Social Influence Analysis in Large-scale Networks

Jie Tang[1], Jimeng Sun[2], Chi Wang[1], and Zi Yang[1]

[1]Dept. of Computer Science and Technology
Tsinghua University
[2]IBM TJ Watson Research Center, USA
June 30th 2009

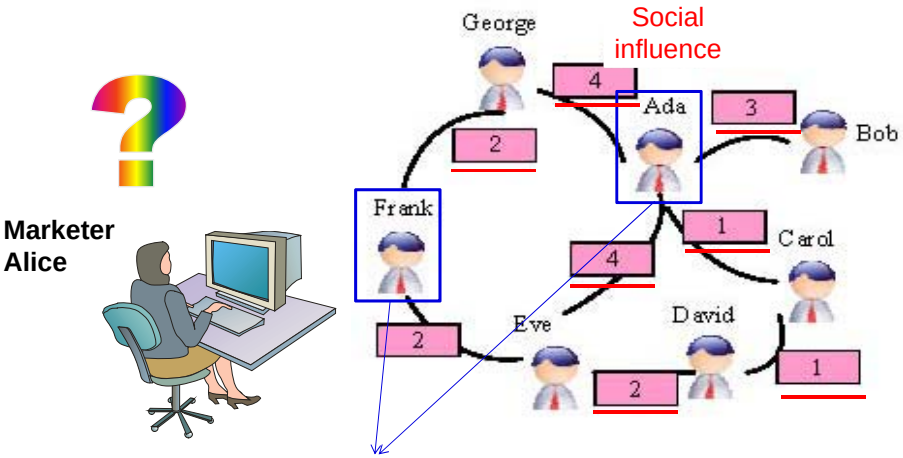
1

Motivation

- Social influence plays a key role in many (online) social networks, e.g., MSN, Flickr, DBLP
- Quantitative measure of the strength of social influence can benefit many real applications
  - Expert finding
  - Social recommendation
  - Influence maximization
  - …

2

Example—Influence Maximization

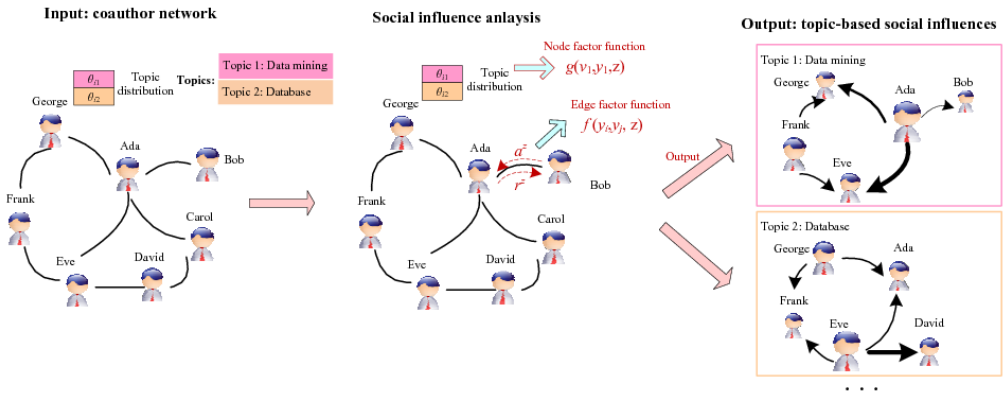


Find a small subset of nodes (users) in a social network that could maximize the spread of influence (Domingos, 01; Richardson, 02; Kempe, 03)

