Models and Algorithms for Social Influence Analysis

Jimeng Sun and Jie Tang

IBM TJ Watson Research Center

Tsinghua University

Social Networks

- Facebook has over **900 million** users
- the **3rd** largest “Country” in the world
- More visitors than Google

- Twitter has over **25 billion** tweets per quarter

- Flickr has more than **5 billion** images

- Pinterest, with higher traffic than Twitter and Google

- Weibo has 2012, 300 million users, 300% yearly increase

- Tencent has over **700 million** users

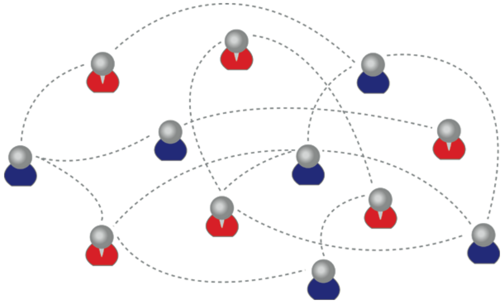
A Trillion Dollar Opportunity

Social networks already become a bridge to connect our daily **physical** life and the **virtual** web space

On2Off **[1]**

[1] Online to Offline is trillion dollar business
http://techcrunch.com/2010/08/07/why-online2offline-commerce-is-a-trillion-dollar-opportunity/

What is a social network?



- A social network is:
  - a **graph** made up of :
  - a set of **individuals**, called “nodes”, and
  - Connected by one or more **relationship**, such as friendship, called “edges”.

新浪微博

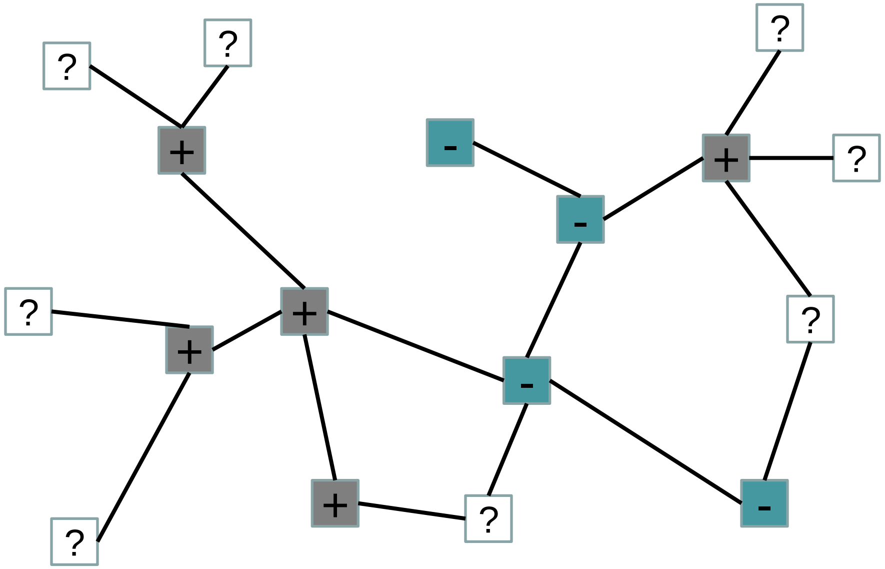
foursquare

facebook

twitter

History of the Web

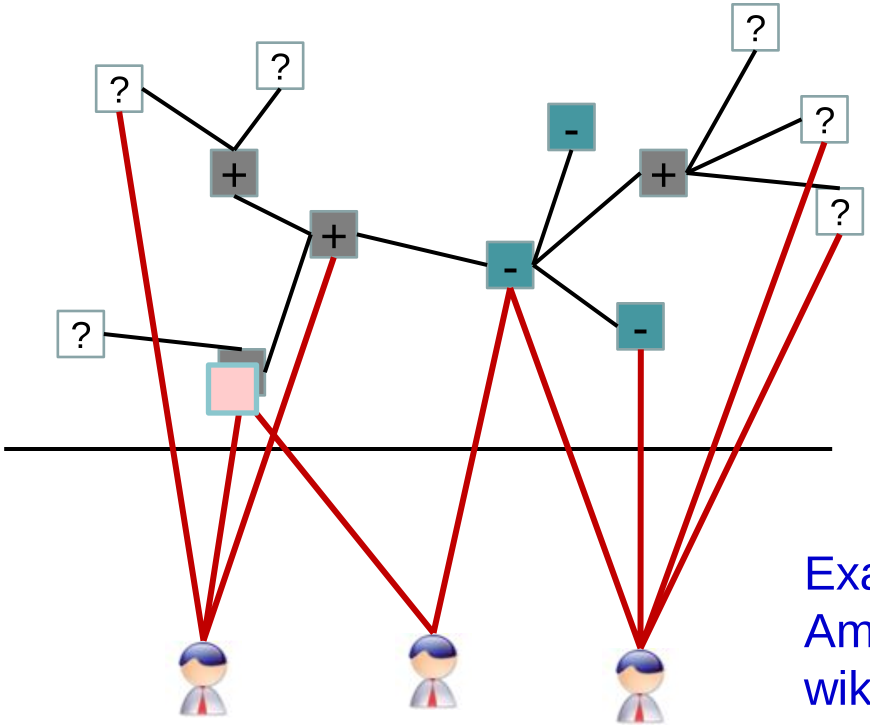
Web 1.0



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

History of the Web

Collaborative Web

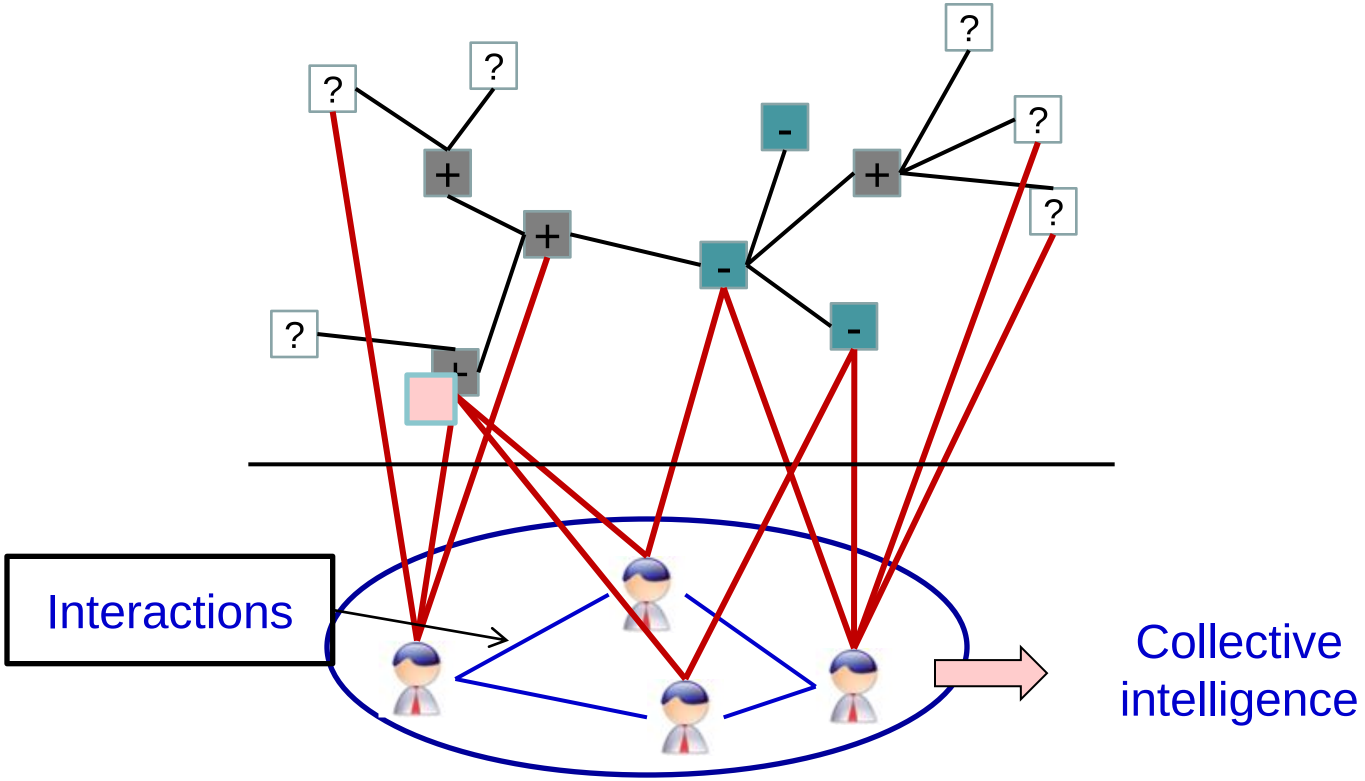


Examples:
Amazon/Netflix
wikipedia

(1) personalized learning
(2) collaborative filtering

History of the Web

Social Web



Interactions

Collective intelligence

(1) interactions
(2) information diffusion

Examples:
facebook, twitters

Social network analysis

- Social influence analysis
- User behavior analysis
- Community structure analysis
- Network properties
- …

Motivation: Social influence

- Social influence refers to the behavior change of individuals affected by others in a network
  - Advertising
  - Social recommendation
  - Marketing
  - Social security
  - …

Agenda

- **<u>Influence related statistics</u>**
- Social similarity and influence
- Influence maximization in viral marketing

Influence Related Statistics

- Edge measures
  - Tie strength
  - Triadic closure
  - Weak ties
  - Edge betweenness
- Node measures
  - Degree
  - Closeness
  - Node betweennes
  - Structural holes

Tie Strength

- Tie Strength [1]

