Community Detection and Behavior Study for Social Computing

Huan Liu[+], Lei Tang[+], and Nitin Agarwal[*]

[+]Arizona State University

[*]University of Arkansas at Little Rock

Updated slides available at http://www.public.asu.edu/~ltang9/
http://www.public.asu.edu/~huanliu/

Acknowledgements

- We would like to express our sincere thanks to Jianping Zhang, John J. Salerno, Sun-Ki Chai, Xufei Wang, Sai Motoru and Reza Zafarani for collaboration, discussion, and valuable comments.
- This work derives from the projects, in part, sponsored by AFOSR and ONR grants.

- Some materials presented here can be found in the following book chapters and references section of this tutorial:
  - Lei Tang and Huan Liu, ***Graph Mining Applications to Social Network Analysis***, in *Managing and Mining Graph Data* (forthcoming)
  - Lei Tang and Huan Liu, ***Understanding Group Structures and Properties in Social Media***, in *Link Mining: Models, Algorithms and Applications* (forthcoming)
- If you wish to use the ppt version of the slides, please contact (or email) us. The ppt version contains more comprehensive materials with additional information and notes and many animations.

Outline

- Social Media
- Data Mining Tasks
- Evaluation

- Principles of Community Detection
- Communities in Heterogeneous Networks
- Evaluation Methodology for Community Detection

- Behavior Prediction via Social Dimensions
- Identifying Influential Bloggers in a Community
  - A related tutorial on Blogosphere

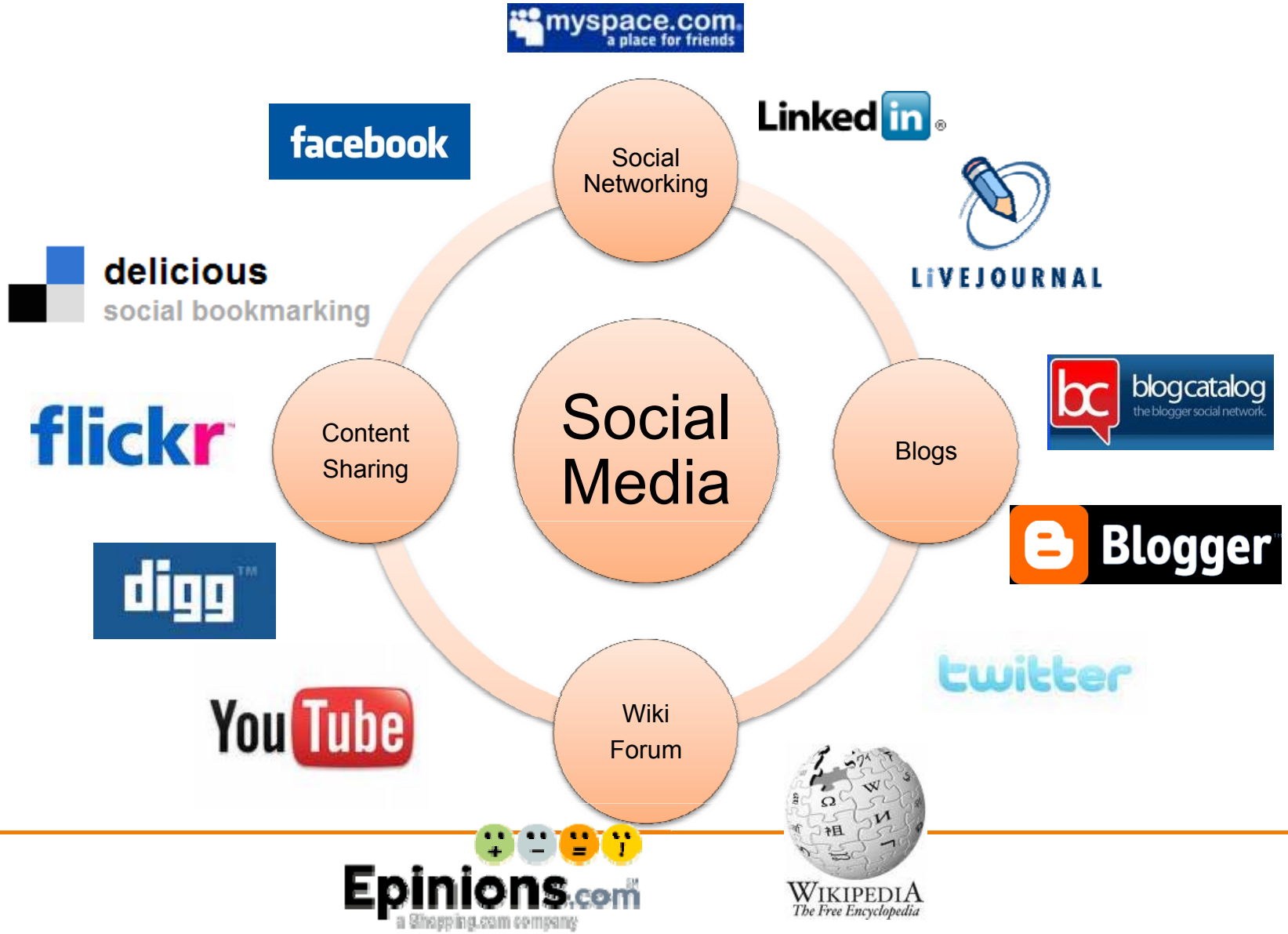
PARTICIPATING WEB AND SOCIAL MEDIA

Traditional Media

Broadcast Media: One-to-Many

Communication Media: One-to-One

Social Media: Many-to-Many



Characteristics of Social Media

- Everyone can be a media outlet
- Disappearing of communications barrier
  - Rich User Interaction
  - User-Generated Contents
  - User Enriched Contents
  - User developed widgets
  - Collaborative environment
  - Collective Wisdom
  - Long Tail

Broadcast Media
Filter, then Publish

Social Media
Publish, then Filter

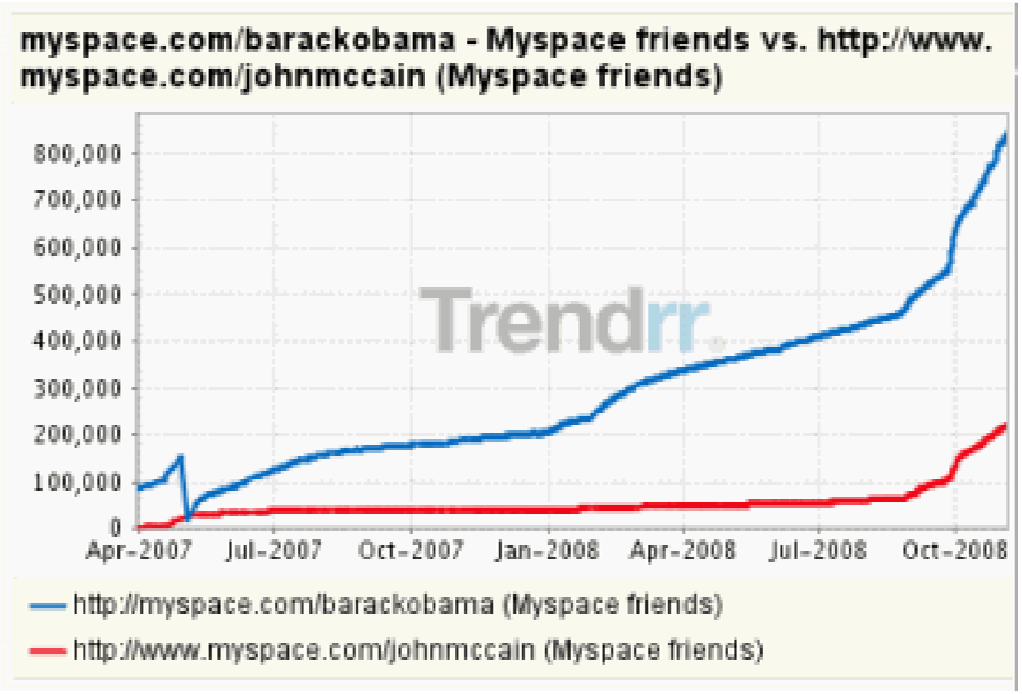
Top 20 Most Visited Websites

- Internet traffic report by Alexa on August 27th, 2009

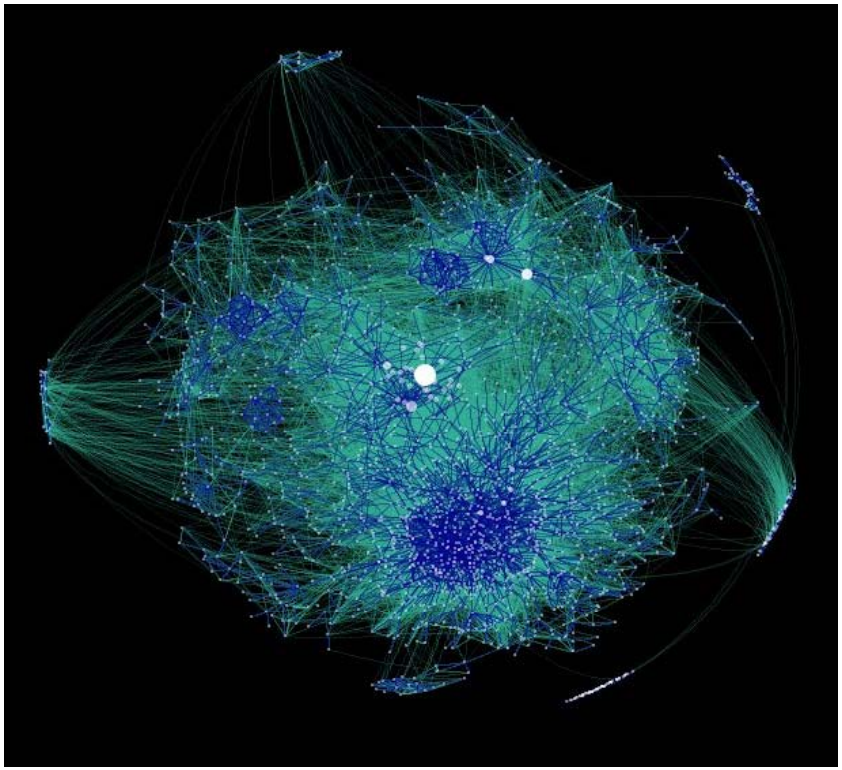
1	Google	11	MySpace
2	Yahoo!	12	Google India
3	Facebook	13	Google Germany
4	YouTube	14	Twitter
5	Windows Live	15	QQ.Com
6	Wikipedia	16	RapidShare
7	Blogger	17	Microsoft Corporation
8	Microsoft Network (MSN)	18	Google France
9	Baidu.com	19	WordPress.com
10	Yahoo! Japan	20	Google UK

- 40% of the top 20 websites are social media sites

Social Media's Important Role



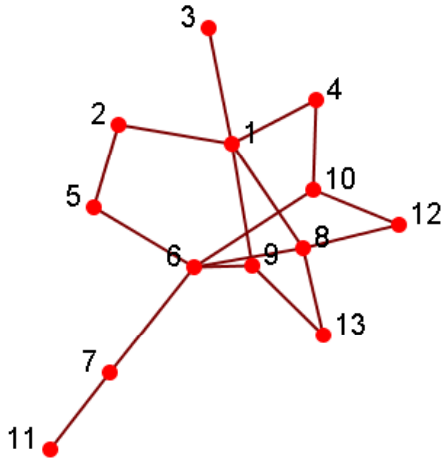
"social networks will complement, and may replace, some government functions,"