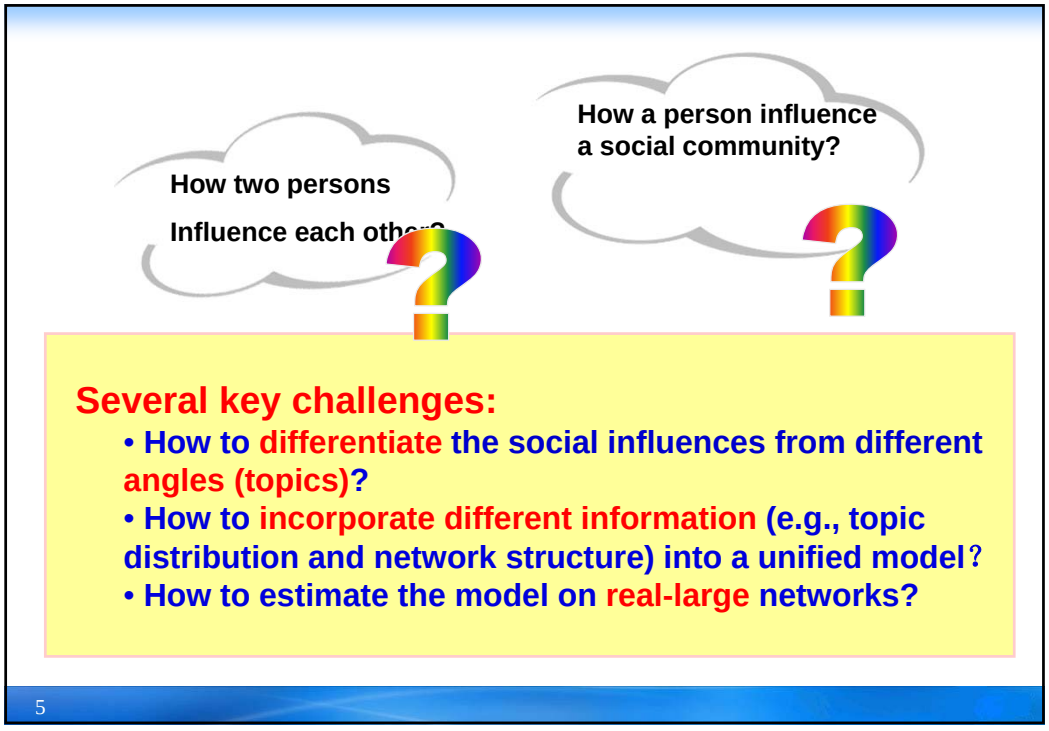
3

Topic-based Social Influence Analysis

- Social network -> Topical influence network



4



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?**

5

Outline

- Related Work
- Topical Affinity Propagation
  - Topical Factor Graph Model
  - Basic TAP Learning
  - Distributed TAP Learning
- Experiments
- Conclusion & Future Work

6

Related Work—Social networks and influences

- Social network
  - Metrics to characterize a social network
  - Web community discovery [Flake,2000]
- Influence in social network
  - The existence of influence. [Singla, 2008] [Anagnostopoulos, 2008]
  - The correlation between social similarity and interactions [Crandall, 2008]

7

Related Work—large-scale mining

- Factor graph models
  - A graph model [Kschischang, 2001]
  - Computing marginal function [Frey, 2006]
  - Message passing/affinity propagation [Frey, 2007]
- Distributed programming model
  - Map-reduce [J. Dean, 2004]

8

Outline

- Related Work
- Topical Affinity Propagation
  - Topical Factor Graph Model
  - Basic TAP Learning
  - Distributed TAP Learning
- Experiments
- Conclusion & Future Work

9

Topical Factor Graph (TFG) Model

- *N* nodes
- A set of observed variables: $\{v_i\}_{i=1}^{N}$
- A set of hidden vectors: $\{\mathbf{y}_i\}_{i=1}^{N}$
- $\mathbf{y}_i \in \{1, \ldots, N\}^T$ models the topic-level influences from other nodes to node v_i
- Each element $y_i^z \in$ {1, … N}, represents the node that has the highest probability to influence node v_i on topic z.

10

Summary of Notations

Table 1: Notations.

SYMBOL	DESCRIPTION
N	number of nodes in the social network
M	number of edges in the social network
T	number of topics
V	the set of nodes in the social network
E	the set of edges
v_i	a single node
y_i^z	node-v_i's representative on topic z
$\mathbf{y}_i$	the hidden vector of representatives for all topics on node v_i
θ_i^z	the probability for topic z to be generated by the node v_i
e_{st}	an edge connecting node v_s and node v_t
w_{st}^z	the similarity weight of the edge e_{st} w.r.t. topic z
μ_{st}^z	the social influence of node v_s on node v_t w.r.t. topic z

11

Feature Functions

- Node feature function ***g(v_i , y_i , z)*** is a feature function defined on node v_i specific to topic z.
- Edge feature function ***f(y_i , y_j , z)*** is a feature function defined on the edge of the input network specific to topic z.
- Global feature function ***h(y_1, …, y_N, k, z)*** is a feature function defined on all nodes of the input network w.r.t. topic z.