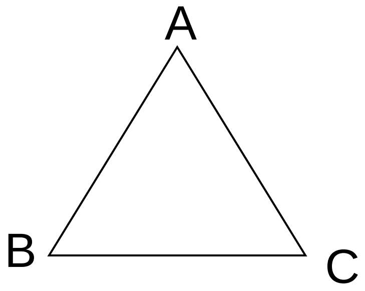
$$S(A, B) = \frac{|n_A \cap n_B|}{|n_A \cup n_B|}$$

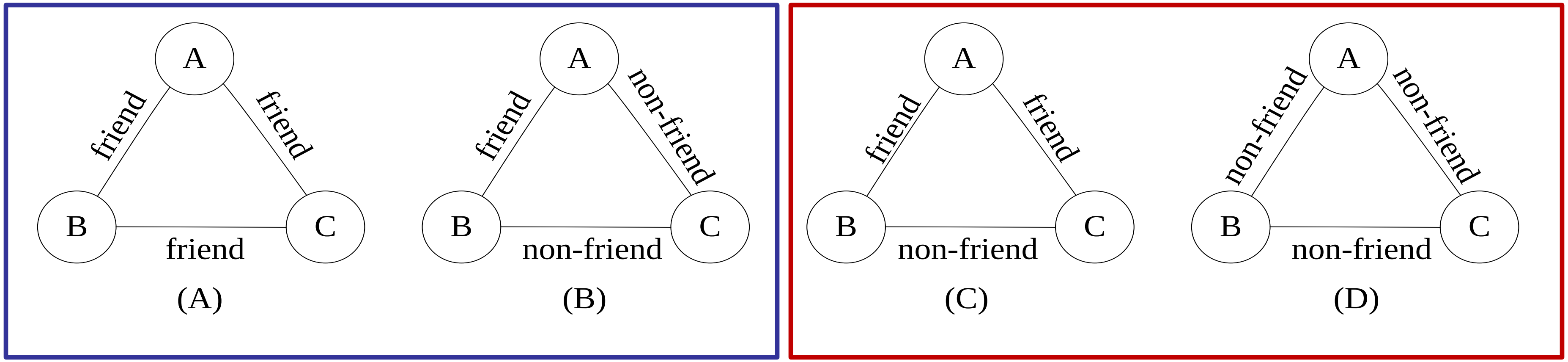
  - n_A,n_B are the neighborhoods of *A* and *B*
  - Depend on the overlap of neighborhoods
  - aka. embeddedness
  - The higher tie strength, the easier for two individuals to trust one another.

[1] M. Granovetter. The strength of weak ties. American Journal of Sociology, 78(6):1360–1380, 1973

Triadic Closure

- Triadic closure can be corollared from tie strength.
  - If A-B and A-C are strong ties, then B-C are likely to be a strong tie.
  - If A-B and A-C are weak ties, then B-C are less likely to be strong tie.
- Related to social balance theory [1]





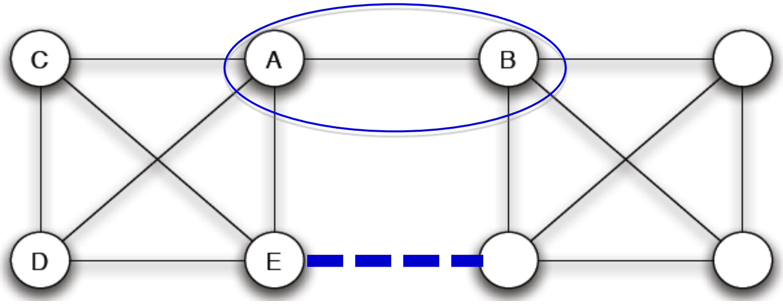
balanced

inbalanced

[1] D. Easley and J. Kleinberg. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, 2010.

Weak Tie

- Weak tie
  - The overlap of neighborhoods is small
- Local bridge
  - No overlapping neighbors



- Global bridge
  - Removal of A-B may result in the disconnection of the connected component containing A and B .
  - In real network, global bridges are rare compared to local bridges
  - The effect of local and global bridges is quite similar

Edge Betweenness

- Edge betweenness [1] [2]
  - The total amount of flow across an edge A-B
  - The information flow between A and B are evenly distributed on the shortest paths between A and B.
- Graph partitioning is an application of edge betweenness
  - Gradually remove edges of high betweenness to turn the network into disconnected components [3].

[1] L. C. Freeman. A set of measure of centrality based on betweenness. Sociometry, 40:35–41, 1977.

[2] L. C. Freeman. Centrality in social networks: Conceptual clarification. Social Networks, 1:215‹239, 1979.

[3] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12):7821–7826, June 2002.

Influence Related Statistics

- Edge measures
  - Tie strength
  - Triadic closure
  - Weak ties
  - Edge betweenness
- Node measures
  - Degree
  - Closeness
  - Node betweennes
  - Structural holes

Node Measures - centrality

- Radial measures
  - computed by random walks that start or end from a given node, including:
  - Degree
  - Katz and its extensions
  - Closeness
- Medial measures
  - computed by the random walks that pass through a given node
  - Node betweenness
  - structural holes

Radial Measures (1)

- Degree

$$c_i^{DEG} = \deg(i)$$

  - deg(i) is the number of edges connecting to node i.

- Katz centrality [1]

$$c_i^{KATZ} = e_i^T \left(\sum_{j=1}^{\infty} (\beta A)^j\right) \mathbf{1}$$

A Adjacency Matrix

  - e_i is a column vector with i^{th} element is 1, and all others are 0
  - β is a positive penalty constant between 0 and 1
  - Counts the number of walks starting from a node, while penalizing the longer walks.

[1] L. Katz. A new index derived from sociometric data analysis. Psychometrika,18:39–43, 1953.

Radial Measures (2)

- Bonacich centrality [1]

A Adjacency Matrix

$$c_i^{BON} = e_i^T(\frac{1}{\beta}\sum_{j=1}^{\infty}(\beta A)^j)\mathbf{1}$$

  - β has negative value, allowing to subtract the even-numbered walks from the odd-numbered walks in exchange networks

- Hubbell centrality [2]

$$c_i^{HUB} = e_i^T(\sum_{j=0}^{\infty} X^j)\mathbf{y}$$

  - General case of Katz and Bonacich centrality where the matrix X and vector y can be assigned different value

- Eigenvector centrality
  - The principle eigenvector of matrix A is Katz centrality

[1] P. Bonacich. Power and centrality: a family of measures. American Journal of Sociology, 92:1170–1182, 1987.

[2] C. Hubbell. An input-output approach to clique identification. Sociometry,28:377–399, 1965.

Radial Measures (3)

- Closeness centrality [Freeman, 1979]

$$c_i^{CLO} = e_i^T S\mathbf{1}.$$

  - An element (*i*, *j*) in matrix *S* contains the shortest path from node *i* to node *j*
  - Compute the average of the shortest distances to all the other nodes.

Medial measures—betweenness

- Betweenness centrality [1]

$$c_i^{BET} = \sum_{j,k} \frac{b_{jik}}{b_{jk}}$$

  - b_{jk} is the number of shortest paths from node *j* to *k*
  - b_{jik} is the number of shortest paths from node *j* to *k* that pass through node *i*
  - Measure how much a given node lies in the shortest paths of other nodes
  - Naïve algorithm involves all-pair shortest paths, requiring $O(n^3)$, [2] introduces a fast algorithm running O(nm) and $O(nm+n^2logn)$
- Random walk based betweenness centrality[3]

$$c_i^{NBE} = \sum_{j \neq i \neq k} R_{jk}^{(i)}.$$

  - $R_{jk}^{(i)}$ is the probability of a random walk from *j* to *k*, which contains *i* as an intermediate node.
  - Instead of considering shortest paths, it considers all possible walks

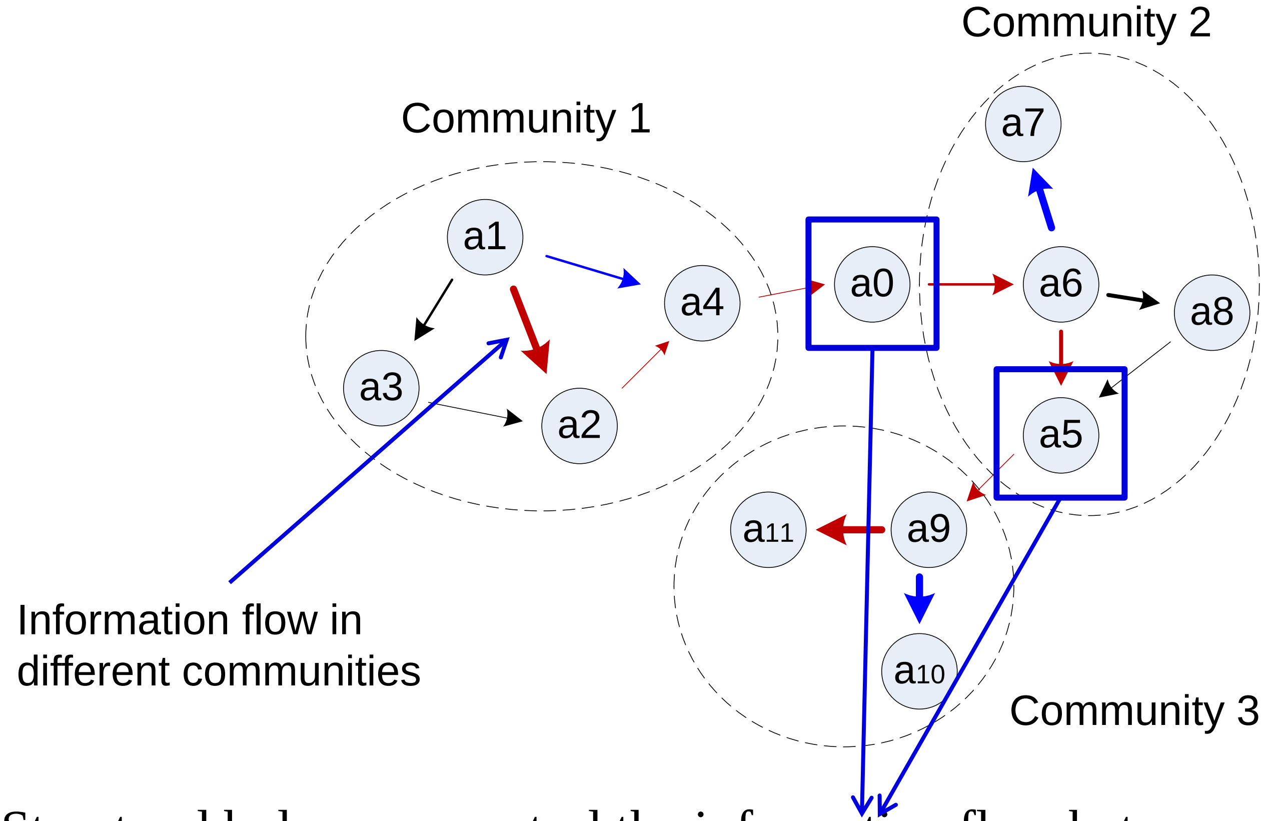
[1] L. C. Freeman. A set of measure of centrality based on betweenness. Sociometry, 40:35–41, 1977.

