Social Influence and Information Diffusion

Jie Tang

Department of Computer Science and Technology
Tsinghua University

Networked World

facebook

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

人人网

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

twitter

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

新浪微博
weibo.com

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

amazon.com

- **79 million** users per month
- **>10 billion** items/year

Alibaba Group
阿里巴巴集团

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

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

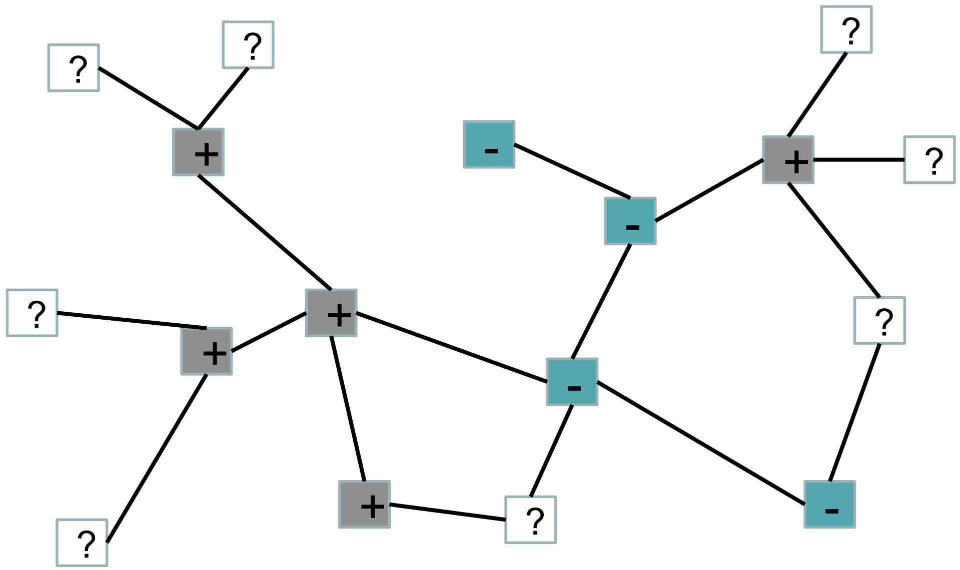
Challenge: Big Social Data

- We generate 2.5×10^{18} byte *big data* per day.

- Big social data:
  - 90% of the data was generated in the past 2 yrs
  - How to mine deep knowledge from the big social data?

15-20 years before…

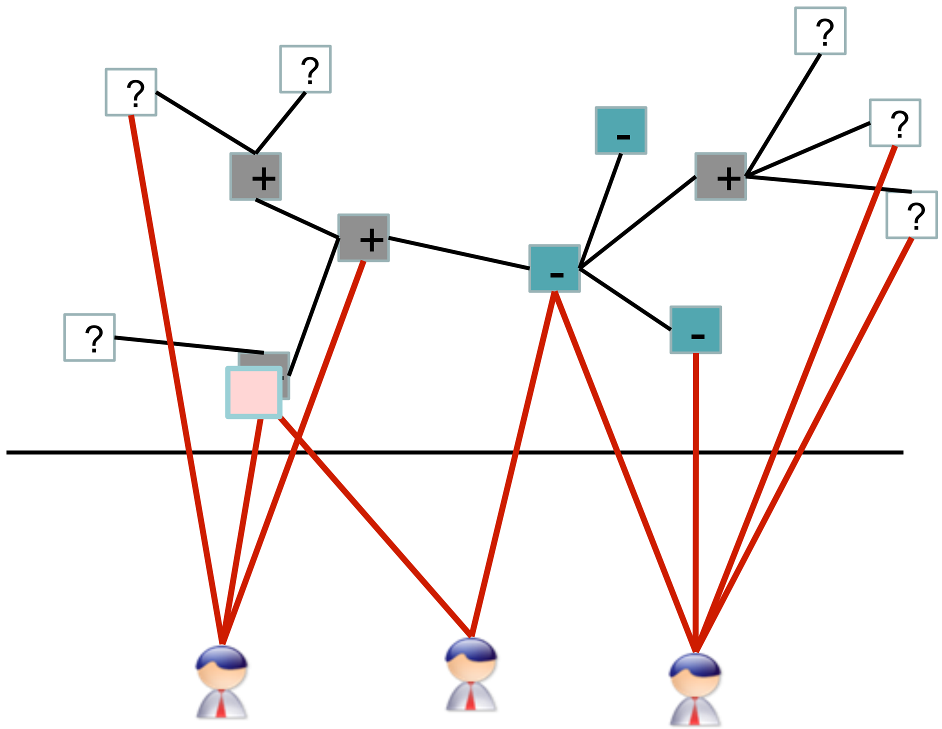
Web 1.0

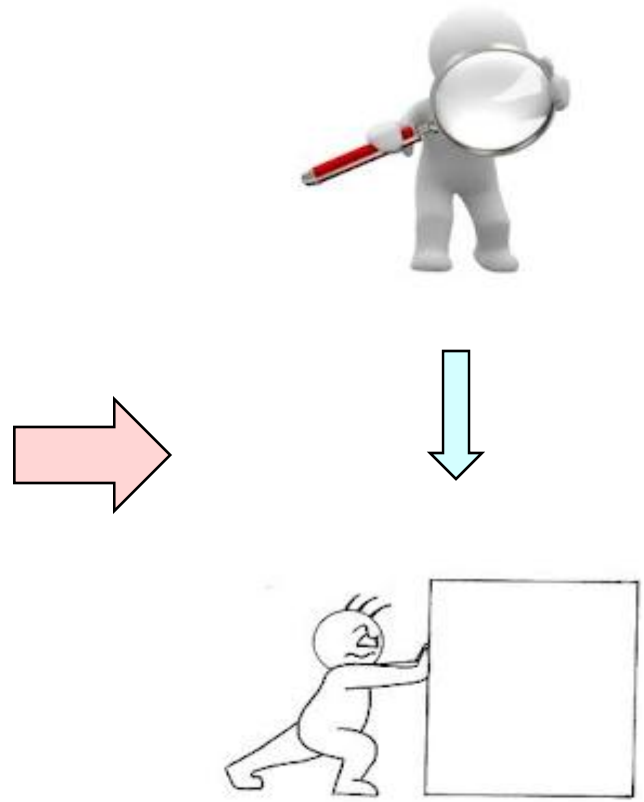


hyperlinks between web pages
Examples:
Google search (information retrieval)

10 years before…

Collaborative Web



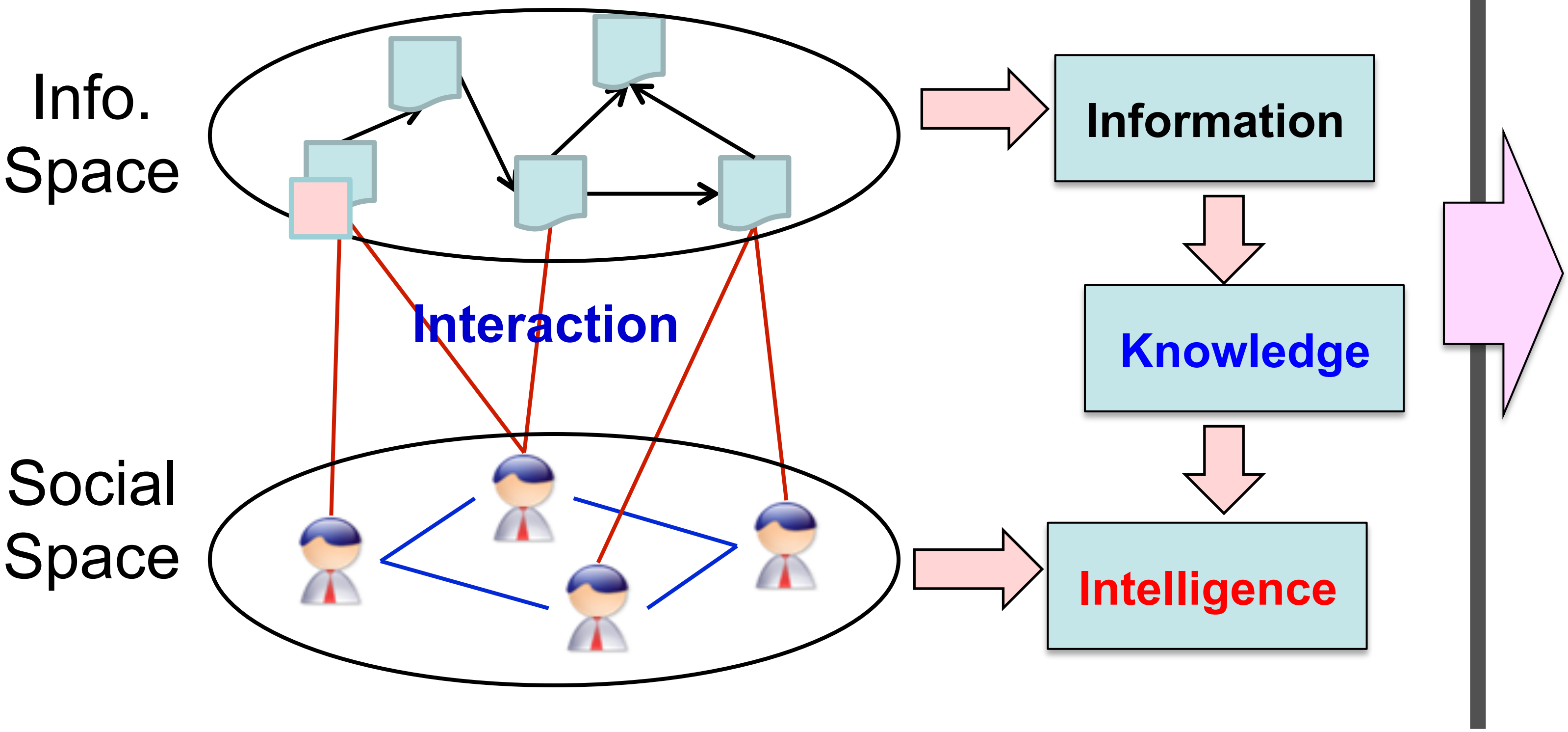


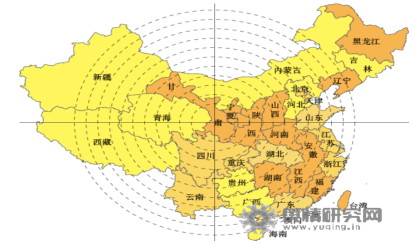
(1) personalized learning
(2) collaborative filtering

Big Social Analytics—In recent 5 years…

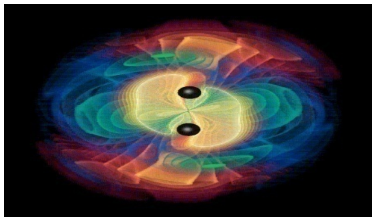
Social Web

Info. Space vs. Social Space

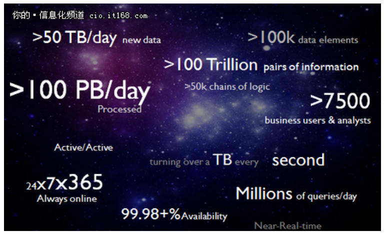




Opinion Mining

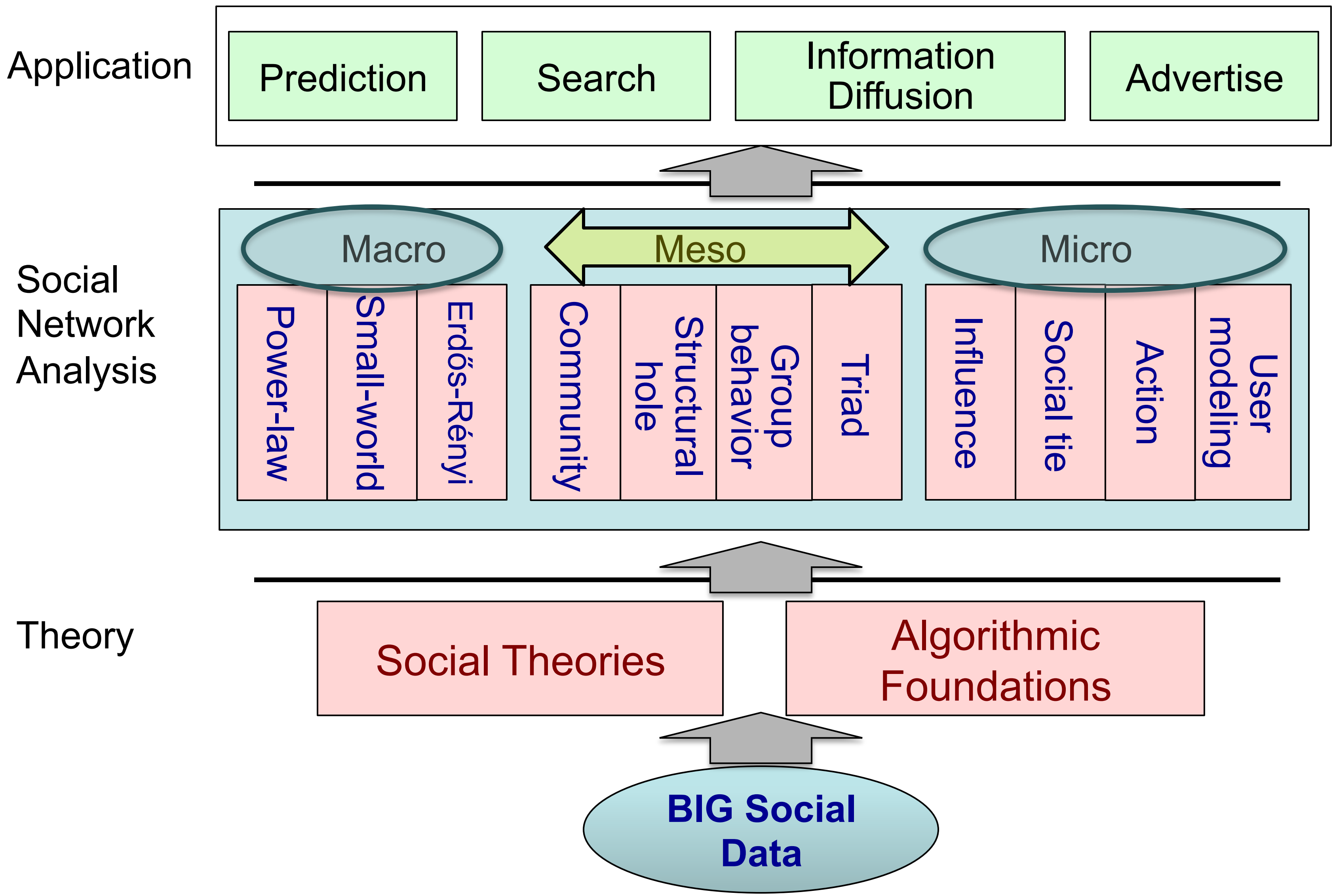


Innovation diffusion

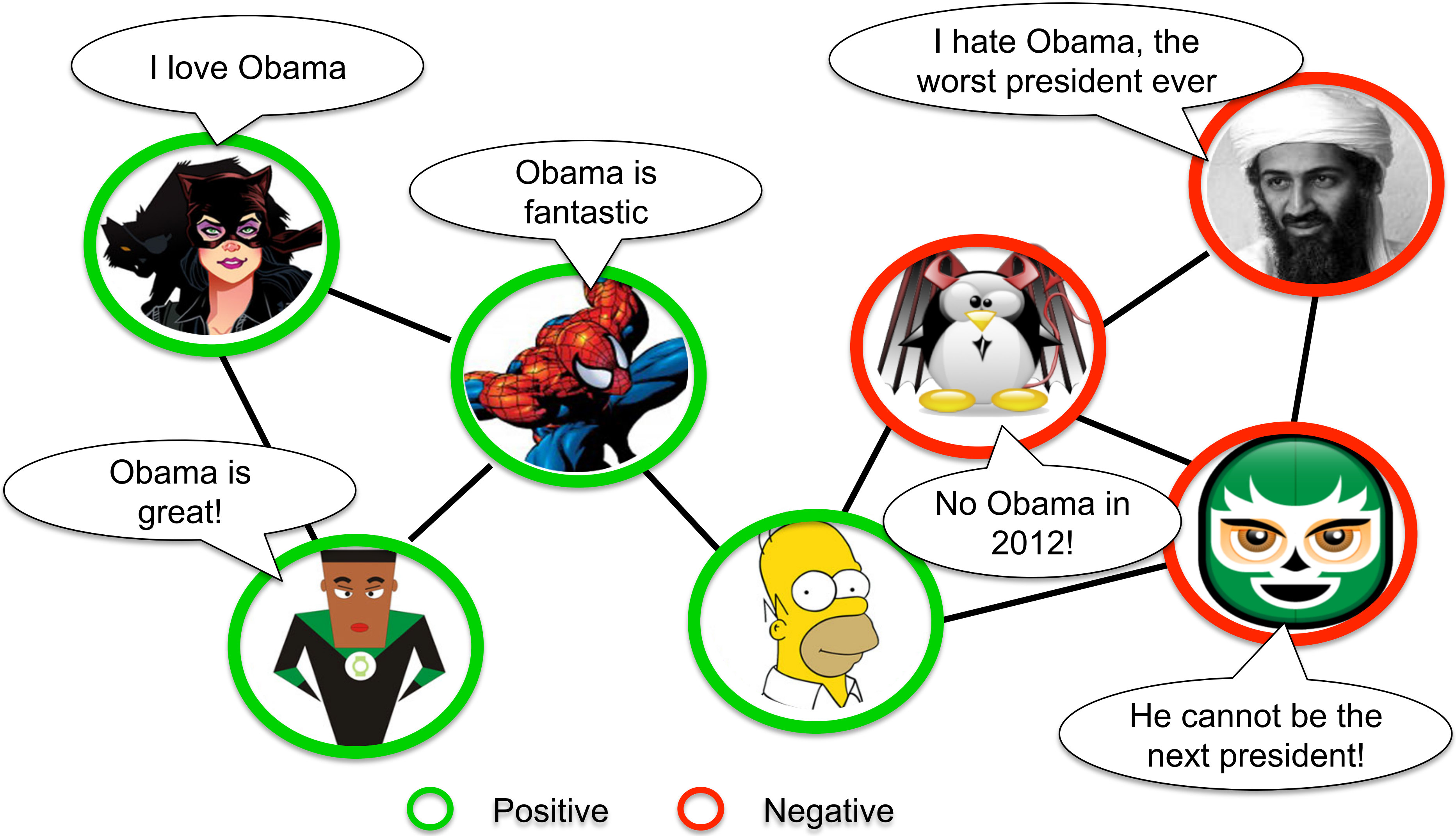


Business intelligence

Core Research in Social Network



“Love Obama”
—social influence in online social networks



What is Social Influence?

- Social influence occurs when one's opinions, emotions, or behaviors are affected by others, intentionally or unintentionally.[1]
  - **Informational social influence**: to accept information from another;
  - **Normative social influence**: to conform to the positive expectations of others.

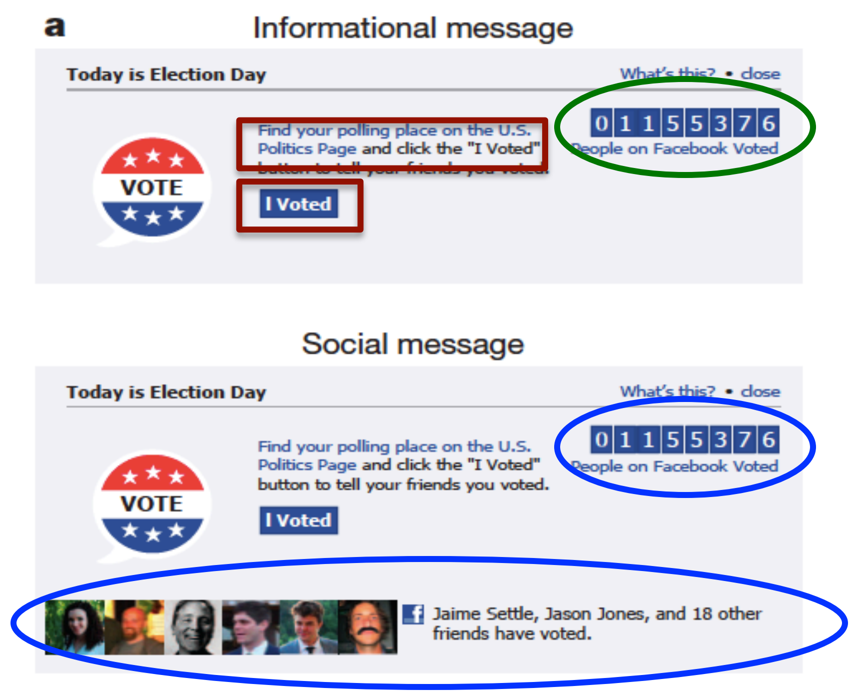
[1] http://en.wikipedia.org/wiki/Social_influence

Does Social Influence really matter?

- Case 1: Social influence and political mobilization[1]
  - Will online political mobilization really work?

A controlled trial (with 61M users on FB)

- **Social msg group:** was shown with msg that indicates one's friends who have made the votes.
- **Informational msg group:** was shown with msg that indicates how many other.
- **Control group:** did not receive any msg.



[1] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. Nature, 489:295-298, 2012.

