynamics and data (features, attributes) on nodes and edges
  - **Computational challenges**
    - Large scale network data
  - **Algorithmic models as vocabulary for expressing complex scientific questions**
    - Social science, physics, biology

Plan for the Talk

- **Plan for the talk:**
  Algorithms for network data
  - **Part 1)** How to we make online social networks more useful
    - **Finding Friends**
    - **Organizing Friends**
  - **Part 2)** Web as sensor into society
    - **Understanding Social Media Content**

Finding Friends

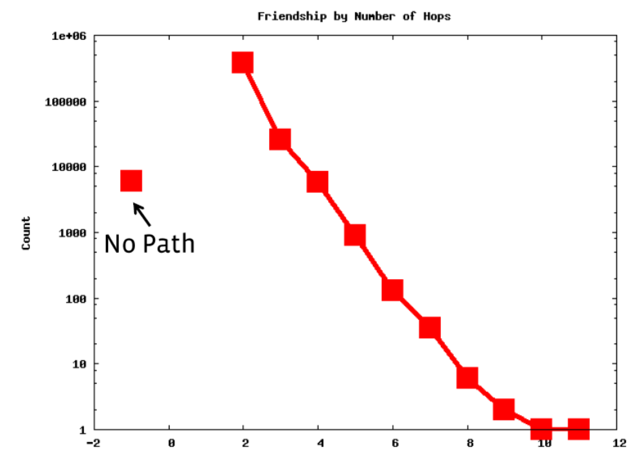
- **Growing body of research captures dynamics of social network graphs**
  [Latanzi, Sivakumar ’08] [Zheleva, Sharara , Getoor ‘09] [Kumar, Novak, Tomkins ‘06] [Kossinets, Watts ‘06] [L., Kleinberg, Faloutsos ’05]

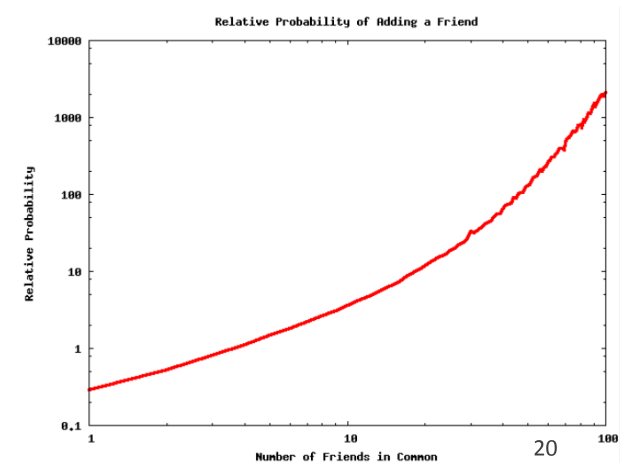
- **What links will occur next?** [LibenNowell, Kleinberg ‘03]
  - **Networks + many other features:**
    Location, School, Job, Hobbies, Interests, etc.

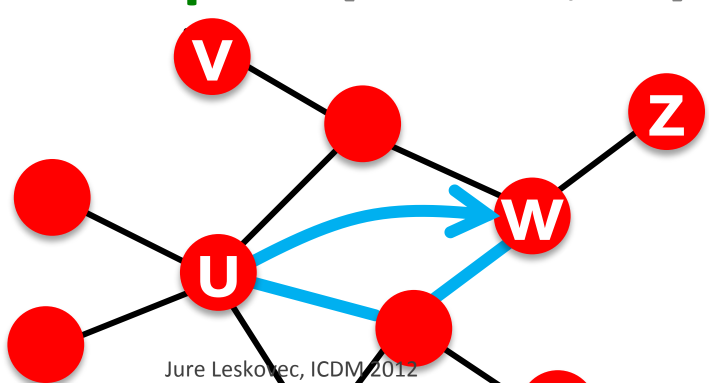
Friend Recommendation

[WSDM '11]

- **Learn to recommend potential friends**
- **Facebook link creation** [Backstrom, L. '11]
  - 92% of new friendships on FB are **friend-of-a-friend**
    - **Triadic closure** [Granovetter, '73]
  - More **common friends** helps:
    - **Social capital** [Coleman, '88]



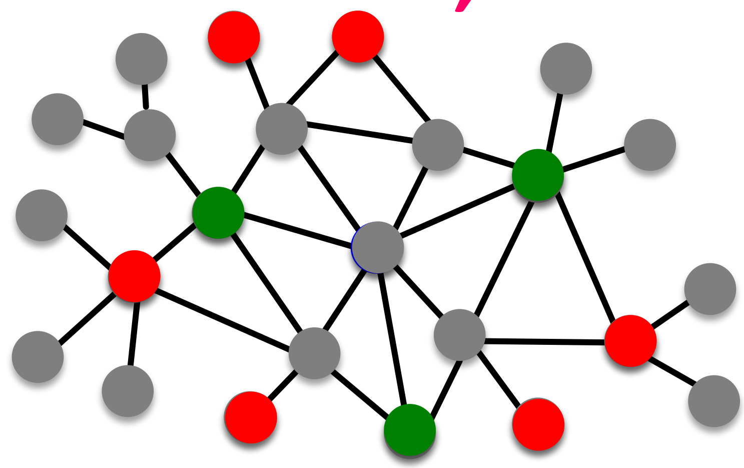




[WSDM '11]

Supervised Link Prediction

- **Goal:** Given a user s, recommend friends

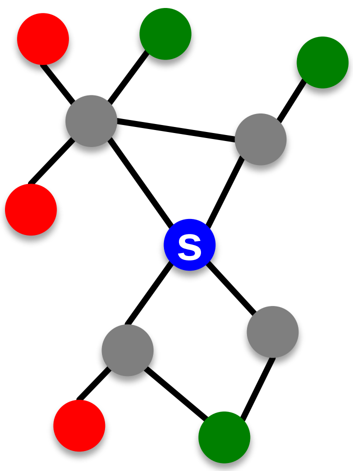


  - **Positive:** Nodes to which s links to in the future
  - **Negative:** Nodes to which s does not link
- **Supervised ranking problem:**
  - **Assign higher scores** to **positive nodes** than to **negative nodes**

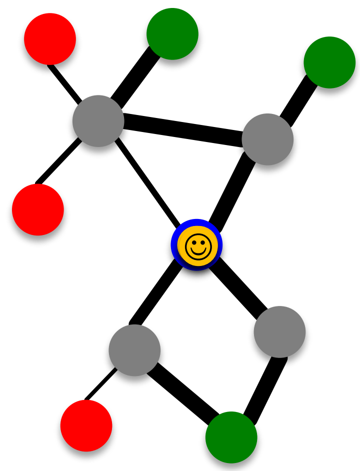
Supervised Link Prediction

- <u>**Q:**</u> How to combine **network structure** and **node and edge features**?
- <u>**A:**</u> Combine **PageRank** with **Supervised learning**
  - **PageRank** is great to capture importances of nodes based on the network structure
  - **Supervised learning** is great with features
- <u>**Idea:**</u> **Use node and edge features to "guide" the random walk**

Supervised Random Walks



Network



Set edge strengths
(want strong edges to point towards **positive** nodes)

Run *Random Walk with Restarts* on the weighted graph

RWR assigns an importance score (visiting probability) to every node

Recommend top *k* nodes with highest score

- **Q: How to set edge strengths?**
- **Idea:** Set edge strengths such that SRW correctly ranks the nodes on the training data

SRW: Learning

- **Goal:** Learn an edge strength function
  $f_\theta(x,y) = \exp(-\sum_i \theta_i \cdot \psi_i(x,y))$
  - $\psi(x,y)$ … features of edge (x,y)
  - θ_i … parameter vector we want to learn
- **Find** $f_\theta(u,v)$ **based on training data:**

$$arg\,min_\theta \sum_{p \in P} \sum_{n \in N} \delta(r_p < r_n) + \lambda ||\theta||^2$$

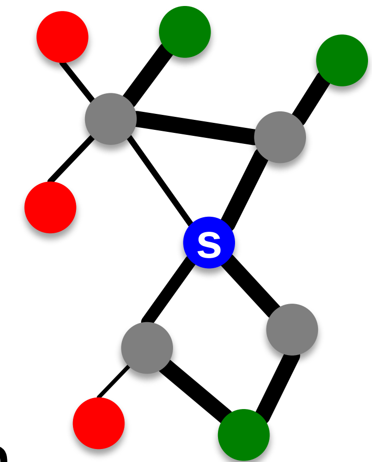
Positive nodes

Negative nodes

Penalty for violating constraint $r_p > r_n$

r_x … score of node x on a weighted graph with edge weights $f_\theta(x,y)$

Data: Facebook

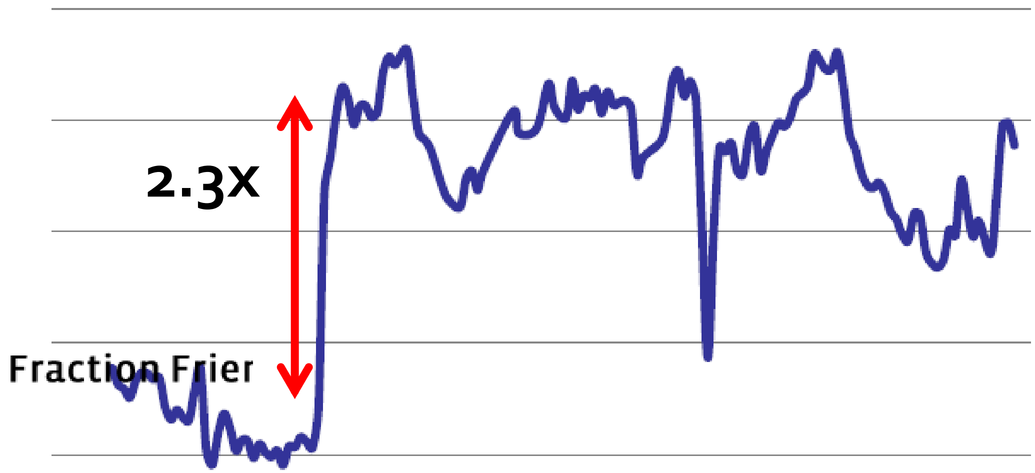


- **Facebook Iceland network**
  - 174,000 nodes (55% of population)
  - Avg. degree 168
  - Avg. person added 26 friends/month
- **Node and edge features:**
  - **Node:** Age, Gender, School
  - **Edge:** Age of an edge, Communication, Profile visits, Co-tagged photos

[WSDM '11]

Link Prediction

- **Results on Facebook Iceland:**
  - Correctly predicts **8** out of **20 (40%)** new friends
  - 2.3x improvement over previous FB-PYMK

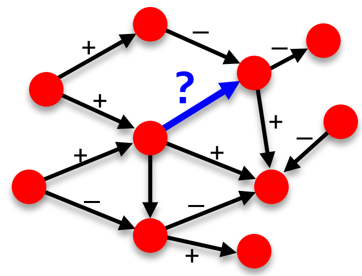


Fraction of friending based on recommendations

SRW: Further Questions

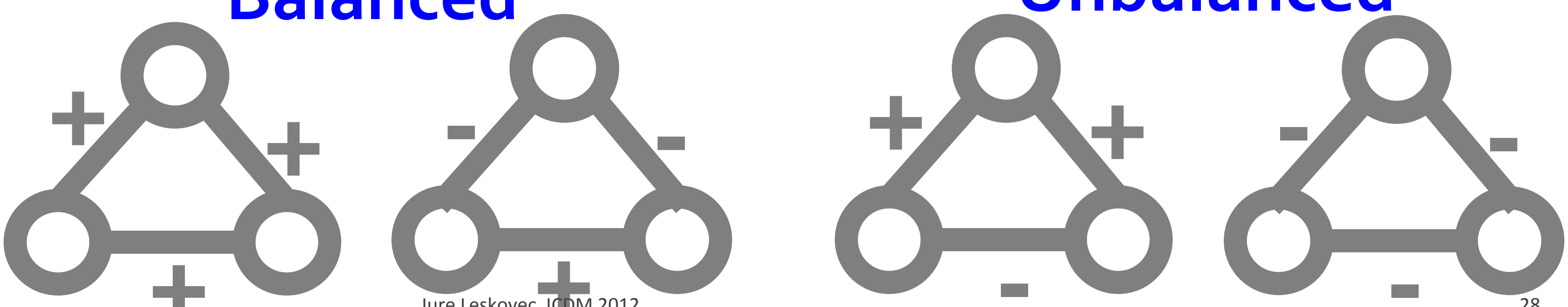
- **Supervised Random Walks are a general framework for ranking nodes on a graph**
  - There is nothing specific to link prediction here
  - Can use any features to learn the ranking
- **Applications: Social recommendations, ranking, filtering**
  - **Friends:** Trust, Homophily
  - **Others:** Experts, People like you
- **Link sentiment: Positive vs. Negative**