[2] U. Brandes. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 2001.

[3] M. E. J. Newman. A measure of betweenness centrality based on random walks. Social Networks, 2005.

Medial measures—structural holes

- Structural holes
  - A node is a structural hole if it is connected to multiple local bridges.
  - The person who serves as a structural hole can interconnect information originating from multiple connectivity of diverse regions of the network.



Information flow in different communities

Structural hole users control the information flow between different communities

Medial measures—structural holes

Pathcount [1]: for each node, the algorithm counts the average number of shortest paths (between each pair of nodes) it lies on, and then selects those nodes with the highest numbers as structural hole nodes.

2-Step Connectivity [2]: for each node, it counts the number of pairs of neighbors who are not directly connected. And then those nodes with the highest numbers are viewed as structural holes.

[1] S. Goyal and F. Vega-Redondo. Structural holes in social networks. *Journal of Economic Theory, 137(1):460–492, 2007.*

[2] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM'12, 2012.*

Agenda

- Influence related statistics
- **Social similarity and influence**
- Influence maximization in viral marketing

Social Similarity and Influence

- Homophily and selection
- Shuffle test for social influence
- Computational social influence model
- Influence and action
- Others

Homophily

- Homophily
  - An actor in the social network tends to be similar to their connected neighbors.
- Originated from different mechanisms
  - Social influence
    - Indicates people tend to follow the behaviors of their friends
  - Selection
    - Indicates people tend to create relationships with other people who are already similar to them
  - Confounding variables
    - Other unknown variables exist, which may cause friends to behave similarly with one another.

Generative models for selection and influence

[Holme et al., 2006]

- A model to characterize how social influence and selection work together to affect users' actions
  - Initial
    - Place *M* edges of the network uniformly at random between vertex pairs, and also assign opinions to vertices uniformly at random.
  - For each step, the simulation either
    - Selection: moves an edge to lie between two individuals whose opinions agree, or
    - Influence: change the opinion of an individual to agree with one of their neighbors.
  - Result
    - Selection tend to generate a large number of small clusters
    - Influence will generate large coherent clusters.

[1] P. Holme and M. E. J. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. Physical Review, 74(056108), 2006.

Multi-dimensional generative models for selection and influence [Crandall et al., 2008]

- Options for generating next action for user *u*
  - Sample from *u*’s own history
  - Sample from history of a friend of *u*
  - Sample from history of any user
  - Perform a new action
- Options for generating next interaction for user *u*
  - Sample from users with similar history of actions
  - Sample a random user to interact with

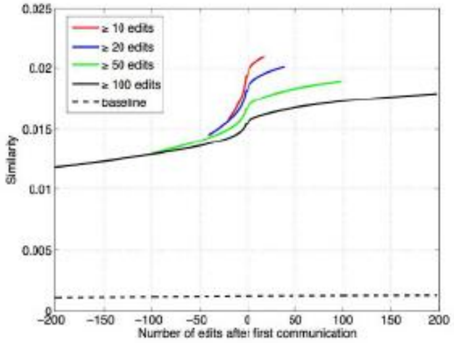


Figure 1: Average cosine similarity of user pairs as a function of the number of edits from time of first interaction, for Wikipedia.

- How does the similarity of two users vary around the time of their first interaction?
  - Sharp increase in similarity immediately before first interaction
  - Continuing but slower increase after first interaction

[1] 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.

Social Similarity and Influence

- Homophily and selection
- Shuffle test for social influence
- Computational social influence
- Influence and action
- Others

Shuffle test for social influence

[Anagnostopoulos et al., 2008]

- Goal
  - to test whether social influence exists in a social network?
- Intuition
  - if social influence does not play a role, then the timing of activation should be independent of the timing of activation of other users.
  - Edge reversal test: if the social correlation is not determined by social influence, then reversing the edge direction will not change the social correlation estimate

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'08), pages 7–15, 2008.

Randomization Test of Social Influence

[Fond et al, 2010]

- Model social network as a time-evolving graph G_t=(V, E_t), and the nodes with attribute at time t is X_t
- Main idea
  - Selection and social influence can be differentiated through the autocorrelation between X_t and G_t
  - Selection process
    - X_{t-1} determines the social network at G_t
  - Social influence
    - G_{t-1} determines the node attributes at time t, i.e., X_t

[1] T. L. Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In Proceeding of the 19th international conference on World Wide Web (WWW'10), 2010.

Social Similarity and Influence

- Homophily and selection
- Shuffle test for social influence
- Computational social influence
- Influence and action
- Others

(a) Quantifying Influence and Selection [1]

$$Selection = \frac{p(a_{ij}^{t} = 1 | a_{ij}^{t-1} = 0, \langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \rangle > \epsilon)}{p(a_{ij}^{t} = 1 | a_{ij}^{t-1} = 0)}$$

–Denominator: the conditional probability that an unlinked pair will become linked

–Numerator: the same probability for unlinked pairs whose similarity exceeds the threshold

$$Influence = \frac{p(\langle \mathbf{x}_i^{t}, \mathbf{x}_j^{tT} \rangle > \langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \rangle | a_{ij}^{t-1} = 0, a_{ij}^{t} = 1)}{p(\langle \mathbf{x}_i^{t}, \mathbf{x}_j^{t} \rangle > \langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \rangle | a_{ij}^{t-1} = 0)}$$

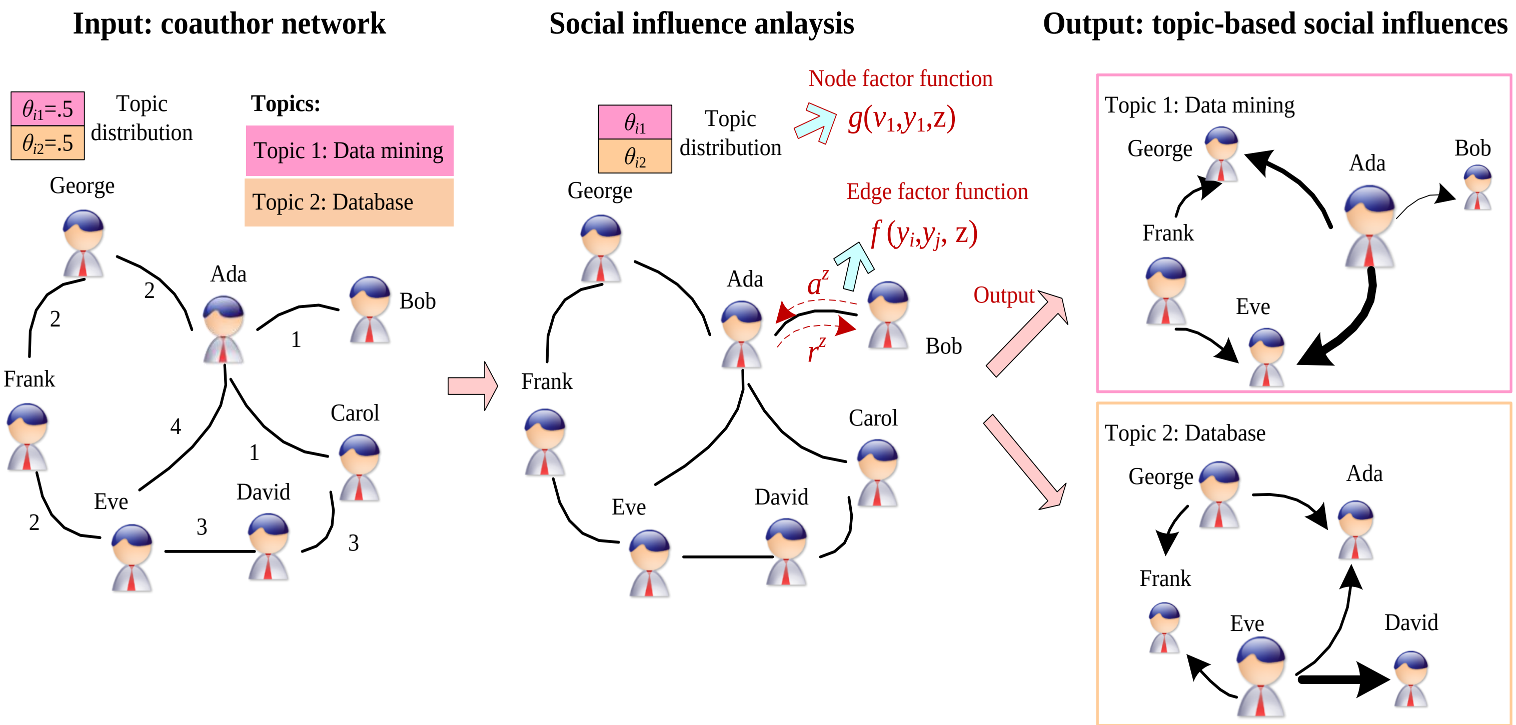
–Denominator: the probability that the similarity increase from time *t-1* to *t* between two nodes that were not linked at time *t-1*

–Numerator: the same probability that became linked at time *t*

- *Model is learned through matrix factorization*

[1] J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09), pages 747–756, 2009.