SOCIAL NETWORKS AND DATA MINING

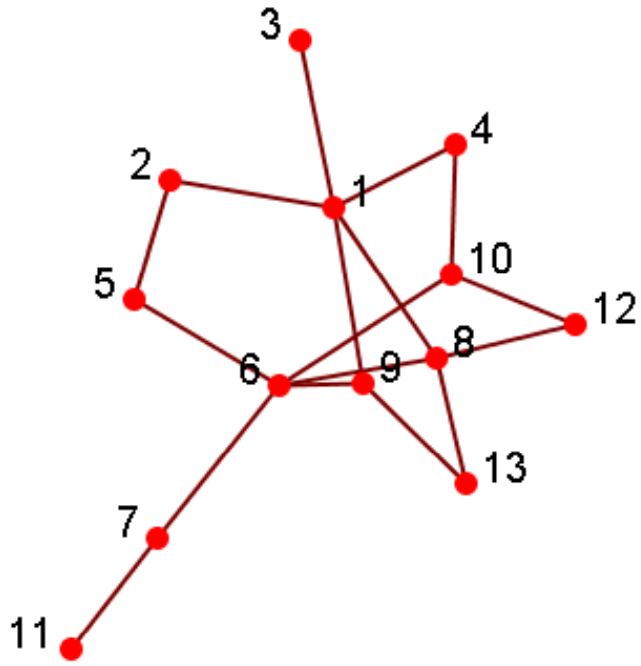
Social Networks

- A social structure made of nodes (individuals or organizations) that are related to each other by various interdependencies like friendship, kinship, etc.
- Graphical representation
  - Nodes = members
  - Edges = relationships
- Various realizations
  - Social bookmarking (Del.icio.us)
  - Friendship networks (facebook, myspace)
  - Blogosphere
  - Media Sharing (Flickr, Youtube)
  - Folksonomies



Sociomatrix

Social networks can also be represented in matrix form



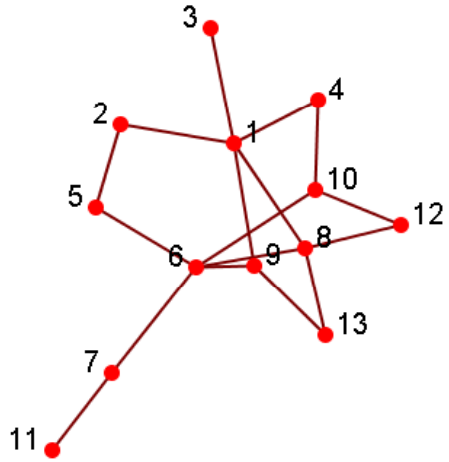
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	1	0	0	0	1	1	0	0	0	0
2	1	0	0	0	1	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0
...													

Social Computing and Data Mining

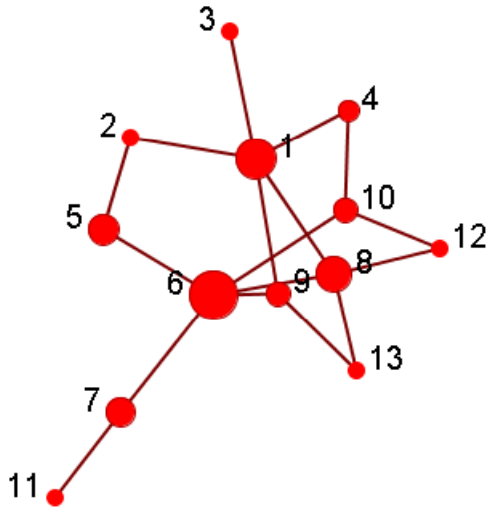
- Social computing is concerned with the study of social behavior and social context based on computational systems.
- Data Mining Related Tasks
  - Centrality Analysis
  - Community Detection
  - Classification
  - Link Prediction
  - Viral Marketing
  - Network Modeling

Centrality Analysis/Influence Study

- Identify the most important actors in a social network
- Given: a social network
- Output: a list of top-ranking nodes



Top 5 important nodes:
6, 1, 8, 5, 10



(Nodes resized by Importance)

Community Detection

- A community is a set of nodes between which the interactions are (relatively) frequent

  a.k.a. group, subgroup, module, cluster

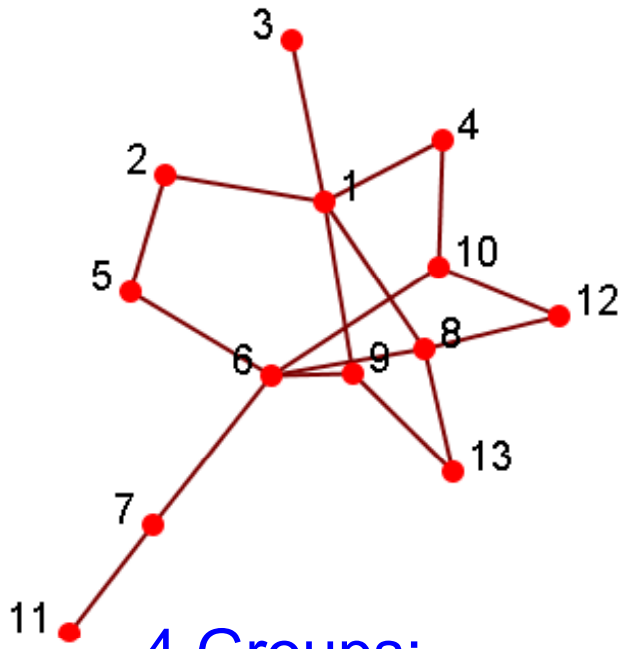
- Community detection

  a.k.a. grouping, clustering, finding cohesive subgroups

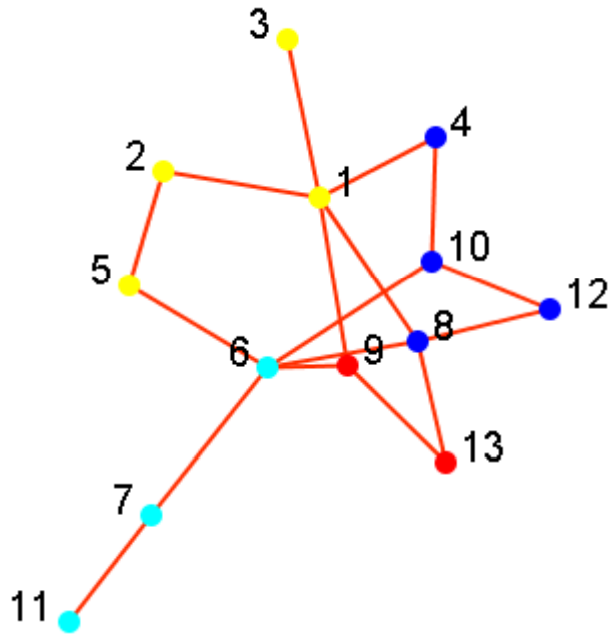
  - Given: a social network
  - Output: community membership of (some) actors

- Applications
  - Understanding the interactions between people
  - Visualizing and navigating huge networks
  - Forming the basis for other tasks such as data mining

Visualization after Grouping



4 Groups:
{1,2,3,5}
{4,8,10,12}
{6,7,11}
{9,13}

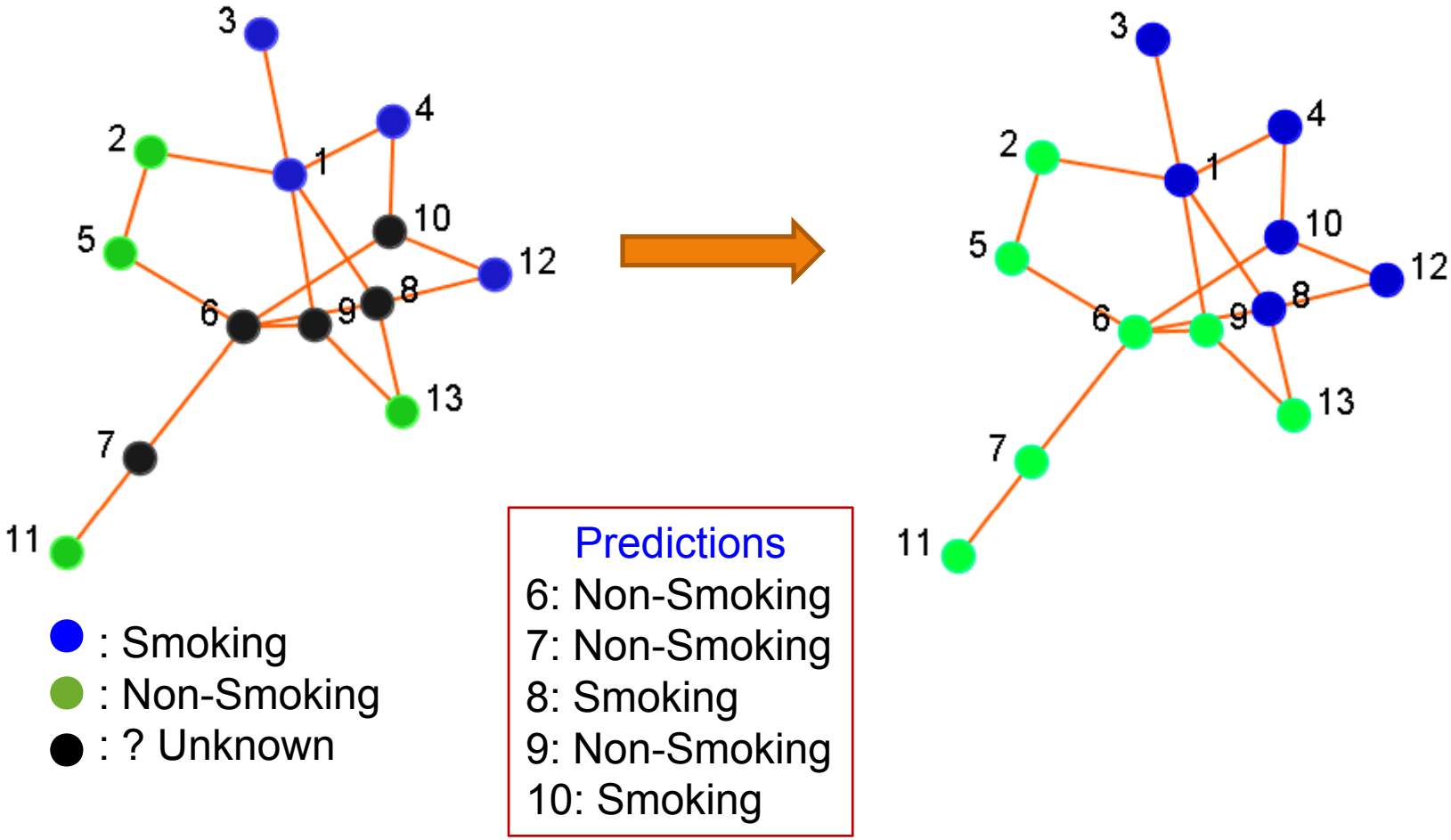


(Nodes colored by Community Membership)

Classification

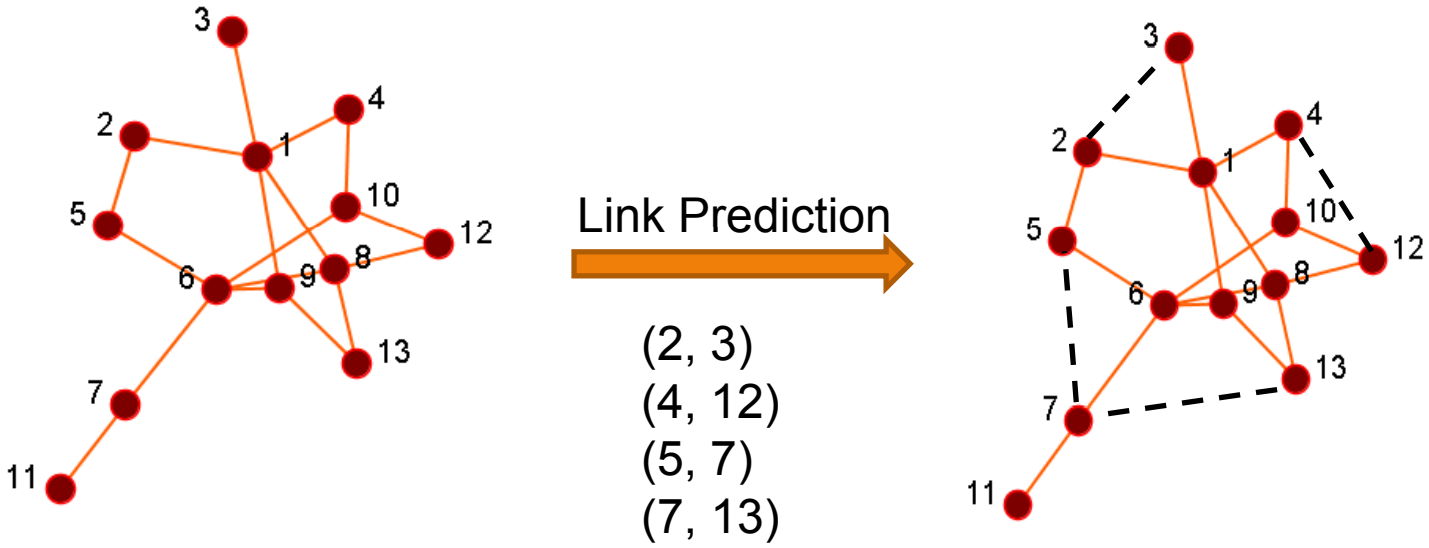
- User Preference or Behavior can be represented as class labels
  - Whether or not clicking on an ad
  - Whether or not interested in certain topics
  - Subscribed to certain political views
  - Like/Dislike a product
- Given
  - A social network
  - Labels of some actors in the network
- Output
  - Labels of remaining actors in the network