12

Feature Functions

- The node feature function is defined as

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

- *NB(i)* is the neighbors of v_i
- $w^z_{ij} = \theta^z_j \alpha_{ij}$ reflects the topical similarity or interaction strength between v_i *and* v_j
- θ^z_j denoting the importance of node-*j* to topic *z*
- α_{ij} denoting the weight of the edge e_{ij}

13

Feature Functions

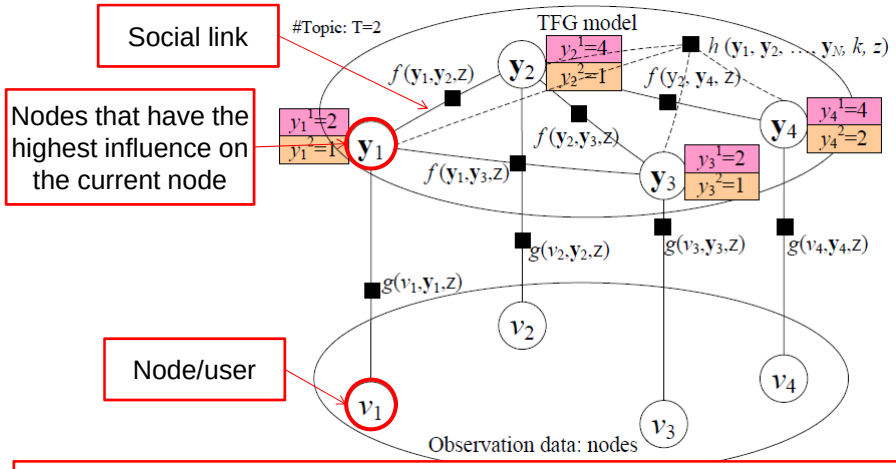
- A binary edge feature function $f(\mathbf{y}_i, \mathbf{y}_j, z) = 1$ if and only if there is an edge e_{ij} between node v_i and node v_j , otherwise 0.
- A global edge feature function *h* on all nodes

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

- *h* constrains the model to bias towards the "true" representative nodes. A representative node on topic *z* must be the representative of itself on topic *z, (i.e.* $y^z_k = k$*)* And it must be a representative of at least another node v_i *(i.e.* $\exists y^z_i = k, i \neq k$

14

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.

15

Topical Factor Graph (TFG)

Objective function:

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

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

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.

16

How to define (topical) feature functions?

– Node feature function

similarity

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

– 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

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

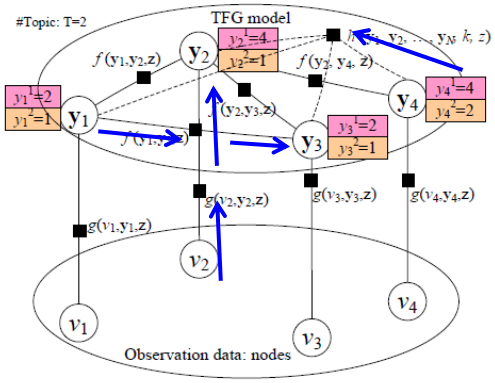
17

Model Learning Algorithm

- Sum-product:

$$m_{y_i \to f_{ij}}(y_i) = \prod_{f' \sim y_i \setminus f_{ij}} m_{f' \to y_i}(y_i)$$

$$m_{f_{ij} \to y_i}(y_i) = \sum_{\sim \{y_i\}} \left(\prod_{y' \sim f \setminus y_i} f(y_i, y') m_{y' \to f_{ij}}(y') \right)$$



- Low efficiency!
- Not easy for distributed learning!

18

New TAP Learning Algorithm

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

2. Design new update rules:

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

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

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

The TAP Learning Algorithm

Input: $G = (V, E)$ and topic distributions $\{\theta_v\}_{v \in V}$
Output: topic-level social influence graphs $\{G_z = (V_z, E_z)\}^T$

1.1 Calculate the node feature function $g(v_i, \mathbf{y}_i, z)$;
1.2 Calculate b_{ij}^z according to Eq. 8;
1.3 Initialize all $\{r_{ij}^z\} \leftarrow 0$;
1.4 **repeat**
1.5 **foreach** *edge-topic pair* (e_{ij}, z) **do**
1.6 Update r_{ij}^z according to Eq. 5;
1.7 **end**
1.8 **foreach** *node-topic pair* (v_j, z) **do**
1.9 Update a_{jj}^z according to Eq. 6;
1.10 **end**
1.11 **foreach** *edge-topic pair* (e_{ij}, z) **do**
1.12 Update a_{ij}^z according to Eq. 7;
1.13 **end**
1.14 **until** *convergence*;
1.15 **foreach** *node* v_t **do**
1.16 **foreach** *neighboring node* $s \in NB(t) \cup \{t\}$ **do**
1.17 Compute μ_{st}^z according to Eq. 9;
1.18 **end**
1.19 **end**
1.20 Generate $G_z = (V_z, E_z)$ for every topic z according to $\{\mu_{st}^z\}$;