Friends and Foes



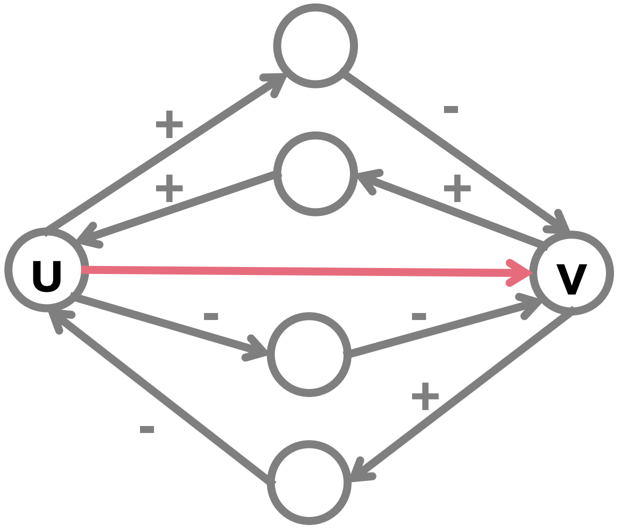
- **Not just if you link to someone but also what do you think of them**
- **Start with the intuition** [Heider '46]
  - The **friend** of my **friend** is my **friend**
  - The **enemy** of **enemy** is my **friend**
  - The **enemy** of **friend** is my **enemy**
  - The **friend** of my **enemy** is my **enemy**

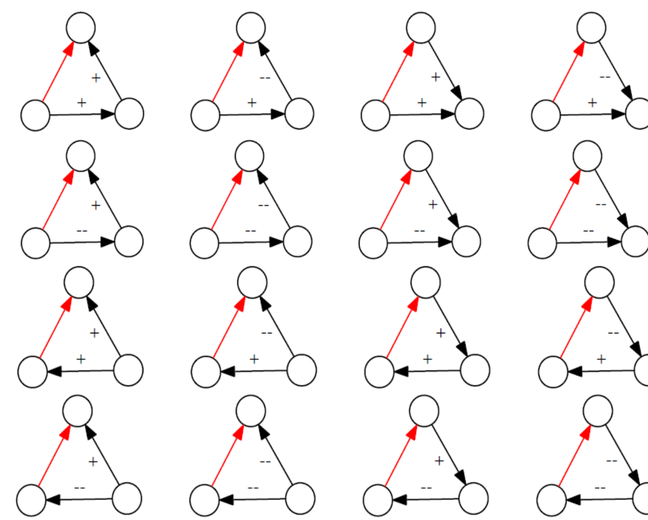
Balanced **Unbalanced**



Friends and Foes

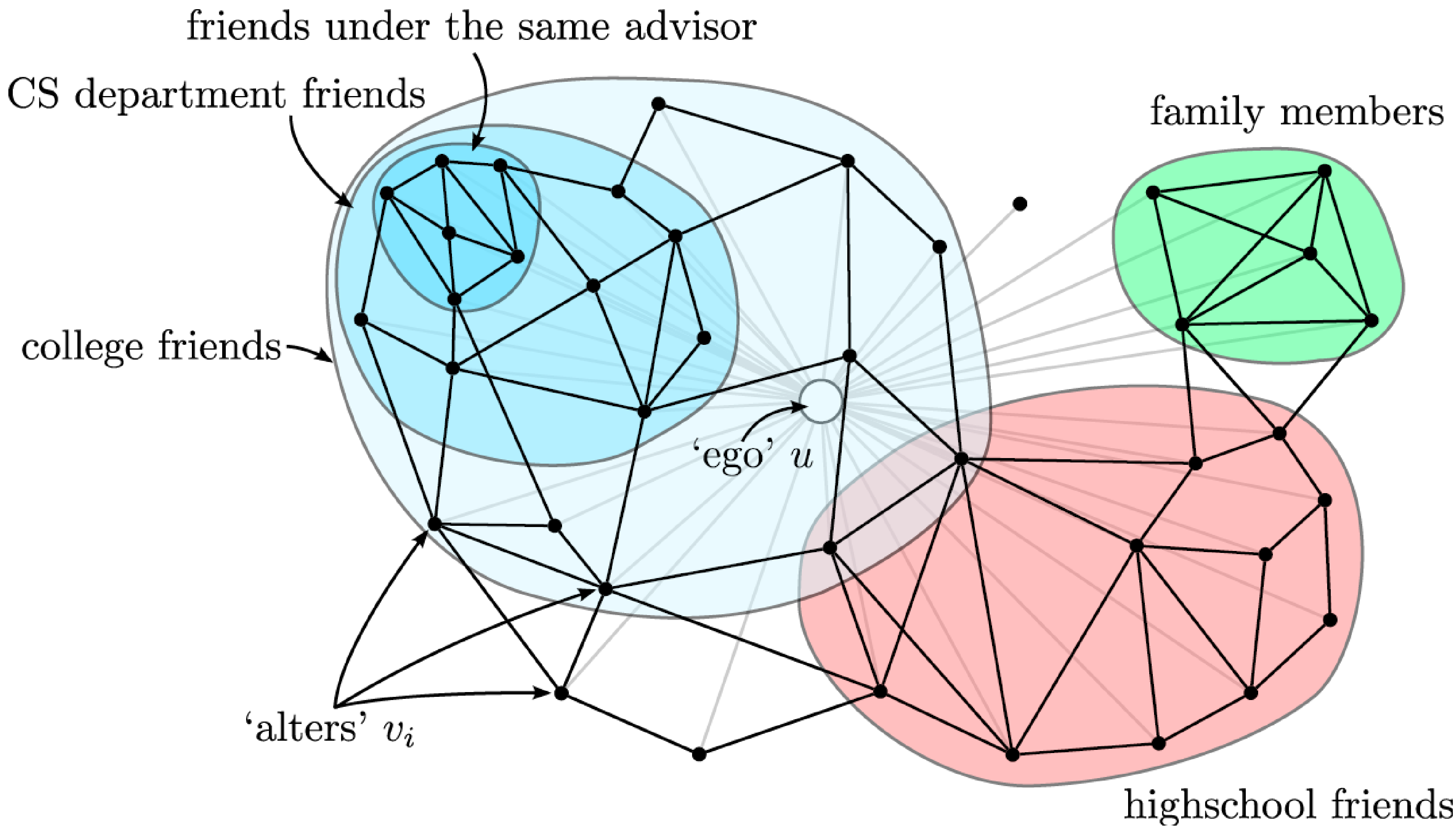
- **Model:**
  - Count the triads in which edge **u** → **v** is embedded: **16 features**
  - Train Logistic Regression
  - **Predictive accuracy: >90%**
- **Signs can be modeled from the local network structure alone!**





[NIPS '12]

Organizing Friends



Discover circles and why they exist

Discovering Social Circles

[NIPS '12]

- **Why is it useful?**
  - Organize friend lists
  - Control privacy and access
  - Filter and organize content

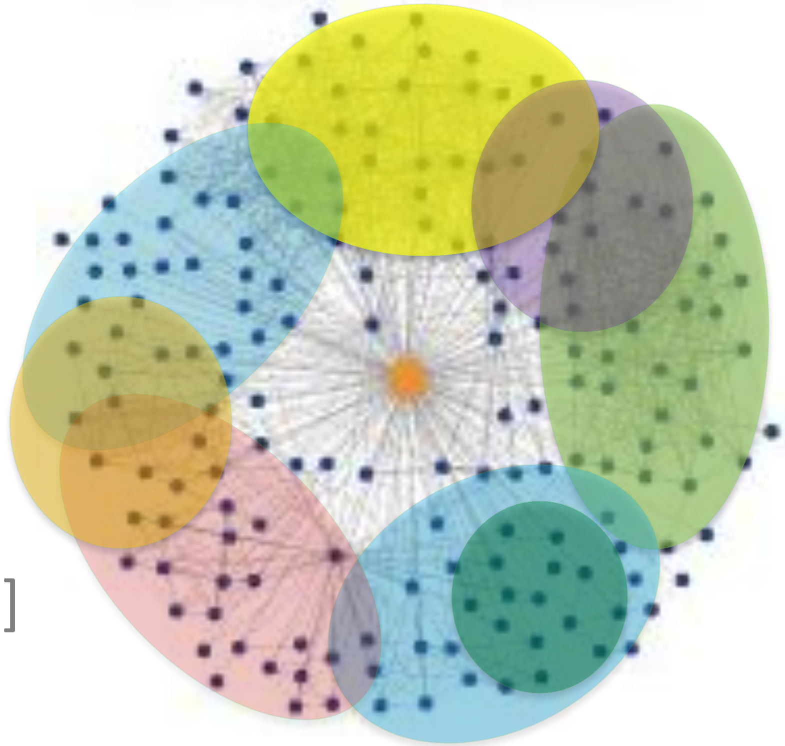
"On Facebook 273 people know I am a dog. The rest can only see my limited profile."

- **All social networks have this feature:**
  - Facebook (groups), Twitter (lists), G+ (circles)
  - **But circles have to be created manually!**

Social Circles: Connections

[NIPS '12]

- **Connections to graph partitioning & community detection**
  [Karypis, Kumar '98]
  [Girvan, Newman '02]
  [Dhillon, Guan, Kulis '07]
  [Yang, Sun, Pandit, Chawla, Han '11]
  ... but we can also use node profile information!



- **Q: How to cluster using network as well as node feature information?**

Model of Social Circles

[NIPS '12]

- Suppose we know all the circles
- For a given circle C model edge prob.:
  $p(x,y) \propto \exp(-\sum_i \theta_{ci} \cdot \psi_i(x,y))$
  - $\psi(x,y)$ … is edge feature vector describing (x,y)
    - Are x and y from same school, same town, same age, …
  - θ_c… parameters that we aim to estimate
    - High θ_{ci} means being similar in i is important for circle c
- Example:

$$\psi(x,y) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{array}{l} work : position : Cryptanalyst \\ work : location : GC\&CS \\ work : location : Royal\ Navy \\ education : name : Cambridge \\ education : type : College \\ education : name : Princeton \\ education : type : Graduate\ School \end{array} \qquad \theta_c = \begin{bmatrix} 1.4 \\ 0 \\ 0 \\ 0.3 \\ 0 \\ 0.2 \\ 1.1 \end{bmatrix}$$

Circle Discovery

[NIPS '12]