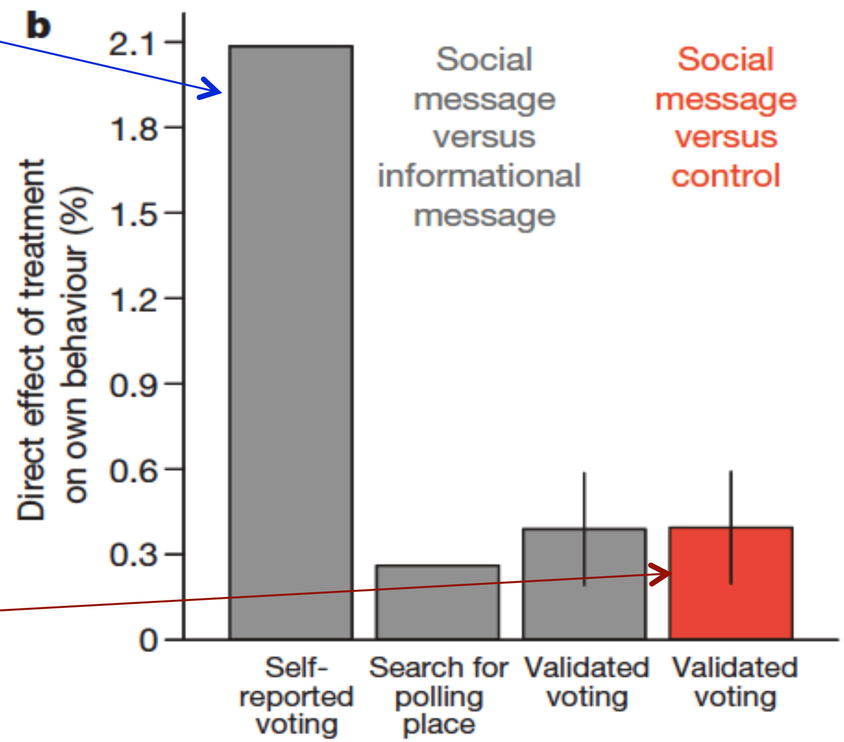
Case 1: Social Influence and Political Mobilization

Social msg group v.s.
Info msg group

Result: The former were 2.08% (*t*-test, *P*<0.01) more likely to click on the “I Voted” button

Social msg group v.s.
Control group

Result: The former were 0.39% (*t*-test, *P*=0.02) more likely to **actually vote** (via examination of public voting records)



[1] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. Nature, 489:295-298, 2012.

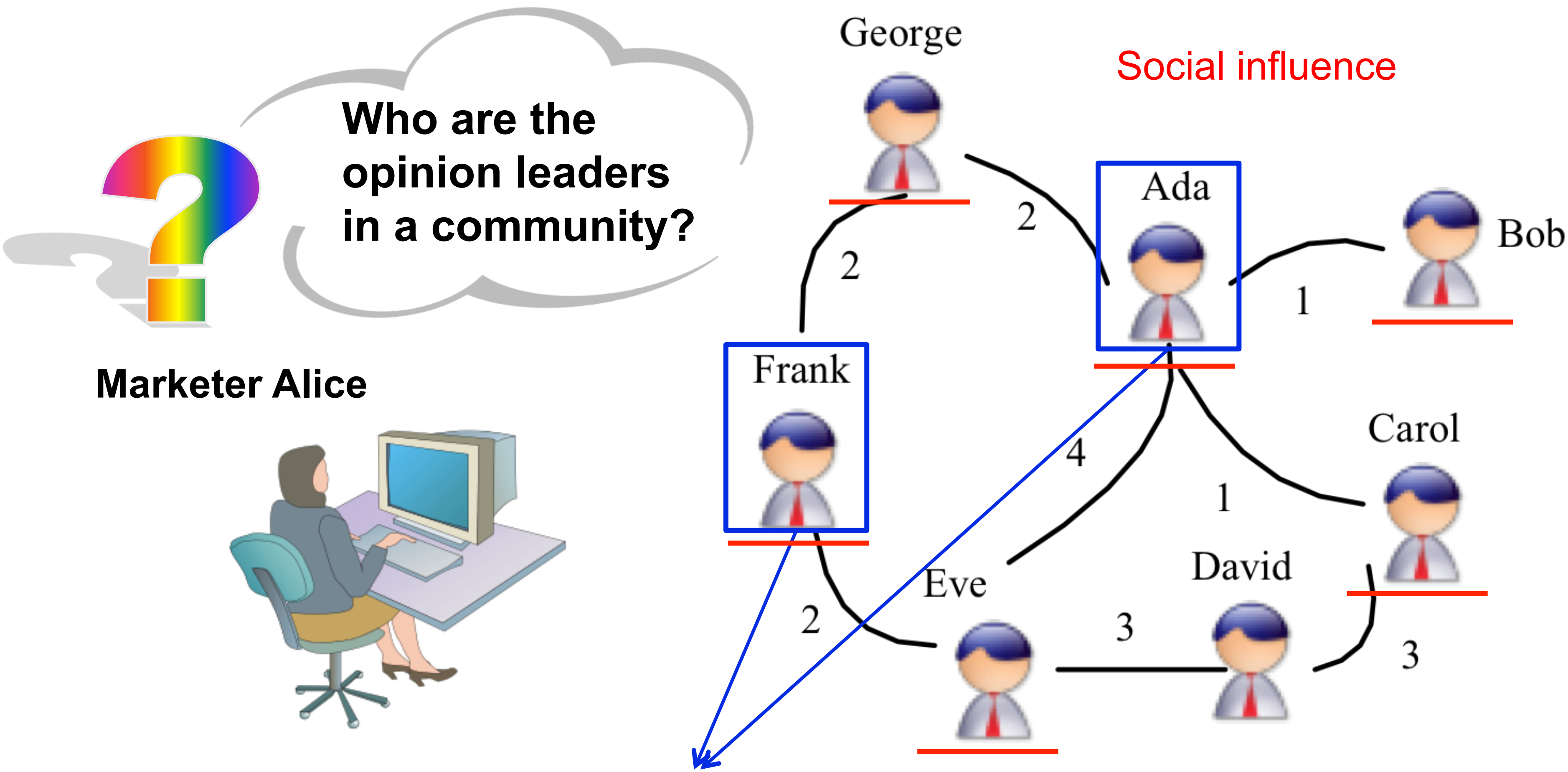
Case 2: Klout[1]—"the standard of influence"

- Toward measuring real-world influence
  - Twitter, Facebook, G+, LinkedIn, etc.
  - Klout generates a score on a scale of 1-100 for a social user to represent her/his ability to engage other people and inspire social actions.
  - Has built 100 million profiles.
- Though controversial[2], in May 2012, Cathay Pacific opens SFO lounge to Klout users
  - A high Klout score gets you into Cathay Pacific's SFO lounge

[1] http://klout.com

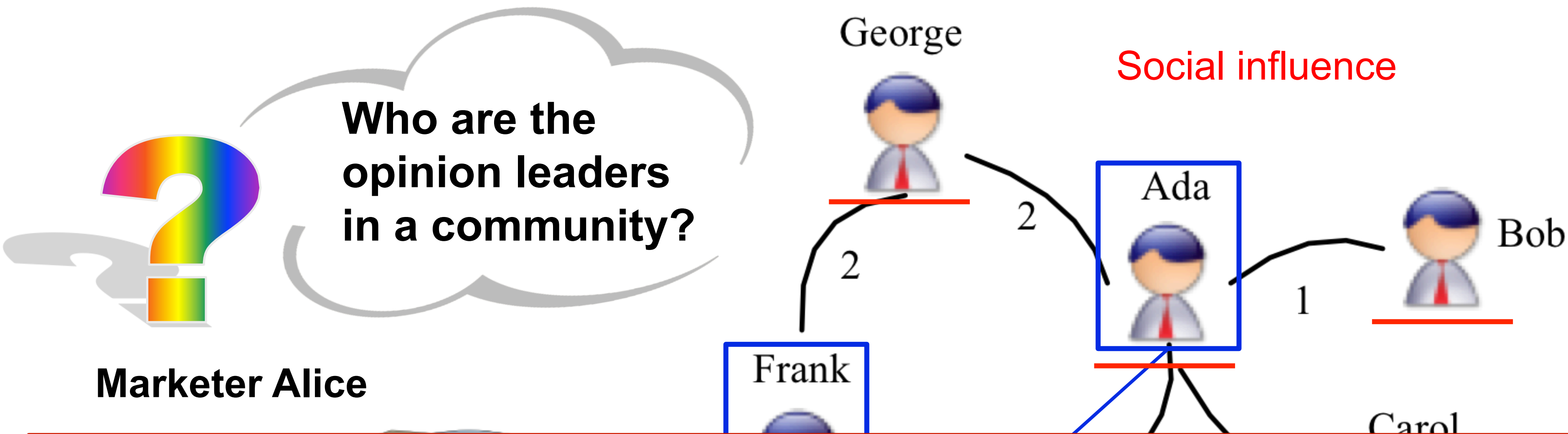
[2] Why I Deleted My Klout Profile, by Pam Moore, at Social Media Today, originally published November 19, 2011; retrieved November 26 2011

Influence Maximization



Find *K* nodes (users) in a social network that could maximize the spread of influence (Domingos, 01; Richardson, 02; Kempe, 03)

Influence Maximization

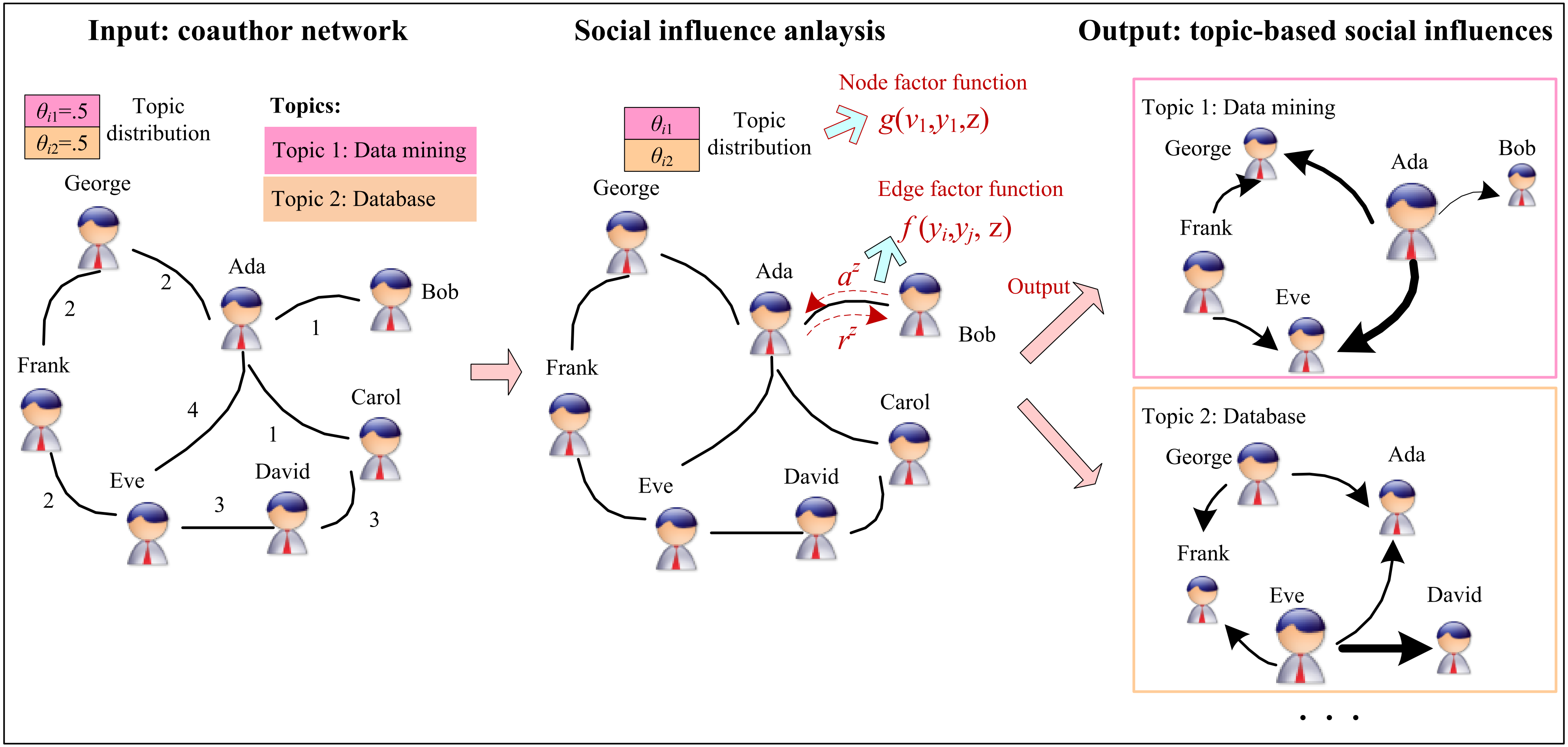


Questions:

- How to quantify the strength of social influence between users?
- How to predict users' behaviors over time?

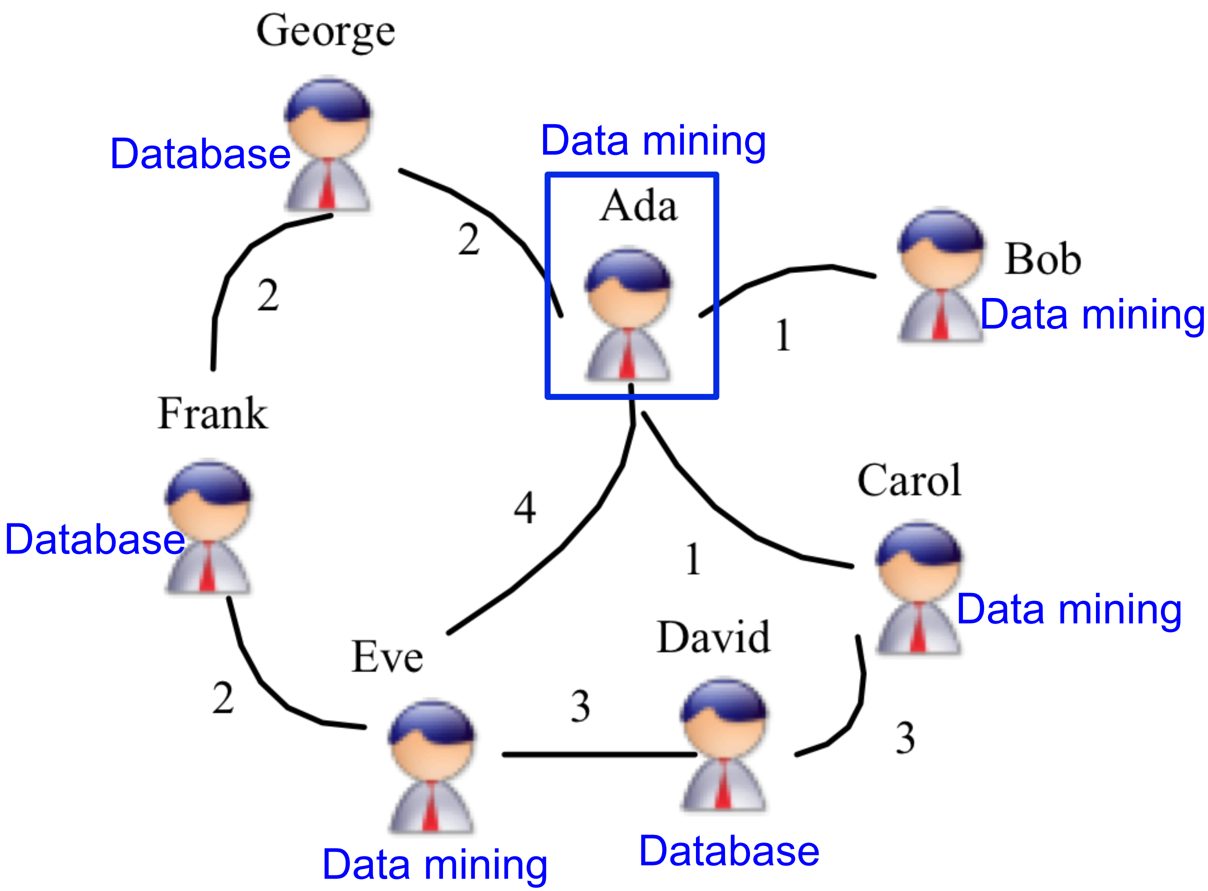
Topic-based Social Influence Analysis

- Social network -> Topical influence network



[1] J. Tang, J. Sun, C. Wang, and Z. Yang. Social Influence Analysis in Large-scale Networks. In KDD'09, pages 807-816, 2009.

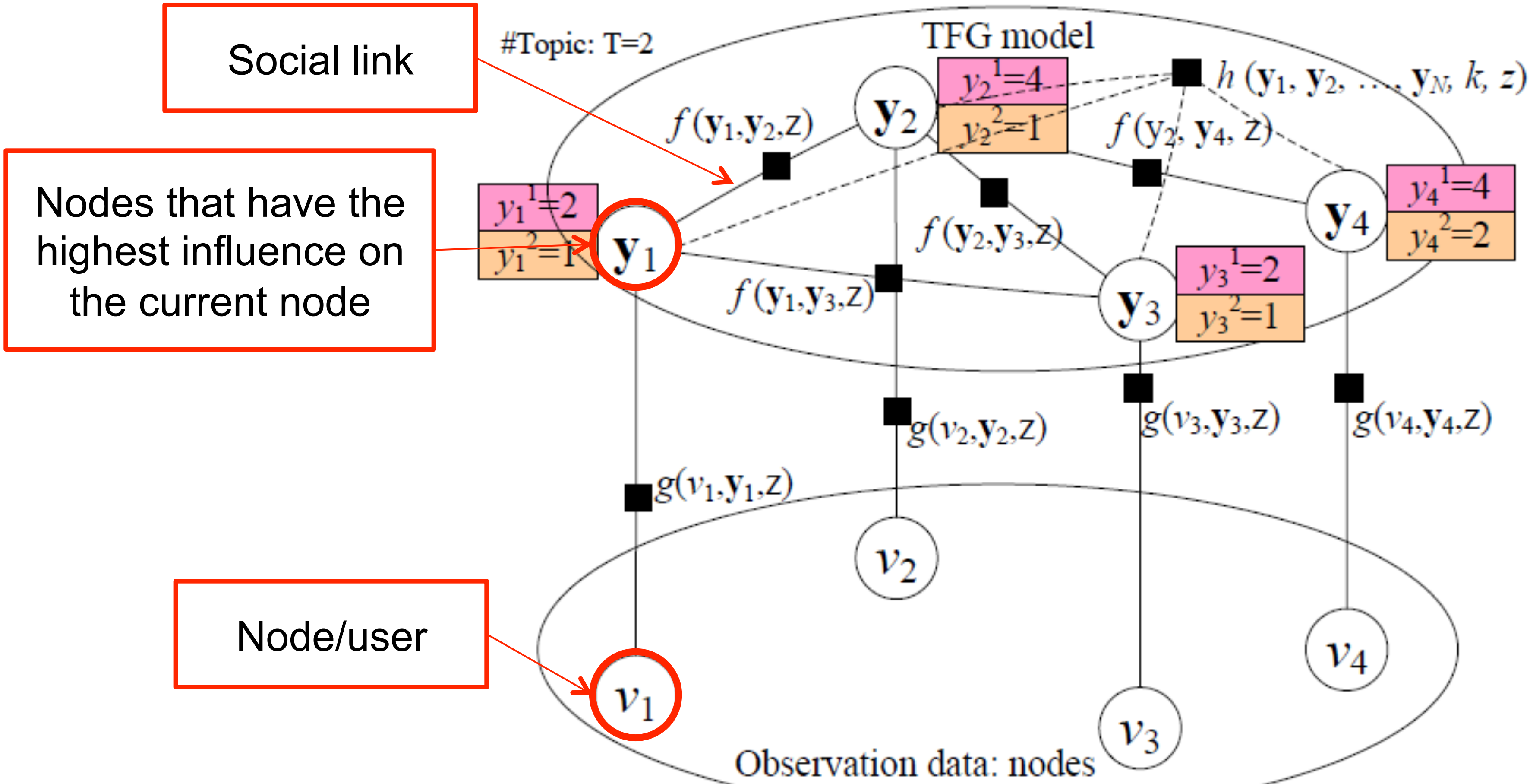
The Solution: Topical Affinity Propagation



Basic Idea:
If a user is located in the center of a "DM" community, then he may have strong influence on the other users.
—Homophily theory

[1] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In KDD, pages 807-816, 2009.

Topical Factor Graph (TFG) Model



The problem is cast as identifying which node has the **highest probability** to **influence** another node on a **specific topic** along with the edge.

Topical Factor Graph (TFG)

Objective function:

$$P(\mathbf{v}, \mathbf{Y}) = \frac{1}{Z} \prod_{k=1}^{N} \prod_{z=1}^{T} \boxed{h(\mathbf{y}_1, \ldots, \mathbf{y}_N, k, z)}$$

$$\prod_{i=1}^{N} \prod_{z=1}^{T} \boxed{g(v_i, \mathbf{y}_i, z)} \prod_{e_{kl} \in E} \prod_{z=1}^{T} \boxed{f(\mathbf{y}_k, \mathbf{y}_l, z)}$$

1. How to define?

2. How to optimize?

- The learning task is to find a configuration for all $\{y_i\}$ to maximize the joint probability.

How to define (topical) feature functions?