(b) Quantifying Influence from Different Topics [1]



Several key challenges:

- **How to differentiate the social influences from different angles (topics)?**
- **How to incorporate different information (e.g., topic distribution and network structure) into a unified model?**
- **How to estimate the model on real-large networks?**

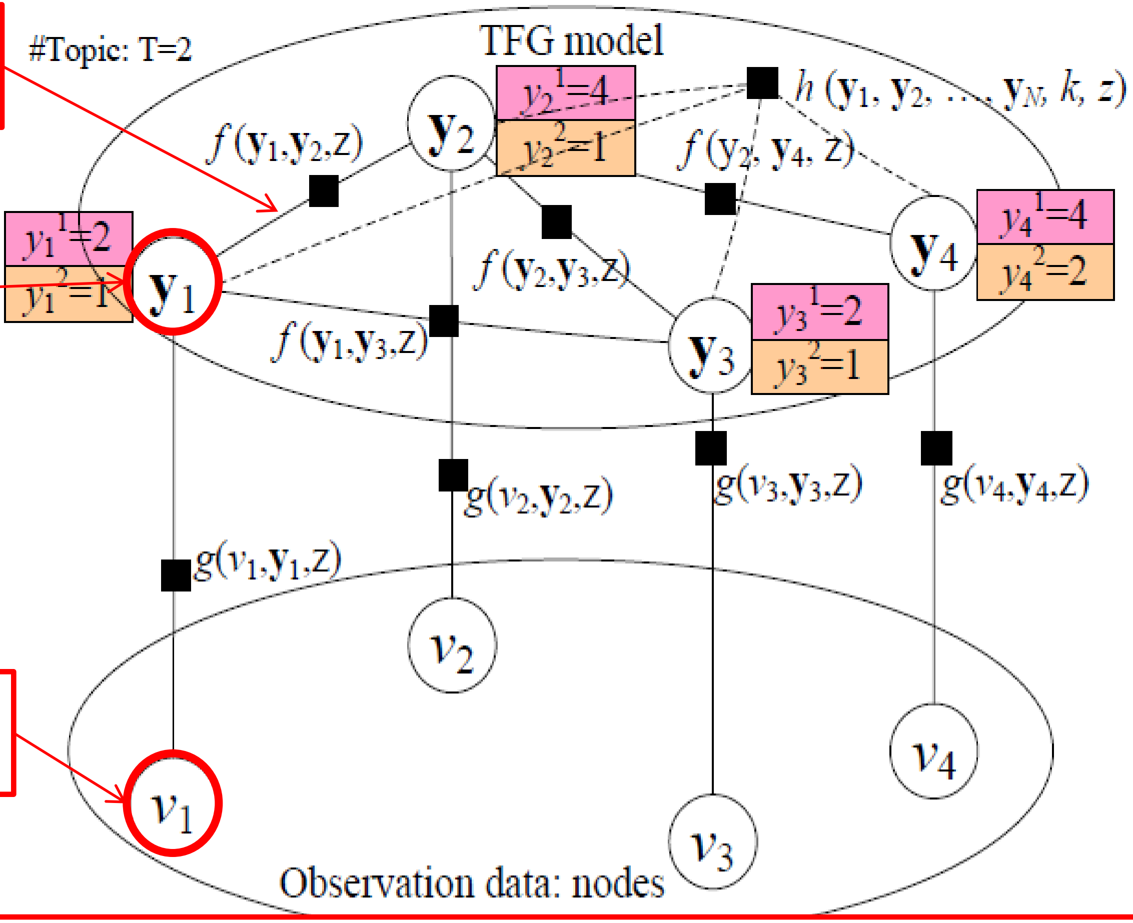
[1] J. Tang, J. Sun, C.Wang, and Z. Yang. Social influence analysis in largescale networks. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'09), pages 807–816, 2009.

Topical Factor Graph (TFG) Model

Social link

Nodes that have the highest influence on the current node

Node/user



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

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

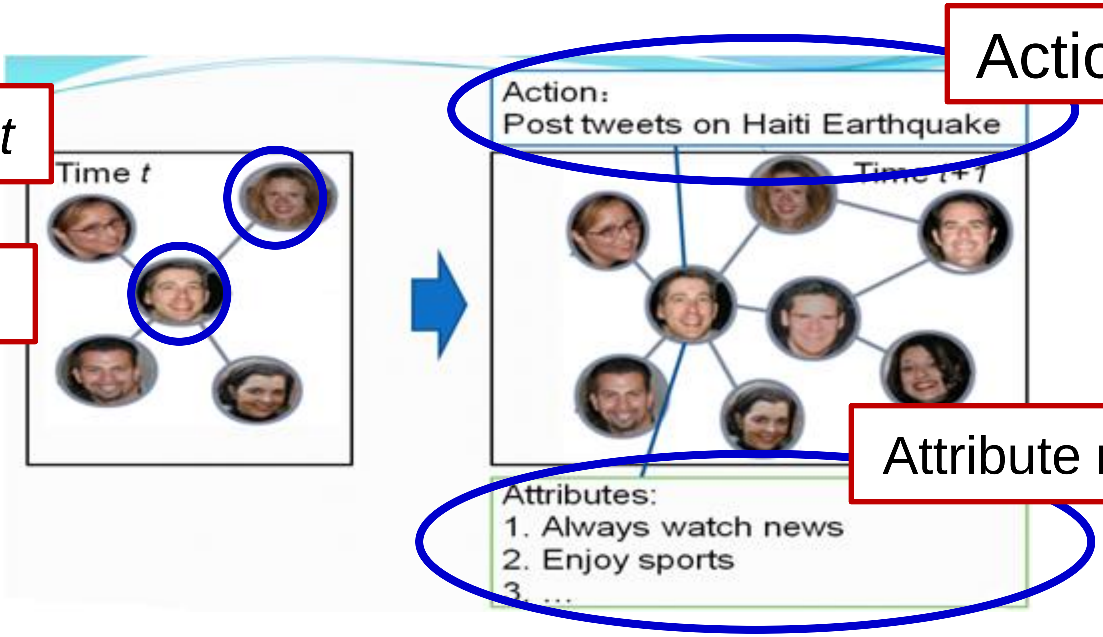
Break & QA

Social Similarity and Influence

- Homophily and selection
- Shuffle test for social influence
- Computational social influence
- Influence and action
- Others

Influence and Action

$$G^t = (V^t, E^t, X^t, Y^t)$$



Actions at time t

Nodes at time t

Edges at time t

Attribute matrix at time t

Input:

$G^t = (V^t, E^t, X^t, Y^t)$

$t = 1,2,\ldots T$

Output:

$F: f(G^t) -> Y^{(t+1)}$

(a) Learning Influence Probabilities [1]

- **Goal:** Learn user influence and action influence from historical actions
- **Assumption**
  - If user v_i performs an action y at time t and later his friend v_j also perform the action, then there is an influence from v_i to v_j

- **User influence probability**

$$infl(v_i) = \frac{\left|\{y \mid \exists v_j, \Delta t : prop(y, v_i, v_j, \Delta t) \wedge \Delta t \geq 0\}\right|}{Y_{v_i}}$$

- Were $\Delta t = t_j - t_i$ is the difference between the time when v_j performing the action and the time when user v_i performing the action, given e_{ij}=1; $prop(y, v_i, v_j, \Delta t)$ represents the action propagation score

[1] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In Proceedings of the 3st ACM International Conference on Web Search and Data Mining (WSDM'10), pages 207–217, 2010.

(a) Learning Influence Probabilities [1]