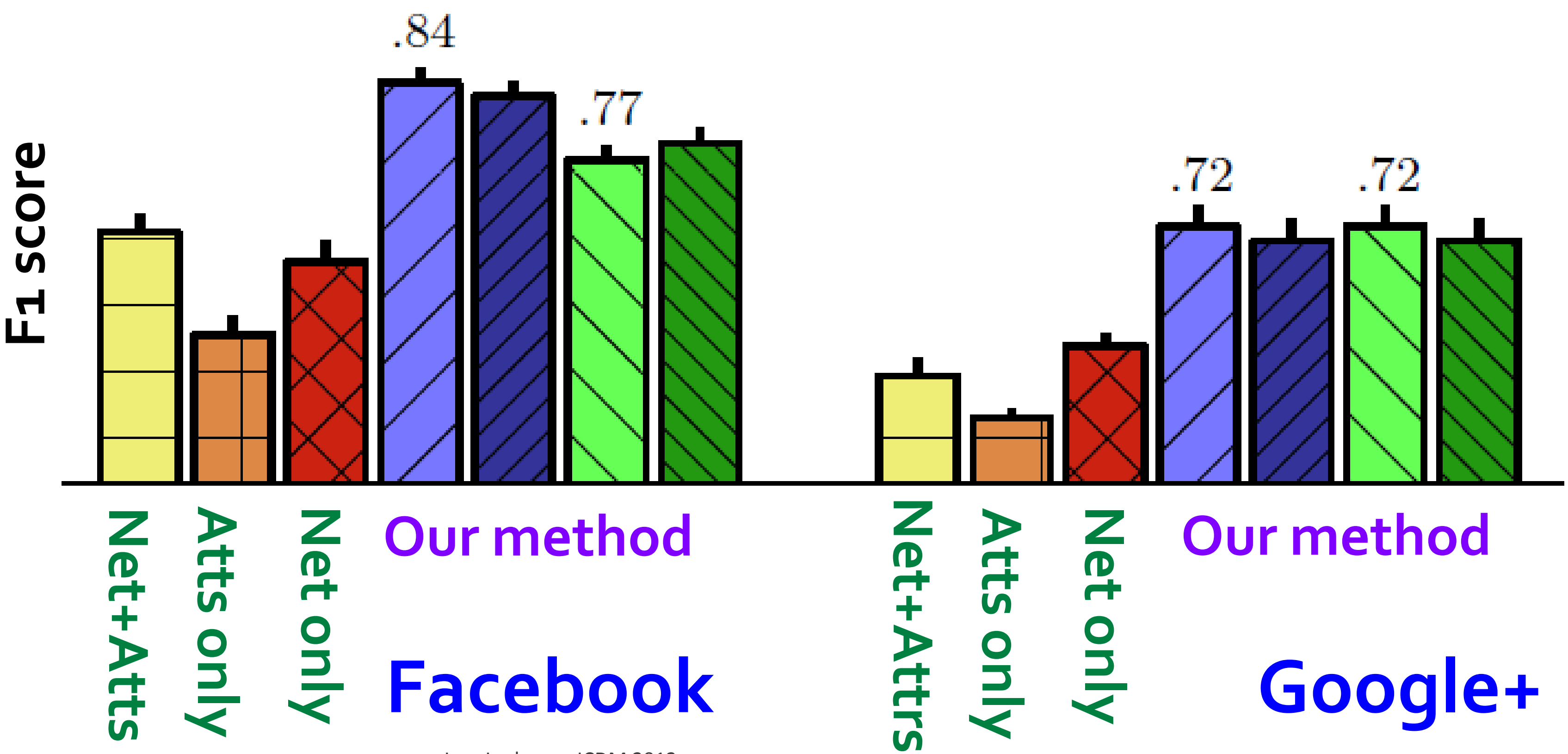
- **Given graph G and edge features $\psi(x, y)$**
- **Want to discover…**
  - Member nodes of each circle C
  - Circle similarity function parameters θ_c

…such that we maximize the likelihood of the observed network:

$$P(G; C) = \prod_{(x,y)\in G} p(x, y) \cdot \prod_{(x,y)\notin G} 1 - p(x, y)$$

Experiments: Facebook

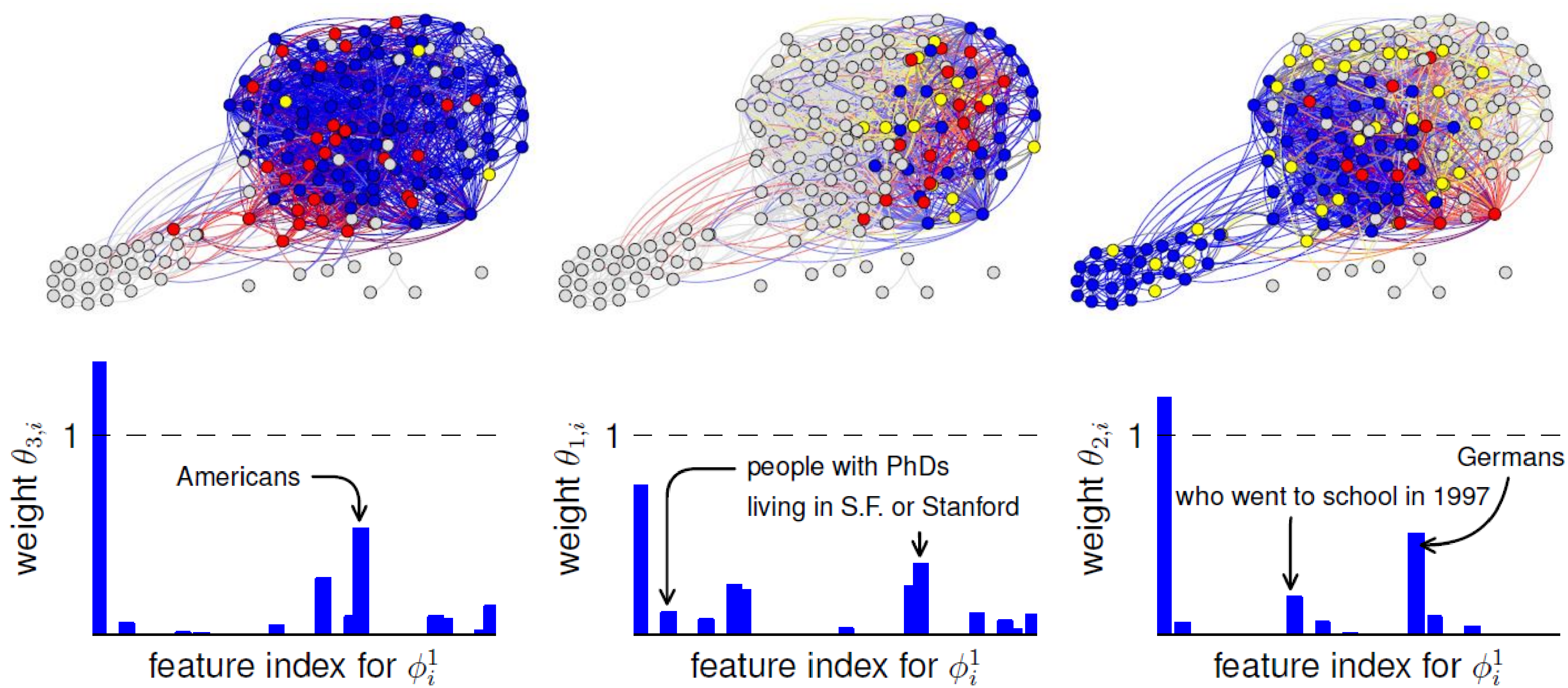
- Given only the network (no labels) try to find the circles. **How well are we doing?**
  - **Ask people to hand label the circles. Compare**



[NIPS '12]

Experiments: Facebook

- **How well do we recover human circles?**
- **Social circles of a particular person:**



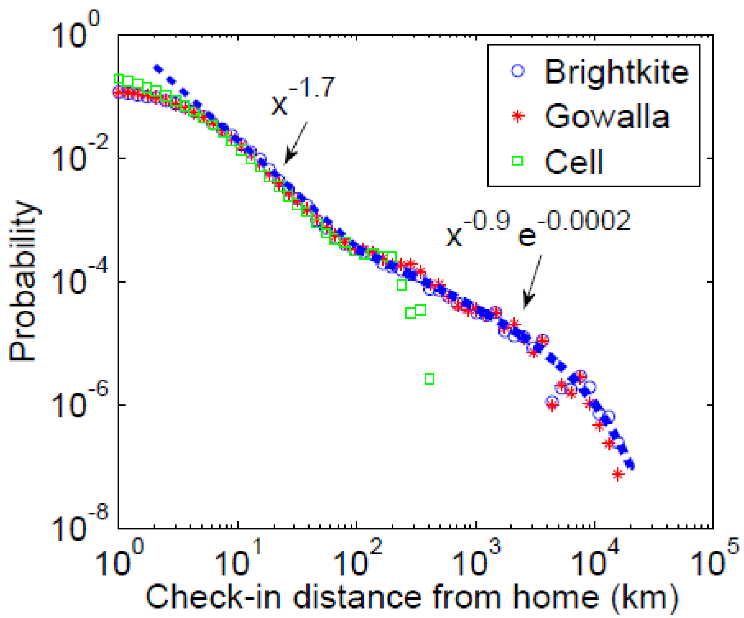
Circles: Further Questions

- **Beyond graph partitioning**
  - **Overlapping clustering of networks with node/edge attributes** [Yoshida ‘10] [McAuley, L. ‘12]
- **Temporal dynamics of circles and groups**
  - **Predict group evolution over time**
    [Kairam, Wang, L. ‘12] [Ducheneaut, Yee, Nickell, Moore ‘07]
- **Modeling circles of non-friends**
  - **Node role discovery in networks**
    [Henderson, Gallagher, Li, Akoglu, Eliassi-Rad, Tong, Faloutsos, ‘11]

[KDD '11]

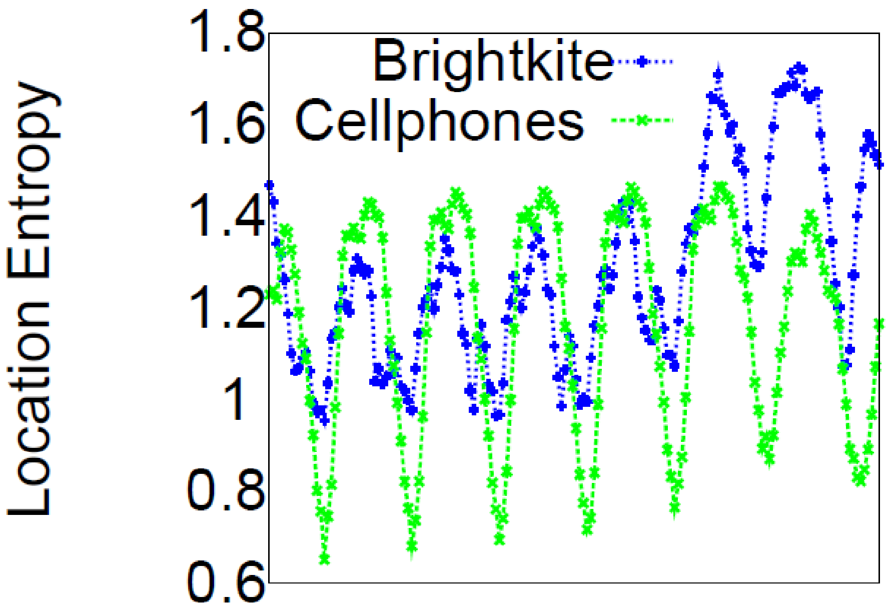
Social Networks & Mobility

- **What's the relation between human mobility and social networks?**
  - **Location-based online social networks**
    - Brightkite, Gowalla: 10m check-ins
  - **Cell phones**
    - Portugal: 500M calls
  - **In terms of mobility the datasets are indistinguishable!**



Modeling Mobility

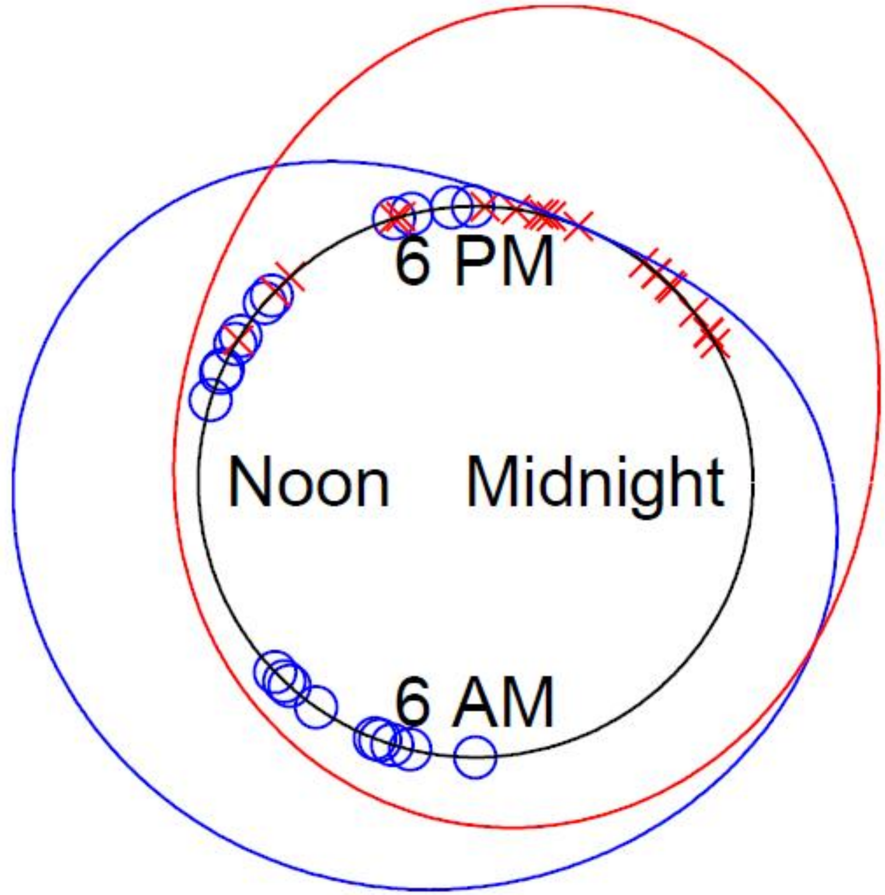
- **Goal:** Model and predict human movement patterns