– Node feature function

similarity

$$g(v_i, \mathbf{y}_i, z) = \begin{cases} \frac{\boxed{w^z_{iy^z_i}}}{\sum_{j \in NB(i)} (w^z_{ij} + w^z_{ji})} & y^z_i \neq i \\ \frac{\sum_{j \in NB(i)} w^z_{ji}}{\sum_{j \in NB(i)} (w^z_{ij} + w^z_{ji})} & y^z_i = i \end{cases}$$

– Edge feature function

$$f(y_i, y_j) = \begin{cases} w[v_i \sim v_j] & y_i = y_j \\ 1 - w[v_i \sim v_j] & y_i \neq y_j \end{cases}$$

or simply binary

– Global feature function

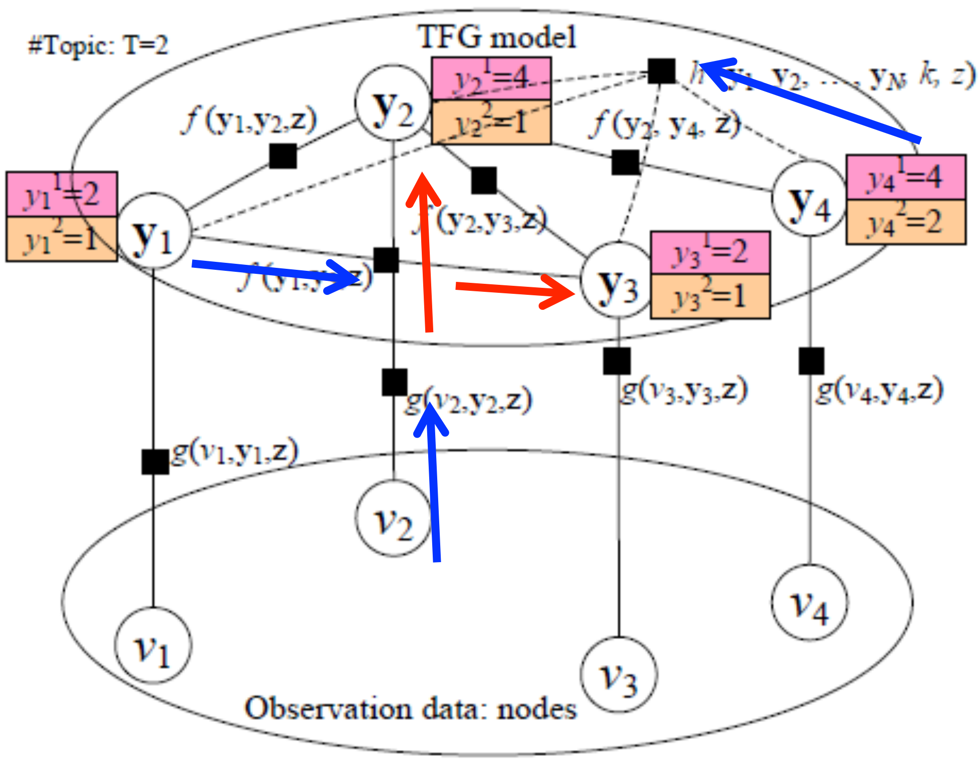
$$h(\mathbf{y}_1, \ldots, \mathbf{y}_N, k, z) = \begin{cases} 0 & \text{if } y^z_k = k \text{ and } y^z_i \neq k \text{ for all } i \neq k \\ 1 & \text{otherwise.} \end{cases}$$

Model Learning Algorithm

Sum-product:

$$\boxed{m_{y\to f}(y,z)} = \prod_{f'\sim y\setminus f} m_{f'\to y}(y,z) \prod_{z'\neq z}\prod_{f'\sim y\setminus f} m_{f'\to y}(y,z')^{(\tau_{z'z})}$$

$$\begin{aligned}\boxed{m_{f\to y}(y,z)} &= \sum_{\sim\{y\}}\Big(f(Y,z)\prod_{y'\sim f\setminus y} m_{y'\to f}(y',z)\Big)\\ &+ \sum_{z'\neq z}\tau_{z'z}\sum_{\sim\{y\}}\Big(f(Y,z')\prod_{y'\sim f\setminus y} m_{y'\to f}(y',z')\Big) \quad (4)\end{aligned}$$



- Low efficiency!
- Not easy for distributed learning!

New TAP Learning Algorithm

1. Introduce two new variables *r* and *a, to* replace the original message *m*.

2. Design new update rules:

$\boxed{m_{ij}}$

$$r_{ij}^{z} = b_{ij}^{z} - \max_{k \in NB(j)} \{b_{ik}^{z} + a_{ik}^{z}\}$$

$$a_{jj}^{z} = \max_{k \in NB(j)} \min\{r_{kj}^{z}, 0\}$$

$$a_{ij}^{z} = \min(\max\{r_{jj}^{z}, 0\}, -\min\{r_{jj}^{z}, 0\} - \max_{k \in NB(j)\setminus\{i\}} \min\{r_{kj}^{z}, 0\}), i \in NB(j)$$

[1] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In KDD, pages 807-816, 2009.

The TAP Learning Algorithm

Input: $G = (V, E)$ and topic distributions $\{\theta_v\}_{v \in V}$

Output: topic-level social influence graphs $\{G_z = (V_z, E_z)\}^T$

Calculate the node feature function $g(v_i, \mathbf{y}_i, z)$;
Calculate b_{ij}^z according to Eq. 8;
Initialize all $\{r_{ij}^z\} \leftarrow 0$;
repeat
  foreach *edge-topic pair* (e_{ij}, z) **do**
    Update r_{ij}^z according to Eq. 5;
  end
  foreach *node-topic pair* (v_j, z) **do**
    Update a_{jj}^z according to Eq. 6;
  end
  foreach *edge-topic pair* (e_{ij}, z) **do**
    Update a_{ij}^z according to Eq. 7;
  end
until *convergence*;
foreach *node* v_t **do**
  foreach *neighboring node* $s \in NB(t) \cup \{t\}$ **do**
    Compute μ_{st}^z according to Eq. 9;
  end
end
Generate $G_z = (V_z, E_z)$ for every topic z according to $\{\mu_{st}^z\}$;

$$b_{ij}^z = \log \frac{g(v_i, \mathbf{y}_i, z)|_{y_i^z = j}}{\sum_{k \in NB(i) \cup \{i\}} g(v_i, \mathbf{y}_i, z)|_{y_i^z = k}}$$

$$r_{ij}^z = b_{ij}^z - \max_{k \in NB(j)} \{b_{ik}^z + a_{ik}^z\}$$

$$a_{jj}^z = \max_{k \in NB(j)} \min\{r_{kj}^z, 0\}$$

$$a_{ij}^z = \min(\max\{r_{jj}^z, 0\}, -\min\{r_{jj}^z, 0\} - \max_{k \in NB(j) \backslash \{i\}} \min\{r_{kj}^z, 0\}), i \in NB(j)$$

$$\mu_{st}^z = \frac{1}{1 + e^{-(r_{ts}^z + a_{ts}^z)}}$$

Experiments

- Data set: (http://arnetminer.org/lab-datasets/soinf/)

Data set	#Nodes	#Edges
Coauthor	640,134	1,554,643
Citation	2,329,760	12,710,347
Film (Wikipedia)	18,518 films 7,211 directors 10,128 actors 9,784 writers	142,426

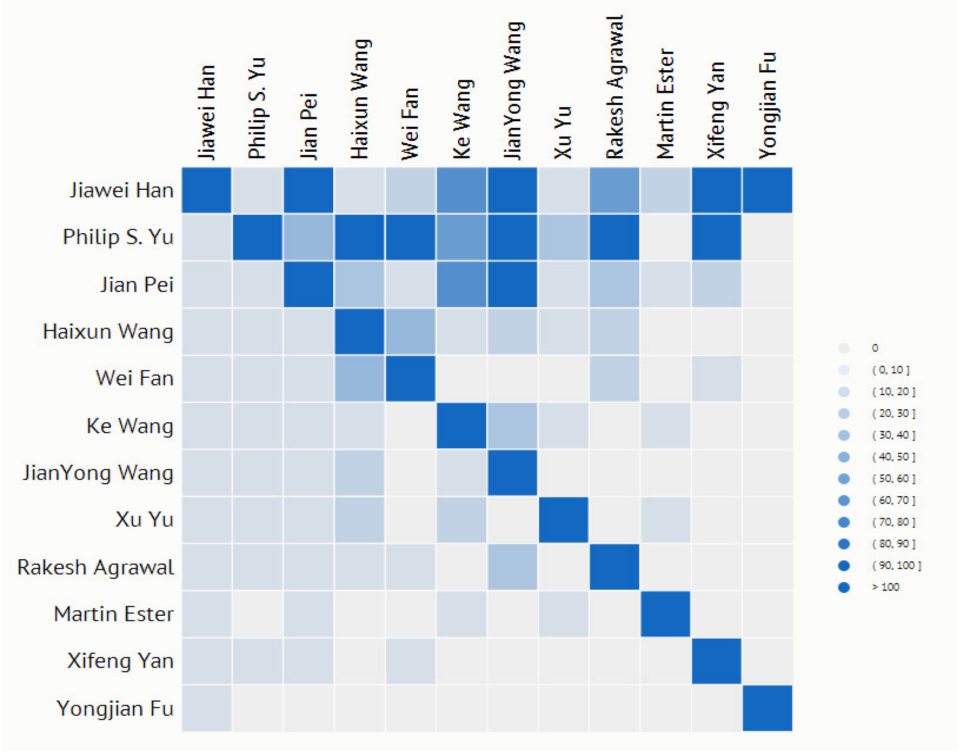
- Evaluation measures
  - CPU time
  - Case study
  - Application

Social Influence Sub-graph on “Data mining”

Table 4: Dynamic influence analysis for Dr. Jian Pei during 2000-2009. Due to space limitation, we only list coauthors who most influence on/by Dr. Pei in each time window.

Year	Pairwise	Influence
2000 - 2001	Influence on Dr. Pei	Jiawei Han (0.4961)
	Influenced by Dr. Pei	Jiawei Han (0.0082)
2002 - 2003	Influence on Dr. Pei	Jiawei Han (0.4045), Ke Wang (0.0418), Jianyong Wang (0.019), Xifeng Yan (0.007), Shiwei Tang (0.0052)
	Influenced by Dr. Pei	Shiwei Tang (0.436), Hasan M.Jamil (0.4289), Xifeng Yan (0.2192), Jianyong Wang (0.1667), Ke Wang (0.0687)
2004 - 2005	Influence on Dr. Pei	Jiawei Han (0.2364), Ke Wang (0.0328), Wei Wang (0.0294), Jianyong Wang (0.0248), Philip S. Yu (0.0156)
	Influenced by Dr. Pei	Chun Tang (0.5929), Shiwei Tang (0.5426), Hasan M.Jamil (0.3318), Jianyong Wang (0.1609), Xifeng Yan (0.1458), Yan Huang (0.1054)
2006 - 2007	Influence on Dr. Pei	Jiawei Han (0.1201), Ke Wang (0.0351), Wei Wang (0.0226), Jianyong Wang (0.018), Ada Wai-Chee Fu (0.0125)
	Influenced by Jian Pei	Chun Tang (0.6095), Shiwei Tang (0.6067), Byung-Won On (0.4599), Hasan M.Jamil (0.3433), Jaewoo Kang (0.3386)
2008 - 2009	Influence on Dr. Pei	Jiawei Han (0.2202), Ke Wang (0.0234), Ada Wai-Chee Fu (0.0208), Wei Wang (0.011), Jianyong Wang (0.0095)
	Influenced by Dr. Pei	ZhaoHui Tang (0.654), Chun Tang (0.6494), Shiwei Tang (0.5923), Zhengzheng Xing (0.5549), Hasan M.Jamil (0.3333), Jaewoo Kang (0.3057)

On “Data Mining” in 2009



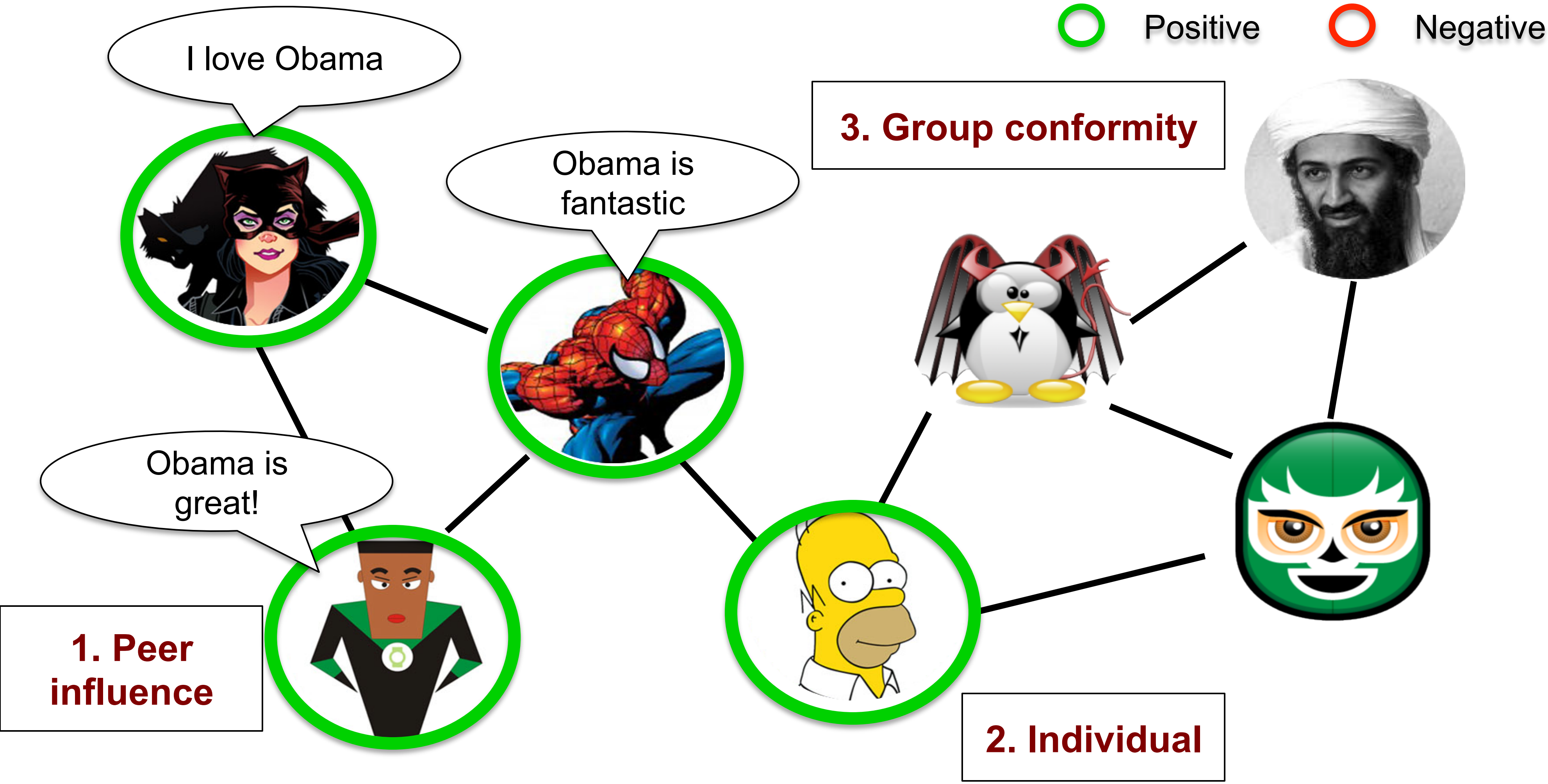
Results on Coauthor and Citation

Dataset	Topic	Representative Nodes
Author	Data Mining	Heikki Mannila, Philip S. Yu, Dimitrios Gunopulos, Jiawei Han, Christos Faloutsos, Bing Liu, Vipin Kumar, Tom M. Mitchell, Wei Wang, Qiang Yang, Xindong Wu, Jeffrey Xu Yu, Osmar R. Zaiane
	Machine Learning	Pat Langley, Alex Waibel, Trevor Darrell, C. Lee Giles, Terrence J. Sejnowski, Samy Bengio, Daphne Koller, Luc De Raedt, Vasant Honavar, Floriana Esposito, Bernhard Scholkopf
	Database System	Gerhard Weikum, John Mylopoulos, Michael Stonebraker, Barbara Pernici, Philip S. Yu, Sharad Mehrotra, Wei Sun, V. S. Subrahmanian, Alejandro P. Buchmann, Kian-Lee Tan, Jiawei Han
	Information Retrieval	Gerard Salton, W. Bruce Croft, Ricardo A. Baeza-Yates, James Allan, Yi Zhang, Mounia Lalmas, Zheng Chen, Ophir Frieder, Alan F. Smeaton, Rong Jin
	Web Services	Yan Wang, Liang-jie Zhang, Schahram Dustdar, Jian Yang, Fabio Casati, Wei Xu, Zakaria Maamar, Ying Li, Xin Zhang, Boualem Benatallah, Boualem Benatallah
	Semantic Web	Wolfgang Nejdl, Daniel Schwabe, Steffen Staab, Mark A. Musen, Andrew Tomkins, Juliana Freire, Carole A. Goble, James A. Hendler, Rudi Studer, Enrico Motta
	Bayesian Network	Daphne Koller, Paul R. Cohen, Floriana Esposito, Henri Prade, Michael I. Jordan, Didier Dubois, David Heckerman, Philippe Smets
Citation	Data Mining	Fast Algorithms for Mining Association Rules in Large Databases, Using Segmented Right-Deep Trees for the Execution of Pipelined Hash Joins, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Discovery of Multiple-Level Association Rules from Large Databases, Interleaving a Join Sequence with Semijoins in Distributed Query Processing
	Machine Learning	Object Recognition with Gradient-Based Learning, Correctness of Local Probability Propagation in Graphical Models with Loops, A Learning Theorem for Networks at Detailed Stochastic Equilibrium, The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length, A Unifying Review of Linear Gaussian Models
	Database System	Mediators in the Architecture of Future Information Systems, Database Techniques for the World-Wide Web: A Survey, The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles, Fast Algorithms for Mining Association Rules in Large Databases
	Web Services	The Web Service Modeling Framework WSMF, Interval Timed Coloured Petri Nets and their Analysis, The design and implementation of real-time schedulers in RED-linux, The Self-Serv Environment for Web Services Composition
	Web Mining	Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Fast Algorithms for Mining Association Rules in Large Databases, The OO-Binary Relationship Model: A Truly Object Oriented Conceptual Model, Distributions of Surfers' Paths Through the World Wide Web: Empirical Characterizations, Improving Fault Tolerance and Supporting Partial Writes in Structured Coterie Protocols for Replicated Objects
	Semantic Web	FaCT and iFaCT, The GRAIL concept modelling language for medical terminology, Semantic Integration of Semistructured and Structured Data Sources, Description of the RACER System and its Applications, DL-Lite: Practical Reasoning for Rich Dls

Still Challenges

How to model influence at different granularities?

Conformity Influence



[1] Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity Influence in Large Social Networks. In **KDD'13**, 2013.

Conformity Influence Definition

- Three levels of conformities
  - Individual conformity
  - Peer conformity
  - Group conformity

Individual Conformity

- The individual conformity represents how easily user v's behavior conforms to her friends

A specific action performed by user v at time t

Exists a friend v' who performed the same action at time t"

$$icf(v) = \frac{\left|\boxed{(a, v, t)} \in A_v \,\middle|\, \exists \boxed{(a, v', t') : e_{vv'} \in E \wedge \epsilon \geq t - t' \geq 0}\right|}{\boxed{|A_v|}}$$

All actions by user v

Peer Conformity

- The peer conformity represents how likely the user v's behavior is influenced by one particular friend v'

A specific action performed by user v' at time t'

User v follows v' to perform the action a at time t

$$pcf(v, v') = \frac{|\boxed{(a, v', t')} \in A_{v'} | \boxed{\exists (a, v, t) : e_{vv'} \in E \wedge \epsilon \geq t - t' \geq 0}|}{\boxed{|A_{v'}|}}$$

All actions by user v'

Group Conformity

- The group conformity represents the conformity of user v's behavior to groups that the user belongs to.