Visualization after Prediction



Link Prediction

- Given a social network, predict which nodes are likely to get connected
- Output a list of (ranked) pairs of nodes
- Example: Friend recommendation in Facebook

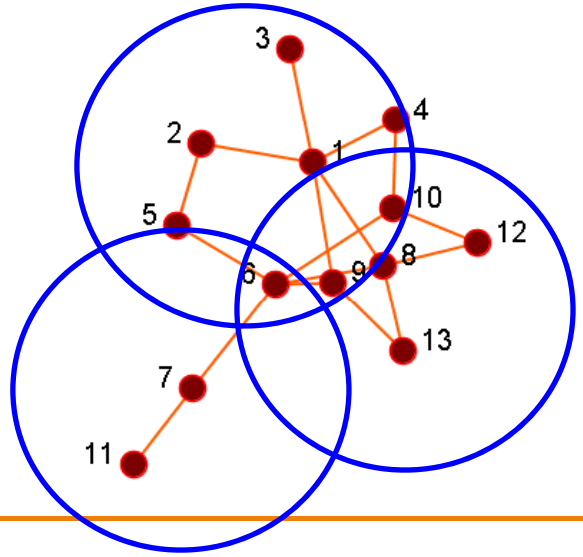


Viral Marketing/Outbreak Detection

- Users have different social capital (or network values) within a social network, hence, how can one make best use of this information?
- Viral Marketing: find out a set of users to provide coupons and promotions to influence other people in the network so my benefit is maximized
- Outbreak Detection: monitor a set of nodes that can help detect outbreaks or interrupt the infection spreading (e.g., H1N1 flu)
- Goal: given a limited budget, how to maximize the overall benefit?

An Example of Viral Marketing

- Find the coverage of the whole network of nodes with the minimum number of nodes
- How to realize it – an example
  - Basic Greedy Selection: Select the node that maximizes the utility, remove the node and then repeat



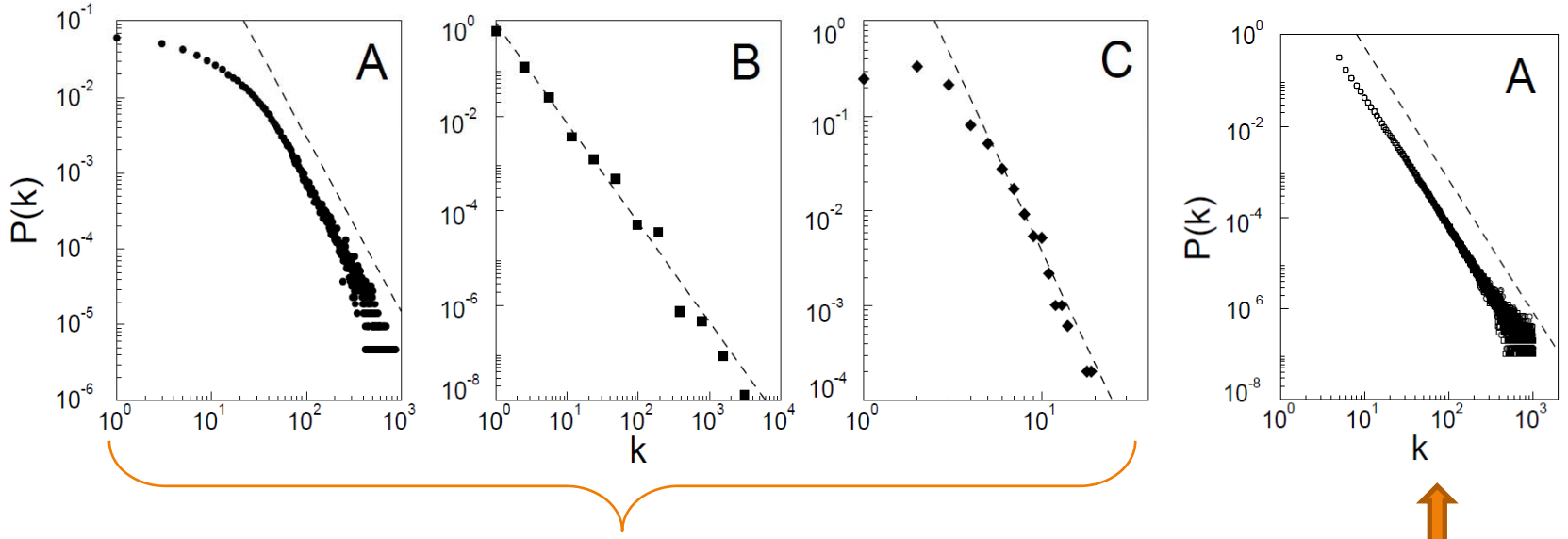
- Select Node 1
- Select Node 8
- Select Node 7

Node 7 is not a node with high centrality!

Network Modeling

- Large Networks demonstrate statistical patterns:
  - Small-world effect (e.g., 6 degrees of separation)
  - Power-law distribution (a.k.a. scale-free distribution)
  - Community structure (high clustering coefficient)
- Model the network dynamics
  - Find a mechanism such that the statistical patterns observed in large-scale networks can be reproduced.
  - Examples: random graph, preferential attachment process
- Used for simulation to understand network properties
  - Thomas Shelling's famous simulation: What could cause the segregation of white and black people
  - Network robustness under attack

Comparing Network Models



observations over various real-word large-scale networks

outcome of a network model

(Figures borrowed from "*Emergence of Scaling in Random Networks*")

Social Computing Applications

- Advertizing via Social Networking
- Behavior Modeling and Prediction
- Epidemic Study
- Collaborative Filtering
- Crowd Mood Reader
- Cultural Trend Monitoring
- Visualization
- Health 2.0

GENERAL EVALUATION MEASURES

Basic Evaluation and Metrics

- Assessment is an essential step
  - Comparing with some ground truth if available
- Obviously, various tasks may require different ways of performance evaluation
  - Ranking
  - Clustering
  - Classification
- An understanding of these concepts will help us to develop more pertinent evaluation methods.

Measuring a Ranked List

- Normalized Discounted Cumulative Gain (NDCG)
- Measuring relevance of returned search result
  - Multi levels of relevance *(r)*: irrelevant (0), borderline (1), relevant (2)
  - Each relevant document contributes some gain to be cumulated
  - Gain from low ranked documents is discounted
  - Normalized by the maximum DCG

$$CG(d_1,...,d_n) = \sum_{i=1}^{n} r_i$$

$$DCG(d_1,...,d_n) = r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2 i}$$

$$MaxDCG = R_1 + \sum_{i=2}^{n} \frac{R_i}{\log_2 i}$$

$$NDCG(d_1,...,d_n) = DCG(d_1,...,d_n) / MaxDCG$$

NDCG - Example

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

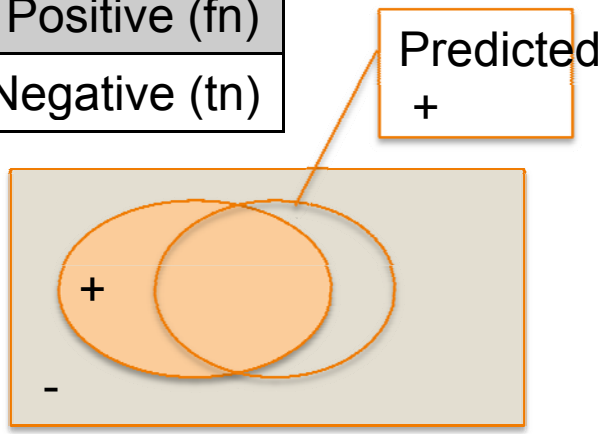
$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

Measuring a Classification Result

- Confusion Matrix

	Prediction (+)	**Prediction (-)**
Truth (+)	True Positive (tp)	False Positive (fn)
Truth (-)	False Positive (fp)	True Negative (tn)



- Measures:

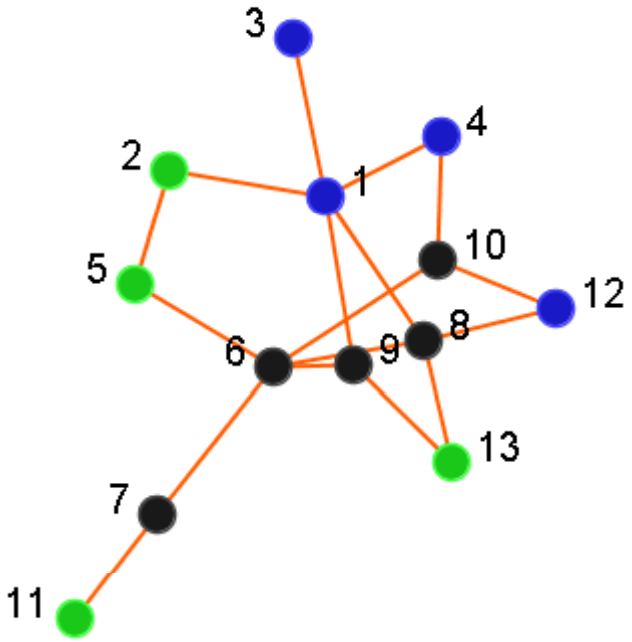
$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

$$precision = \frac{tp}{\Pr ediction(+)} = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{Truth(+)} = \frac{tp}{tp + fn}$$

$$F - measure = \frac{2 \bullet precision \bullet recall}{precision + recall}$$

F-measure Example



● : Smoking
● : Non-Smoking
● : ? Unknown

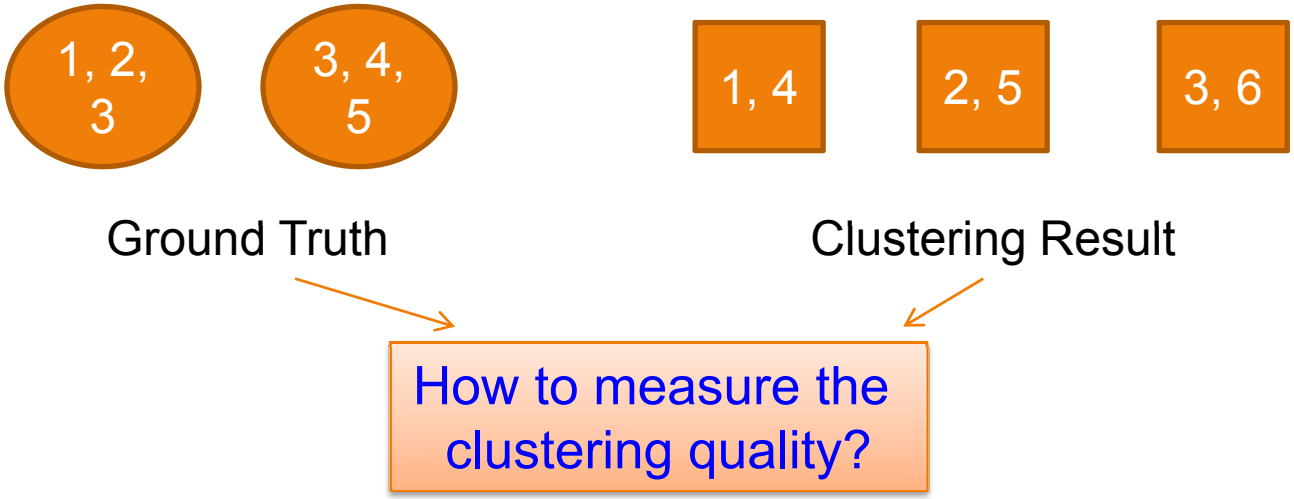
Predictions
6: Non-Smoking
7: Non-Smoking
8: Smoking
9: Non-Smoking
10: Smoking