- **Observation:**
  - Low location entropy at night/morning
  - Higher entropy over the weekend
- **3 ingredients of the model:**
  - **Spatial, Temporal, Social**

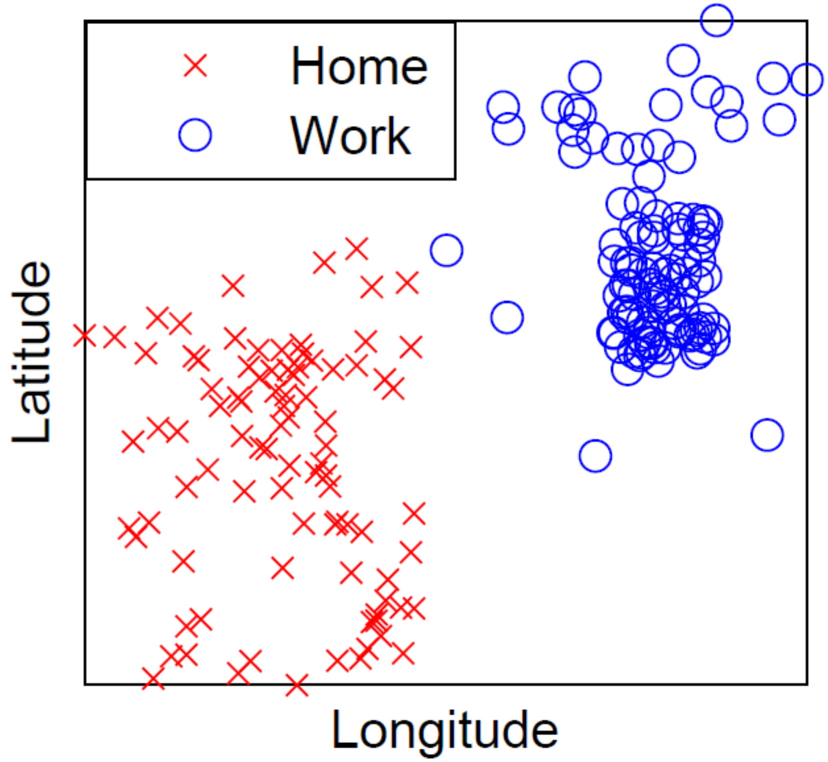
Modeling Mobility

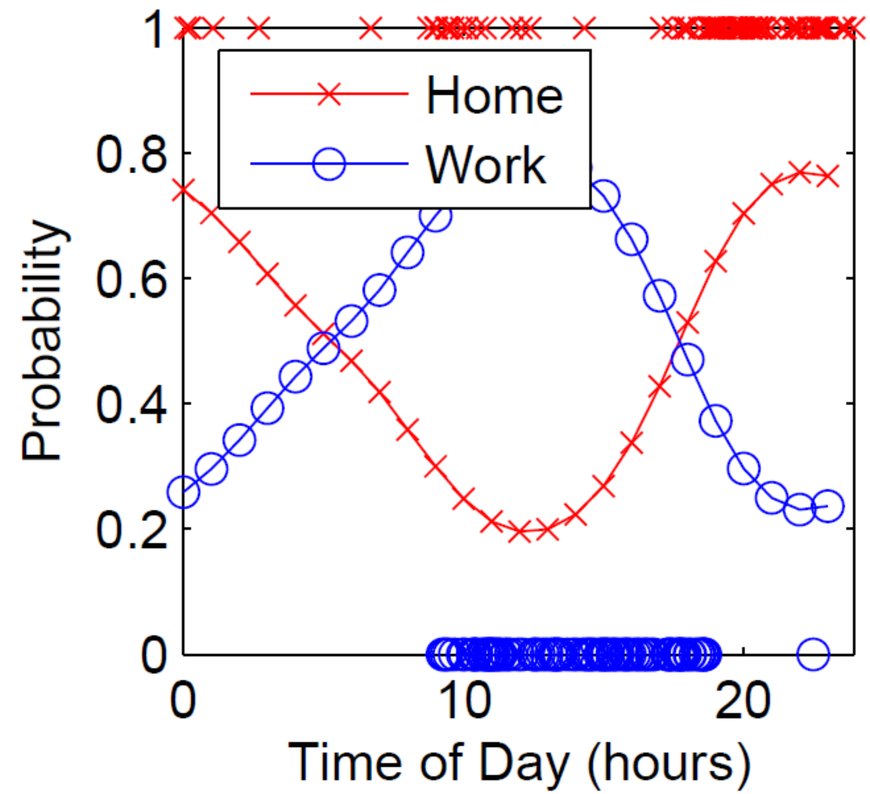
Spatial model:
Home vs. Work Location

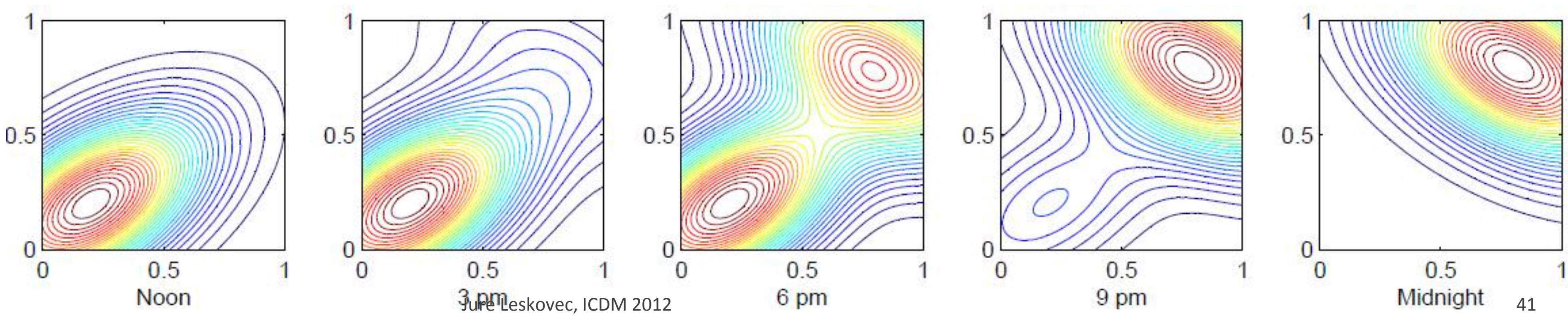


Temporal model:
Mobility Home vs. Work

Example: Gowalla User

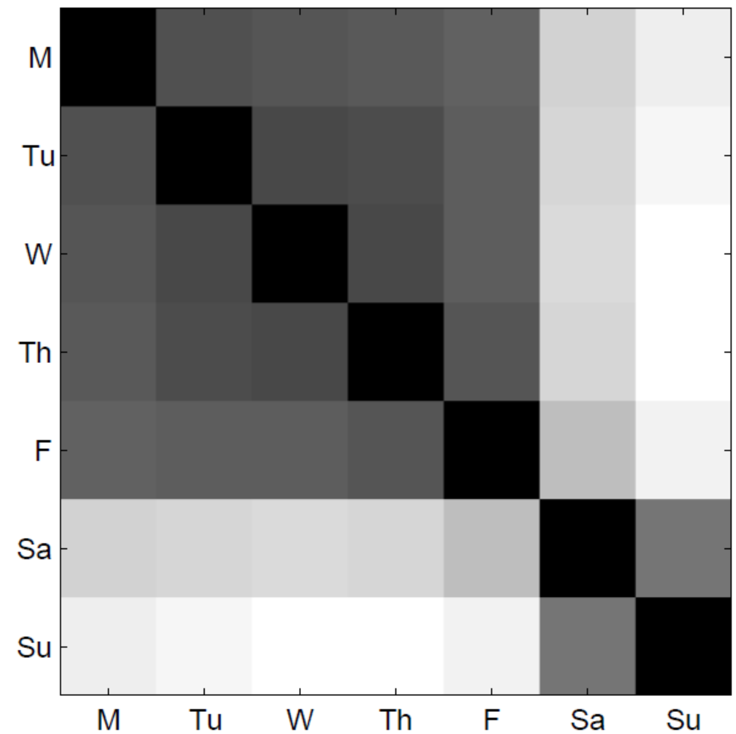






Weekend Mobility

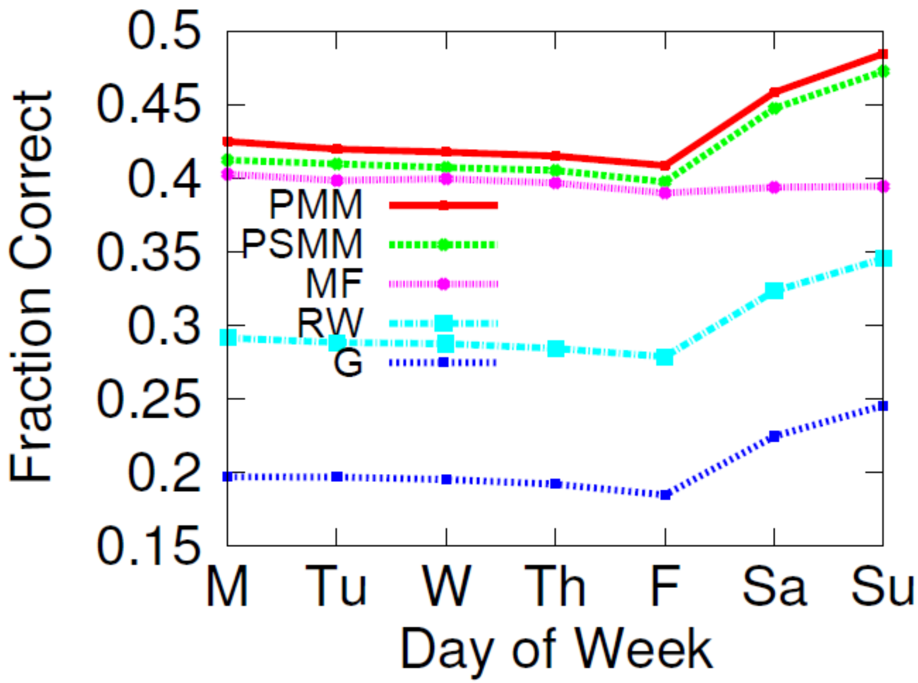
- **Social network plays particularly important role on weekends**
- **Include social network into the model**
  - Prob. that user visits location ***X*** depends on:
    - Distance(***X***, ***F***)
    - Time since a friend was at location ***F***
      - ***F*** = Friend's last known location