***τ-group action:* an action performed by more than a percentage τ of all users in the group C_k**

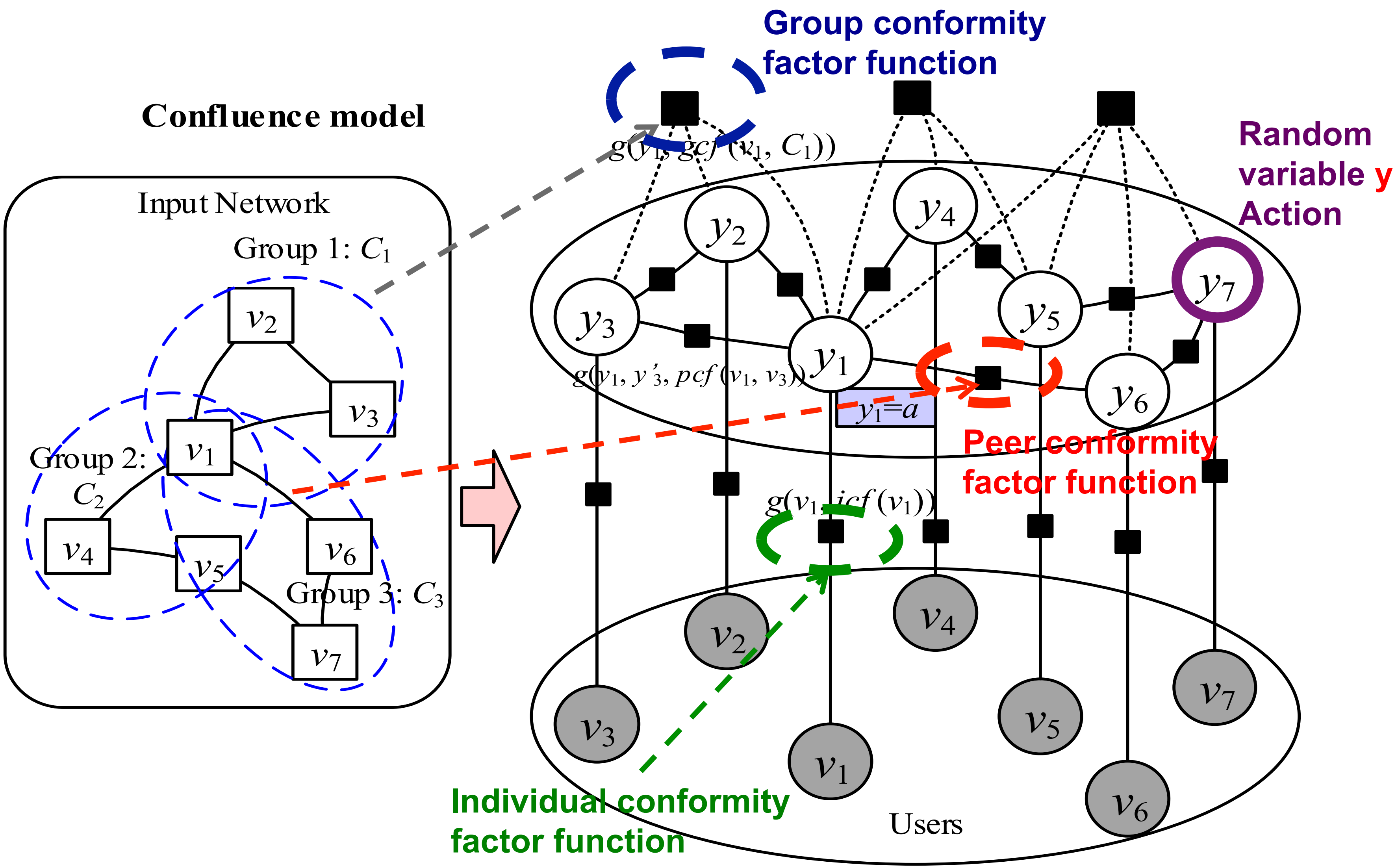
A specific τ-group action

User v conforms to the group to perform the action a at time t

$$gcf^{\tau}(v, C_{vk}) = \frac{\left| \boxed{(a, v', t')} \in A^{\tau}_{C_k} \boxed{\left| \exists (a, v, t) : \mathbb{I}[c_{ik}] \wedge \epsilon \geq t - t' \geq 0 \right|} \right.}{\boxed{\left| A^{\tau}_{C_k} \right|}}$$

All τ-group actions performed by users in the group C_k

Confluence
—A conformity-aware factor graph model



[1] Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity Influence in Large Social Networks. In **KDD'13**, 2013.

Model Instantiation

$$
\begin{aligned}
\mathcal{O}(\theta) &= \log P_\theta(Y|G,A) \\
&= \sum_{i=1}^{N}\left[\sum_{j=1}^{d}\alpha_j f(y_i, x_{ij}) + \beta_i \boxed{g(y_i, icf(v_i))}\right] \\
&\quad + \sum_{e_{ij}\in E} \mathbb{I}[y'] \gamma_{ij} \boxed{g(y_i, y'_j, pcf(v_i, v_j))} \\
&\quad + \sum_{i=1}^{N}\sum_{k=1}^{m} \mathbb{I}[c_{ik}] \mu_{ik} \boxed{g(y_i, gcf(v_i, C_k))} - \log Z
\end{aligned}
$$

Individual conformity factor function

$$\boxed{g(y_i, y'_j, pcf(v_i, v_j)) = (\frac{1}{2})^{\frac{t-t'}{\lambda}} pcf(v_i, v_j)}$$

Peer conformity factor function

$$\boxed{g(y_i, gcf^\tau(v_i, C_k)) = (\frac{1}{2})^{\frac{t-t'}{\lambda}} gcf^\tau(v_i, C_k)}$$

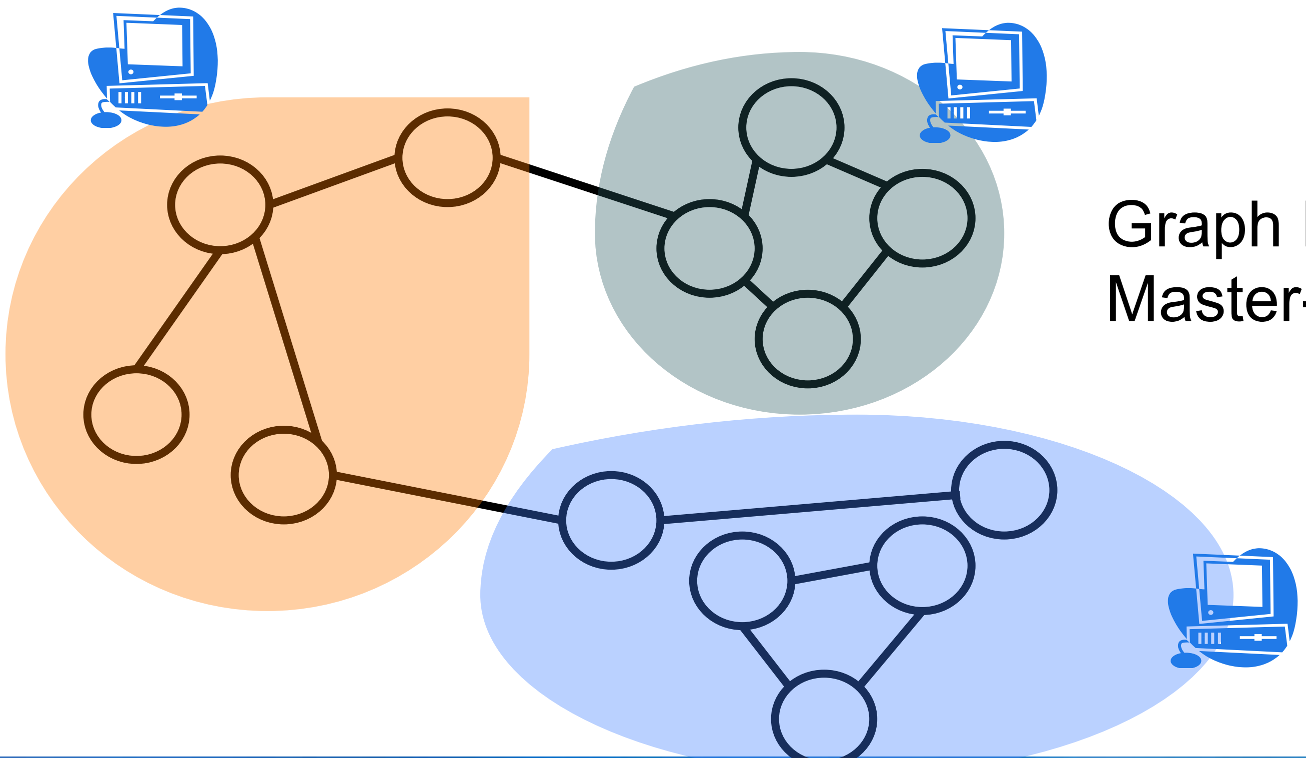
Group conformity factor function

$$\boxed{g(y_i, icf(v_i)) = \frac{\sum_{k=1}^{|A_{v_i}|} (\frac{1}{2})^{\frac{t-t'}{\lambda}} \mathbb{I}[y'_j \wedge e_{ij} \in E]}{|A_v|}}$$

Distributed Learning

Master
Global update

Slave
Compute local gradient via random sampling



Graph Partition by Metis
Master-Slave Computing

Distributed Model Learning

Input: network G, action history A, and learning rate η;
Output: learned parameters $\theta = (\{\alpha\}, \{\beta\}, \{\gamma\}, \{\mu\})$;

Unknown parameters to estimate

Initialize $\alpha, \beta, \gamma, \mu$;
Construct the graphical structure G in the Confluence model;
Partition the graph G into M subgraphs $[G_1, \cdots, G_M]$;
repeat

%Distribute the parameter to calculate local belief ;
Master broadcasts θ to all Slaves;

(1) Master

for $l = 1$ *to* M **do**
Each Slave calculates local belief for each marginal probability according to Eqs. 6 and 7 on subgraph G_l;
Slave send back the obtained local belief;
end

(2) Slave

$$m^l_{ij}(y_i) = \sigma \sum_{y_i} \psi^l_{ij}(y_i, y_j) \psi^l_i(y_i) \prod_{k \in NB(i) \backslash j} m^l_{ki}(y_i)$$

$$P(y_i|.) = \sigma \sum_{l=1}^{M} b^l_i(y_i)$$

...ulate the marginals an...
...r calculates the margin...
...r calculates the gradien...
Master updates all parameter...

$$b^l_i(y_i) = \psi^l_i(y_i) \prod_{k \in NB(i)} m^l_{ki}(y_i)$$

(3) Master

$$\alpha_j^{new} = \alpha_j^{old} + \eta \frac{\mathcal{O}(\theta)}{\alpha_j}$$

until *convergence*;

Algorithm 1: Distributed model learning.

Results with Conformity Influence — Four Datasets

Network	#Nodes	#Edges	Behavior	#Actions
Weibo	1,776,950	308,489,739	Post a tweet	6,761,186
Flickr	1,991,509	208,118,719	Add comment	3,531,801
Gowalla	196,591	950,327	Check-in	6,442,890
ArnetMiner	737,690	2,416,472	Publish paper	1,974,466

- Baselines
  - *Support Vector Machine (SVM)*
  - *Logistic Regression (LR)*
  - *Naive Bayes (NB)*
  - *Gaussian Radial Basis Function Neural Network (RBF)*
  - *Conditional Random Field (CRF)*
- Evaluation metrics
  - *Precision, Recall, F1, and Area Under Curve (AUC)*

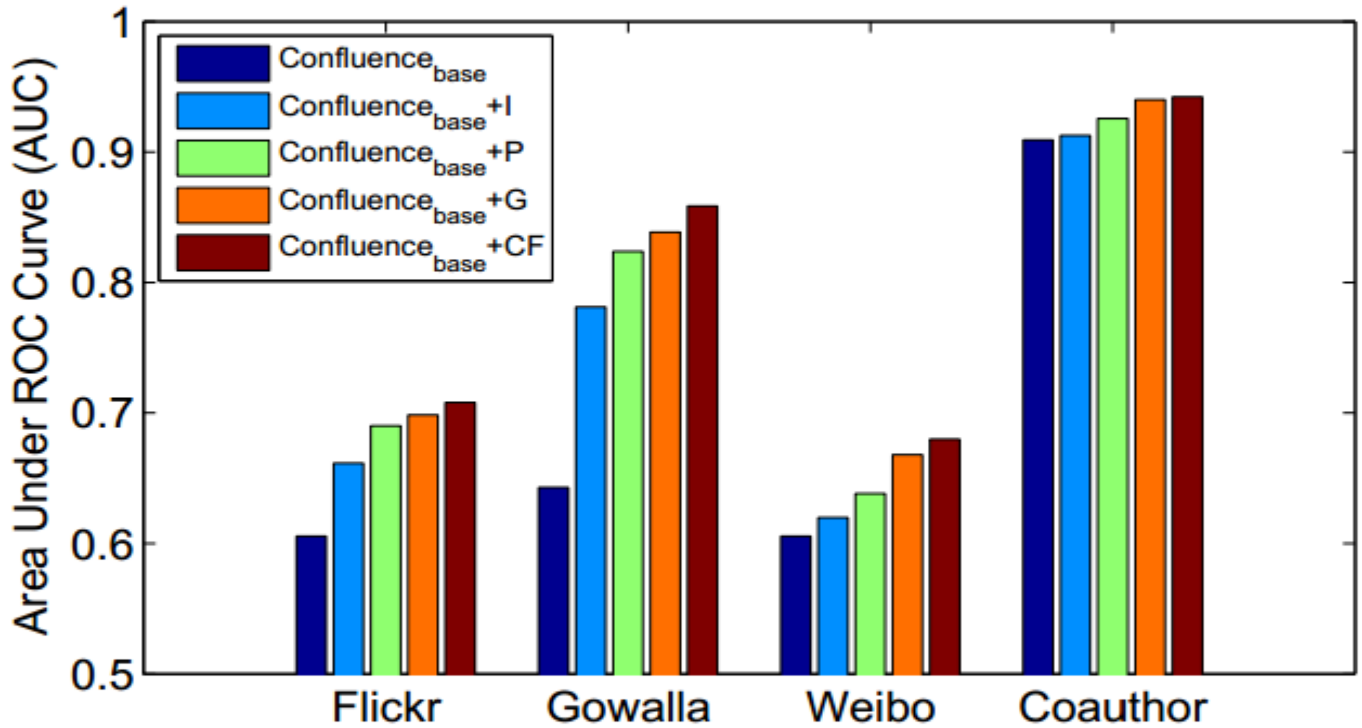
** All the datasets are publicly available for research.

Prediction Accuracy

Data	Method	Precision	Recall	F1-Measure	AUC
Flickr	SVM	0.5921 (±0.0036)	0.5905 (±0.0031)	0.5802 (±0.0012)	0.6473 (±0.0004)
	LR	0.6010 (±0.0052)	0.5900 (±0.0057)	0.5770 (±0.0018)	0.6510 (±0.0008)
	NB	0.6170 (±0.0071)	0.6040 (±0.0083)	0.5920 (±0.0031)	0.6520 (±0.0019)
	RBF	**0.6250 (±0.0039)**	0.5960 (±0.0010)	0.5720 (±0.0024)	0.6700 (±0.0010)
	CRF	0.5474 (±0.0030)	**0.8002 (±0.0009)**	0.6239 (±0.0016)	0.6722 (±0.0010)
	Confluence	0.5472 (±0.0025)	0.7770(±0.0010)	**0.6342 (±0.0010)**	**0.7383 (±0.0006)**
Gowalla	SVM	0.9290 (±0.0212)	0.9310 (±0.0121)	0.9295 (±0.0105)	0.9280 (±0.0042)
	LR	0.9320 (±0.0234)	0.9290 (±0.0234)	0.9310 (±0.0155)	0.9500 (±0.0054)
	NB	0.9310 (±0.0197)	0.9290 (±0.0335)	0.9300 (±0.0223)	0.9520 (±0.0030)
	RBF	0.9320 (±0.0254)	0.9280 (±0.0284)	0.9300 (±0.0182)	0.9540 (±0.0022)
	CRF	0.9330 (±0.0100)	0.9320 (±0.0291)	0.9330 (±0.0164)	0.9610 (±0.0019)
	Confluence	**0.9372 (±0.0097)**	**0.9333 (±0.0173)**	**0.9352 (±0.0101)**	**0.9644 (±0.0140)**
Weibo	SVM	0.5060 (±0.0381)	0.5060 (±0.0181)	0.5060 (±0.0157)	0.5070 (±0.0053)
	LR	0.5190 (±0.0461)	0.6450 (±0.0104)	0.5750 (±0.0281)	0.5390 (±0.0133)
	NB	0.5120 (±0.0296)	0.6700 (±0.0085)	0.5810 (±0.0165)	0.5390 (±0.0132)
	RBF	**0.5240 (±0.0248)**	0.5690 (±0.0098)	0.5460 (±0.0159)	0.5450 (±0.0103)
	CRF	0.5150 (±0.0353)	0.6310 (±0.0121)	0.5720 (±0.0209)	0.6320 (±0.0139)
	Confluence	0.5185 (±0.0296)	**0.9967 (±0.0085)**	**0.6816 (±0.0156)**	**0.7572 (±0.0077)**
Co-Author	SVM	0.7672 (±0.0338)	0.8671 (±0.0145)	0.8256 (±0.0129)	0.8562 (±0.0115)
	LR	0.8700 (±0.0261)	0.7640 (±0.0346)	0.8140 (±0.0221)	0.8500 (±0.0030)
	NB	0.7640 (±0.0177)	0.8510 (±0.0185)	0.8050 (±0.0048)	0.8720 (±0.0074)
	RBF	0.7720 (±0.0182)	0.8830 (±0.0191)	0.8240 (±0.0145)	0.8790 (±0.0031)
	CRF	0.8081 (±0.0252)	0.8771 (±0.0249)	0.8360 (±0.0087)	0.9025 (±0.0025)
	Confluence	**0.8818 (±0.0105)**	**0.9089 (±0.0130)**	**0.8818 (±0.0084)**	**0.9579 (±0.0022)**

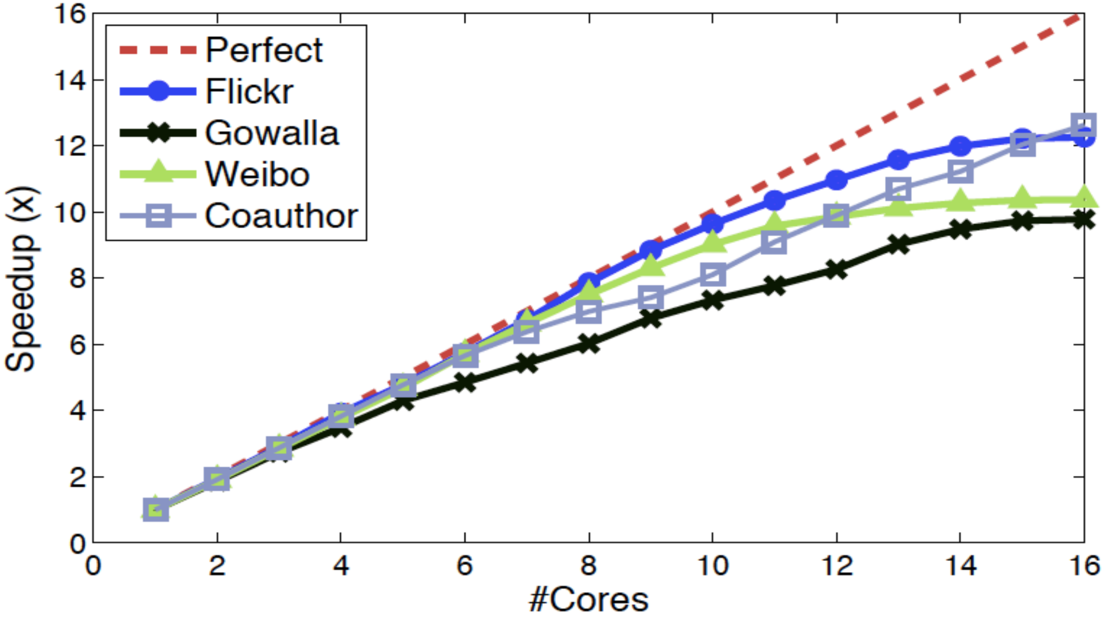
t-test, $p << 0.01$

Effect of Conformity



Confluence$_{base}$ stands for the Confluence method without any social based features
Confluence$_{base}$+I stands for the Confluence$_{base}$ method plus only individual conformity features
Confluence$_{base}$+P stands for the Confluence$_{base}$ method plus only peer conformity features
Confluence$_{base}$+G stands for the Confluence$_{base}$ method plus only group conformity

Scalability performance

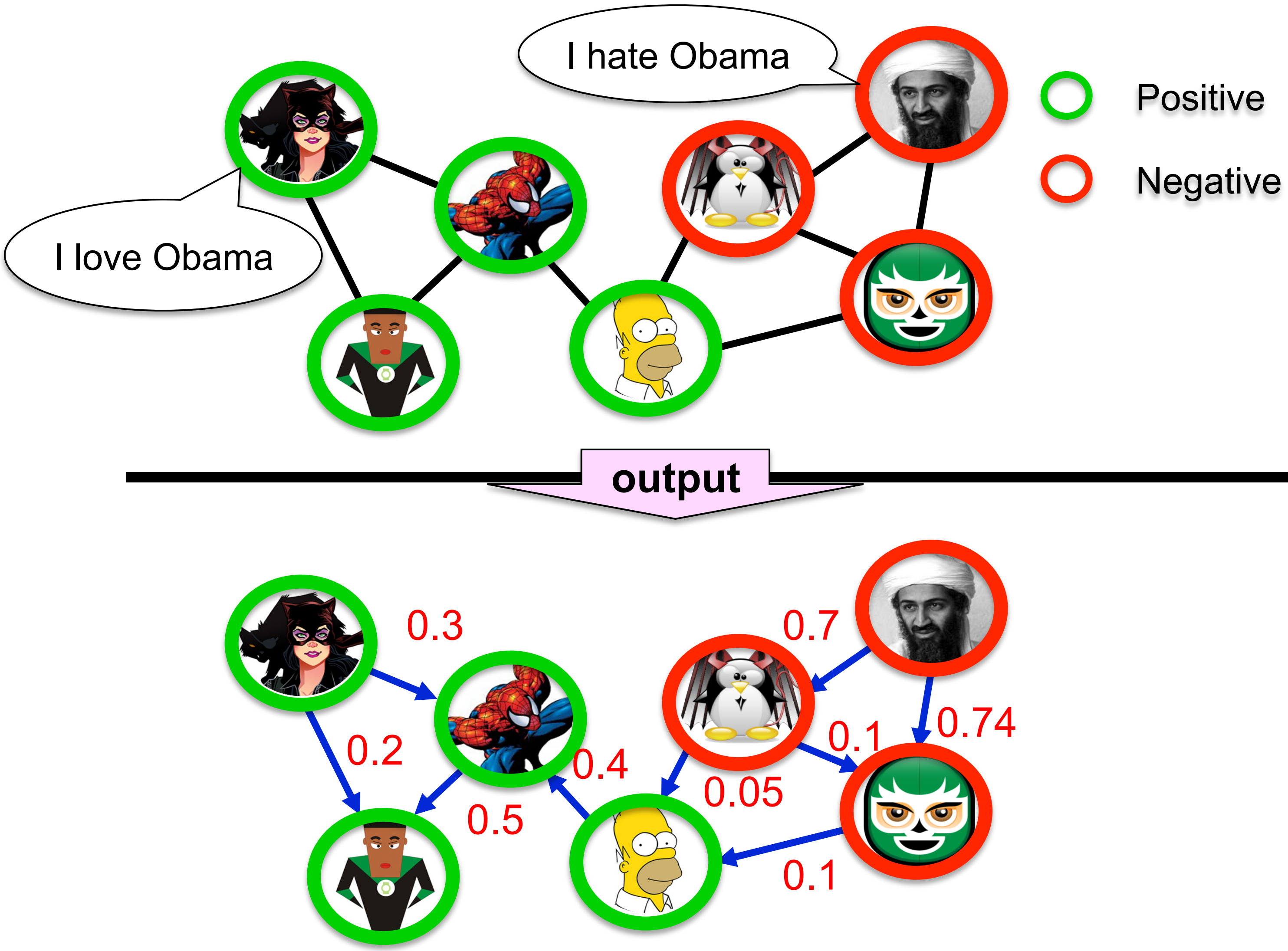


Achieve ~ 9×speedup with 16 cores

Table 4: Running time of the proposed algorithm (hour).