Truth
6: Smoking
7: Non-Smoking
8: Smoking
9: Smoking
10: Smoking

	Truth (+)	Truth (-)
Prediction (+)	2 (node 8, 10)	0
Prediction (-)	2 (node 6, 9)	1 (node 7)

Accuracy = (2+1)/ 5 = 60%
Precision = 2/(2+0)= 100%
Recall = 2/(2+2) = 50%
F-measure= 2*100% * 50% / (100% + 50%) = 2/3

Measuring a Clustering Result



- The number of communities after grouping can be different from the ground truth
- No clear community correspondence between clustering result and the ground truth
- Normalized Mutual Information can be used

Normalized Mutual Information

- Entropy: the information contained in a distribution

$$H(X) = \sum_{x \in X} p(x) \log p(x)$$

- Mutual Information: the shared information between two distributions

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right)$$

- Normalized Mutual Information (between 0 and 1)

$$NMI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

- Consider a partition as a distribution (probability of one node falling into one community), we can compute the matching between two clusterings

NMI

$$H(X) = \sum_{x \in X} p(x) \log p(x)$$

$$H(\pi^a) = \sum_{h}^{k^{(a)}} \frac{n_h^a}{n} \log(\frac{n_h^a}{n})$$

$$H(\pi^b) = \sum_{\ell}^{k^{(b)}} \frac{n_\ell^b}{n} \log(\frac{n_\ell^b}{n})$$

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right)$$

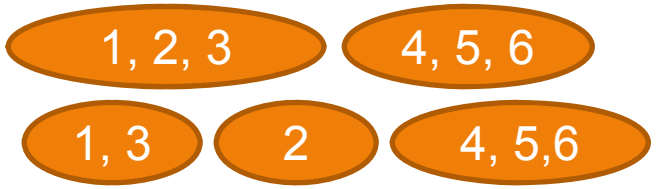
$$I(\pi^a, \pi^b) = \sum_{h} \sum_{\ell} \frac{n_{h,\ell}}{n} \log \left(\frac{\frac{n_{h,\ell}}{n}}{\frac{n_h^a}{n} \frac{n_\ell^b}{n}} \right)$$

$$NMI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

$$NMI(\pi^a, \pi^b) = \frac{\sum_{h=1}^{k^{(a)}} \sum_{\ell=1}^{k^{(b)}} n_{h,\ell} \log \left(\frac{n \cdot n_{h,l}}{n_h^{(a)} \cdot n_\ell^{(b)}} \right)}{\sqrt{\left(\sum_{h=1}^{k^{(a)}} n_h^{(a)} \log \frac{n_h^a}{n} \right) \left(\sum_{\ell=1}^{k^{(b)}} n_\ell^{(b)} \log \frac{n_\ell^b}{n} \right)}}$$

NMI-Example

- Partition a: [1, 1, 1, 2, 2, 2]
- Partition b: [1, 2, 1, 3, 3, 3]



$n = 6$

$k^{(a)} = 2$

$k^{(b)} = 3$

	n_h^a
h=1	3
h=2	3

	n_l^b
l=1	2
l=2	1
l=3	3

$n_{h,l}$	l=1	l=2	l=3
h=1	2	1	0
h=2	0	0	3

$$NMI(\pi^a, \pi^b) = \frac{\sum_{h=1}^{k^{(a)}} \sum_{\ell=1}^{k^{(b)}} n_{h,\ell} \log\left(\frac{n \cdot n_{h,l}}{n_h^{(a)} \cdot n_\ell^{(b)}}\right)}{\sqrt{\left(\sum_{h=1}^{k^{(a)}} n_h^{(a)} \log\frac{n_h^a}{n}\right)\left(\sum_{\ell=1}^{k^{(b)}} n_\ell^{(b)} \log\frac{n_\ell^b}{n}\right)}} \quad = 0.8278$$

Outline

- Social Media
- Data Mining Tasks
- Evaluation

- Principles of Community Detection
- Communities in Heterogeneous Networks
- Evaluation Methodology for Community Detection

- Behavior Prediction via Social Dimensions
- Identifying Influential Bloggers in a Community
  - A related tutorial on Blogosphere

PRINCIPLES OF COMMUNITY DETECTION

Communities

- Community: “subsets of actors among whom there are relatively strong, direct, intense, frequent or positive ties.”

  -- Wasserman and Faust, *Social Network Analysis, Methods and Applications*
- Community is a set of actors interacting with each other *frequently*
  - e.g. people attending this conference
- A set of people without interaction is NOT a community
  - e.g. people waiting for a bus at station but don’t talk to each other
- People form communities in Social Media

Example of Communities

Communities from Facebook

Name: **Social Computing**
Type: Organizations
Members: 14 members

Name: **Social Computing**
Type: Internet & Technology
Members: 12 members

Name: **Social Computing Magazine**
Type: Internet & Technology
Members: 34 members

Name: **Trustworthy Social Computing**
Type: Internet & Technology
Members: 28 members

Name: **Social Computing for Business**
Type: Internet & Technology
Members: 421 members

Name: **UCLA Social Sciences Computing**
Type: Internet & Technology
Members: 22 members

Name: **Social Media and Computing**
Type: Organizations
Members: 6 members

Communities from Flickr

! * Urban LIFE in Metropolis ////
4,286 members | 31 discussions | 89,645 items | Created 46 months ago | Join?
UrbanLIFE, People, Parties, Dance, Musik, Life, Love, Culture, Food and Everything what we could imagine by hearing that word URBANLIFE! Have some FUN! Please add... (more)

Islam Is The Way Of Life (Muslim World)
619 members | 13 discussions | 2,685 items | Created 23 months ago | Join?
The word islām is derived from the Arabic verb aslama, which means to accept, surrender or submit. Thus, Islam means submission to and acceptance of God, and believers must... (more)

*** THE CELEBRATION OF ~LIFE~ (Post1~Award1) [only living things]**
4,871 members | 22 discussions | 40,519 items | Created 21 months ago | Join?
WELCOME to THE CELEBRATION OF ~LIFE~ (Post1~Award1) PLEASE INVITE & COMMENT USING only THE CODES FOUND BELOW! ☆ ☆ This group is for sharing BEAUTIFUL, TOP QUALITY images... (more)

"Enjoy Life!"
2,027 members | 10 discussions | 39,916 items | Created 23 months ago | Join?
There are lovely moments and adorable scenes in our lives. Some are in front of you, and some are just waiting to be discovered. A gaze from someone we love, might touch the... (more)

Baby's life
2,047 members | 185 discussions | 30,302 items | Created 32 months ago | Join?
This group is designed to highlight milestones and important events in your baby's life (ie 1st time smiling/crawling/sitting in a high chair/reading/playing etc). It can also be... (more)

Only group members s
pool

Pond Life
903 members | 20 discussions | 6,877 items | Created 32 months ago | Join?
Pic of the week: chosen from the pool by the group admins. Nuphar by guus timpers Pond Life is a group for all aquatic flora and fauna. Koi ponds, wildlife ponds, garden ponds,... (more)

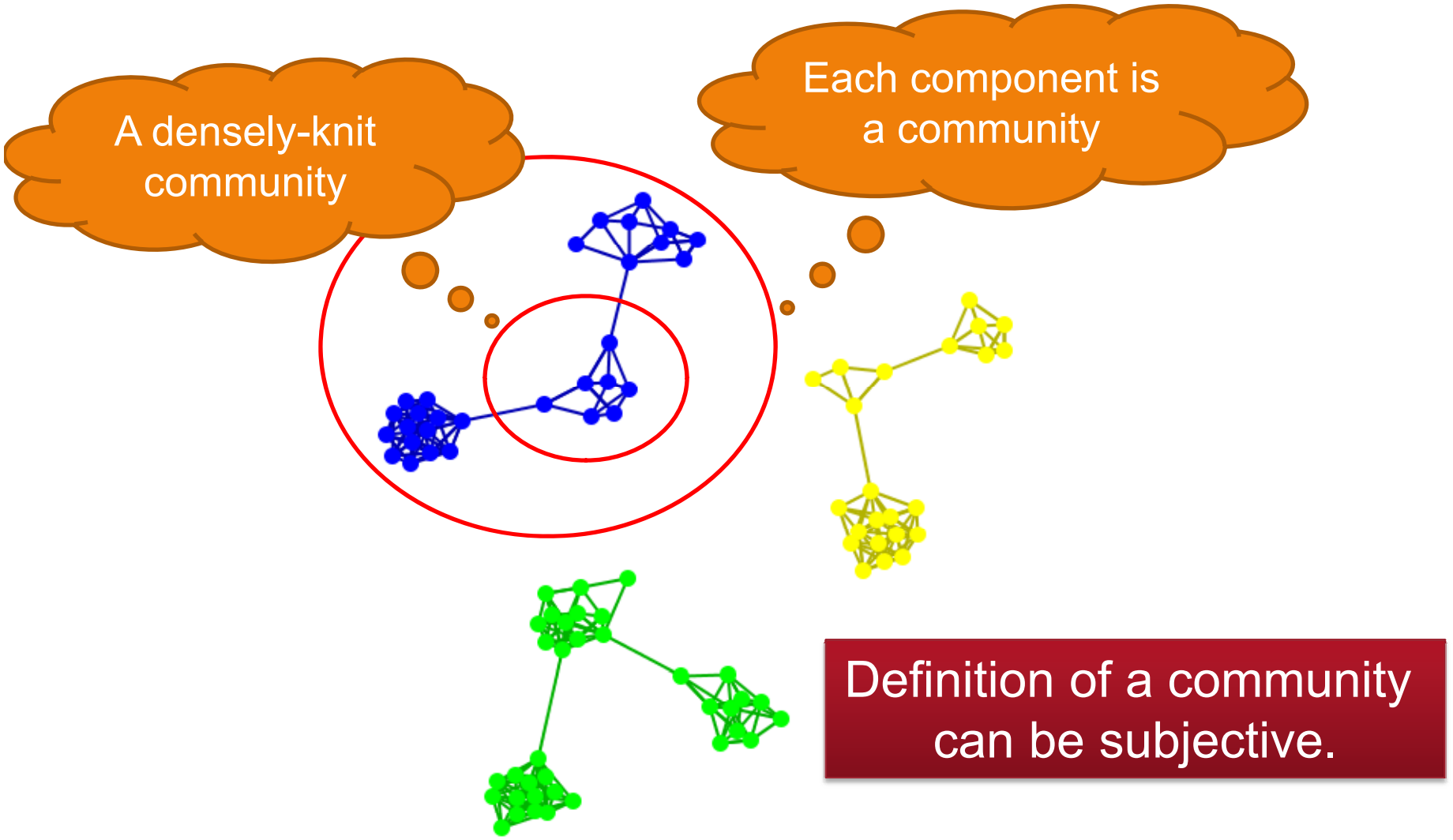
Why Communities in Social Media?