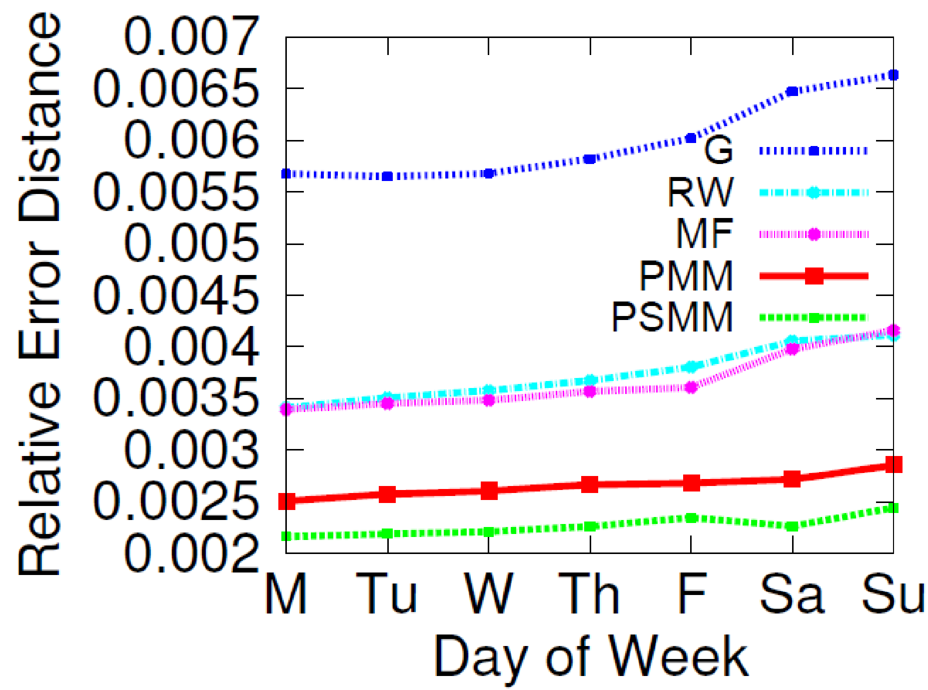


Mobility similarity

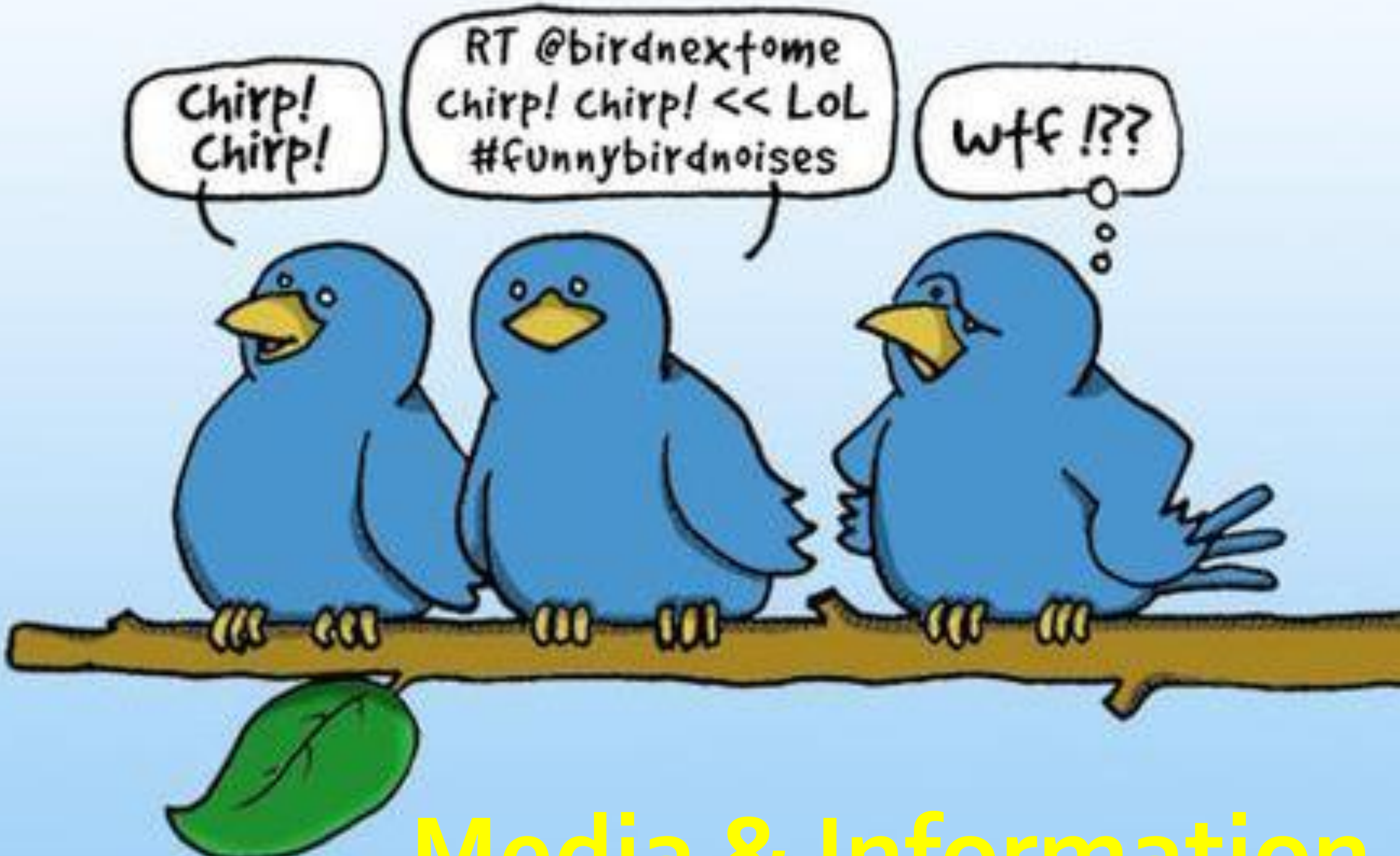
Mobility: Results

- **Cellphones:** Whenever user receives or makes a call predict her location





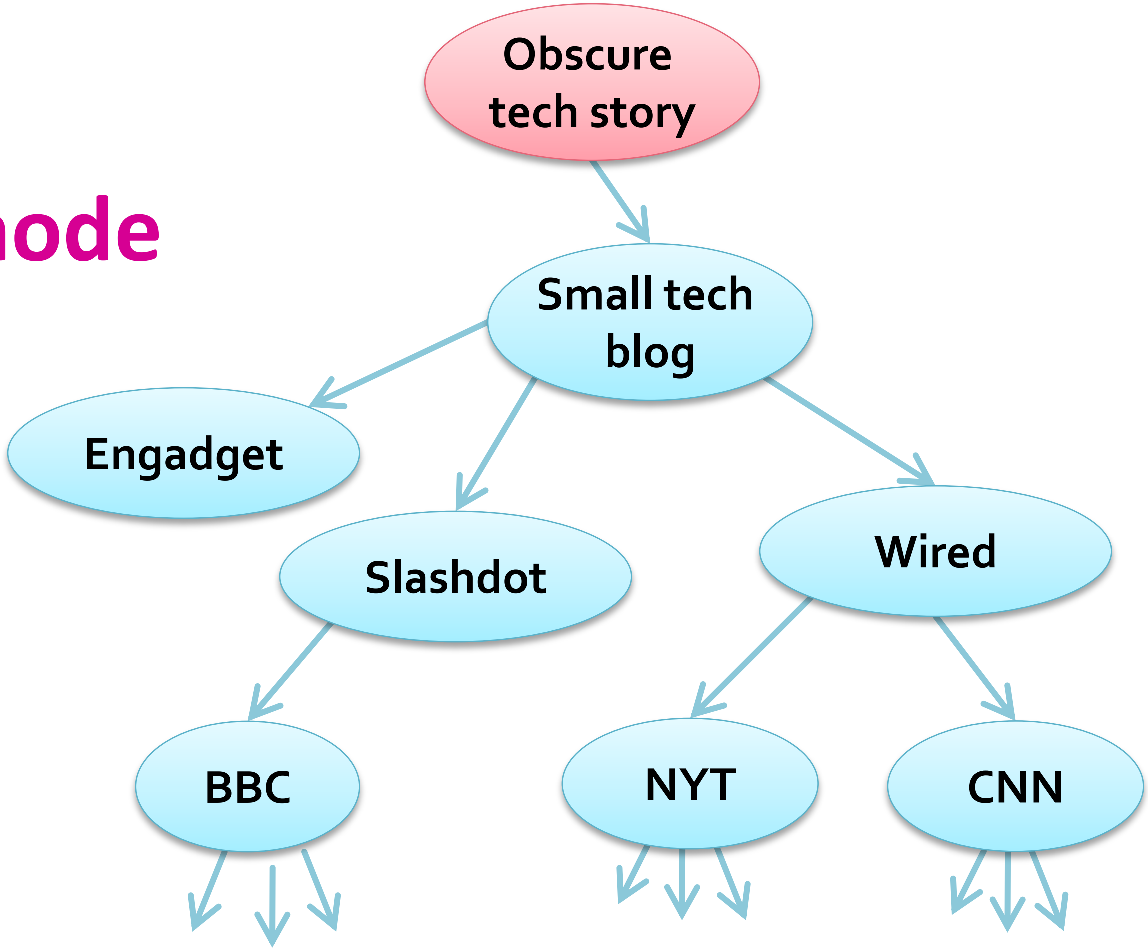
G … model by Gonzalez&Barabasi
RW… predict last known location
MF… predict most frequent location
PMM… periodic mobility model
PSMM… periodic social mobility model



Media & Information

Diffusion in Networks

- Information flows from a node to node like an epidemic
- How does information transmitted by mainstream media interact with social networks?



Diffusion in Online Media

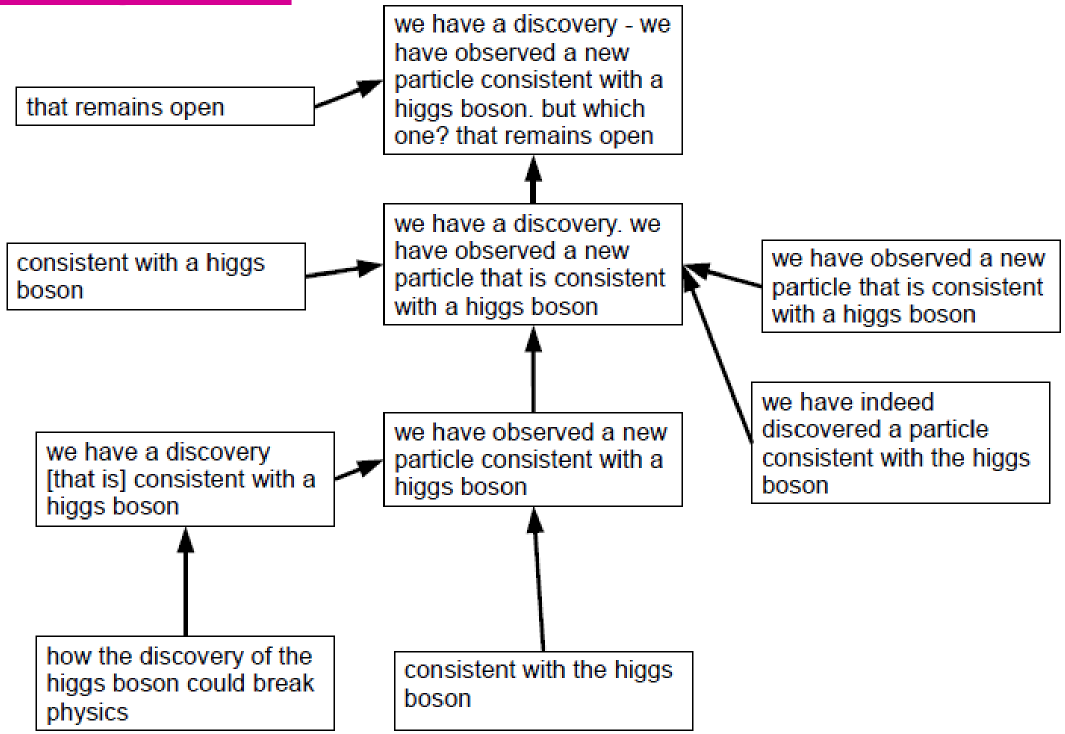
- **Since August 2008 we have been collecting 30M articles/day:** 6B articles, 20TB of data
- **Challenge:**

  How to track information as it spreads?

Meme-tracking

[WWW '13]

- **Goal:** **Trace textual phrases that spread through many news articles**
- **Challenge 1: Phrases mutate!**

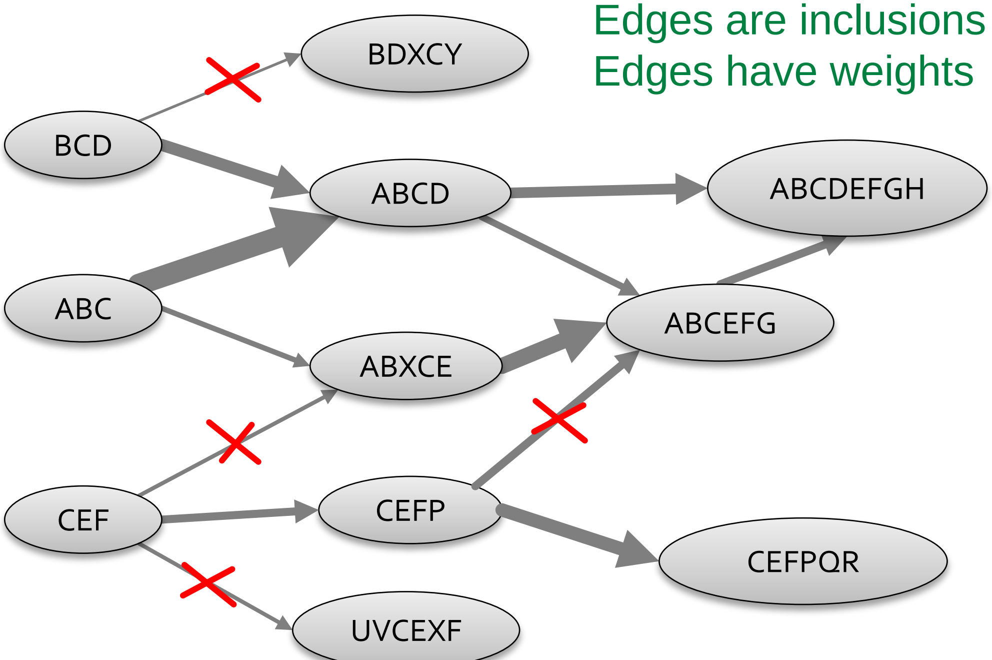


Mutations of a meme about the **Higgs boson particle**.