- **Action influence probability**

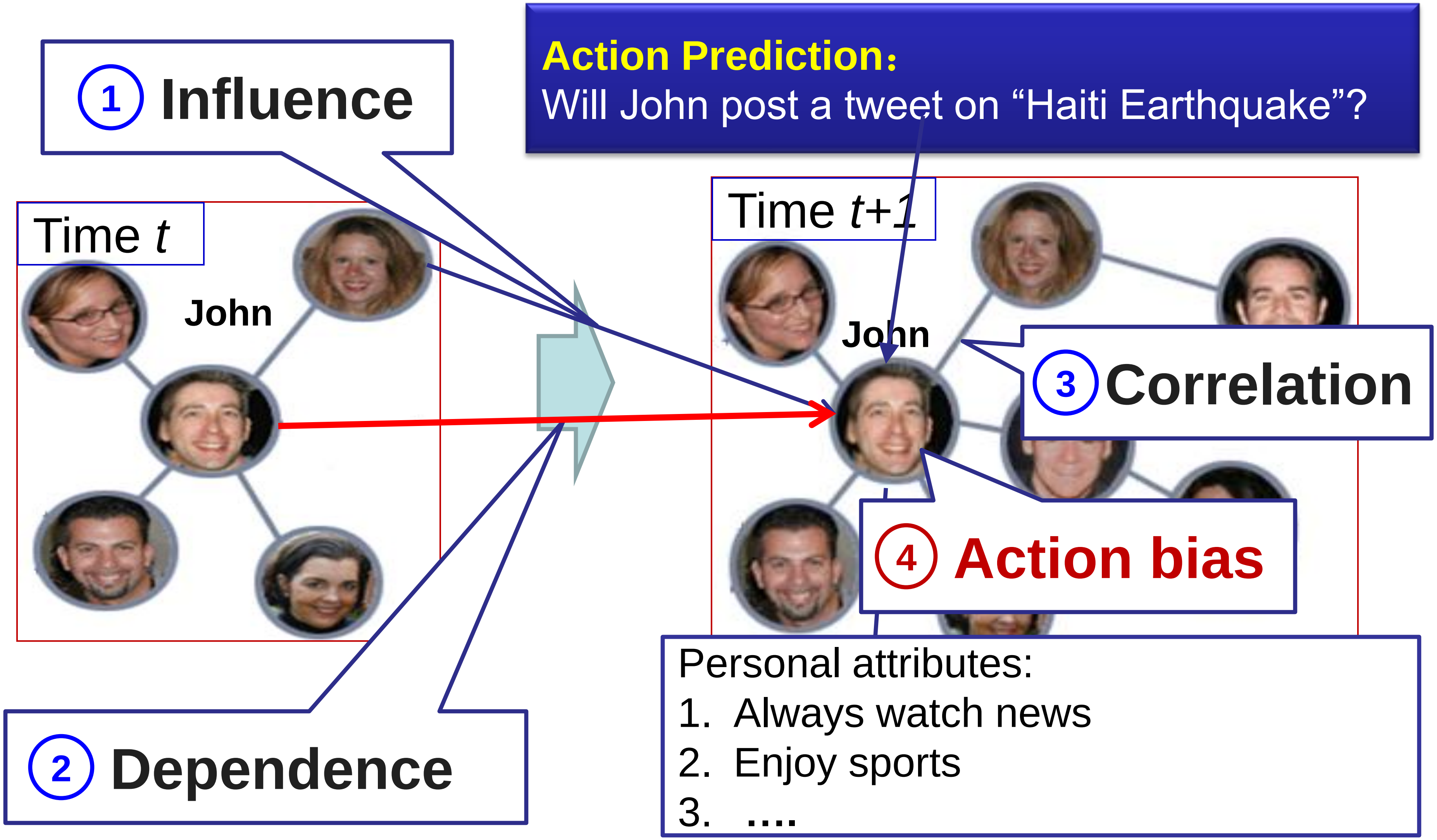
$$infl(y) = \frac{\left|\{v_j \mid \exists v_i, \Delta t : \boxed{prop(y, v_i, v_j, \Delta t)} \wedge \Delta t \geq 0\}\right|}{\text{number of users performing } y}$$

- were $\Delta t = t_j - t_i$ is the difference between the time when v_j performing the action and the time when user v_i performing the action, given e_{ij}=1; $prop(y, v_i, v_j, \Delta t)$ represents the action propagation score

[1] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In Proceedings of the 3st ACM International Conference on Web Search and Data Mining (WSDM'10), pages 207–217, 2010.

(b) Social Action Modeling and Prediction

[Tang et al., 2010]



Action Prediction:
Will John post a tweet on "Haiti Earthquake"?

① **Influence**

Time *t*

John

Time *t+1*

John

③ **Correlation**

④ **Action bias**

② **Dependence**

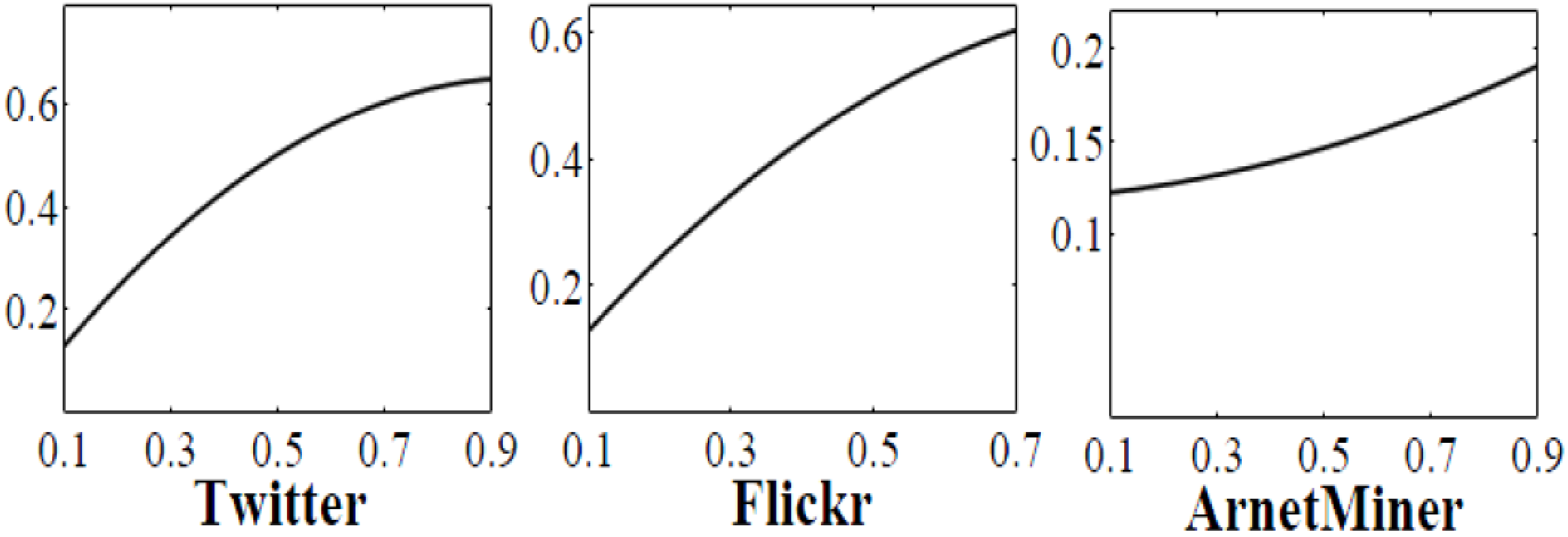
Personal attributes:

1. Always watch news
2. Enjoy sports
3. ….

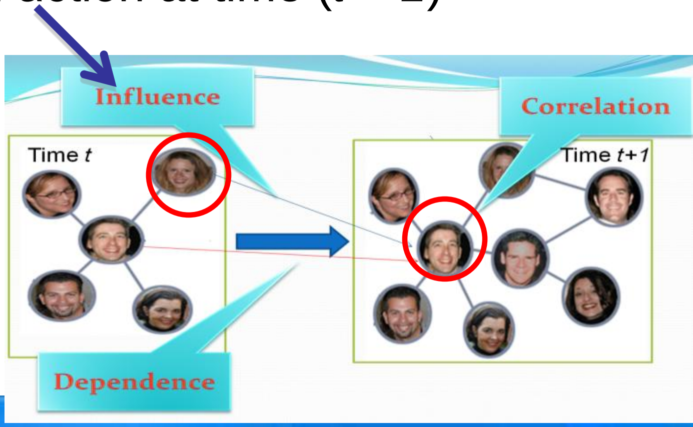
[1] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'10), pages 807–816, 2010.

Statistical Study: Influence

Y-axis: the likelihood that the user also performs the action at time *t*

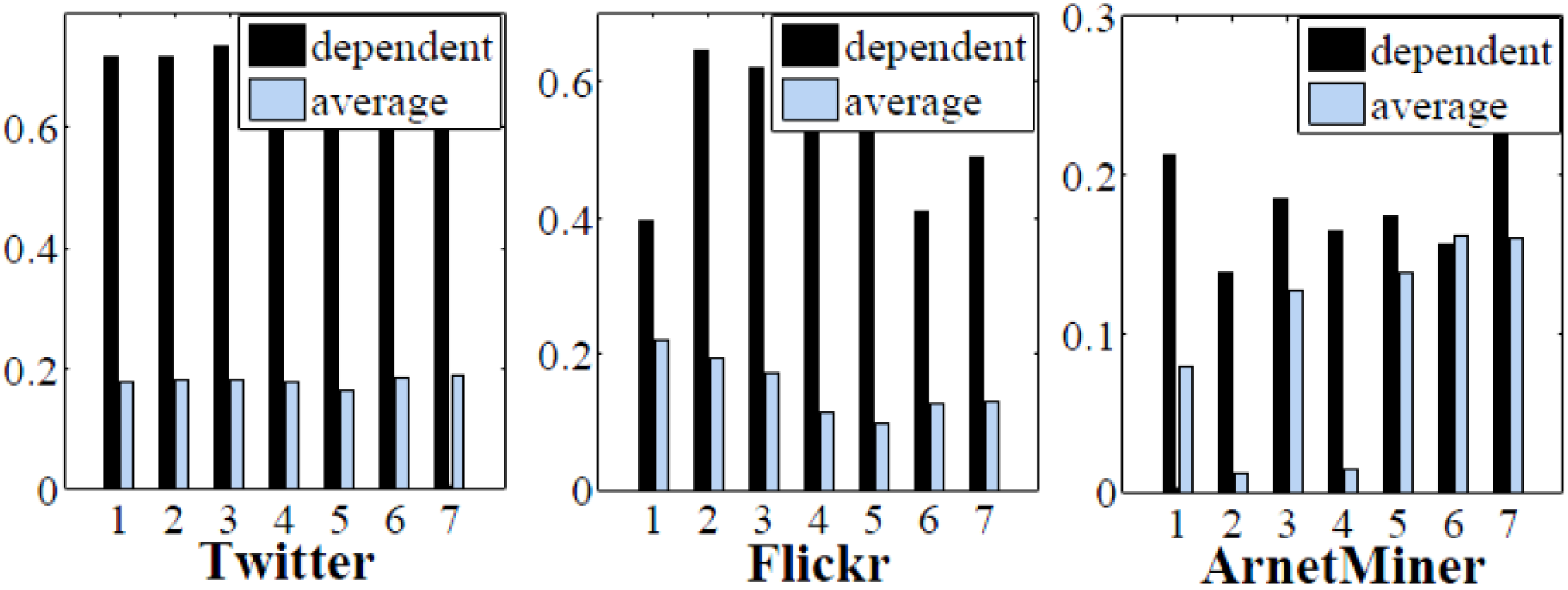


X-axis: the percentage of one’s friends who perform an action at time (*t* − *1*)