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

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

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

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

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

Distributed TAP Learning

- Map-Reduce
  - Map: (key, value) pairs
    - $e_{ij}/a_{ij} \rightarrow e_{i*}/a_{ij}$; $e_{ij}/b_{ij} \rightarrow e_{i*}/b_{ij}$; $e_{ij}/r_{ij} \rightarrow e_{*j}/r_{ij}$.
  - Reduce: (key, value) pairs
    - e_{ij} / * → new r_{ij}; e_{ij}/* → new a_{ij}
- For the global feature function

THEOREM 1. *If the global feature function h can be factorized into $h = \prod_{k=1}^{N} h_k$, for every $i \in \{1, \ldots, N\}, y_i \neq k, y_i' \neq k, h_k(y_1, \ldots, y_i, \ldots, y_N) = h_k(y_1, \ldots, y_i', \ldots, y_N)$, then the message passing update rules can be simplified to influence update rules.* ■

21

Outline

- Related Work
- Topical Affinity Propagation
  - Topical Factor Graph Model
  - Basic TAP Learning
  - Distributed TAP Learning
- Experiments
- Conclusion & Future Work

22

Experiment

- Data set: (ArnetMiner.org and Wikipedia)
  - **Coauthor** dataset：640,134 authors and 1,554,643 coauthor relations
  - **Citation** dataset: 2,329,760 papers and 12,710,347 citations between these papers
  - **Film** dataset: 18,518 films, 7,211 directors, 10,128 actors, and 9,784 writers
- Evaluation measures
  - CPU time
  - Case study
  - Application

23

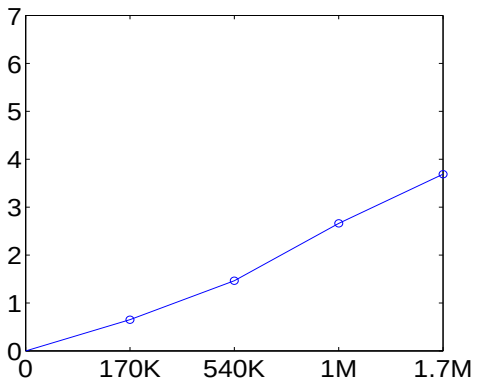
Scalability Performance

Table 2: Scalability performance of different methods on real data sets. >10hr means that the algorithm did not terminate when the algorithm runs more than 10 hours.

Methods	Citation	Coauthor	Film
Sum-Product	N/A	>10hr	1.8 hr
Basic TAP Learning	>10hr	369s	**57s**
Distributed TAP Learning	**39.33m**	**104s**	148s

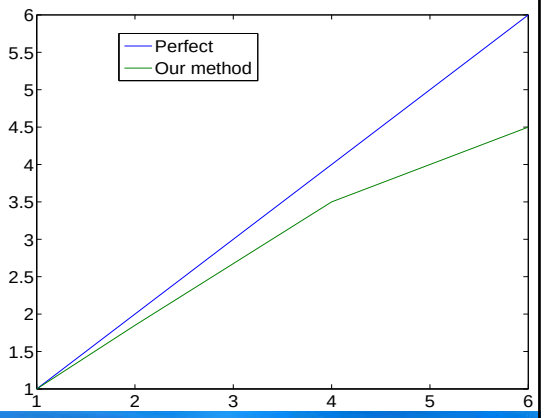
24

Speedup results



Speedup vs. Dataset size

Speedup vs. #Computer nodes



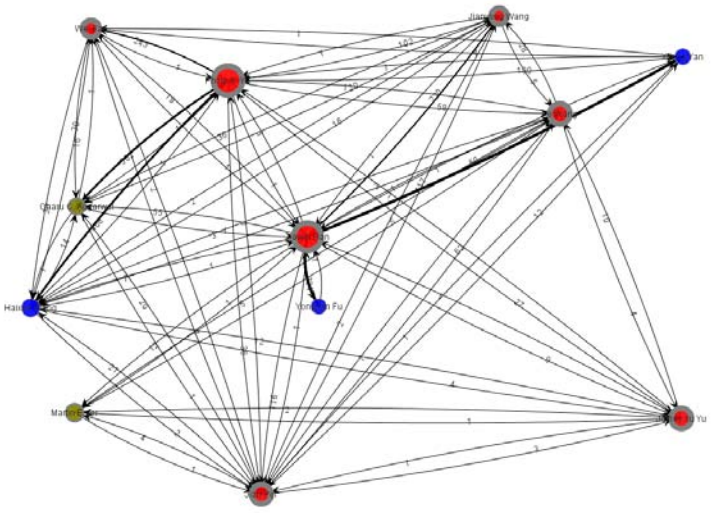
25

Influential nodes on different topics

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 DIs

26

Social Influence Sub-graph on “Data mining”



27

Application—Expert Finding

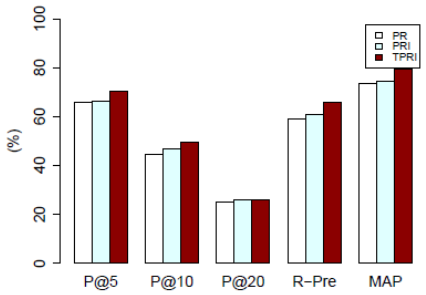


Table 7: Performance of expert finding with different approaches.