Finding Mutational Variants

- **Goal:** Find mutational variants of a phrase
- **Objective:**
  - In a DAG of approx. phrase inclusion, **delete min total edge weight** such that **each component has a single "sink"**

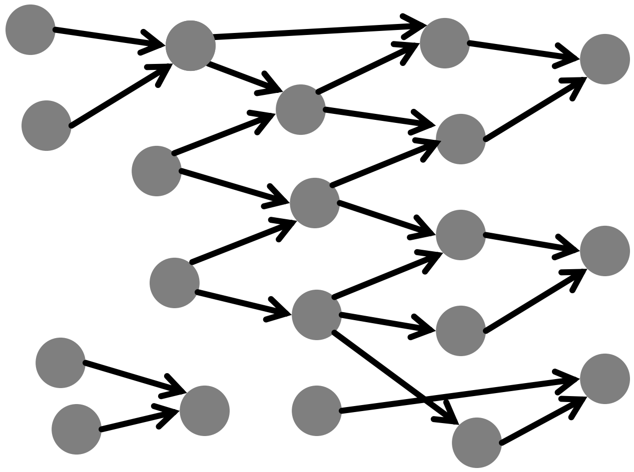
Nodes are phrases
Edges are inclusions
Edges have weights



The Algorithm

[WWW '13]

- Challenge 2: 20TB of data!
- Solution: **Incremental phrase clustering**
  - Phrases arrive in a stream
  - Simultaneously cluster the graph and attach new phrases to the graph
  - Dynamically remove completed clusters
- **Overall, it takes 1 server, 60GB memory and 4 days to process 6B documents**

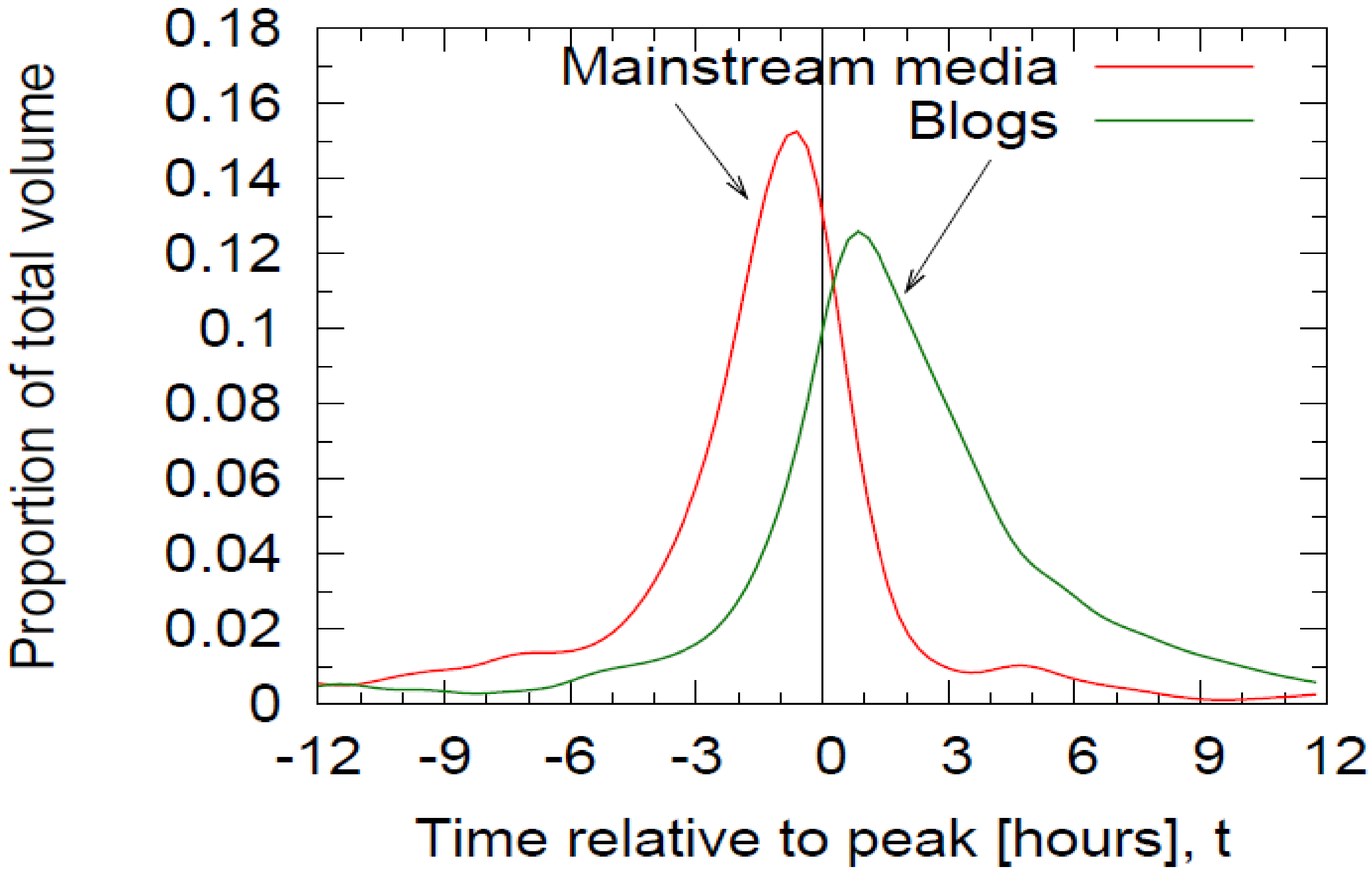


Memes over Time

Visualization of 1 month of data from October 2012

- **Browse all 4 years of data at** http://snap.stanford.edu/nifty

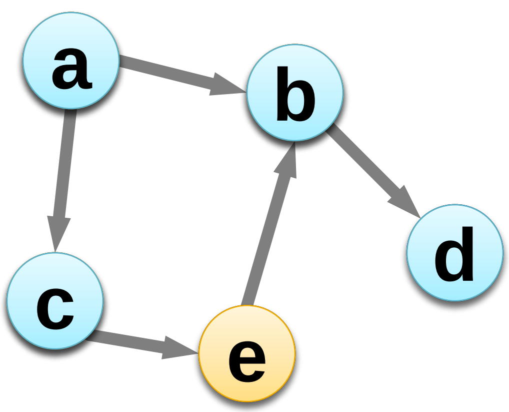
Do blogs lead mass media?



Do blogs lead mass media in reporting news? Blogs trail for 2.5h

Inferring Diffusion Networks

- **Challenge 3: Information network is hidden**
- **Goal: Infer the information diffusion network**
  - There is a **hidden** network, and
  - We only see **times** when nodes get "infected"



  - **Yellow** info: (a,1), (c,2), (b,3), (e,4)
  - **Blue** info: (c,1), (a,4), (b,5), (d,6)

[KDD '10]

Inferring Networks

	Virus propagation	Word of mouth & Viral marketing
Process	Viruses propagate through the network	Recommendations and influence propagate
We observe	We only observe when people get sick	We only observe when people buy products
It's hidden	But NOT who **infected** them	But NOT who **influenced** them

Can we infer the underlying network?

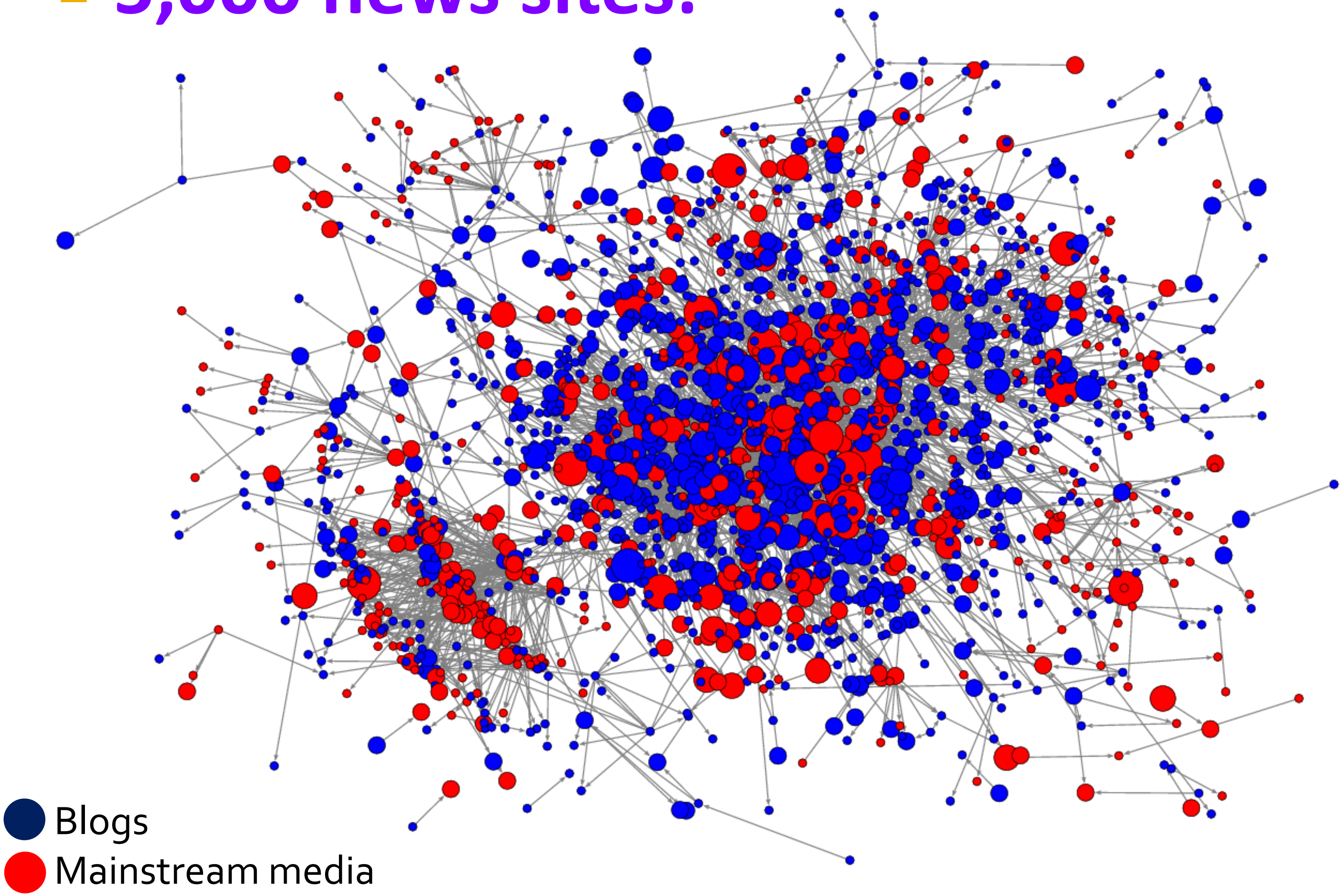
Yes, convex optimization problem!