- Human beings are social
- Part of Interactions in social media is a glimpse of the physical world
- People are connected to friends, relatives, and colleagues in the real world as well as online
- Easy-to-use social media allows people to extend their social life in unprecedented ways
  - Difficult to meet friends in the physical world, but much easier to find friend online with similar interests

Community Detection

- Community Detection: “formalize the strong social groups based on the social network properties”
- Some social media sites allow people to join groups, is it necessary to extract groups based on network topology?
  - Not all sites provide community platform
  - Not all people join groups
- Network interaction provides rich information about the relationship between users
  - Groups are *implicitly* formed
  - Can complement other kinds of information
  - Help network visualization and navigation
  - Provide basic information for other tasks

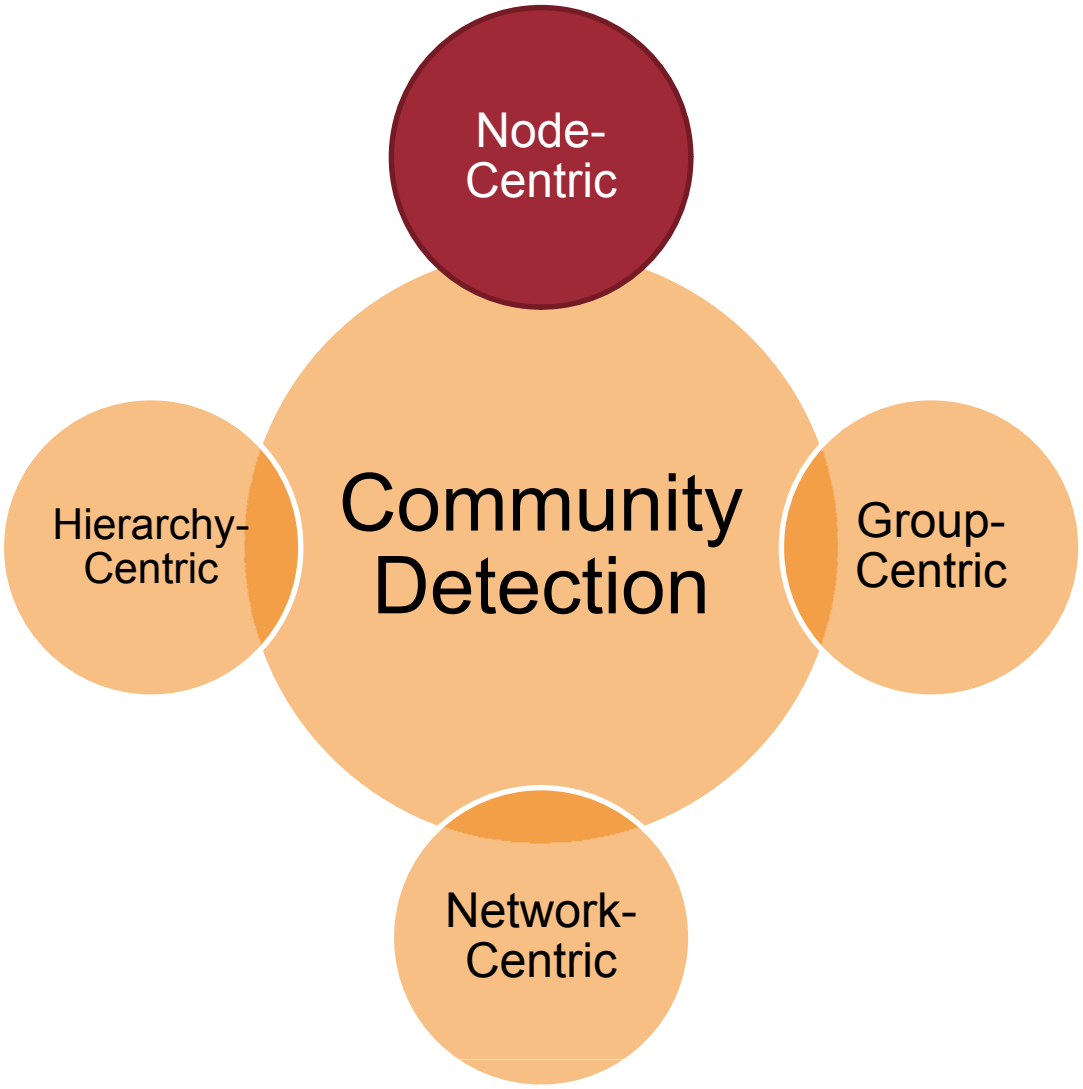
Subjectivity of Community Definition



Taxonomy of Community Criteria

- Criteria vary depending on the tasks
- Roughly, community detection methods can be divided into 4 categories (not exclusive):
- Node-Centric Community
  - Each node in a group satisfies certain properties
- Group-Centric Community
  - Consider the connections within a group as a whole. The group has to satisfy certain properties without zooming into node-level
- Network-Centric Community
  - Partition the whole network into several disjoint sets
- Hierarchy-Centric Community
  - Construct a hierarchical structure of communities

Node-Centric Community Detection

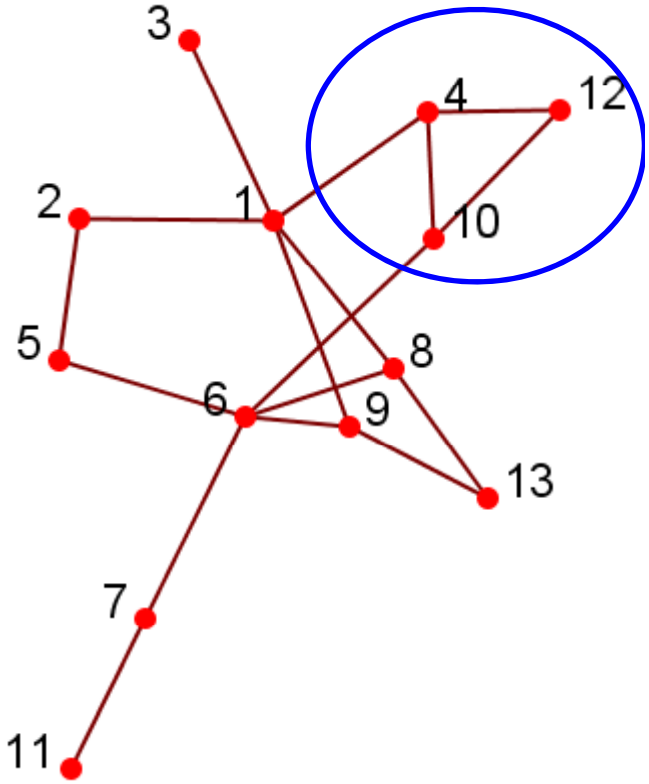


Node-Centric Community Detection

- Nodes satisfy different properties
  - Complete Mutuality
    - cliques
  - Reachability of members
    - k-clique, k-clan, k-club
  - Nodal degrees
    - k-plex, k-core
  - Relative frequency of Within-Outside Ties
    - LS sets, Lambda sets
- Commonly used in traditional social network analysis
- Here, we discuss some representative ones

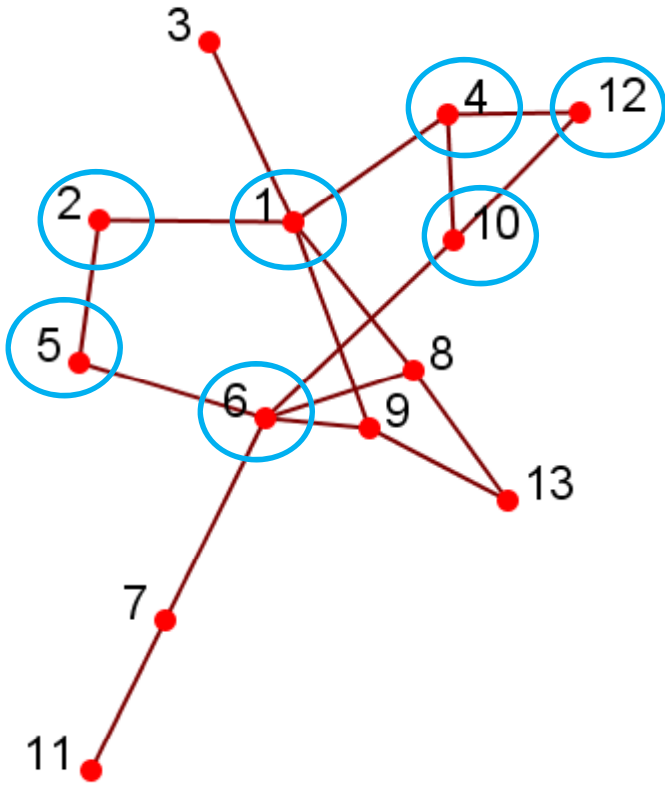
Complete Mutuality: Clique

- A maximal complete subgraph of three or more nodes all of which are adjacent to each other
  - NP-hard to find the maximal clique
  - Recursive pruning: To find a clique of size k, remove those nodes with less than k-1 degrees
  - Very strict definition, unstable
  - Normally use cliques as a core or seed to explore larger communities



Geodesic

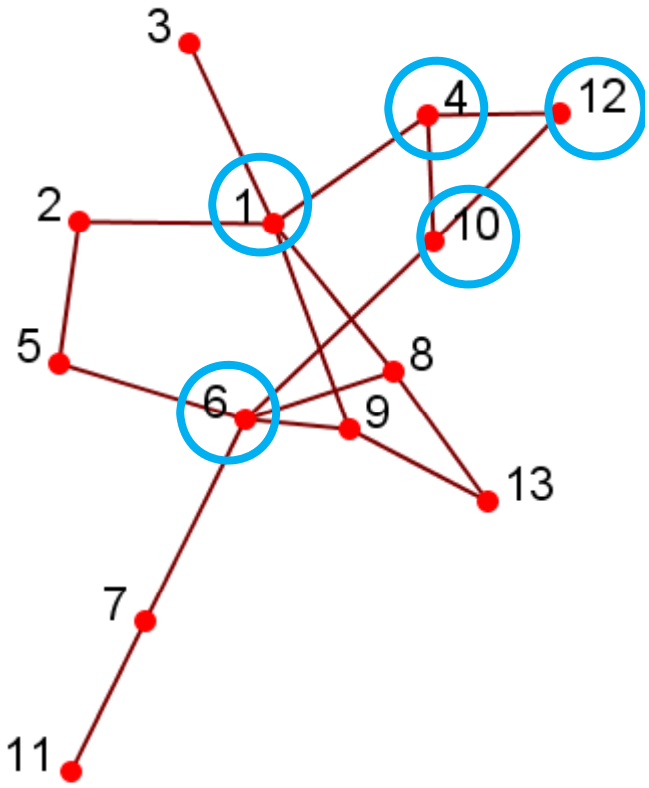
- Reachability is calibrated by the Geodesic distance
- Geodesic: a shortest path between two nodes (12 and 6)
  - Two paths: 12-4-1-2-5-6, 12-10-6
  - 12-10-6 is a geodesic
- Geodesic distance: #hops in geodesic between two nodes
  - e.g., d(12, 6) = 2, d(3, 11)=5
- Diameter: the maximal geodesic distance for any 2 nodes in a network
  - #hops of the longest shortest path