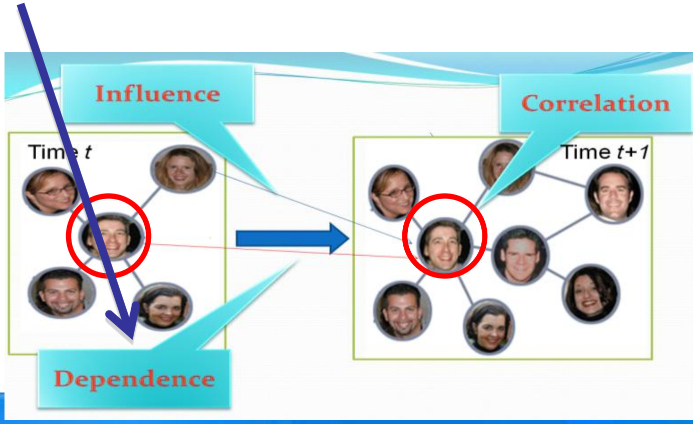


Statistical Study: Dependence

Y-axis: the likelihood that a user performs an action

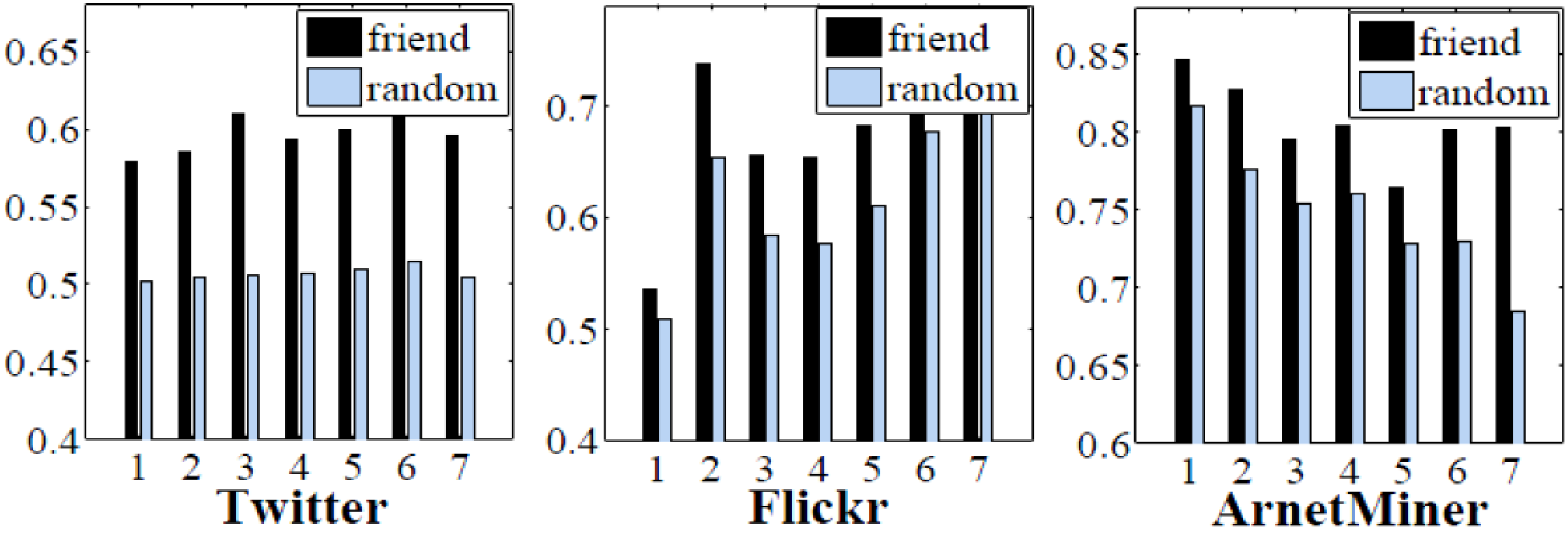


X-axis: different time windows (1-7)

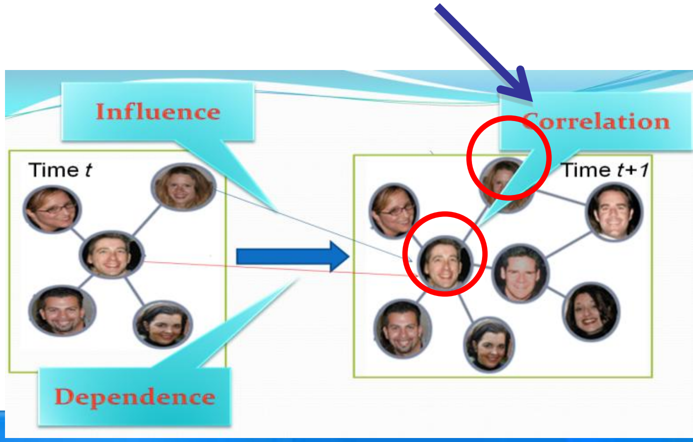


Statistical Study: Correlation

Y-axis: the likelihood that two friends(random) perform an action together

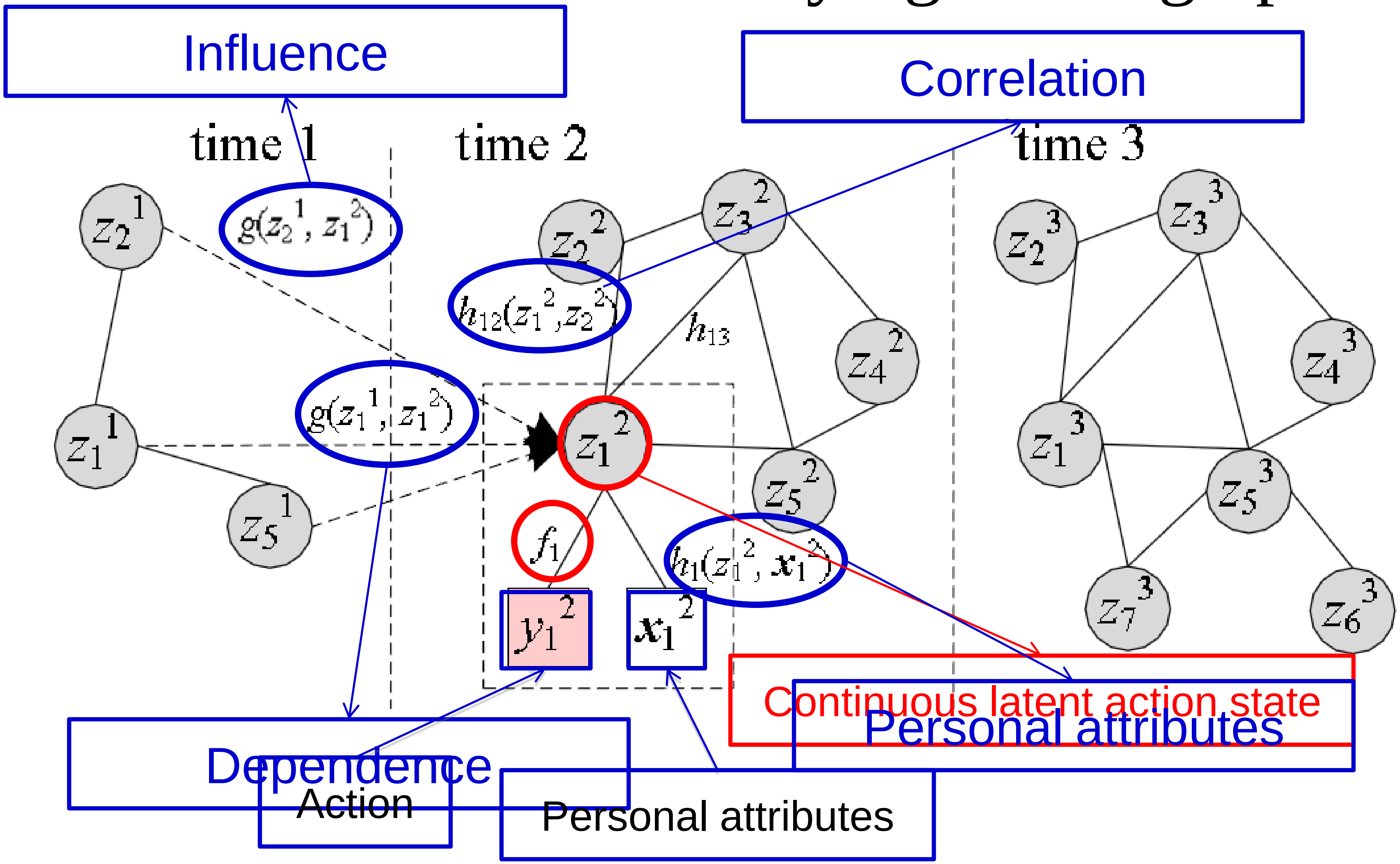


X-axis: different time windows (1-7)



NTT-FGM Model[1]