[Gomez-Rodriguez, L., Krause, '10, Myers, L., '10]

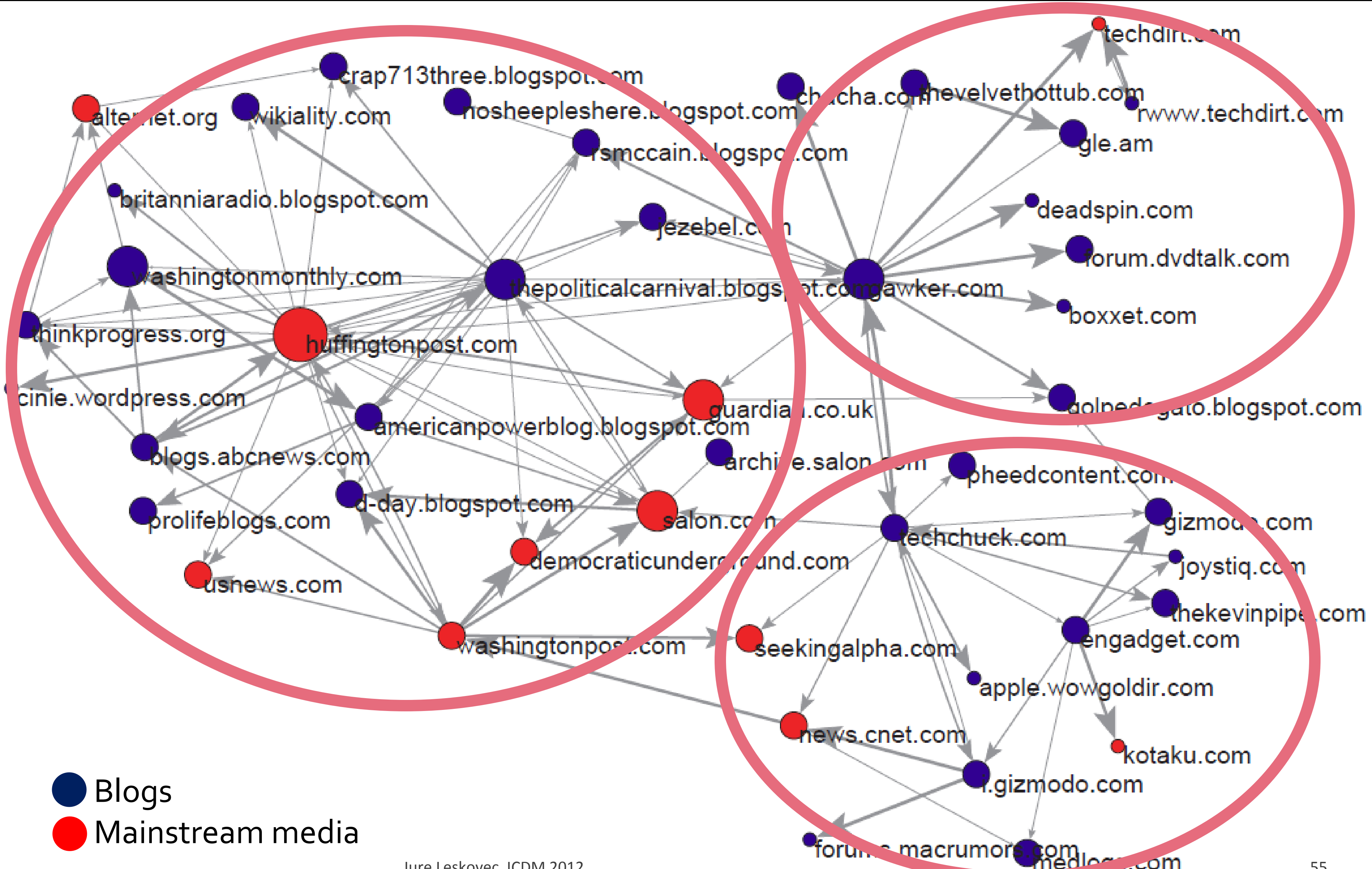
News Diffusion Network

[KDD '10]

- **5,000 news sites:**

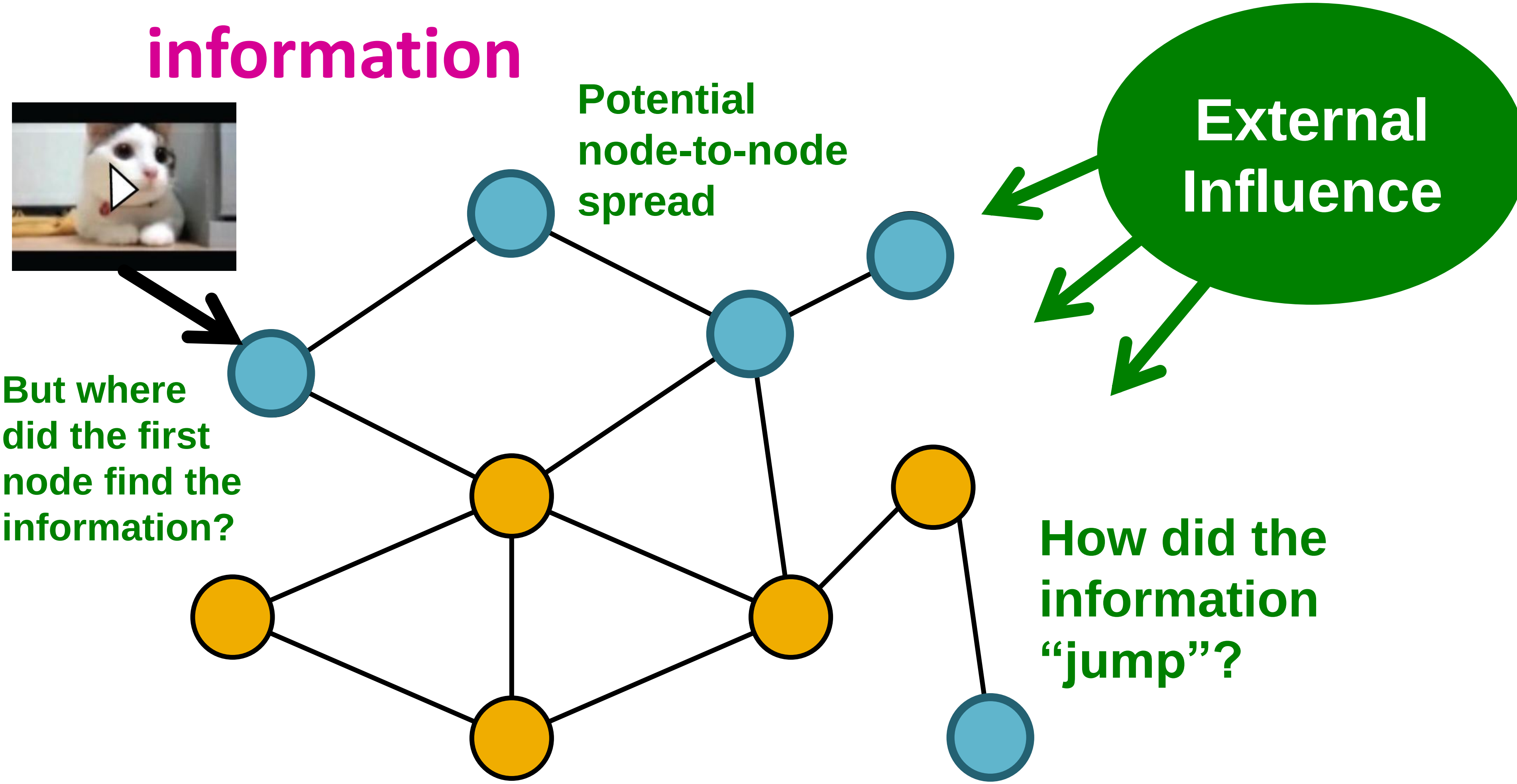


News Diffusion Network



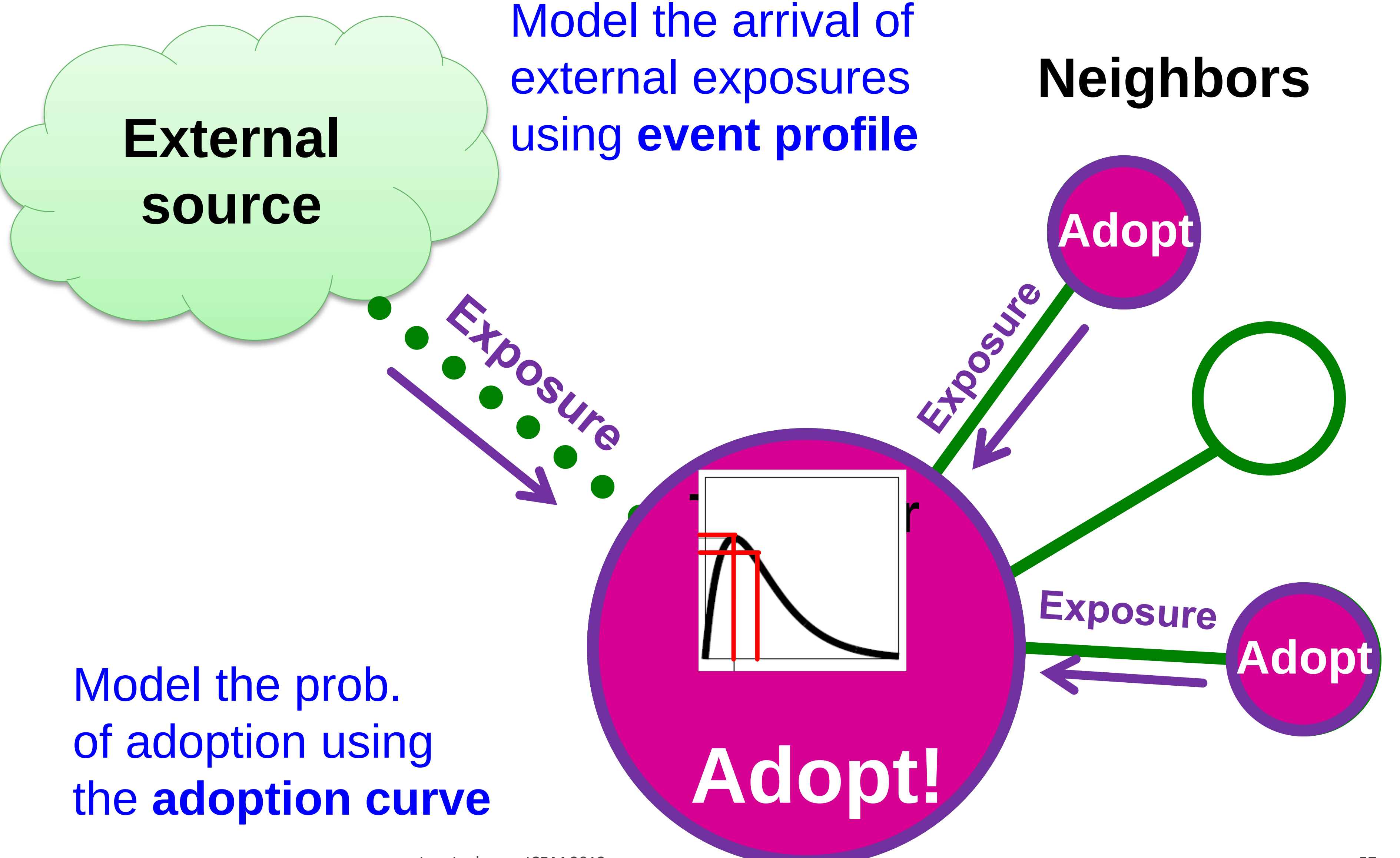
Information Diffusion

- Observe times when nodes adopt the information



Towards the Model

Model the arrival of external exposures using **event profile**

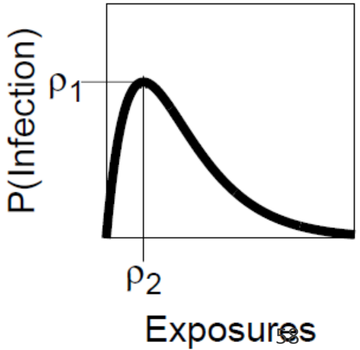


Model the prob. of adoption using the **adoption curve**

[KDD '12]