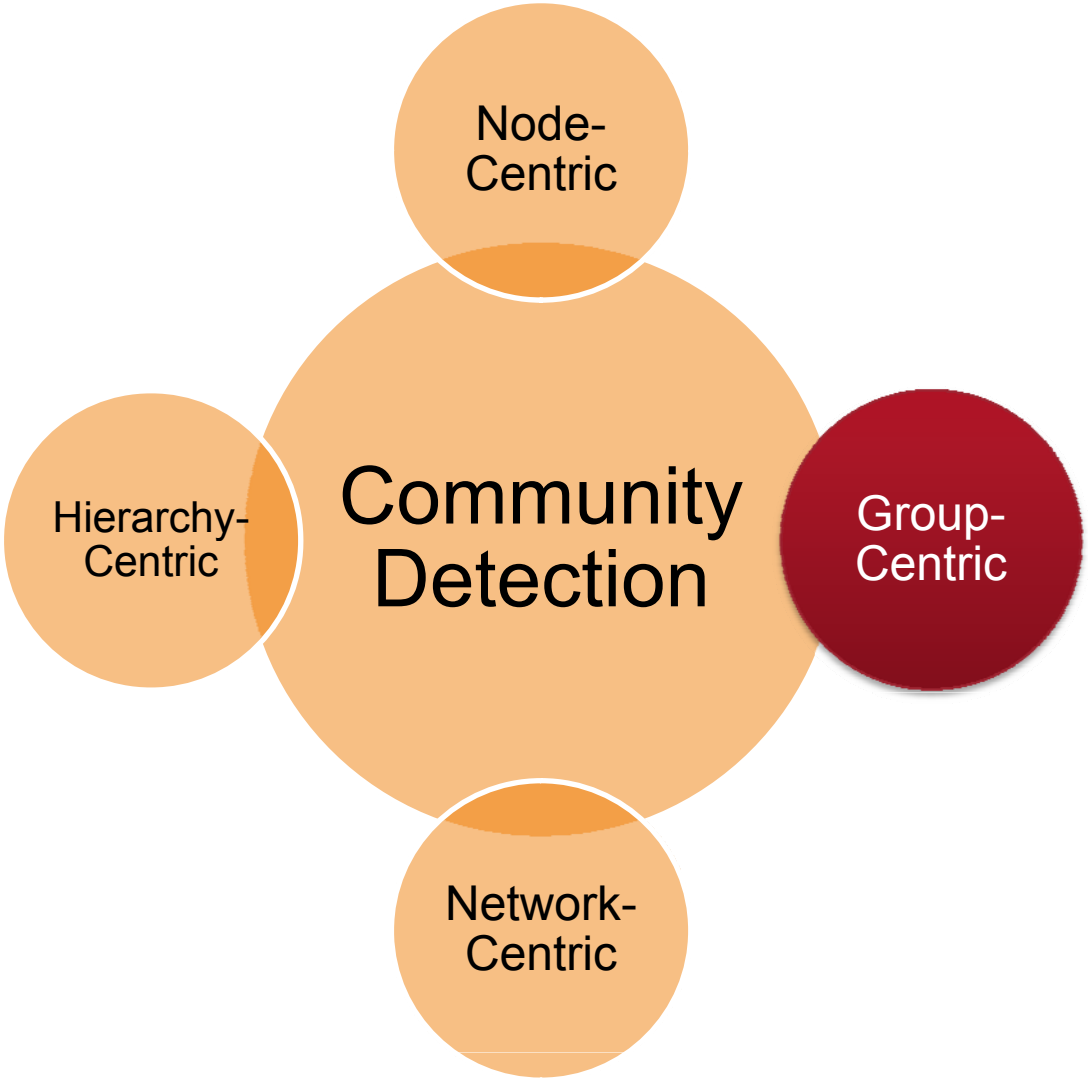
Diameter = 5

Reachability: k-clique, k-club

- Any node in a group should be reachable in k hops
- k-clique: a maximal subgraph in which the largest geodesic distance between any nodes <= k
- A k-clique can have diameter larger than k within the subgraph
  - e.g., 2-clique {12, 4, 10, 1, 6}
  - Within the subgraph d(1, 6) = 3
- k-club: a substructure of diameter <= k
  - e.g., {1,2,5,6,8,9}, {12, 4, 10, 1} are 2-clubs



Group-Centric Community Detection



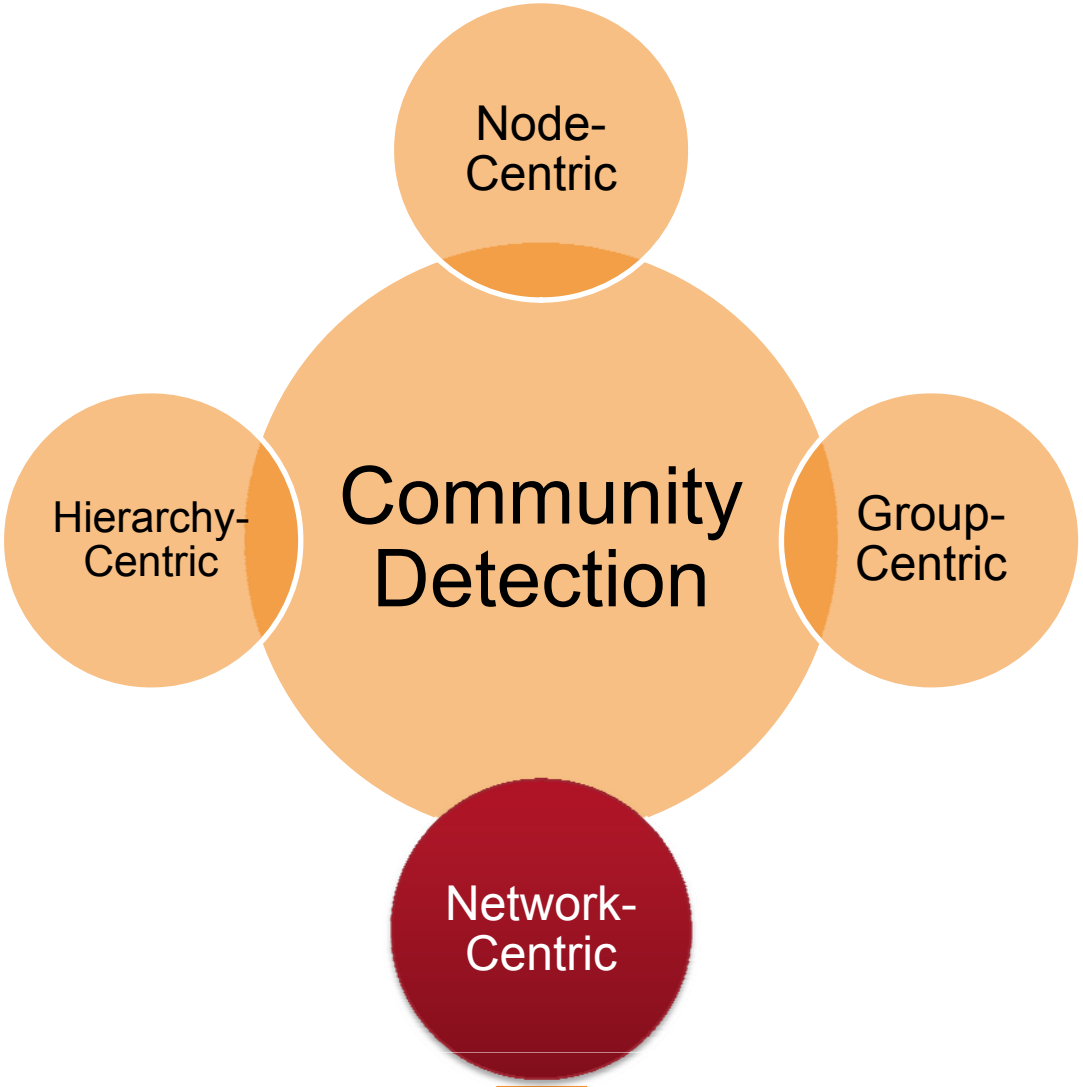
Group-Centric Community Detection

- Consider the connections within a group as whole,
- OK for some nodes to have low connectivity
- A subgraph with V_s nodes and E_s edges is a γ-dense quasi-clique if

$$\frac{E_s}{V_s(V_s-1)/2} \geq \gamma$$

- Recursive pruning:
  - Sample a subgraph, find a maximal γ-dense quasi-clique (the resultant size = k)
  - Remove the nodes that
    - whose degree < kγ
    - all their neighbors with degree < kγ

Network-Centric Community Detection



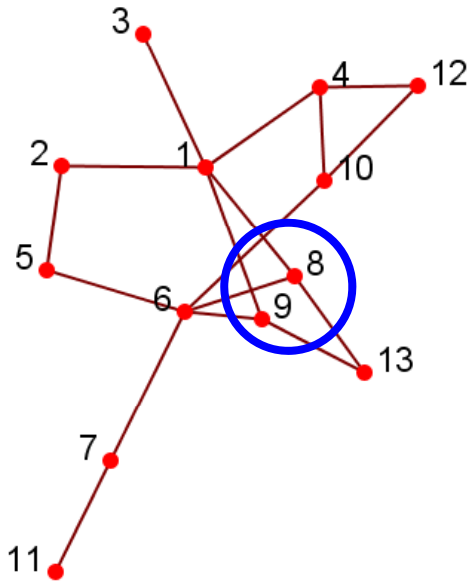
Network-Centric Community Detection

- To form a group, we need to consider the connections of the nodes globally.

- Goal: partition the network into disjoint sets
  - Groups based on Node Similarity
  - Groups based on Latent Space Model
  - Groups based on Block Model Approximation
  - Groups based on Cut Minimization
  - Groups based on Modularity Maximization

Node Similarity

- Node similarity is defined by how similar their interaction patterns are
- Two nodes are structurally equivalent if they connect to the same set of actors
  - e.g., nodes 8 and 9 are structurally equivalent
- Groups are defined over equivalent nodes
  - Too strict
  - Rarely occur in a large-scale
  - Relaxed equivalence class is difficult to compute
- In practice, use vector similarity
  - e.g., cosine similarity, Jaccard similarity



Vector Similarity

a vector

structurally equivalent

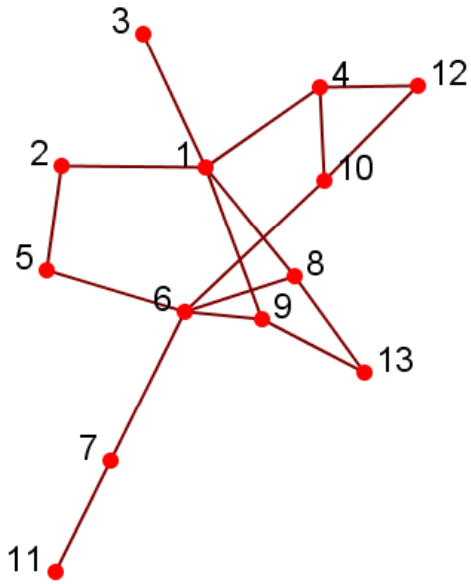
	1	2	3	4	5	6	7	8	9	10	11	12	13
5		1				1							
8	1					1							1
9	1					1							1

Cosine Similarity: $\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$

$$sim(5,8) = \frac{1}{\sqrt{2} \times \sqrt{3}} = \frac{1}{\sqrt{6}}$$

Jaccard Similarity: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$

$$J(5,8) = \frac{|\{6\}|}{|\{1,2,6,13\}|} = 1/4$$



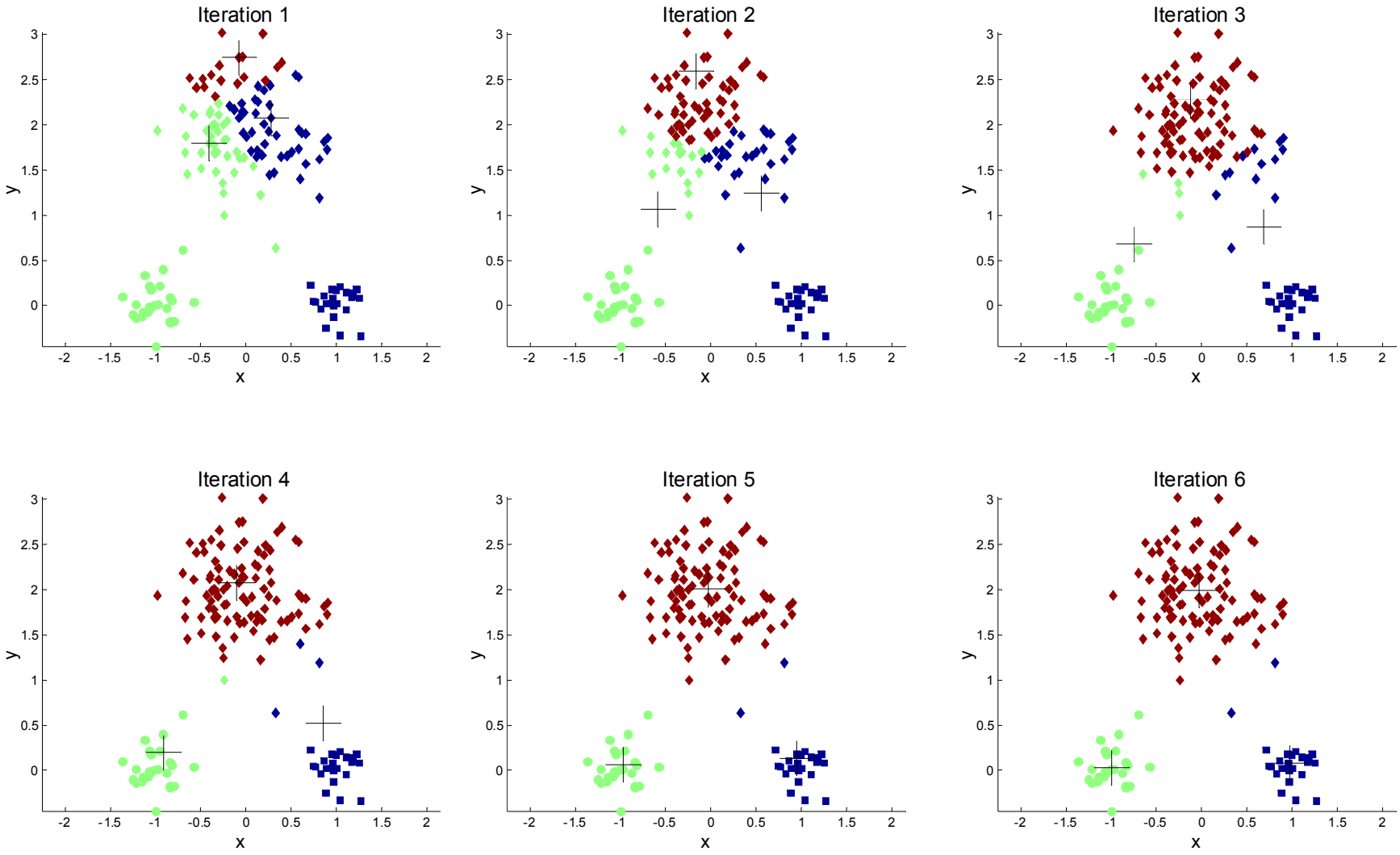
Clustering based on Node Similarity

- For practical use with huge networks:
  - Consider the connections as features
  - Use Cosine or Jaccard similarity to compute vertex similarity
  - Apply classical k-means clustering Algorithm
- K-means Clustering Algorithm
  - Each cluster is associated with a centroid (center point)
  - Each node is assigned to the cluster with the closest centroid