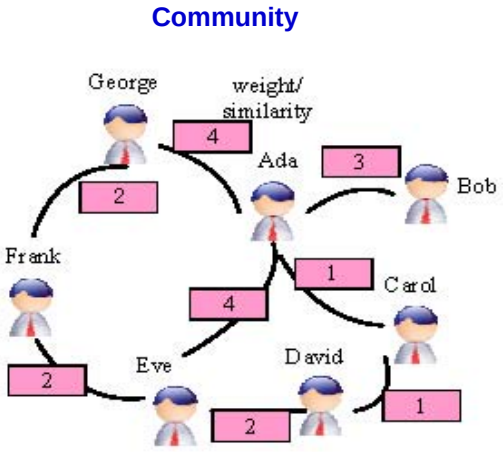
Expert finding data from (Tang, KDD08; ICDM08)

http://arnetminer.org/lab-datasets/expertfinding/

28

Application—Influence Maximization

Marketer Alice



[Domingos, 01; Richardson, 02; Kempe, 03]

29

Outline

- Related Work
- Topical Affinity Propagation
  - Topical Factor Graph Model
  - Basic TAP Learning
  - Distributed TAP Learning
- Experiments
- Conclusion & Future Work

30

Conclusion

- Formalize a novel problem of topic-based social influence analysis.
- Propose a Topical Factor Graph model to describe the problem using a graphical probabilistic model.
- Present an algorithm and its distributed version to efficiently train the TFG model.
- Experimental results on three different types of data sets demonstrate the effectiveness and efficiency of the proposed approach.

31

Future Work

- Model:
  - Jointly learn topic distribution and social influence
  - Semi-supervised learning
- Many other social analysis tasks:
  - Influence maximization
  - Community influence
  - Personality
  - …

32

Thanks!

Q&A

Online resource: (data, codes, tools)
http://arnetminer.org/lab-datasets/soinf/

HP: http://keg.cs.tsinghua.edu.cn/persons/tj/

For more information,
please come to our poster tonight!

33

Background

- Social influence plays a key role in many (online) social networks:
  - Instant messaging: MSN, Jaber, Skype
  - Sharing site: Flickr, Picasa, Youtube
  - Wikis, blogs, microblogs
  - Collaboration network: DBLP
- Social network analysis has mainly focused on macro-level models
  - E.g., degree distribution, community, diameter, etc.
- Social influence analysis
  - Qualitatively measure the existence of social influence
  - No quantitative measure of the strength of social influence

34

Influence between individuals

- Coauthor data

Topic: Data Mining		Topic: Database		Topic: Machine Learning	
Jiawei Han	Heikki Mannila	Jiawei Han	Heikki Mannila	Jiawei Han	Heikki Mannila
David Clutter	Arianna Gallo	David Clutter	Vladimir Estivill-Castro	Hasan M. Jamil	Vladimir Estivill-Castro
Hasan M. Jamil	Marcel Holsheimer	Shiwei Tang	Marcel Holsheimer	K. P. Unnikrishnan	Marcel Holsheimer
K. P. Unnikrishnan	Robert Gwadera	Hasan M. Jamil	Robert Gwadera	Shiwei Tang	Mika Klemettinen
Ramasamy Uthurusamy	Vladimir Estivill-Castro	Ramasamy Uthurusamy	Mika Klemettinen	Ramasamy Uthurusamy	Robert Gwadera
Shiwei Tang	Mika Klemettinen	K. P. Unnikrishnan	Arianna Gallo	David Clutter	Arianna Gallo

- On Citation data

Fast Algorithms for Mining Association Rules in Large Databases	Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data
Mining Large Itemsets for Association Rules	Mining Web Site?s Clusters from Link Topology and Site Hierarchy
A New Framework For Itemset Generation	Predictive Algorithms for Browser Support of Habitual User Activities on the Web
Efficient Mining of Partial Periodic Patterns in Time Series Database	A Fine Grained Heuristic to Capture Web Navigation Patterns
A New Method for Similarity Indexing of Market Basket Data	A Road Map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining
A General Incremental Technique for Maintaining Discovered Association Rules	

35