Noise tolerant time-varying factor graphs



Influence

Correlation

Continuous latent action state

Personal attributes

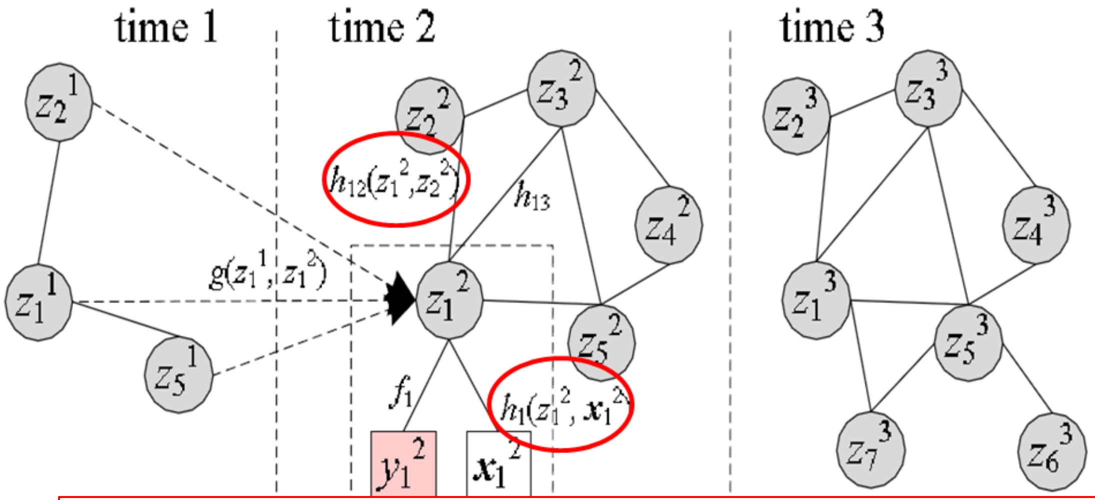
Dependence

Action

Personal attributes

[1] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'10), pages 807–816, 2010.

Model Instantiation



$$
\begin{aligned}
g_{ji}(z_i^t, z_j^{t-1}) &= -(z_i^t - z_j^{t-1})^2 \\
h_{ij}(z_i^t, z_j^t) &= -(z_i^t - z_j^t)^2 \\
h_k(z_i^t, x_{ik}^t) &= -(z_i^t - x_{ik}^t)^2
\end{aligned}
$$

How to estimate the parameters?

$$
p(\mathbf{Y}|\mathbf{G}) = \frac{1}{Z}\exp\{\sum_{t=1}^{T}\sum_{i=1}^{N} -\frac{(y_i^t - z_i^t)^2}{2\sigma^2} + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} \boxed{\lambda_{ij}} m_{ji}^{t-1} g(z_i^t, z_j^{t-1})
$$

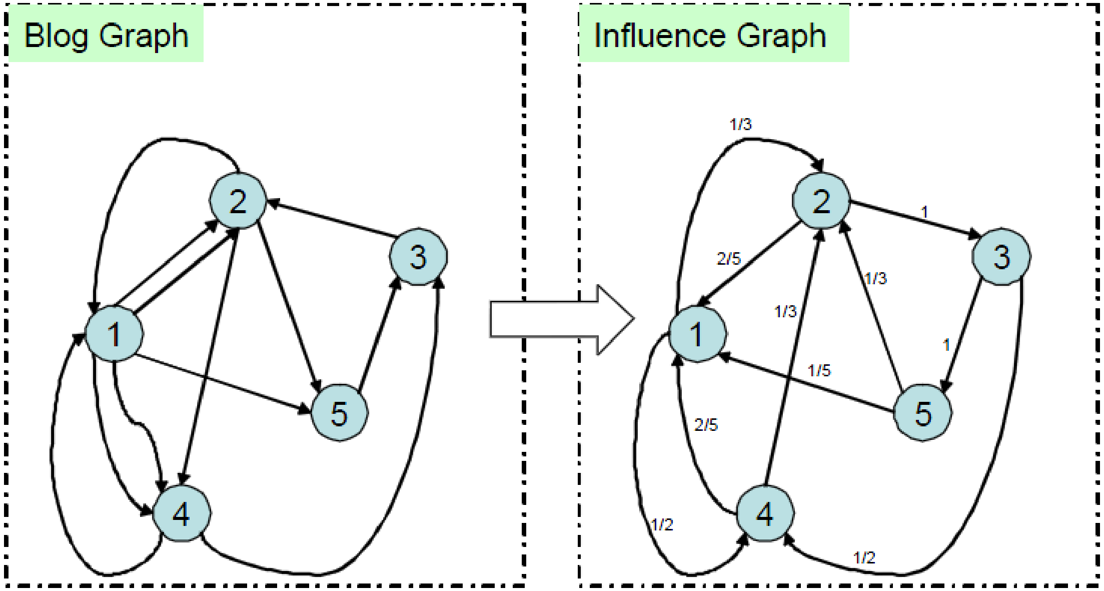
$$
+ \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} \boxed{\beta_{ij}} m_{ij}^t h_{ij}(z_i^t, z_j^t) + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{d} \boxed{\alpha_k} h_k(z_i^t, x_{ik}^t)\}
$$

Social Similarity and Influence

- Homophily and selection
- Shuffle test for social influence
- Computational social influence
- Influence and action
- Others

(a) Influence and Interaction

- Influence can also be reflected by the interactions between users
  - Facebook: one can post messages on friends’ wall page to influence her/him.
  - Twitter: one can use follower/following members on Twitter to infer the influence
- [Xiang et al., 2010] use a latent variable model to infer relationship strength based on profile similarity and interaction activity
  - Considering the relationship strength to be the hidden effect of user profile similarities, as well as the hidden cause of the interaction s between users.
  - The basic idea is to model social influence using a link analysis method with random walk



[1] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In Proceeding of the 19th international conference on World Wide Web (WWW’10), pages 981–990, 2010.

(b) Influence and Autocorrelatoin

- Autocorrelation
  - A set of linked users e_{ij}=1 and an attribute matrix X associated with these users, as the correlation between the values of X on these instance pairs
  - Social influence, diffusion processes, and the principle of homophily give rise to auto correlated observations
- Behavior correlation
  - [1] utilize the observed behavior correlation to predict the collective behavior in social media through
    - Social dimension extraction
    - Discriminative learning

[1] L. Tang and H. Liu. Scalable learning of collective behavior based on sparse social dimensions. In Proceeding of the 18th ACM conference on Information and knowledge management(CIKM'09), pages 1107–1116, New York, NY, USA, 2009. ACM.

(c) Influence and Grouping Behavior

- Grouping behavior
  - E.g., user's participation behavior into a forum
- [1] analyzes four factors in online forums that potentially influence people's behavior in joining communities
  - Friends of Reply Relationship.
    - Describe how users are influenced by the number of neighbors with whom they have ever had any reply relationship.
  - Community Sizes
    - Quantify the 'popularity' of information
  - Average Ratings of Top Posts
    - Measure how the authority of information impacts user behavior.
  - Similarities of Users
    - If two users are 'similar' in a certain way, what is the correlation of the sets of communities they join?

[1] X. Shi, J. Zhu, R. Cai, and L. Zhang. User grouping behavior in online forums. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'09), pages 777–786, New York, NY, USA, 2009. ACM.

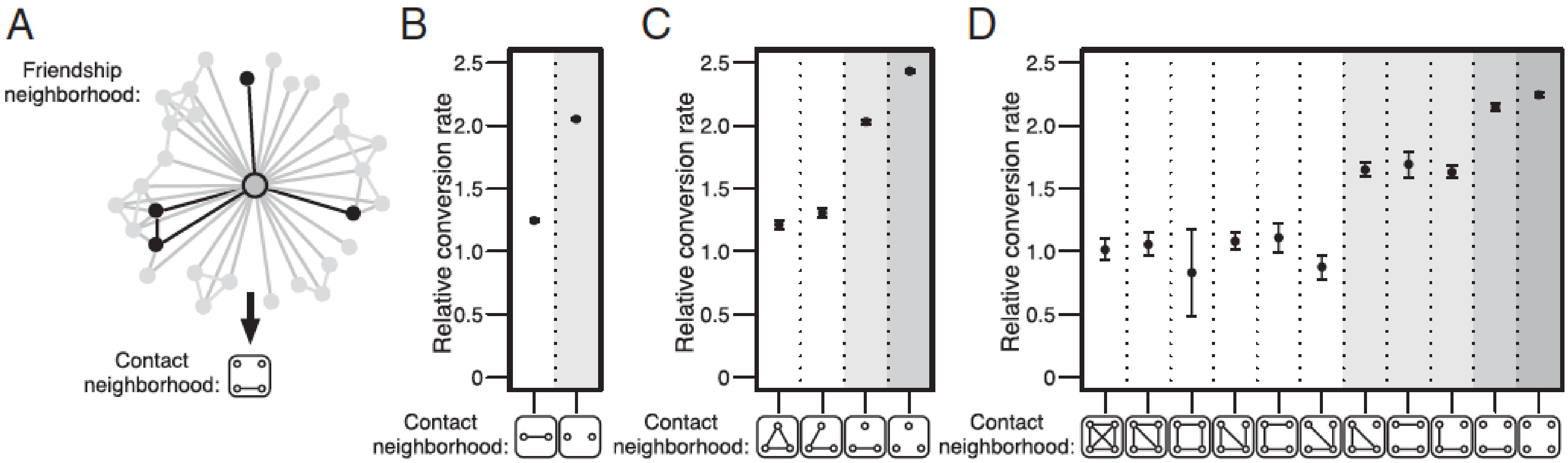
(c) Influence and Grouping Behavior (cont.)

- [1]construct a bipartite graph with two sets of nodes, users and communities to study the impact of factors
  - They build a bipartite Markov Random Field model to evaluate how much each feature affects the grouping behavior in online forums
  - The advantage is that it can explicitly investigate how a user’s joining behavior is affected by her friends’ joining behavior.