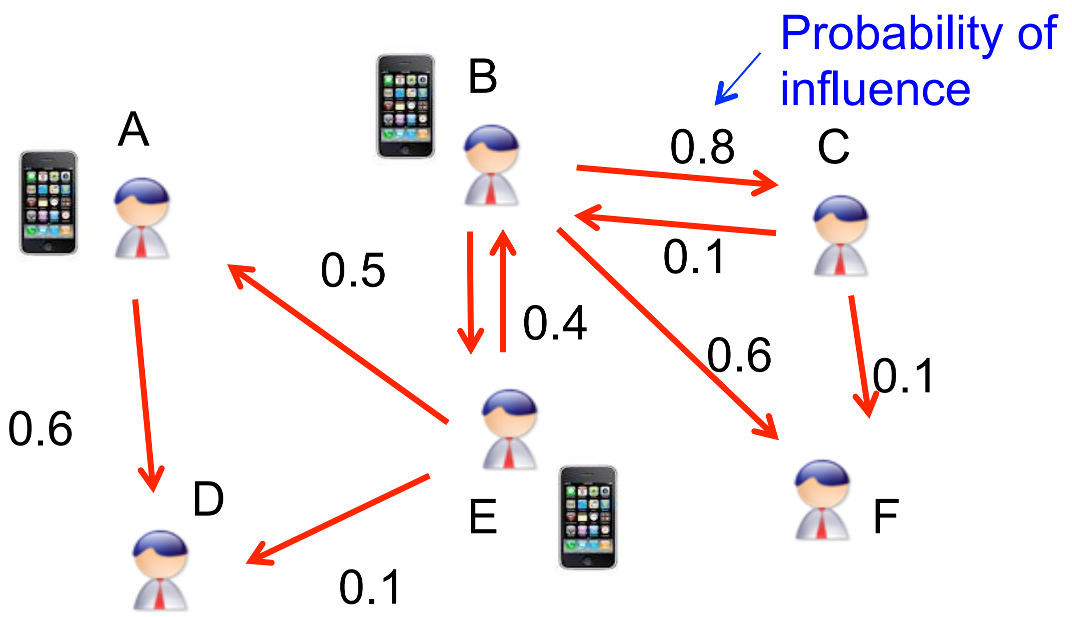
Data Set	Flickr	Gowalla	Weibo	Co-Author
Confluence	1.602	0.245	1.083	0.512
Confluence (single)	19.637	2.395	11.229	6.464
CRF	3.864	0.387	2.547	1.823

Output of social influence learning



Influence Maximization

- Influence maximization
  - Minimize marketing cost and more generally to maximize profit.
  - E.g., to get a small number of influential users to adopt a new product, and subsequently trigger a large cascade of further adoptions.



[1] P. Domingos and M. Richardson. Mining the network value of customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'01), pages 57–66, 2001.

Problem Abstraction

- We associate each user with a status:
  - Active or Inactive
  - The status of the chosen set of users (seed nodes) to market is viewed as active
  - Other users are viewed as inactive
- Influence maximization
  - Initially all users are considered inactive
  - Then the chosen users are activated, who may further influence their friends to be active as well

Diffusion Influence Model

- Linear Threshold Model
- Cascade Model

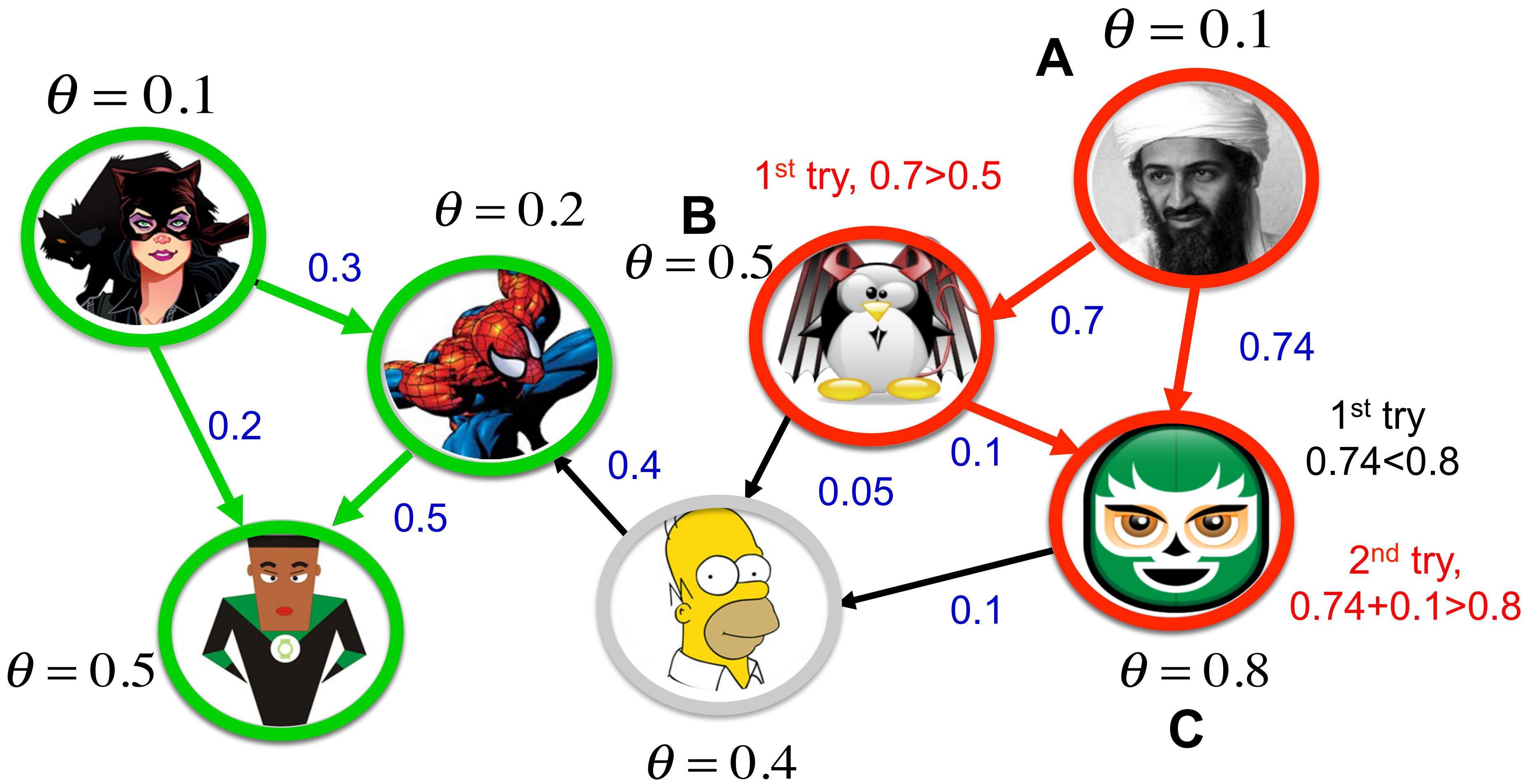
Linear Threshold Model

- General idea
  - Whether a given node will be active can be based on an arbitrary monotone function of its neighbors that are already active.
- Formalization
  - f_v : map subsets of v's neighbors' influence to real numbers in [0,1]
  - θ_v : a threshold for each node
  - S: the set of neighbors of v that are active in step t-1
  - Node v will turn active in step t if $f_v(S) > \theta_v$
- Specifically, in [Kempe, 2003], f_v is defined as $\sum_{u \in S} b_{v.u}$, where $b_{v,u}$ can be seen as a fixed weight, satisfying

$$\sum_{v \in N(u)} b_{u,v} \leq 1$$

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03), pages 137–146, 2003.

Linear Threshold Model: An example



Cascade Model

- Cascade model
  - $p_v(u,S)$: the success probability of user u activating user v
  - User u tries to activate v and finally succeeds, where S is the set of v's neighbors that have already attempted but failed to make v active
- Independent cascade model
  - $p_v(u,S)$ is a constant, meaning that whether v is to be active does not depend on the order v's neighbors try to activate it.
  - Key idea: Flip coins c in advance -> live edges
  - $F_c(A)$: People influenced under outcome c (set cover)
  - $F(A)$ = Sum $_c$P(c) $F_c(A)$ is submodular as well

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03), pages 137–146, 2003.

Theoretical Analysis

- NP-hard[1]
  - Linear threshold model
  - General cascade model
- Kempe Prove that approximation algorithms can guarantee that the influence spread is within(1-1/e) of the optimal influence spread.
  - Verify that the two models can outperform the traditional heuristics
- Recent research focuses on the efficiency improvement
  - [2] accelerates the influence procedure by up to 700 times
- It is still challenging to extend these methods to large data sets

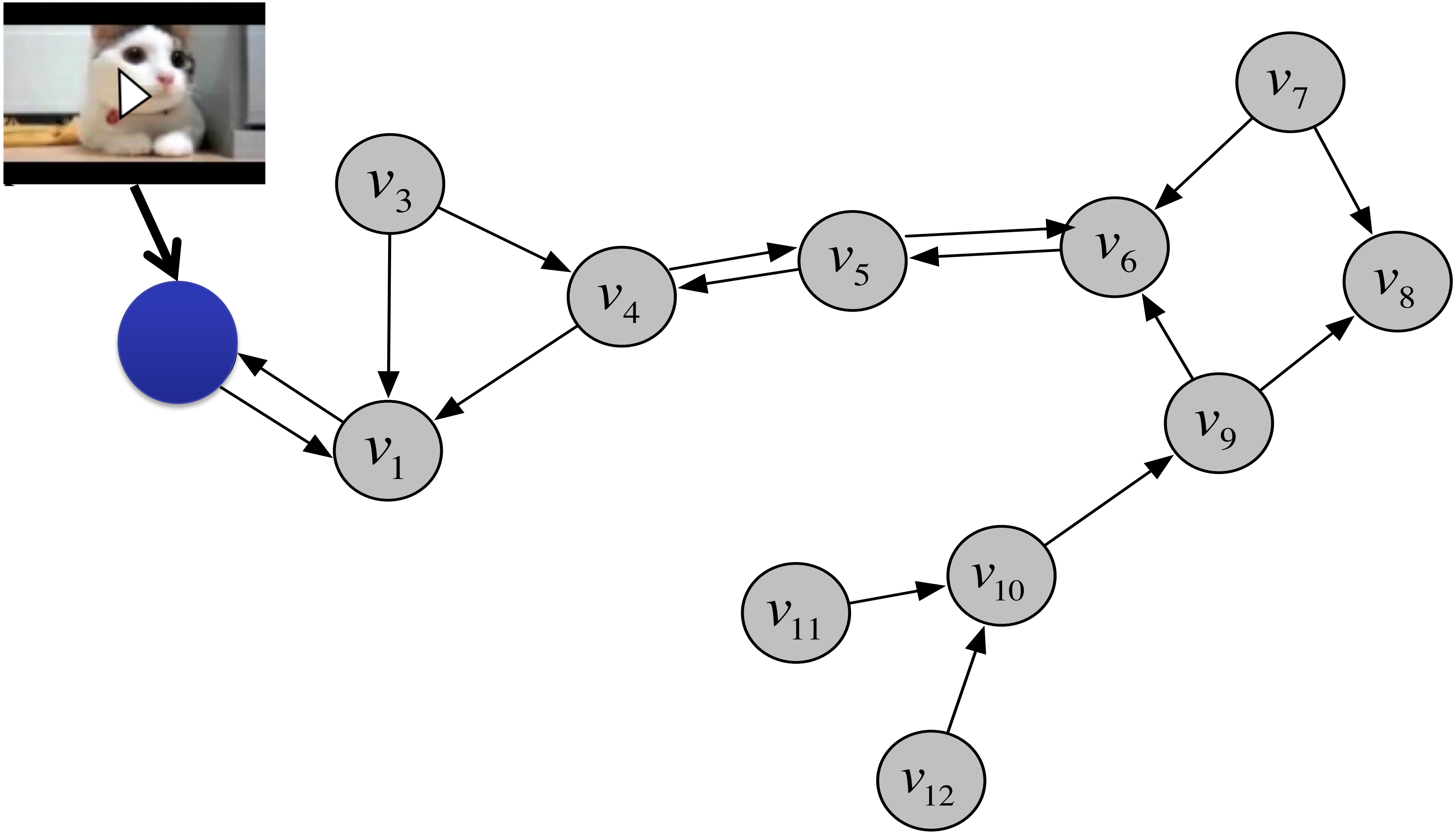
[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'03), pages 137–146, 2003.
[2] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'07), pages 420–429, 2007.

Social Role vs. Information Diffusion

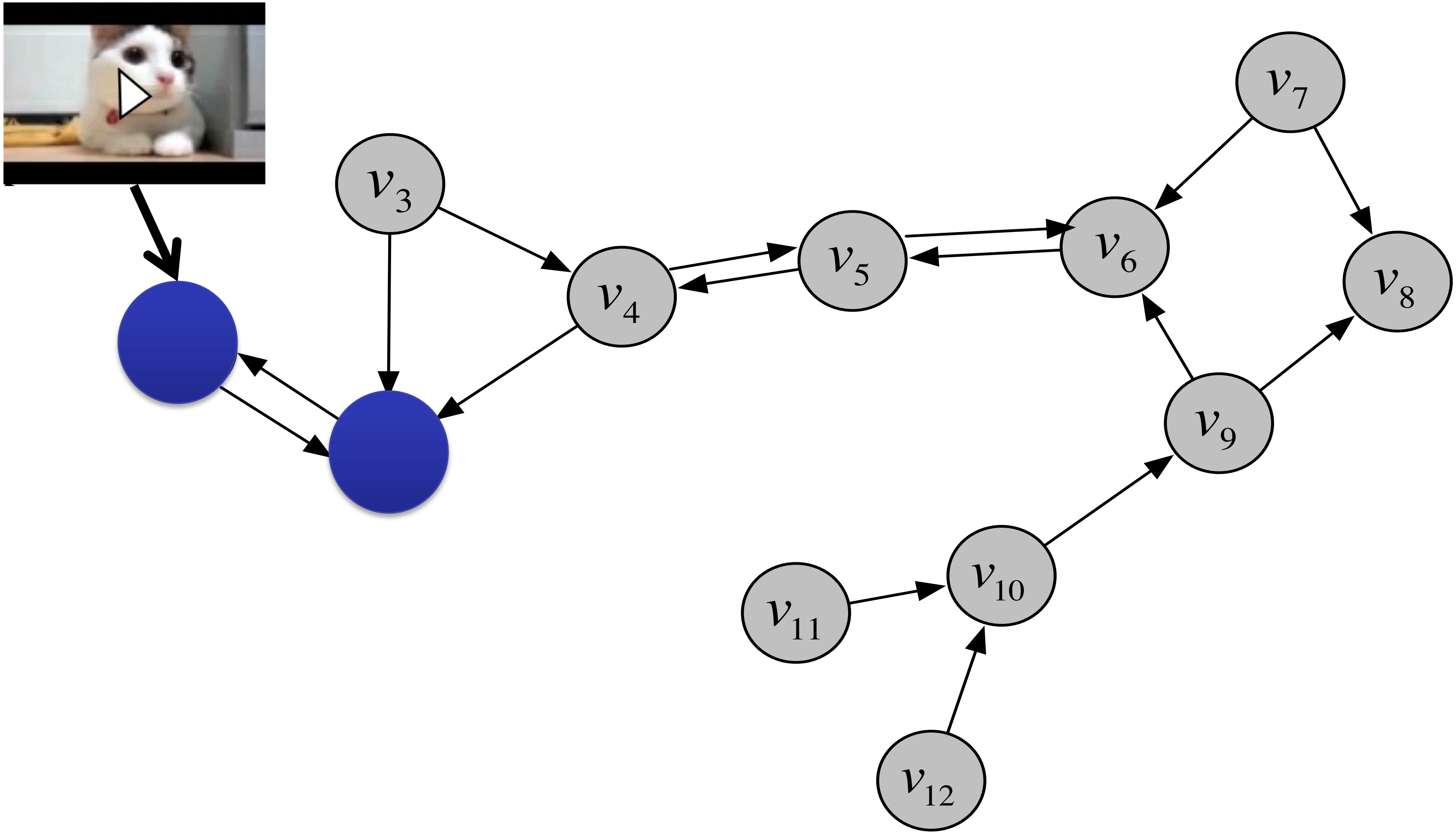
- In practice, the diffusion process is very complex.
  - The diffusion influences the structure of the network and user's position in the network in turn affects the influence they may have on other users
- Social role vs. information diffusion
  - Study on Twitter reveals that ***50%*** of Twitter contents are produced by less than ***1%*** of users who act as ***opinion leaders***[1]
  - Another study reveals that ***25%*** of information diffusion in Twitter is controlled by ***1%*** users serving as ***structural hole spanners***[2]

[1] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In **WWW'11**, pages 705–714, 2011.
[2] T. Lou and J. Tang. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. In **WWW'13**, pages 837-848, 2013.