Results: Different Topics

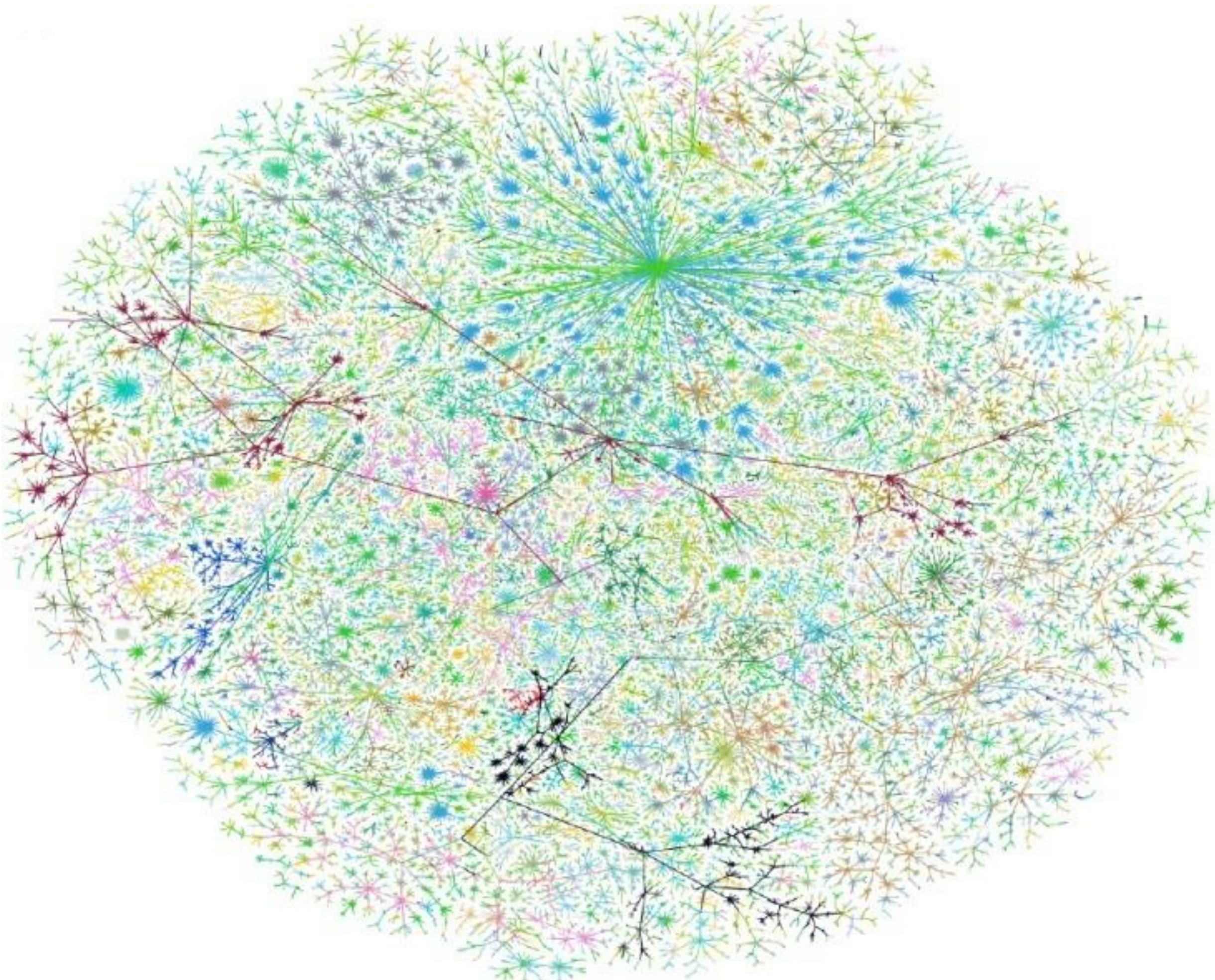
	max P(k)	k at max P(k)	Duration (hours)	% Ext. Exposures
Politics (25)	0.0007 +/- 0.0001	4.59 +/- 0.76	51.24 +/- 16.66	47.38 +/- 6.12
World (824)	0.0013 +/- 0.0000	2.97 +/- 0.10	43.54 +/- 2.94	26.07 +/- 1.19
Entertain. (117)	0.0015 +/- 0.0002	3.52 +/- 0.28	89.89 +/- 16.13	17.87 +/- 2.51
Sports (24)	0.0010 +/- 0.0003	4.76 +/- 0.83	87.85 +/- 38.03	43.88 +/- 6.97
Health (81)	0.0016 +/- 0.0002	3.25 +/- 0.30	100.09 +/- 17.57	18.81 +/- 3.33
Tech. (226)	0.0013 +/- 0.0001	3.00 +/- 0.16	83.05 +/- 8.73	18.36 +/- 1.80
Business (298)	0.0015 +/- 0.0001	3.18 +/- 0.16	49.61 +/- 5.14	22.27 +/- 1.79
Science (106)	0.0012 +/- 0.0002	4.06 +/- 0.30	135.28 +/- 16.19	20.53 +/- 2.78
Travel (16)	0.0005 +/- 0.0001	2.33 +/- 0.29	151.73 +/- 39.70	39.99 +/- 6.60
Art (32)	0.0006 +/- 0.0001	5.26 +/- 0.66	188.55 +/- 48.17	27.54 +/- 5.30
Edu. (31)	0.0009 +/- 0.0001	3.77 +/- 0.51	130.53 +/- 38.63	21.45 +/- 6.40



More details: Myers, Zhu, L. : Information diffusion and external influence in networks, *KDD* 2012.

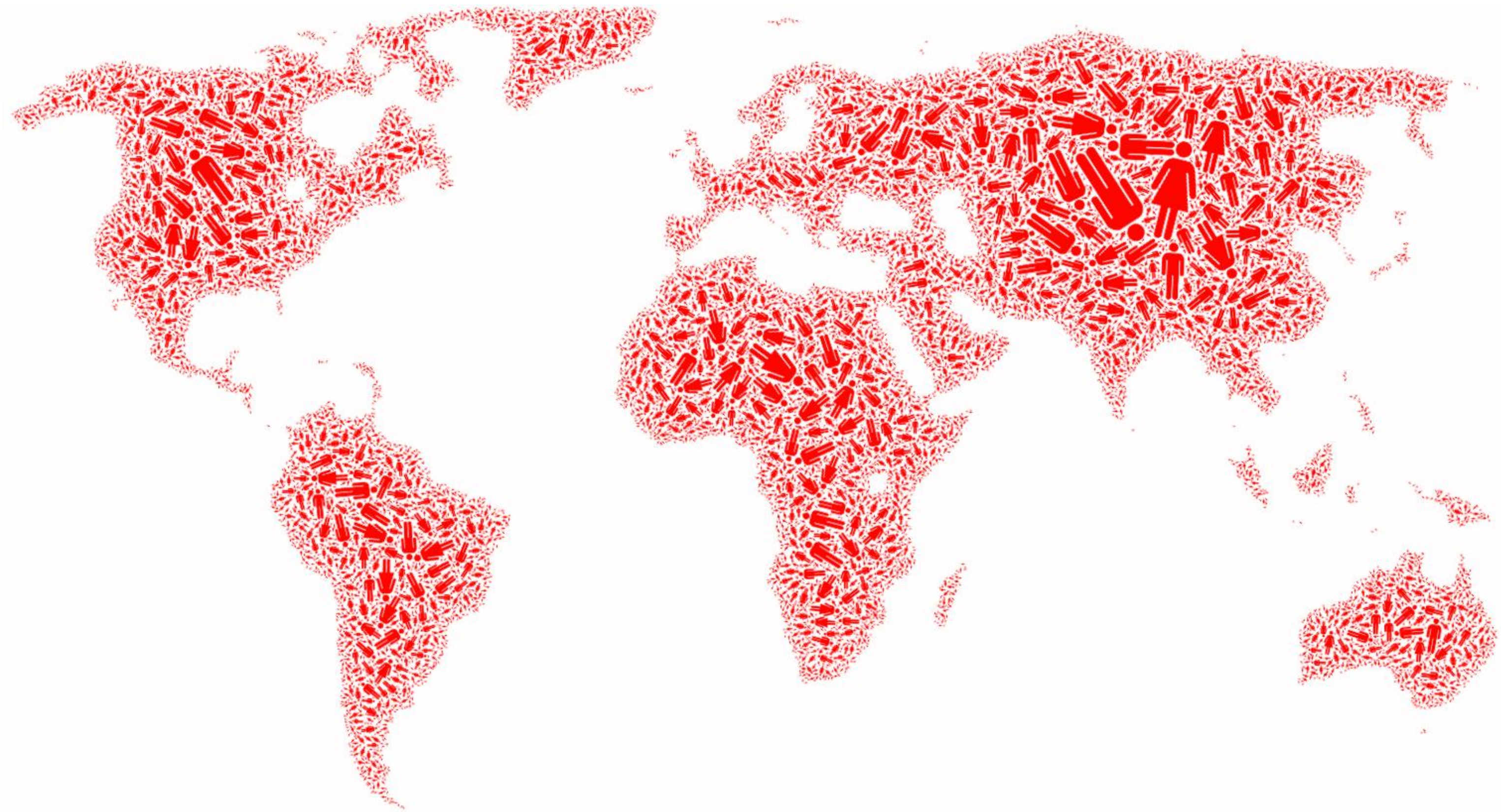
Diffusion: Further Questions

- **Can we recognize fundamental patterns of human behavior from raw digital traces?**
- **Can such analysis help identify dynamics of polarization?** [Adamic, Glance '05]
- **Connections to mutation of information:**
  - How does **attitude** and **sentiment change** in different parts of the network?
  - How does **information change** in different parts of the network?



Networks: What's beyond?

What's beyond?



Networks are a natural language for reasoning about problems spanning society, technology and information

Conclusion & Reflections

- **Only recently has large scale network data become available**
  - Opportunity for large scale analyses
  - **Benefits of working with massive data**
    - Observe "invisible" patterns
- **Lots of interesting networks questions both in CS as well as in general science**
  - Need scalable algorithms & models

Towards the Model of You

- Social networks — implicit for millenia — are being recorded in our information systems
- Software has a complete trace of your activities — and increasingly knows more about your behavior than you do
- Models based on algorithmic ideas will be crucial in understanding these developments

Towards the Model of You

- **From models of populations to models of individuals**
  - Distributions over millions of people leave open several possibilities:
    - **Individual are highly diverse, and the distribution only appears in aggregate**, or
    - **Each individual personally follows (a version of) the distribution**
  - Recent studies suggests that sometimes the second option may in fact be true [Barabasi '05]

Network Data & Code

- **Research on networks is both algorithmic and empirical**
- Need to network data:
  - **Stanford Large Network Dataset Collection**
    - Over 60 large online networks with metadata
    - http://snap.stanford.edu/data
  - **SNAP: Stanford Network Analysis Platform**
    - A general purpose, high performance system for dynamic network manipulation and analysis
    - Can process 1B nodes, 10B edges
    - http://snap.stanford.edu

THANKS!

Data + Code:

http://snap.stanford.edu

References

- Supervised Random Walks: Predicting and Recommending Links in Social Networks by L. Backstrom, J. Leskovec. *ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.
- Predicting Positive and Negative Links in Online Social Networks by J. Leskovec, D. Huttenlocher, J. Kleinberg. *ACM WWW International conference on World Wide Web (WWW)*, 2010.
- Learning to Discover Social Circles in Ego Networks by J. McAuley, J. Leskovec. *Neural Information Processing Systems (NIPS)*, 2012.
- Defining and Evaluating Network Communities based on Ground-truth by J. Yang, J. Leskovec. *IEEE International Conference On Data Mining (ICDM)*, 2012.
- The Life and Death of Online Groups: Predicting Group Growth and Longevity by S. Kairam, D. Wang, J. Leskovec. *ACM International Conference on Web Search and Data Mining (WSDM)*, 2012.

References

- Meme-tracking and the Dynamics of the News Cycle by J. Leskovec, L. Backstrom, J. Kleinberg. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.
- Inferring Networks of Diffusion and Influence by M. Gomez-Rodriguez, J. Leskovec, A. Krause. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.
- On the Convexity of Latent Social Network Inference by S. A. Myers, J. Leskovec. *Neural Information Processing Systems (NIPS)*, 2010.
- Structure and Dynamics of Information Pathways in Online Media by M. Gomez-Rodriguez, J. Leskovec, B. Schoelkopf. *ACM International Conference on Web Search and Data Mining (WSDM)*, 2013.
- Modeling Information Diffusion in Implicit Networks by J. Yang, J. Leskovec. *IEEE International Conference On Data Mining (ICDM)*, 2010.
- Information Diffusion and External Influence in Networks by S. Myers, C. Zhu, J. Leskovec. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.
- Clash of the Contagions: Cooperation and Competition in Information Diffusion by S. Myers, J. Leskovec. *IEEE International Conference On Data Mining (ICDM)*, 2012.