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Illustration of k-means clustering



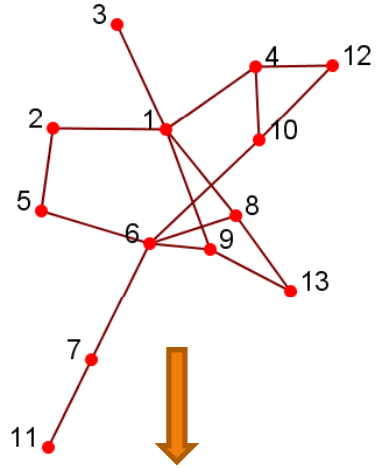
Groups on Latent-Space Models

- Latent-space models: Transform the nodes in a network into a lower-dimensional space such that the distance or similarity between nodes are kept in the Euclidean space
- Multidimensional Scaling (MDS)
  - Given a network, construct a proximity matrix to denote the distance between nodes (e.g. geodesic distance)
  - Let D denotes the *square distance* between nodes
  - $S \in R^{n \times k}$ denotes the coordinates in the lower-dimensional space

$$SS^T = -\frac{1}{2}(I - \frac{1}{n}ee^T)D(I - \frac{1}{n}ee^T) = \Delta(D)$$

  - Objective: minimize the difference $\min \| \Delta(D) - SS^T \|_F$
  - Let $\Lambda = diag(\lambda_1, \cdots, \lambda_k)$ (the top-k eigenvalues of Δ), V the top-k eigenvectors
  - Solution: $S = V\Lambda^{1/2}$
- Apply k-means to S to obtain clusters

MDS-example



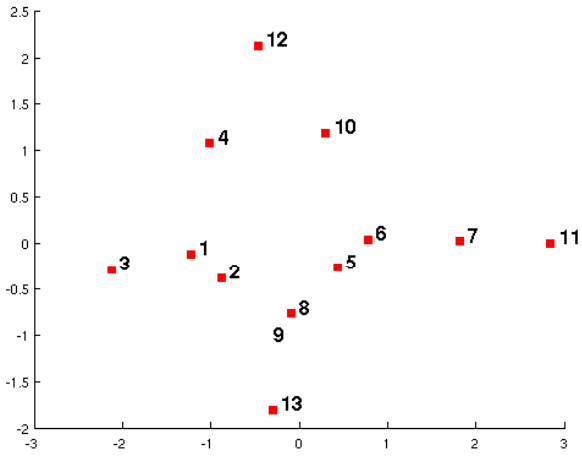
Geodesic Distance Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	1	2	2	3	1	1	2	4	2	2
2	1	0	2	2	1	2	3	2	2	3	4	3	3
3	1	2	0	2	3	3	4	2	2	3	5	3	3
4	1	2	2	0	3	2	3	2	2	1	4	1	3
5	2	1	3	3	0	1	2	2	2	2	3	3	3
6	2	2	3	2	1	0	1	1	1	1	2	2	2
7	3	3	4	3	2	1	0	2	2	2	1	3	3
8	1	2	2	2	2	1	2	0	2	2	3	3	1
9	1	2	2	2	2	1	2	2	0	2	3	3	1
10	2	3	3	1	2	1	2	2	2	0	3	1	3
11	4	4	5	4	3	2	1	3	3	3	0	4	4
12	2	3	3	1	3	2	3	3	3	1	4	0	4
13	2	3	3	3	3	2	3	1	1	3	4	4	0

MDS

S

-1.22	-0.12
-0.88	-0.39
-2.12	-0.29
-1.01	1.07
0.43	-0.28
0.78	0.04
1.81	0.02
-0.09	-0.77
-0.09	-0.77
0.30	1.18
2.85	0.00
-0.47	2.13
-0.29	-1.81

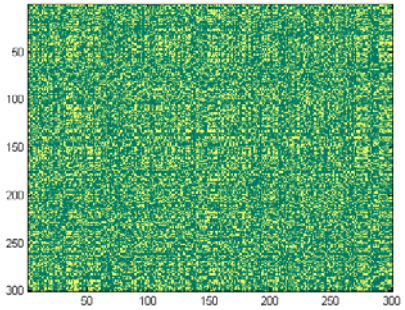


k-means

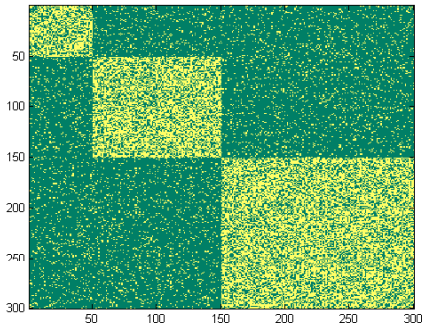
1, 2, 3, 4, 10, 12

5, 6, 7, 8, 9, 11, 13

Block-Model Approximation



After Reordering



Network Interaction Matrix

Block Structure

- Objective: Minimize the difference between an interaction matrix and a block structure

$$\min_{S,\Sigma} \|A - S\Sigma S^T\|_F$$

$$s.t. \quad S \in \{0,1\}^{n\times k}, \Sigma \in R^{k\times k} \text{ is diagonal}$$

S is a community indicator matrix

- Challenge: S is discrete, difficult to solve
- Relaxation: Allow S to be continuous satisfying $S^T S = I_k$
- Solution: the top eigenvectors of A
- Post-Processing: Apply k-means to S to find the partition

Cut-Minimization

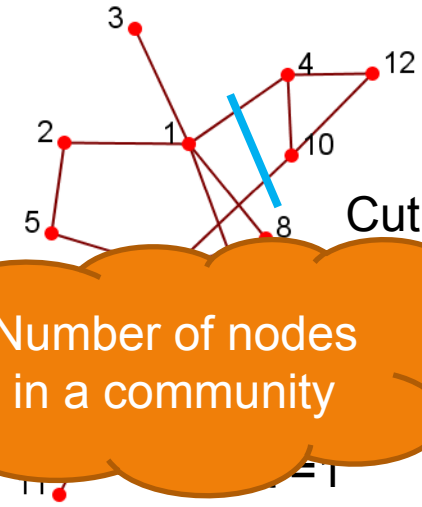
- Between-group interactions should be infrequent
- Cut: number of edges between two sets of nodes
- Objective: minimize the cut

$$cut(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} cut(C_i, \overline{C_i})$$

  - Limitations: often find communities of only one node
  - Need to consider the group size
- Two commonlv-used variants:

$$\text{Ratio-cut}(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C_i})}{|V_i|}$$

$$\text{Normalized-cut}(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C_i})}{vol(V_i)}$$



Number of nodes in a community

Number of within-group Interactions

Graph Laplacian

- Can be relaxed into the following min-trace problem

$$\min_{S \in R^{n \times k}} Tr(S^T L S) \quad s.t.\ S^T S = I$$

- L is the (normalized) Graph Laplacian

$$\begin{aligned} L &= D - A \\ \text{normalized-}L &= I - D^{-1/2} A D^{-1/2} \end{aligned}$$

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$$

- Solution: S are the eigenvectors of L with smallest eigenvalues (except the first one)
- Post-Processing: apply k-means to S
- a.k.a.Spectral Clustering

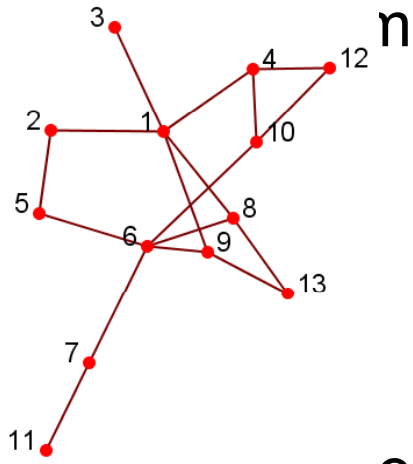
Modularity Maximization

- Modularity measures the group interactions compared with the expected random connections in the group
- In a network with m edges, for two nodes with degree d_i and d_j, the expected random connections are $d_i d_j / 2m$
- The interaction utility in a group:

$$\sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m$$

- To partition the group into multiple groups, we maximize

$$\max \frac{1}{2m} \sum_{C} \sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m$$



Expected Number of edges between 6 and 9 is 5*3/(2*17) = 15/34

Modularity Matrix

- The modularity maximization can also be formulated in matrix form

$$Q = \frac{1}{2m} Tr(S^T B S)$$

- B is the modularity matrix

$$B_{ij} = A_{ij} - d_i d_j / 2m$$

- Solution: top eigenvectors of the modularity matrix

Matrix Factorization Form

- For latent space models, block models, spectral clustering and modularity maximization
- All can be formulated as

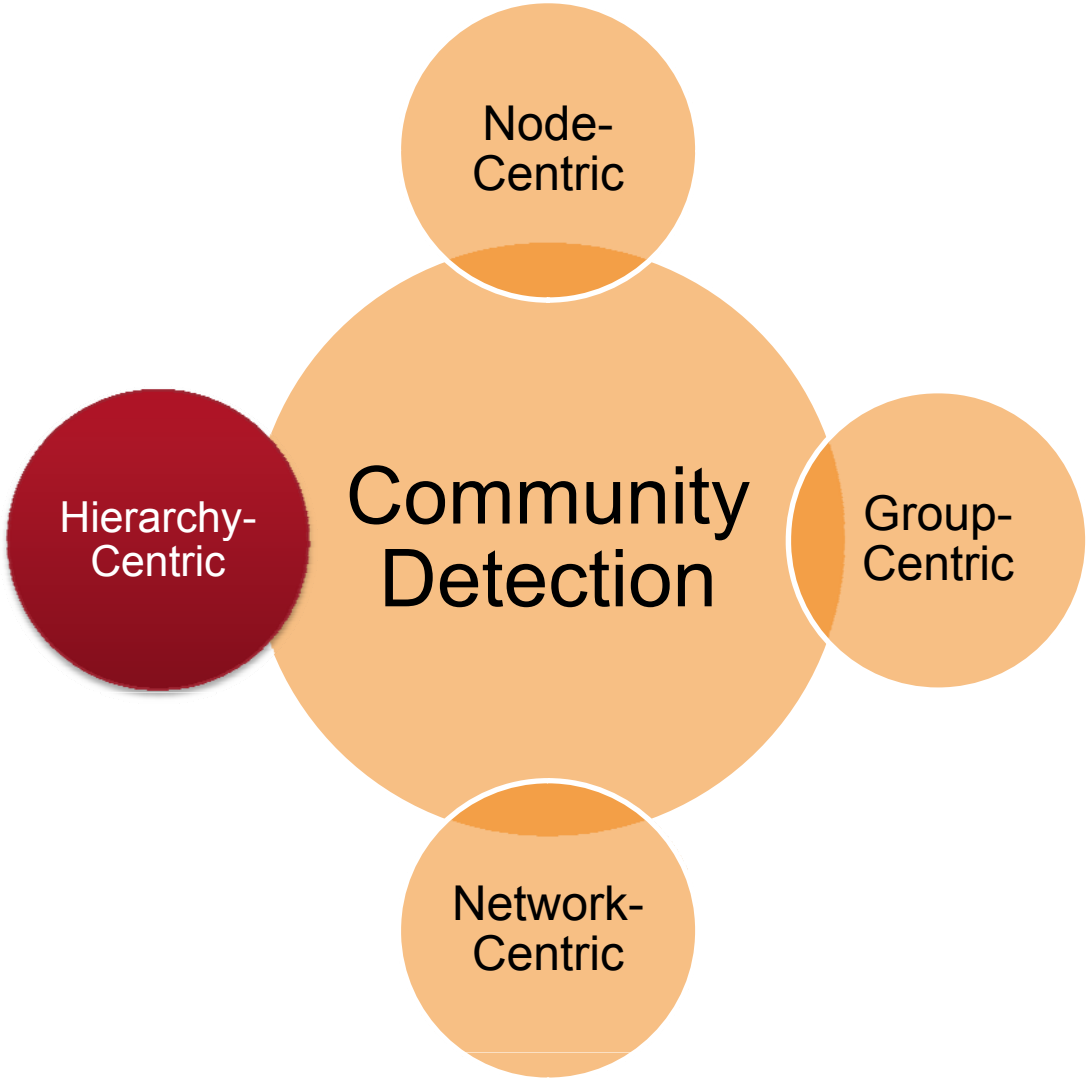
$$\max(\min)_S \quad Tr(S^T X S)$$
$$s.t. \quad S^T S = I$$