[1] X. Shi, J. Zhu, R. Cai, and L. Zhang. User grouping behavior in online forums. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD’09), pages 777–786, New York, NY, USA, 2009. ACM.

(d) Quantifying influence by structural diversity [1]

- They find that the probability of influence is tightly controlled by the number of connected components in an individual's contact neighborhood, rather than by the actual size of the neighborhood.
- They analyze Facebook email invitation that find that for contact neighborhoods consisting of 2-4nodes, the probability an invitation is accepted is much higher when the nodes in the neighborhood are not connected by a link.



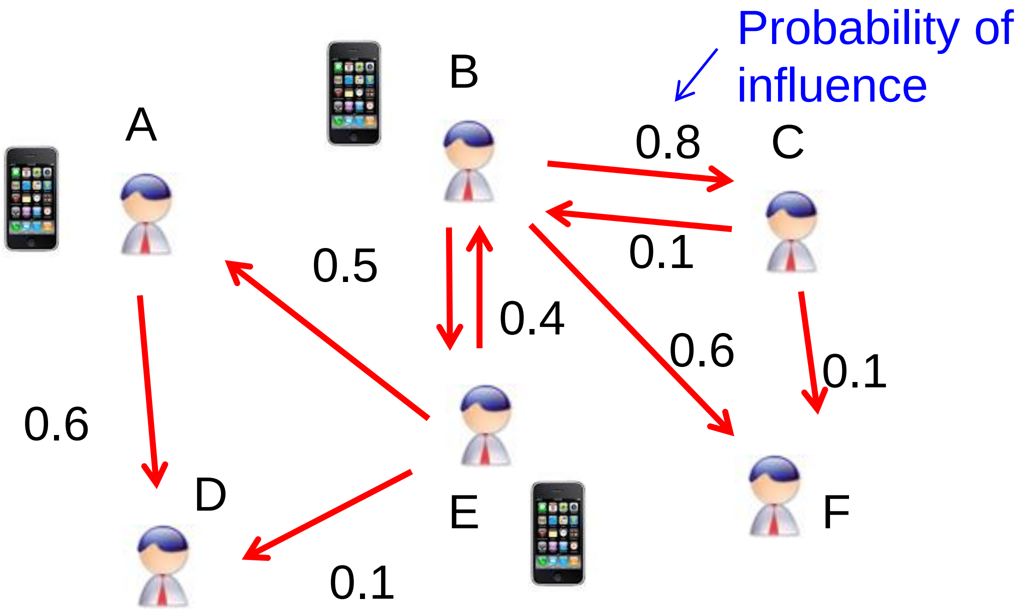
[1] J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg. Structural Diversity in Social Contagion. Proc. National Academy of Sciences, 109(16) 5962-5966, 17 April 2012.

Agenda

- Influence related statistics
- Social similarity and influence
- **Influence maximization in viral marketing**

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

Alg1: High-degree heuristic

- Choose the seed nodes according to their degree.
- Intuition
  - The nodes with more neighbors would arguably tend to impose more influence upon its direct neighbors.
  - Know as “degree centrality”

Alg2: Low-distance Heuristic

- Consider the nodes with the shortest paths to other nodes as seed nodes
- Intuition
    - Individuals are more likely to be influenced by those who are closely related to them.

Alg3: Degree Discount Heuristic

- General idea
  - If *u* has been selected as a seed, then when considering selecting *v* as a new seed based on its degree, we should not count the edge *v*->*u*
- Referred to SingleDiscount
- Specifically, for a node *v* with d_v neighbors of which t_v are selected as seeds, we should discount *v*'s degree by $2t_v + (d_v - t_v)t_v p$

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 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

$$\sum_{u neighbors of v} b_{v,u} \leq 1$$

[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.

Cascade Model

- Cascade model
  - $p_v(u,S)$: the success probability of user u activating user v
    - i.e., 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
  - Fc(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.

Evaluation

- 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] accelerate the influence procedure by up to 700 times
- It is still challenging to extend these methods to large data sets

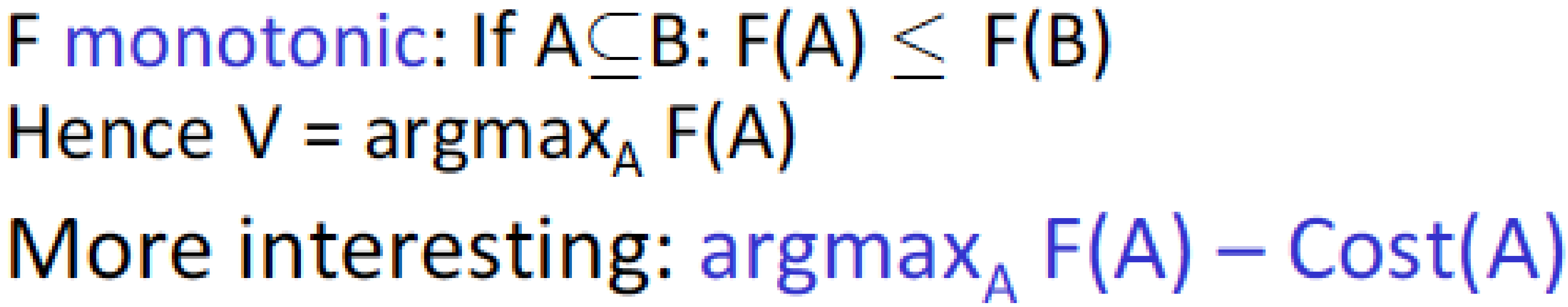
[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'03), pages 137–146, 2003.
[2] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen,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

Implementation

- **Objective function:**

  - F(S) = Expected #people influenced when targeting a set of users S

  - (1-1/e) approximation ratio



F monotonic: If $A \subseteq B$: $F(A) \leq F(B)$

Hence $V = \text{argmax}_A F(A)$

More interesting: $\text{argmax}_A F(A) – \text{Cost}(A)$

[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.

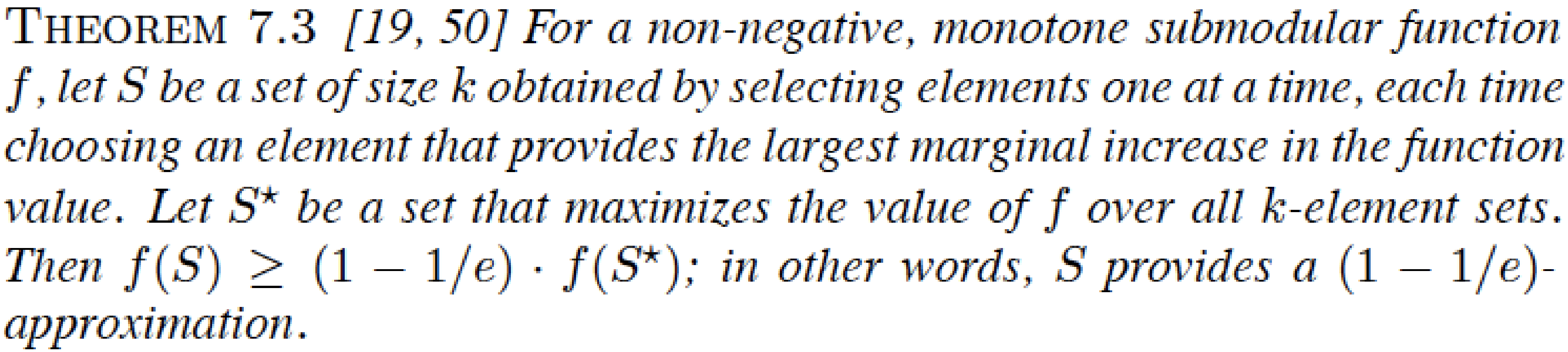
[2] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD’03), pages 137–146, 2003.

Maximizing the Spread of Influence

- Solution
  - Use a submodular function to approximste the influence function
- Given a function *f* that is submodular, taking only non-negative values, we have

$$f(S \cup \{v\}) \geq f(S)$$

  - Then the problem can be transformed into finding a *k*-element set *S* for which *f*(*S*) is maximized.



THEOREM 7.3 *[19, 50] For a non-negative, monotone submodular function f, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let $S^\star$ be a set that maximizes the value of f over all k-element sets. Then $f(S) \geq (1 - 1/e) \cdot f(S^\star)$; in other words, S provides a $(1 - 1/e)$-approximation.*

[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.

Application: Online Advertising

[Goyal et al., 2008]

- Propose viral marketing through frequent pattern mining.
- Assumption
  - Users can see their friends actions.
- Basic formation of the problem
  - Actions take place in different time steps, and the actions which come up later could be influenced by the earlier taken actions.
- Approach
  - Define leaders as people who can influence a sufficient number of people in the network with their actions for a long enough period of time.
  - Finding leaders in a social network makes use of action logs.

[1] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08), pages 499–508, 2008.

Application: Influential Blog Discovery

- Influential Blog Discovery
  - In the web 2.0 era, people spend a significant amount of time on user-generated content web sites, like blog sites.
  - Opinion leaders bring in new information, ideas, and opinions, and disseminate them down to the masses.
- Four properties for each bloggers
  - **Recognition**: A lot of inlinks to the article.
  - **Activity generation**: A large number of comments indicates that the blog is influential.
  - **Novelty**: with less outgoing links.
  - **Eloquence**: Longer articles tend to be more eloquent, and can thus be more influential.

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM'08), pages 207–217, 2008.

Conclusion

- Social influence analysis aims at qualitatively and quantitatively measuring the influence
- This tutorial includes
  - Basic statistical measures of network
  - Social influence analysis
  - Influence maximization and applications
- Potential research directions
  - Scalable social influence analysis
  - More applications using social influence

Thanks!

Jimeng Sun and Jie Tang

IBM TJ Watson Research Center

Tsinghua University