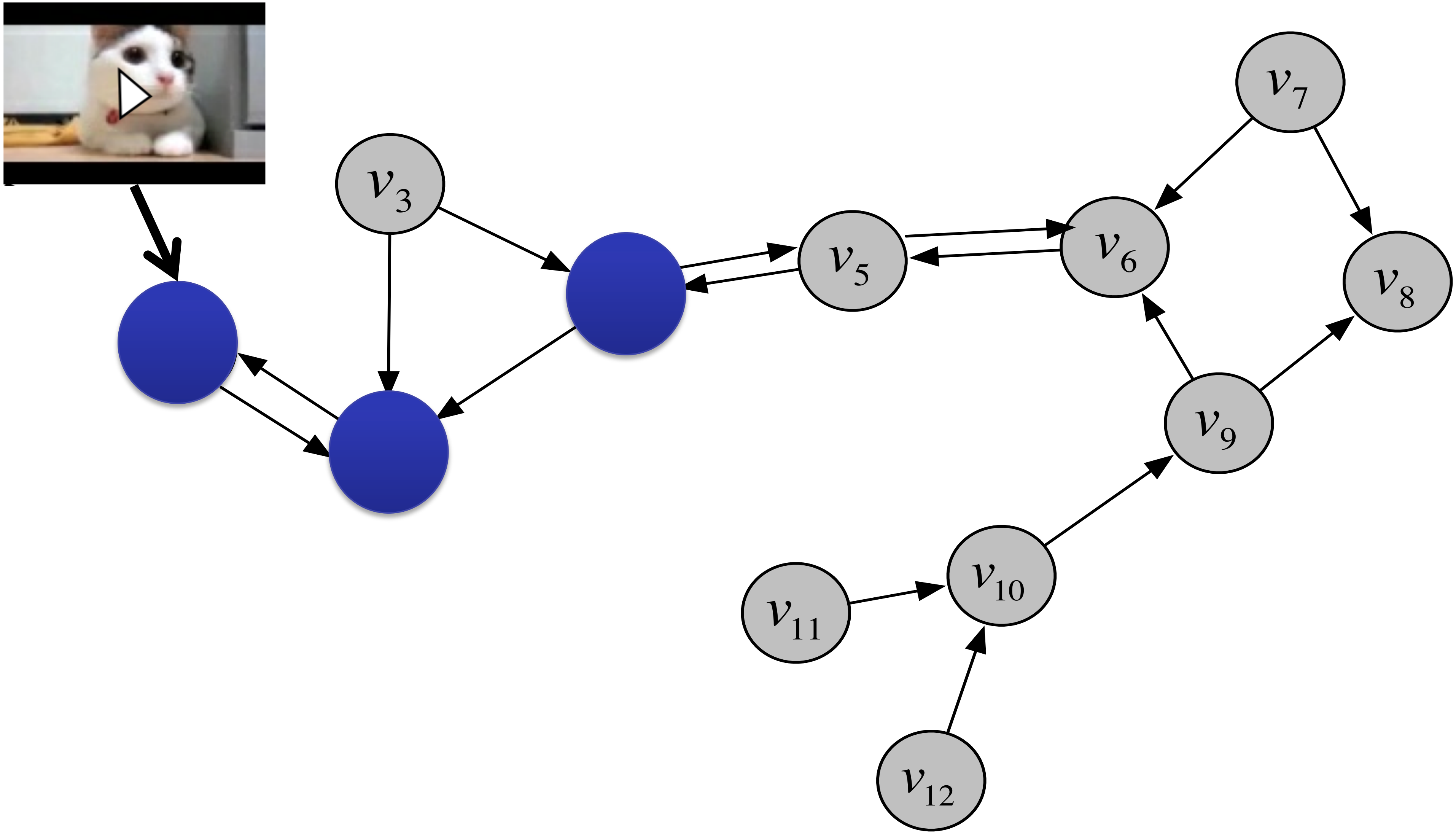
Information Diffusion Example



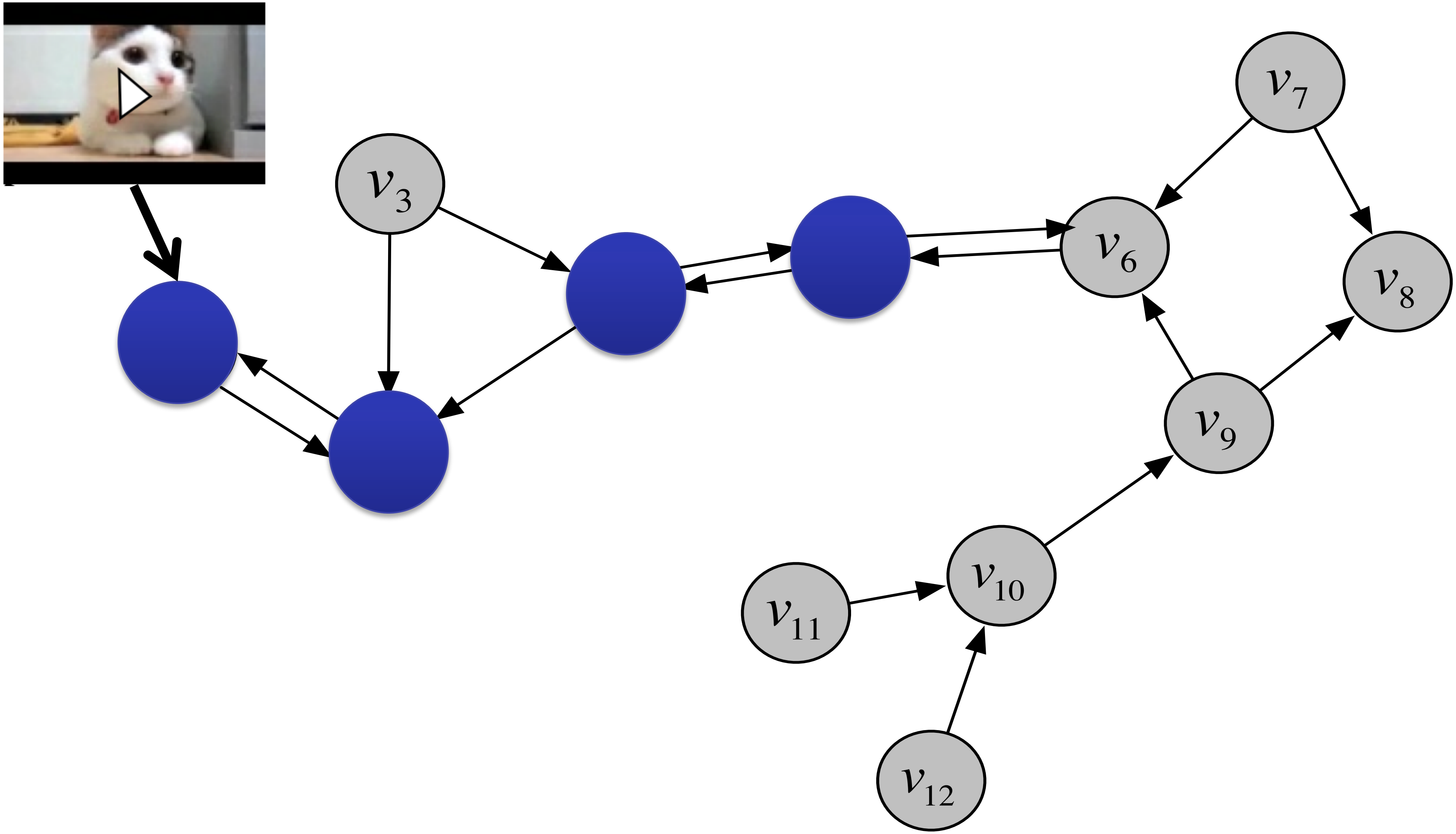
Information Diffusion Example



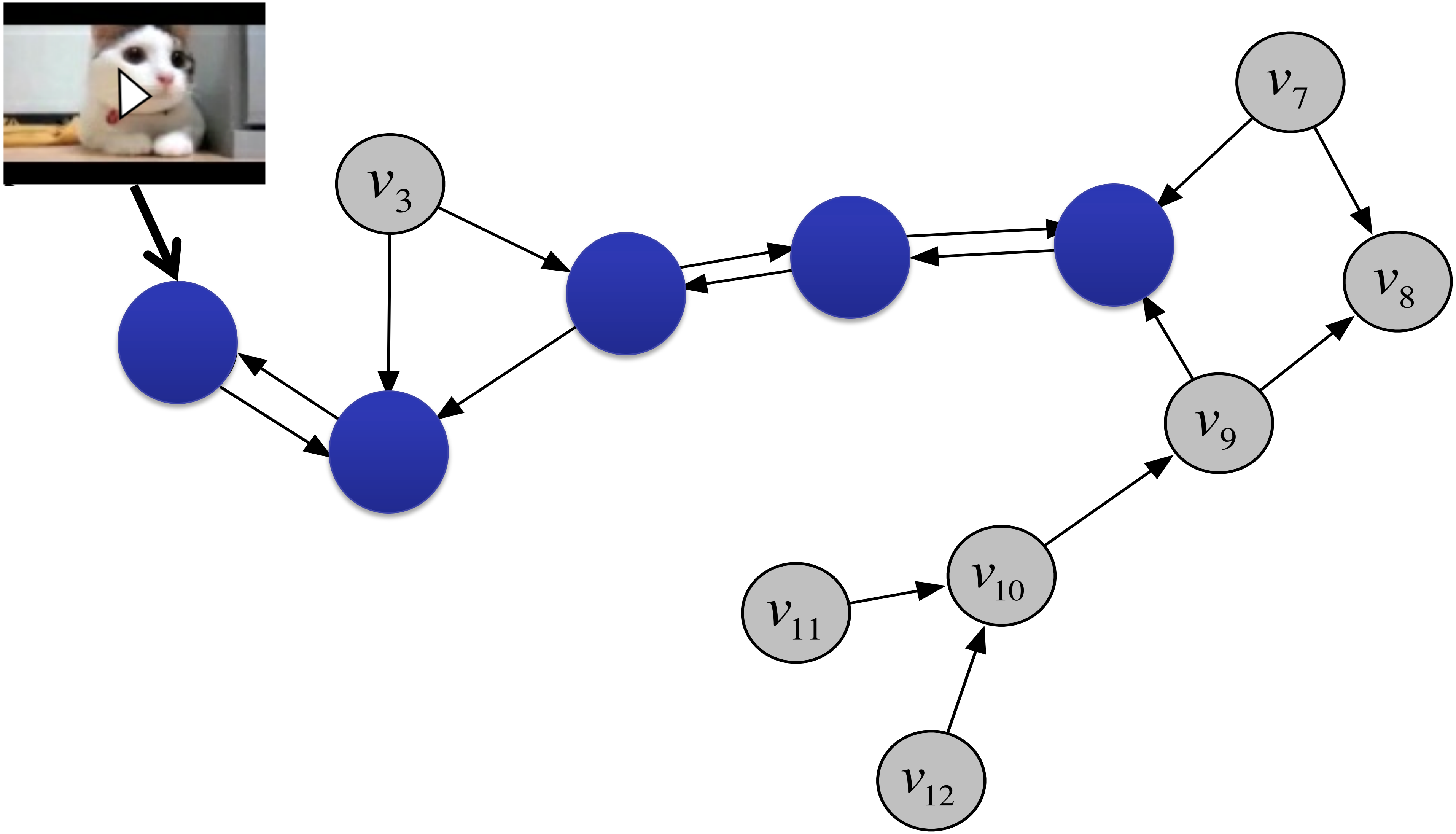
Information Diffusion Example



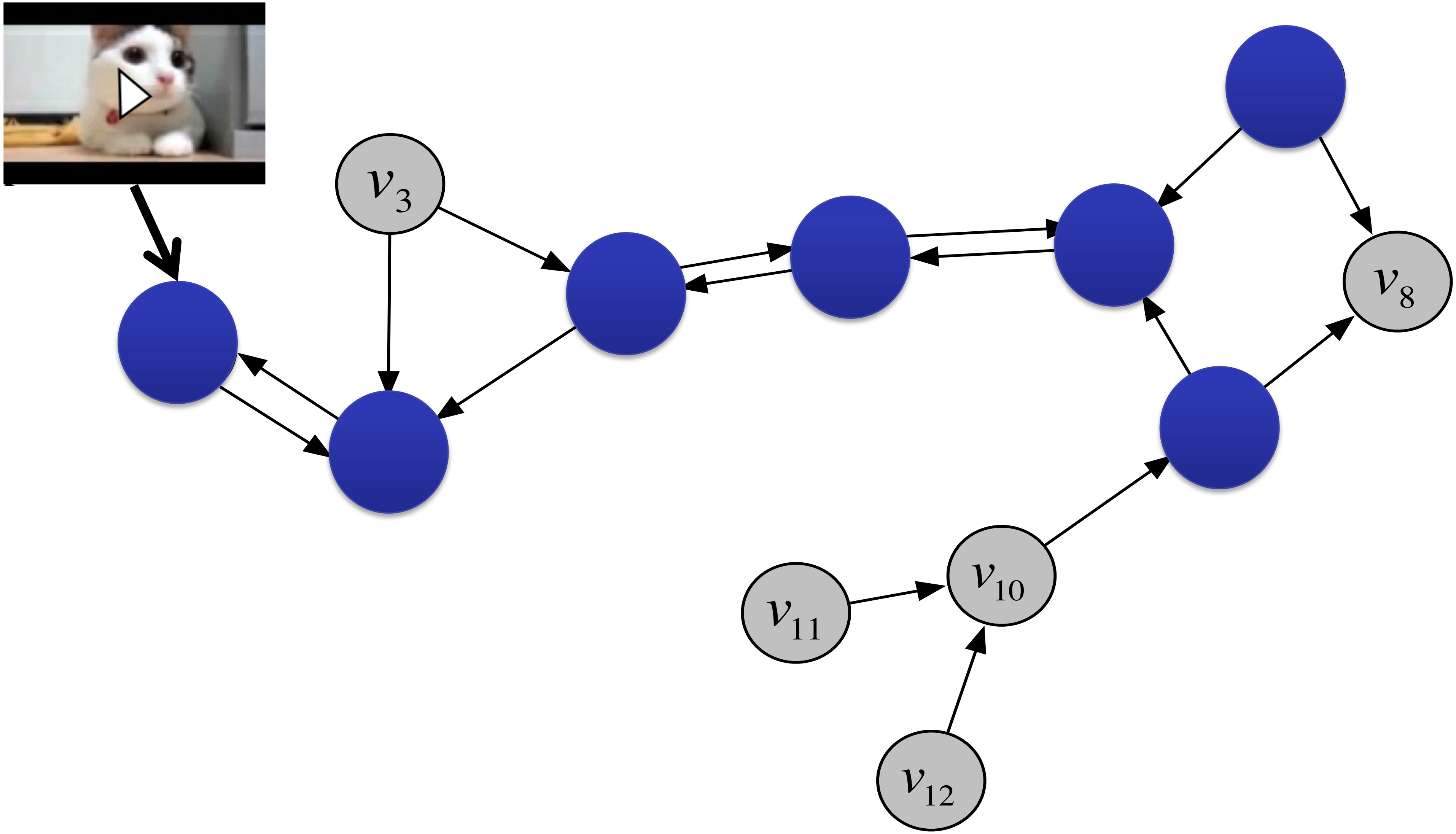
Information Diffusion Example



Information Diffusion Example

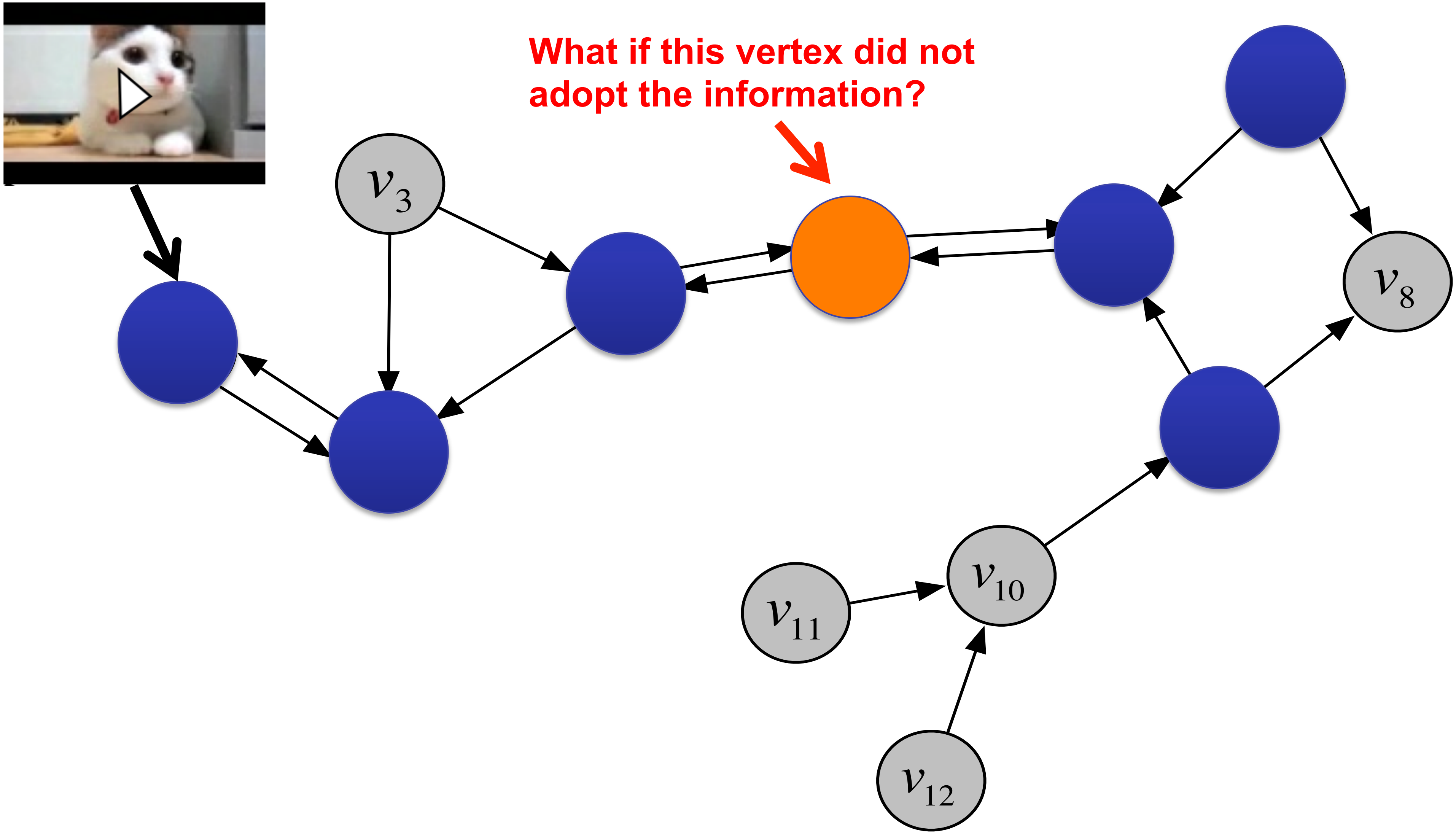


Information Diffusion Example

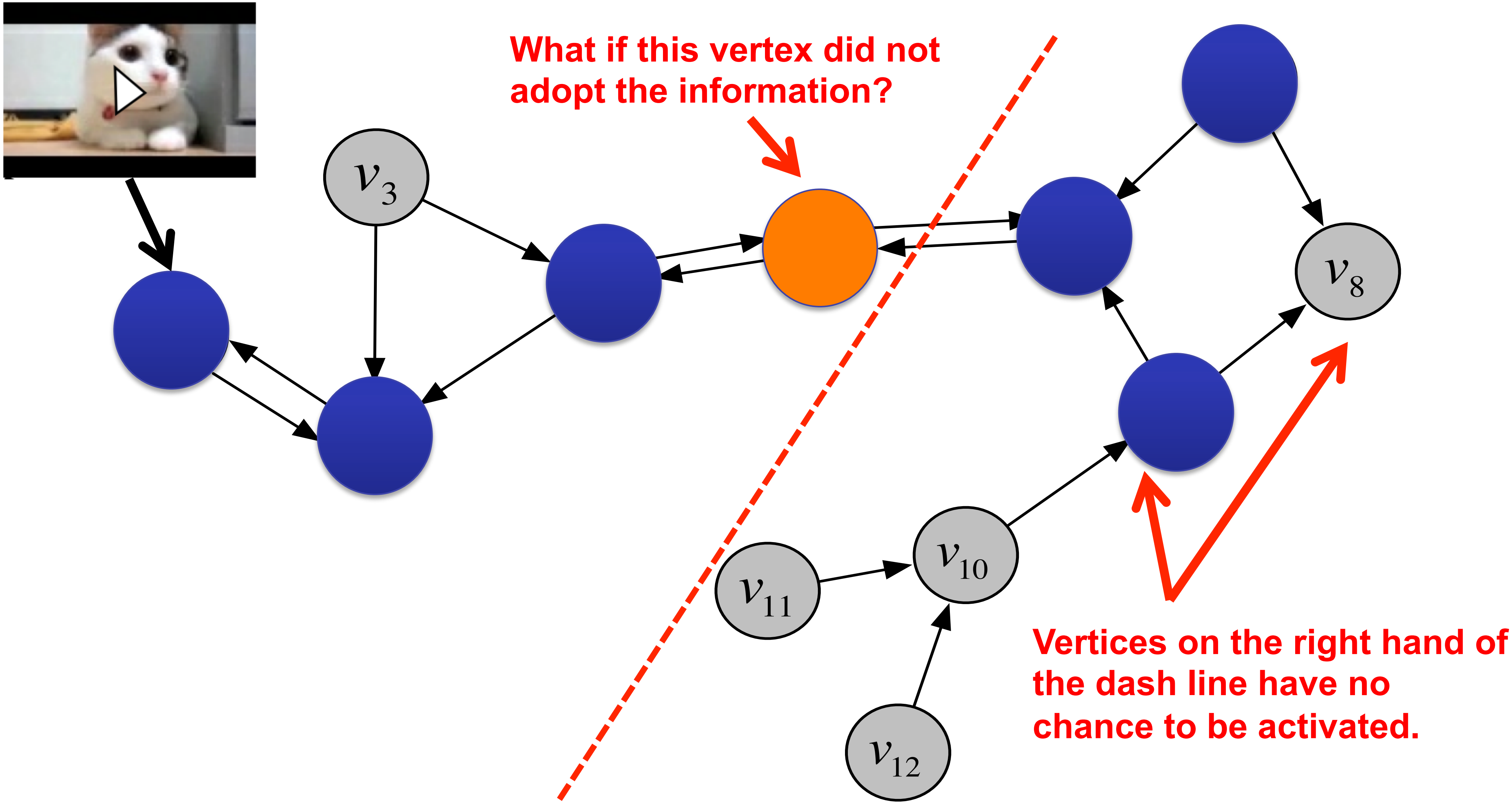


Role-aware: Information Diffusion Example

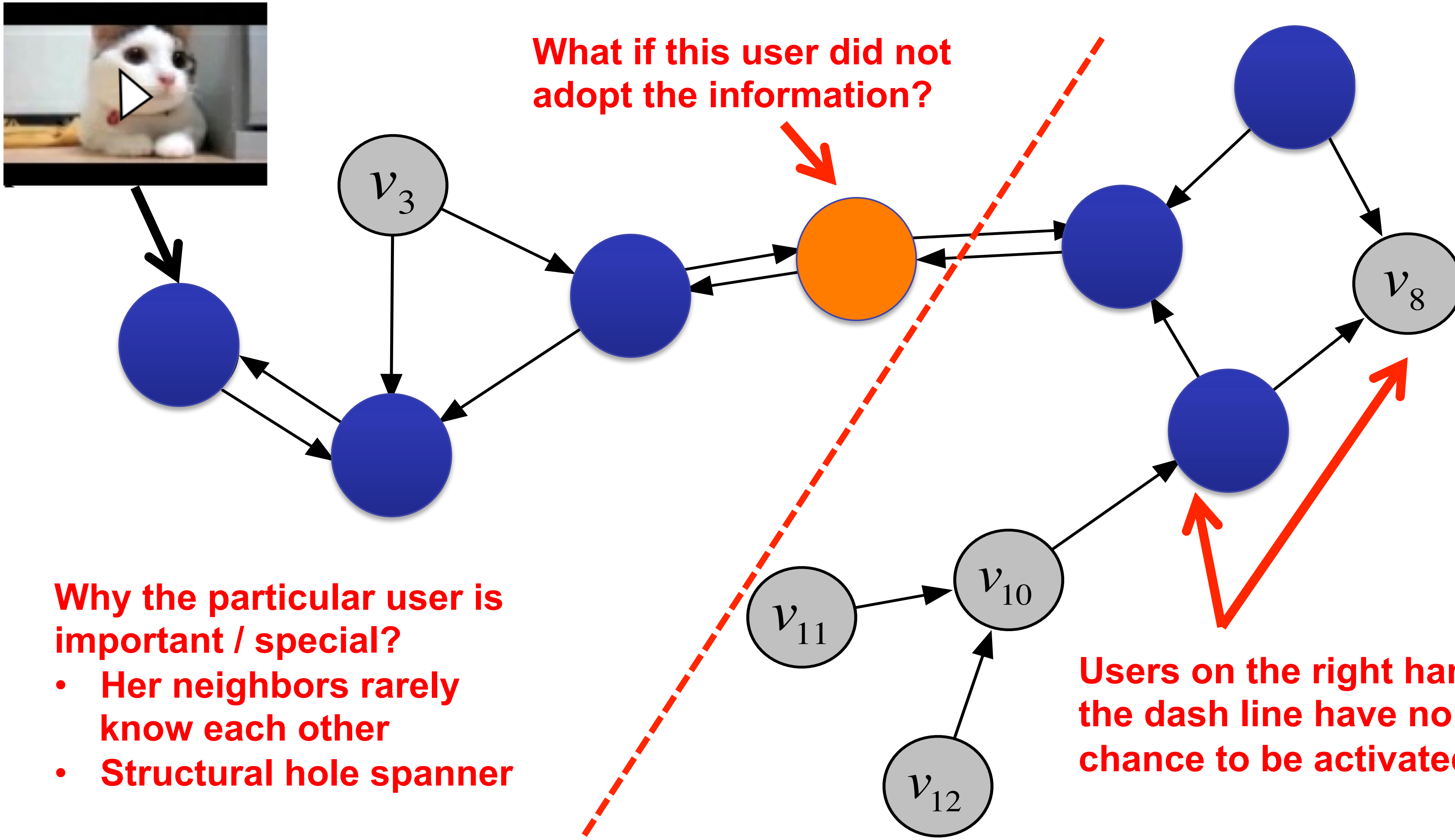
What if this vertex did not adopt the information?



Role-aware: Information Diffusion Example



Role-aware: Information Diffusion Example



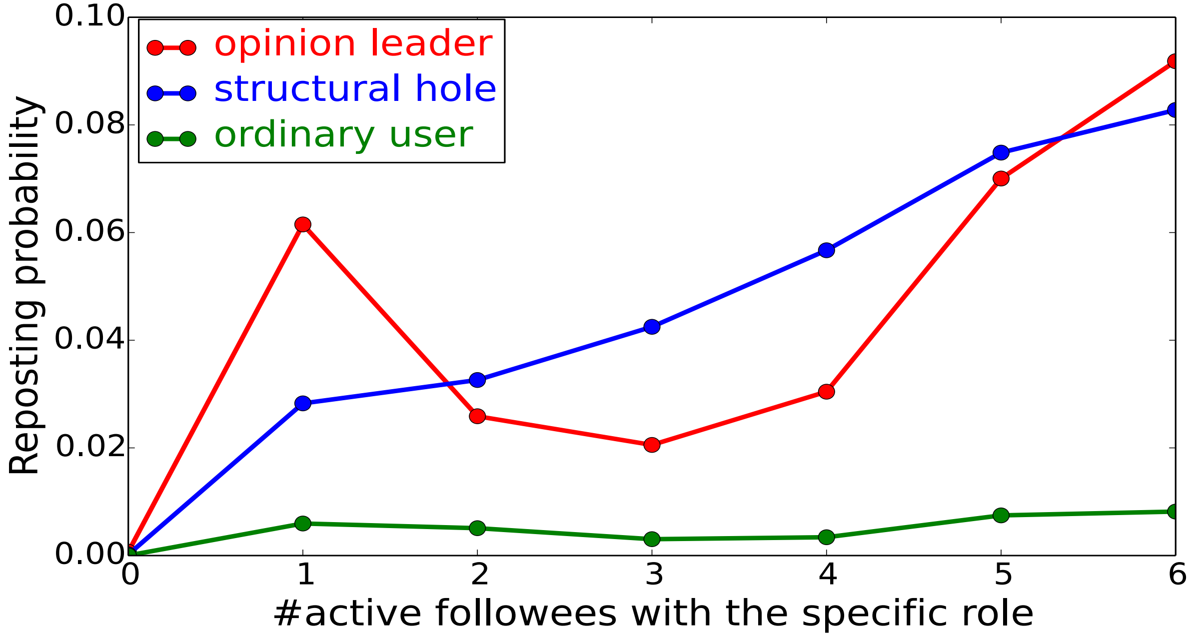
What if this user did not adopt the information?

Why the particular user is important / special?

- **Her neighbors rarely know each other**
- **Structural hole spanner**

Users on the right hand of the dash line have no chance to be activated.

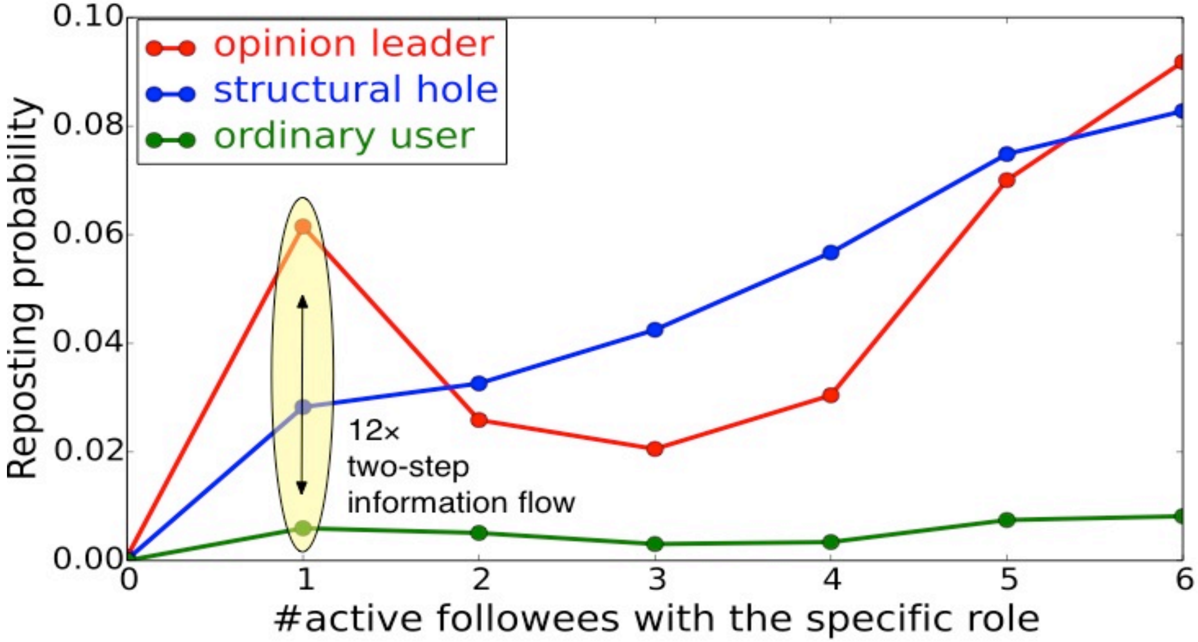
Preliminary Results on Weibo



X: number of v's active followees with different social roles.

Y: the probability of v being activated.

Preliminary Results on Weibo (2)

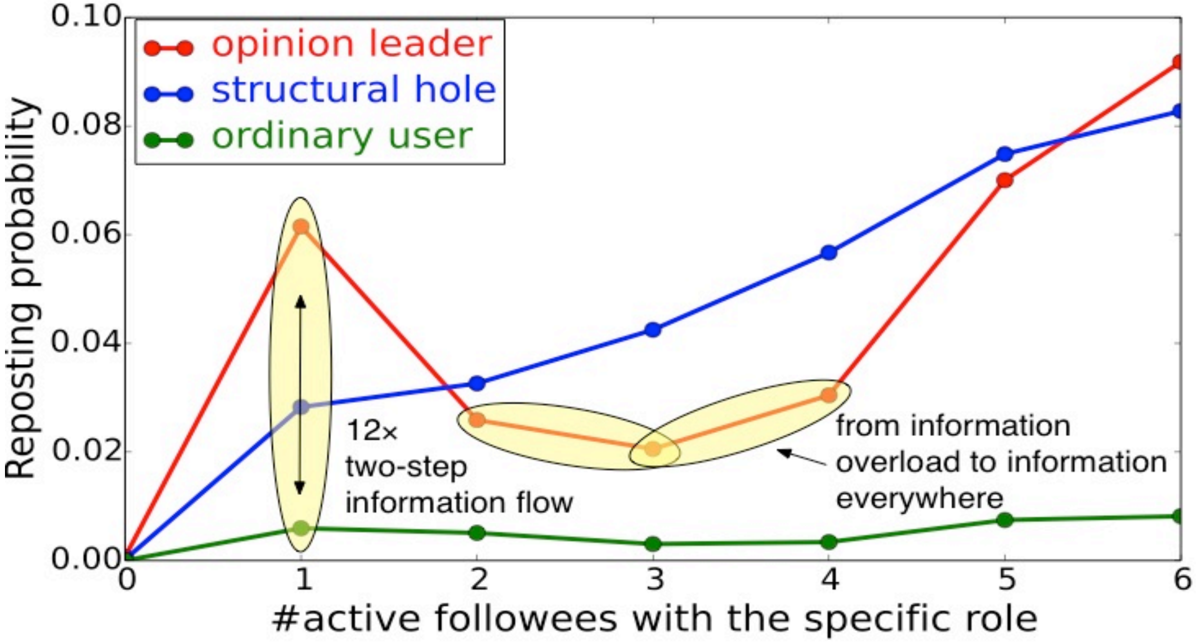


X: number of v's active followees with different social roles.

Y: the probability of v being activated.

[1] Lazarsfeld, P. F.; Berelson, B.; and Gaudet, H. 1944. The peoples choice: How the voter makes up his mind in a presidential election. New York: Duell, Sloan and Pearce .

Preliminary Results on Weibo (3)



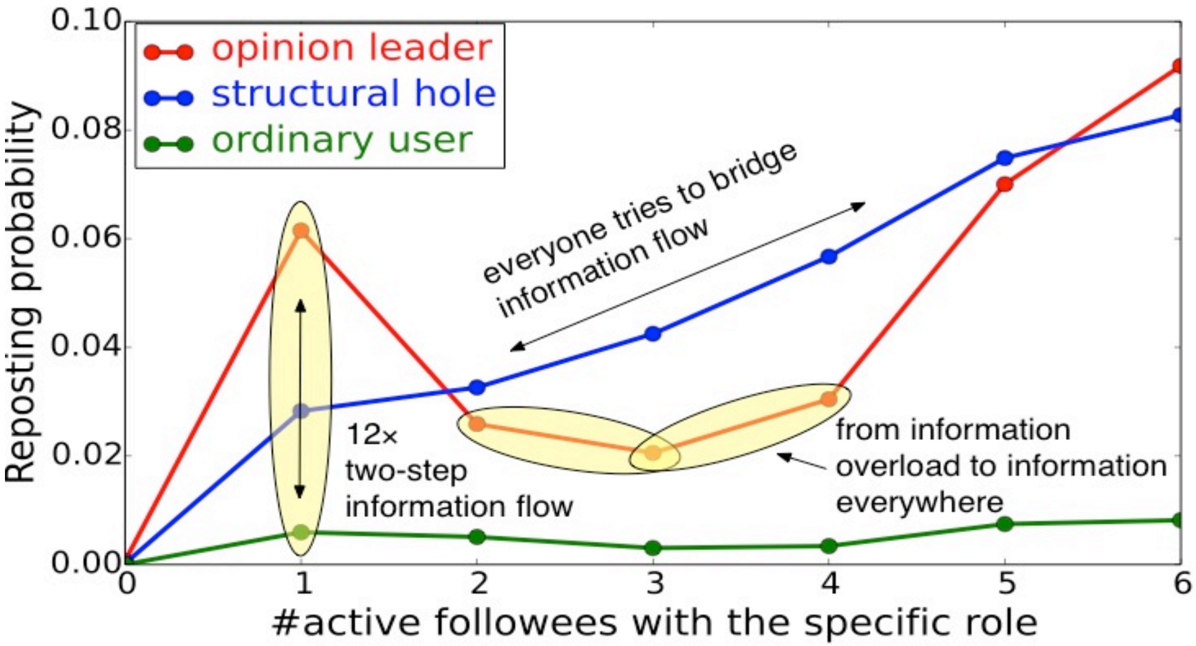
X: number of v's active followees with different social roles.

Y: the probability of v being activated.

- Information overload: 2-3 opinion leaders are sufficient to spread a piece of information throughout a community
- Information everywhere: spreading the information becomes a social norm to adopt

[2] Burt, R. S. 2001. Structural holes versus network closure as social capital. Social capital: Theory and research 31–56.
[3] Burt, R. S. 2009. Structural holes: The social structure of competition . Harvard University Press.

Preliminary Results on Weibo (4)



X: number of v's active followees with different social roles.

Y: the probability of v being activated.

- Structural hole spanners tend to bring information that a certain community is rarely exposed to.

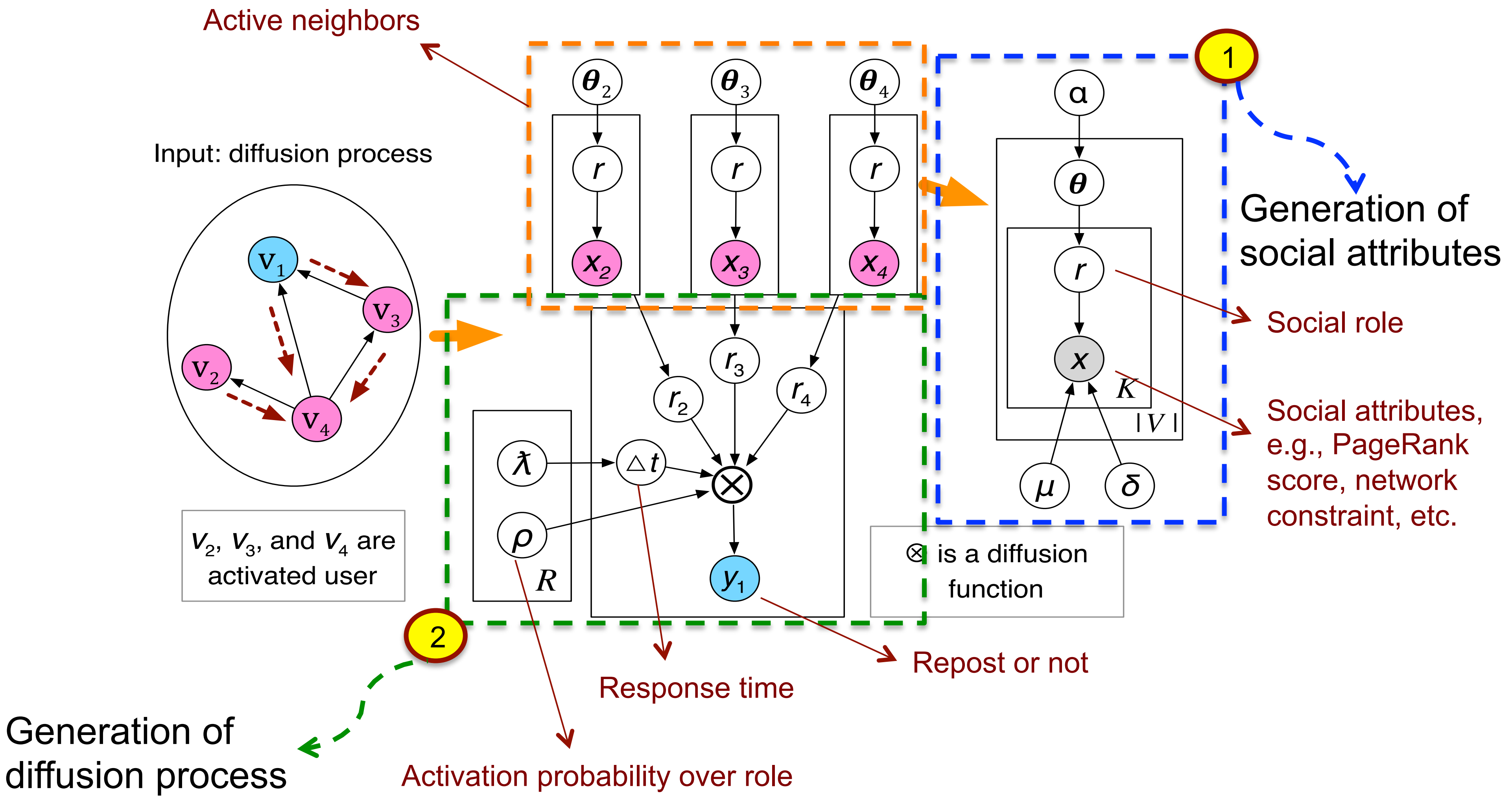
Problem Formulation

- Input:
  - Social Network – which users are connected
  - Diffusion Tree – which comprises a set of 4-tuples: {(u,v,i,t)} indicating user v re-tweet the message i from u at time t
- Output:
  - Predict the diffusion tree in future
  - The social role distribution of each user

Definition 2. **Social Role Distribution.** The social role distribution of a user $v \in V$ is denoted by $\theta_{\mathbf{v}}$, which is a R-dimensional vector and satisfies $\sum_r \theta_{vr} = 1$. θ_{vr} is the probability that v plays the role r when diffusing a certain message.

[1] Y. Yang, J. Tang, C. W.-K. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. RAIN: Social Role-Aware Information Diffusion. In **AAAI'15**.

RAIN: social Role-Aware INformation diffusion



[1] Y. Yang, J. Tang, C. W.-K. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. RAIN: Social Role-Aware Information Diffusion. In **AAAI'15**.

Modeling Diffusion Process

- The probability that the user u will succeed in activating one of her followers v at time t

$$\varphi_{iuv}^{t} = P(\boxed{z_{iuv}^{t} = 1}) = \sum_{r} \boxed{\rho_r}\boxed{\lambda_r}(1 - \boxed{\lambda_r})^{t-t_{iu}-1}\boxed{\theta_{ur}}.$$

A latent variable indicate u activates v at time t successfully

Modeling the response time (diffusion delay)

Social role distribution

Activation probability over role r

Modeling Diffusion Process

- The probability that user v is not activated by user u within the time period [t_{iu}+1, t]

$$
\begin{aligned}
\varepsilon_{iuv}^{t} &= \boxed{P(z_{iuv}^{t} = 0)} \\
&= \sum_{r} \theta_{ur}(1 - \rho_r[\sum_{t'=t_{iu}+1}^{t} \lambda_r(1-\lambda_r)^{t'-t_{iu}-1}]) \\
&= \sum_{r} \theta_{ur}[\rho_r(1-\lambda_r)^{t-t_{iu}} + 1 - \rho_r].
\end{aligned}
$$

A latent variable indicate u fails to activates v within time period [t_{iu}+1,t]

Modeling Diffusion Process

- The probability user v is active at time t

$$P(v \in A_{it}) = \boxed{\sum_{\mathbf{z}^t_{i*v}} P(\mathbf{z}_{i*v})} \boxed{\prod_{u \in B(v) \cap D_{it-1}} P(z^t_{iuv} = 0)}$$
$$= \prod_{u \in B(v) \cap D_{it-1}} (\varphi^t_{iuv} + \varepsilon^t_{iuv}) - \prod_{u \in B(v) \cap D_{it-1}} \varepsilon^t_{iuv}.$$

All adoption results

All users fails to activate user v

Modeling Diffusion Process

- The probability that user v is never activated by the last timestamp T

$$P(v \notin D_{iT}) = \prod_{u \in B(v) \cap D_{iT}} \sum_{r} (1 - \rho_r)\theta_{ur}.$$

Assumption here:
T >> the last observed timestamp

Modeling Social Attributes

- We assume each attribute of a user u is sampled according to a Gaussian distribution w.r.t. the social role of u

$$P(x_{uk}) = \sum_{r} \sqrt{\frac{\boxed{\delta_{rk}}}{2\pi}} \exp\{-\frac{\boxed{\delta_{rk}}(x_{uk} - \boxed{\mu_{rk}})^2}{2}\}\theta_{ur}.$$

Gaussian parameters over role

Modeling Learning with Gibbs Sampling

- Initialize the proposed model to default parameter settings
- Sample latent variable r for each social attribute of a user u according to

$$P(r_{uk}|\mathbf{r}_{\neg uk},\mathbf{x}) = \frac{P(\mathbf{x},\mathbf{r})}{P(\mathbf{x}_{\neg uk},\mathbf{r}_{\neg uk})} = \frac{n_{ur_{uk}}^{\neg uk}+\alpha}{\sum_r (n_{ur}^{\neg uk}+\alpha)} \frac{\Gamma(\tau_2+\frac{n_{r_{uk}k}}{2})}{\Gamma(\tau_2+\frac{n_{r_{uk}k}^{\neg uk}}{2})} \times \frac{\sqrt{(\tau_1+n_{r_{uk}k}^{\neg uk})}\eta(n_{r_{uk}k}^{\neg uk},\bar{x}_{r_{uk}k}^{\neg uk},s_{r_{uk}}^{\neg uk})}{\sqrt{(\tau_1+n_{r_{uk}k})}\eta(n_{r_{uk}k},\bar{x}_{r_{uk}k},s_{r_{uk}})},$$

- Sample r, \delta t, and z for each diffusion tree node according to

$$\begin{aligned}
&P(r_{iuv},\Delta t_{iuv},z_{iuv}|\mathbf{r}_{\neg iuv},\Delta\mathbf{t}_{\neg iuv},\mathbf{z}_{\neg iuv},\mathbf{y}) \\
&= \frac{P(\mathbf{r},\Delta\mathbf{t},\mathbf{z},\mathbf{y})}{P(\mathbf{r}_{\neg iuv},\Delta\mathbf{t}_{\neg iuv},\mathbf{z}_{\neg iuv},\mathbf{y}_{\neg iuv})} \\
&= \frac{n_{ur_{iuv}}^{\neg iuv}+\alpha}{\sum_r (n_{ur}^{\neg iuv}+\alpha)} \times \frac{n_{z_{iuv}r_{iuv}}^{\neg iuv}+\beta_1^{z_{iuv}}\beta_0^{1-z_{iuv}}}{n_{1r_{iuv}}^{\neg iuv}+\beta_1+n_{0r_{iuv}}^{\neg iuv}+\beta_0} \\
&\times \frac{(n_{r_{iuv}}^{\neg iuv}+\gamma_1)\prod_{t=0}^{\Delta t-2}(s_{r_{iuv}}^{\neg iuv}-n_{r_{iuv}}^{\neg iuv}+\gamma_0+t)}{\prod_{t=0}^{\Delta t-1}(\gamma_1+s_{r_{iuv}}^{\neg iuv}+\gamma_0+t)} \times \Phi,
\end{aligned}$$

Gibbs Sampling (cont.)

- Update parameters

$$\theta_{ur} = P(\tilde{r} = r | \mathbf{r}, \Delta\mathbf{t}, \mathbf{z}, \mathbf{y}) = \frac{n_{ur} + \alpha}{\sum_r (n_{ur} + \alpha)}$$
$$\lambda_r = P(\Delta\tilde{t} = 1 | \tilde{r} = r, \mathbf{r}, \Delta\mathbf{t}, \mathbf{z}, \mathbf{y}) = \frac{n_r + \gamma_1}{\gamma_1 + s_r + \gamma_0}$$
$$\rho_r = P(\tilde{z} = 1 | \tilde{r} = r, \mathbf{r}, \Delta\mathbf{t}, \mathbf{z}, \mathbf{y}) = \frac{n_{1r} + \beta_1}{n_{1r} + \beta_1 + n_{0r} + \beta_0}$$

- Approximate Gaussian parameters by their expectations

$$\mu_{rk} \approx E(\mu_{rk}) = \frac{\tau_0 \tau_1 + n_{rk}\bar{x}_{rk}}{\tau_1 + n_{rk}},$$
$$\delta_{rk} \approx E(\delta_{rk}) = \frac{2\tau_2 + n_{rk}}{2\tau_3 + n_{rk}s_{rk} + \frac{\tau_1 n_{rk}(\bar{x}_{rk} - \tau_0)^2}{\tau_1 + n_{rk}}}.$$

[1] Y. Yang, J. Tang, C. W.-K. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. RAIN: Social Role-Aware Information Diffusion. In **AAAI'15**.

Dataset

- We employ a dataset from Tencent Weibo, which consists of ***4,588,559*** original posts, and ***184,491*** relevant users
  - We remove original posts reposted **< *5*** times which remains ***242,831*** original posts
  - We use data on Nov. 1 to train the model and Nov. 2 to test
- We categorize the posts based on their topics extracted by LDA and labeled manually: *campus, constellation, movie, history, society, health, political and travel.*

Micro-level Prediction

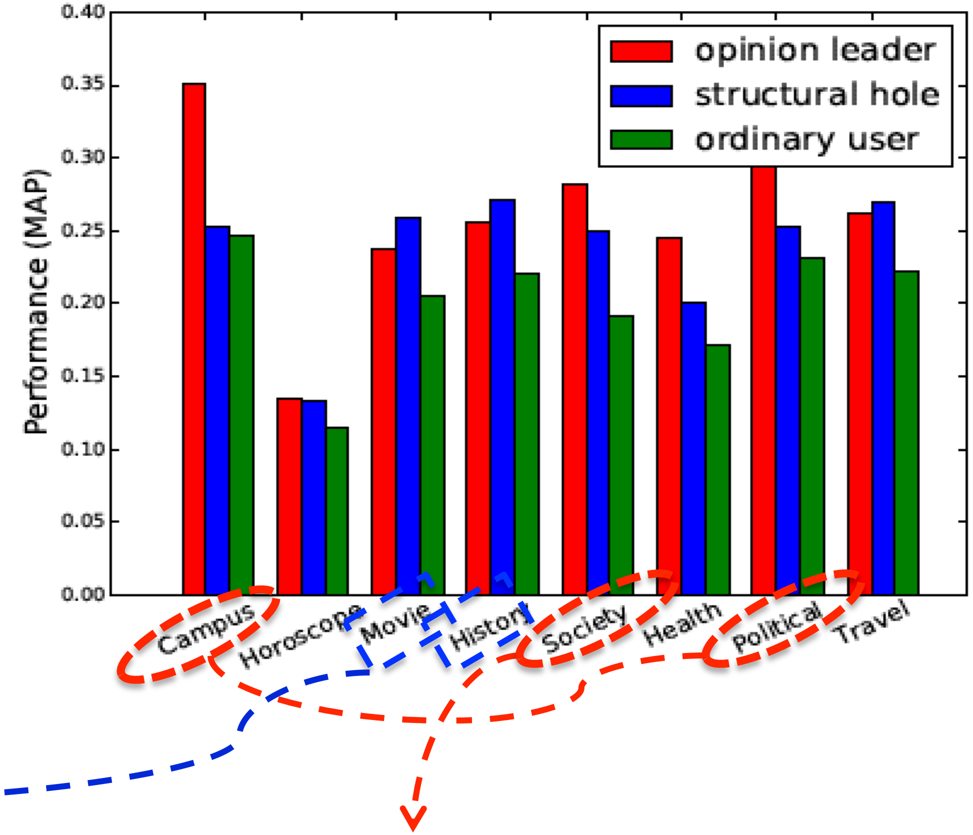
- Predict whether a user will repost a given message.
- Count
  - ranks users by the number of active followees
  - performs worst due to the lack of supervised information
- SVM
  - employs three features to train a classifier
    - #active followers
    - #active followees
    - #whether the user have reposted any similar messages before
  - neglects the diffusion mechanism
- IC Model
  - traditional IC model with fitted parameters
  - suffers from data sparseness and model complexity
- RAIN
  - improves the performance +32.6% in terms of MAP

Table 2: Performance of repost prediction on several topics.

Topic	Method	P@10	P@50	P@100	MAP
Campus	Count	0.028	0.010	0.006	0.068
	SVM	0.098	0.045	0.032	0.127
	IC Model	**0.231**	0.142	0.102	0.259
	RAIN	0.228	**0.145**	**0.106**	**0.263**
Horoscope	Count	0.019	0.010	0.006	0.005
	SVM	0.124	**0.162**	0.088	**0.263**
	IC Model	0.149	0.111	0.098	0.125
	RAIN	**0.171**	0.121	**0.102**	0.130
Movie	Count	0.015	0.007	0.004	0.009
	SVM	0.094	0.111	0.060	0.199
	IC Model	0.227	0.147	**0.147**	0.236
	RAIN	**0.229**	**0.173**	0.144	**0.238**
History	Count	0.191	0.056	0.033	0.096
	SVM	0.154	0.051	0.030	0.221
	IC Model	0.206	0.134	**0.135**	0.230
	RAIN	**0.225**	**0.171**	0.134	**0.262**
Society	Count	0.245	0.058	0.029	0.156
	SVM	0.100	0.023	0.012	0.122
	IC Model	0.171	0.131	**0.109**	0.198
	RAIN	**0.176**	**0.140**	0.106	**0.204**
Health	Count	0.041	0.008	0.005	0.035
	SVM	0.164	0.064	0.039	**0.197**
	IC Model	0.169	0.113	0.096	0.162
	RAIN	**0.175**	**0.134**	**0.115**	0.185
Political	Count	0.019	0.005	0.003	0.007
	SVM	0.104	0.077	0.039	0.176
	IC Model	0.209	0.132	0.102	0.224
	RAIN	**0.216**	**0.164**	**0.130**	**0.239**
Travel	Count	0.142	0.056	0.031	0.103
	SVM	0.094	0.048	0.032	0.128
	IC Model	**0.206**	0.120	0.098	0.254
	RAIN	0.194	**0.159**	**0.126**	**0.260**

Social Role Analysis

RAIN can better predict opinion leaders and structural hole spanners, as ordinary users tend to behave more randomly



Structural hole spanners can be better predicted on more general topics, which tend to propagate from one community to another

Opinion leaders can be better predicted on more regional and specialized topics

Macro-level Prediction

- We predict the ***scale*** of a diffusion process
  - X-axis: the number of reposts
  - Y-axis: the proportion of original posts with particular number of reposts

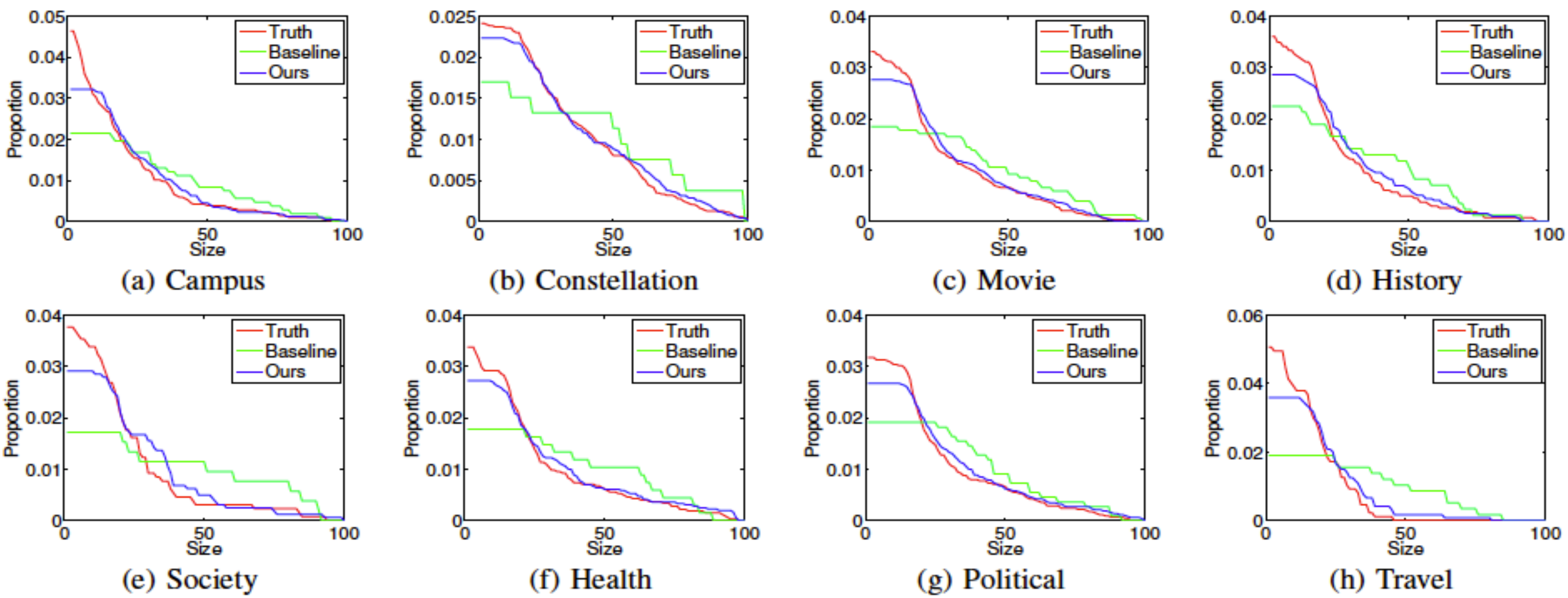


Figure 8: Diffusion scale distributions of the different topics in the test set.

Macro-level Prediction

- We predict the ***duration*** of a diffusion process
  - X-axis: the time interval between the first and last posts
  - Y-axis: the proportion of original posts with particular time interval

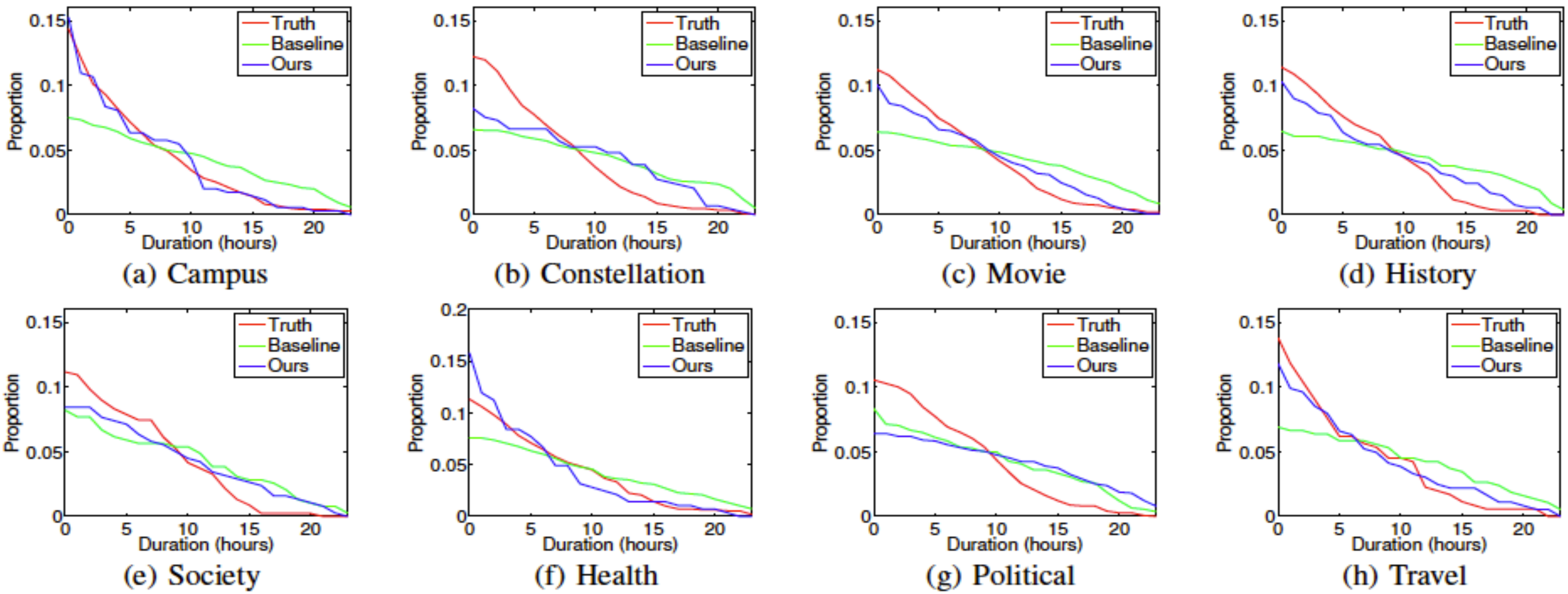


Figure 9: Diffusion duration distributions of the different topics in the test set.

Summary

- Big social data provides unprecedented opportunities to study interactions between users
- Social Influence
  - Learning social influence
  - Influence maximization
- Information Diffusion
  - Linear threshold (LT)
  - Independent cascaded (IC)
  - Role-aware diffusion (RAIN)

Related Publications

- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In **KDD'09**, pages 807-816, 2009.
- Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social action tracking via noise tolerant time-varying factor graphs. In **KDD'10**, pages 807–816, 2010.
- Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In **KDD'11**, pages 1397–1405, 2011.
- Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity Influence in Large Social Networks. In **KDD'13**, pages 347-355, 2013.
- Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla. Inferring User Demographics and Social Strategies in Mobile Social Networks. In **KDD'14**, 2014.
- Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. Social Influence Locality for Modeling Retweeting Behaviors. In **IJCAI'13**, pages 2761-2767, 2013.
- Jing Zhang, Jie Tang, Honglei Zhuang, Cane Wing-Ki Leung, and Juanzi Li. Role-aware Conformity Influence Modeling and Analysis in Social Networks. In **AAAI'14**, 2014.
- Yang Yang, Jie Tang, Cane Wing-Ki Leung, Yizhou Sun, Qicong Chen, Juanzi Li, and Qiang Yang. RAIN: Social Role-Aware Information Diffusion. In **AAAI'15**, 2015.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In **KDD'08**, pages 990-998, 2008.
- Tiancheng Lou and Jie Tang. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. In **WWW'13**, pages 837-848, 2013.
- Lu Liu, Jie Tang, Jiawei Han, and Shiqiang Yang. Learning Influence from Heterogeneous Social Networks. In **DMKD**, 2012, Volume 25, Issue 3, pages 511-544.
- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. In **TKDD**, Vol 7(2), 2013.
- Jimeng Sun and Jie Tang. A Survey of Models and Algorithms for Social Influence Analysis. Social Network Data Analytics, Aggarwal, C. C. (Ed.), Kluwer Academic Publishers, pages 177–214, 2011.

References

- S. Milgram. The Small World Problem. **Psychology Today**, 1967, Vol. 2, 60–67
- J.H. Fowler and N.A. Christakis. The Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study. **British Medical Journal** 2008; 337: a2338
- R. Dunbar. Neocortex size as a constraint on group size in primates. **Human Evolution**, 1992, 20: 469–493.
- R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. **Nature**, 489:295-298, 2012.
- http://klout.com
- Why I Deleted My Klout Profile, by Pam Moore, at **Social Media Today**, originally published November 19, 2011; retrieved November 26 2011
- S. Aral and D Walker. Identifying Influential and Susceptible Members of Social Networks. **Science**, 337:337-341, 2012.
- J. Ugandera, L. Backstromb, C. Marlowb, and J. Kleinberg. Structural diversity in social contagion. **PNAS**, 109 (20):7591-7592, 2012.
- S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. **PNAS**, 106 (51):21544-21549, 2009.
- J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In **KDD'09**, pages 747–756, 2009.
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. **Journal of Educational Psychology** 66, 5, 688–701.
- http://en.wikipedia.org/wiki/Randomized_experiment

References(cont.)

- A. Anagnostopoulos, R. Kumar, M. Mahdian. Influence and correlation in social networks. In **KDD’08**, pages 7-15, 2008.
- L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.
- G. Jeh and J. Widom. Scaling personalized web search. In **WWW '03**, pages 271-279, 2003.
- G. Jeh and J. Widom, SimRank: a measure of structural-context similarity. In **KDD’02**, pages 538-543, 2002.
- A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In **WSDM’10**, pages 207–217, 2010.
- P. Domingos and M. Richardson. Mining the network value of customers. In **KDD’01**, pages 57–66, 2001.
- D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In **KDD’03**, pages 137–146, 2003.
- J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In **KDD’07**, pages 420–429, 2007.
- W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In **KDD'09**, pages 199-207, 2009.
- E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: evidence from field experiments. In **EC'12**, pages 146-161, 2012.
- A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In **CIKM’08**, pages 499–508, 2008.
- N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In **WSDM’08**, pages 207–217, 2008.

References(cont.)

- E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In **EC ’09**, pages 325–334, New York, NY, USA, 2009. ACM.
- P. Bonacich. Power and centrality: a family of measures. **American Journal of Sociology**, 92:1170–1182, 1987.
- R. B. Cialdini and N. J. Goldstein. Social influence: compliance and conformity. **Annu Rev Psychol**, 55:591–621, 2004.
- D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In **KDD’08**, pages 160–168, 2008.
- P. W. Eastwick and W. L. Gardner. Is it a game? evidence for social influence in the virtual world. **Social Influence**, 4(1):18–32, 2009.
- S. M. Elias and A. R. Pratkanis. Teaching social influence: Demonstrations and exercises from the discipline of social psychology. **Social Influence**, 1(2):147–162, 2006.
- T. L. Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In **WWW’10**, 2010.
- M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In **KDD’10**, pages 1019–1028, 2010.
- M. E. J. Newman. A measure of betweenness centrality based on random walks. **Social Networks**, 2005.
- D. J. Watts and S. H. Strogatz. Collective dynamics of ’small-world’ networks. **Nature**, pages 440–442, Jun 1998.
- J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In **ICDM’05**, pages 418–425, 2005.

Thank you！

Collaborators: John Hopcroft, Jon Kleinberg, Chenhao Tan (**Cornell**)

Jiawei Han and Chi Wang (**UIUC**)

Jimeng Sun (**IBM**) Tiancheng Lou (**Google**)

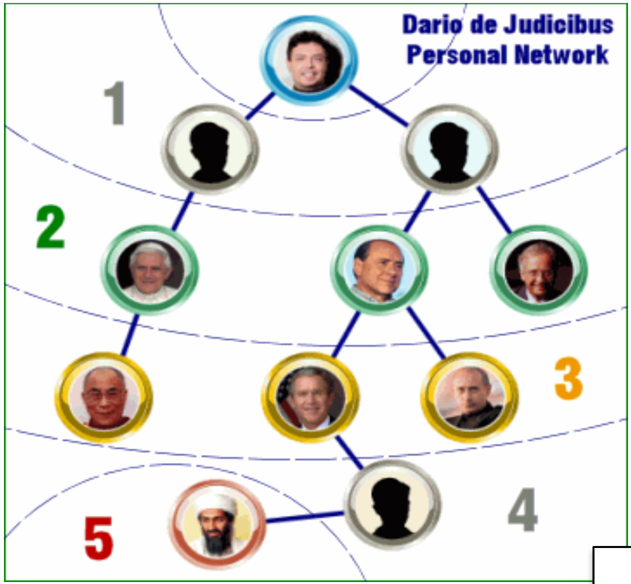
Wei Chen, Ming Zhou, Long Jiang (**Microsoft**)

Jing Zhang, Zhanpeng Fang, Zi Yang, Sen Wu, Jia Jia (**THU**)

Jie Tang, KEG, Tsinghua U,	http://keg.cs.tsinghua.edu.cn/jietang
Download all data & Codes,	http://arnetminer.org/download

The theory of “Three Degree of Influence”

Six degree of separation[1]



Three degree of Influence[2]

You are able to **influence** up to >1,000,000 persons in the world, according to the Dunbar’s number[3].

[1] S. Milgram. The Small World Problem. Psychology Today, 1967, Vol. 2, 60–67

[2] J.H. Fowler and N.A. Christakis. The Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study. British Medical Journal 2008; 337: a2338

[3] R. Dunbar. Neocortex size as a constraint on group size in primates. Human Evolution, 1992, 20: 469–493.