X=
- $\Delta(D)$ (Latent Space Models)
- Sociomatrix (Block Model Approximation)
- Graph Laplacian (Cut Minimization)
- Modularity Matrix (Modularity maximization)

Recap of Network-Centric Community

- Network-Centric Community Detection
  - Groups based on Node Similarity
  - Groups based on Latent Space Models
  - Groups based on Cut Minimization
  - Groups based on Block-Model Approximation
  - Groups based on Modularity maximization
- Goal: Partition network nodes into several disjoint sets
- Limitation: Require the user to specify the number of communities beforehand

Hierarchy-Centric Community Detection

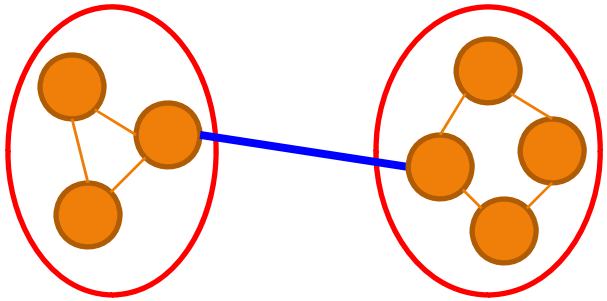


Hierarchy-Centric Community Detection

- Goal: Build a hierarchical structure of communities based on network topology
- Facilitate the analysis at different resolutions
- Representative Approaches:
  - Divisive Hierarchical Clustering
  - Agglomerative Hierarchical Clustering

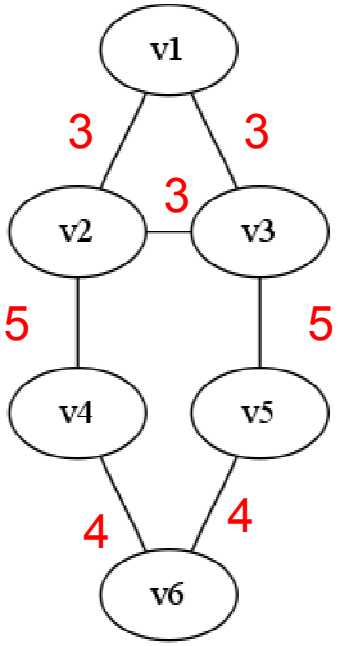
Divisive Hierarchical Clustering

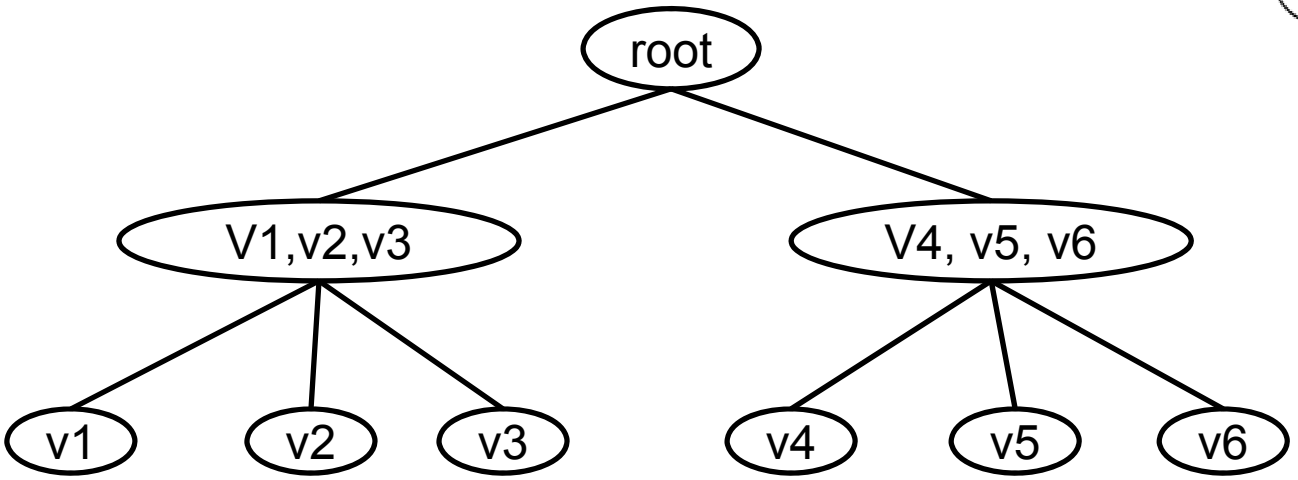
- Divisive Hierarchical Clustering
  - Partition the nodes into several sets
  - Each set is further partitioned into smaller sets
- Network-centric methods can be applied for partition
- One particular example is based on edge-betweenness
- Edge-Betweenness: Number of shortest paths between any pair of nodes that pass through the edge
- Between-group edges tend to have larger edge-betweenness



Divisive clustering on Edge-Betweenness

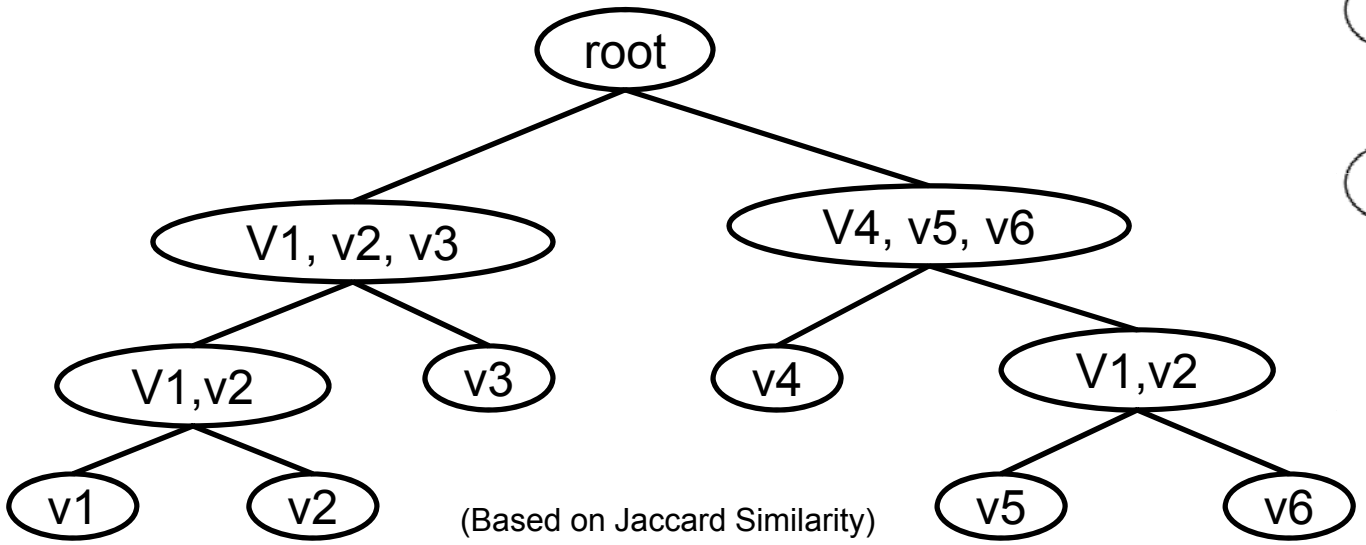
- Progressively remove edges with the highest betweenness
  - Remove e(2,4), e(3, 5)
  - Remove e(4,6), e(5,6)
  - Remove e(1,2), e(2,3), e(3,1)



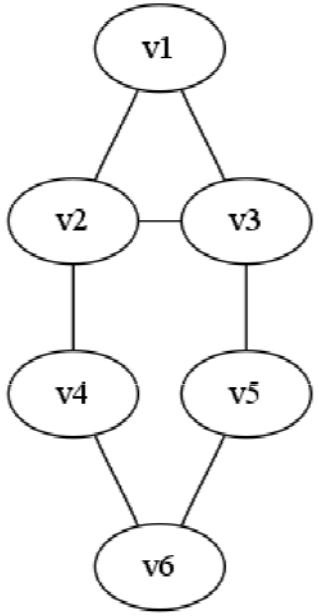


Agglomerative Hierarchical Clustering

- Initialize each node as a community
- Choose two communities satisfying certain criteria and merge them into larger ones
  - Maximum Modularity Increase
  - Maximum Node Similarity

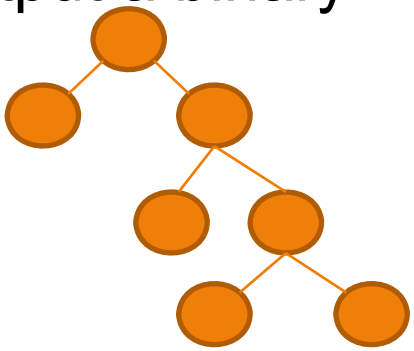


(Based on Jaccard Similarity)



Recap of Hierarchical Clustering

- Most hierarchical clustering algorithm output a binary tree
  - Each node has two children nodes
  - Might be highly imbalanced



- Agglomerative clustering can be very sensitive to the nodes processing order and merging criteria adopted.

- Divisive clustering is more stable, but generally more computationally expensive

Summary of Community Detection

- The Optimal Method?
- It varies depending on applications, networks, computational resources etc.
- Scalability can be a concern for networks in social media
- Other lines of research
  - Communities in directed networks
  - Overlapping communities
  - Community evolution
  - Group profiling and interpretation

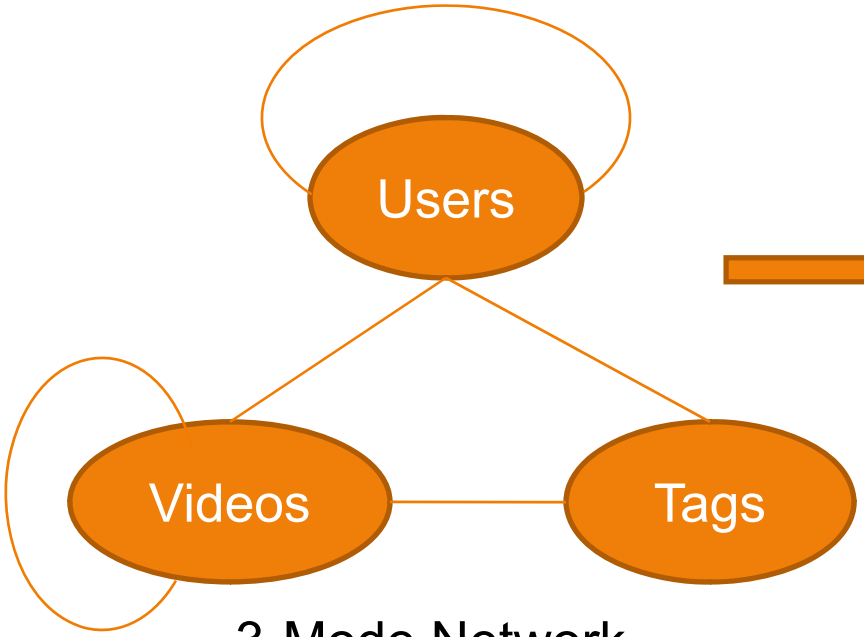
COMMUNITIES IN HETEROGENEOUS NETWORKS

Heterogeneous Network

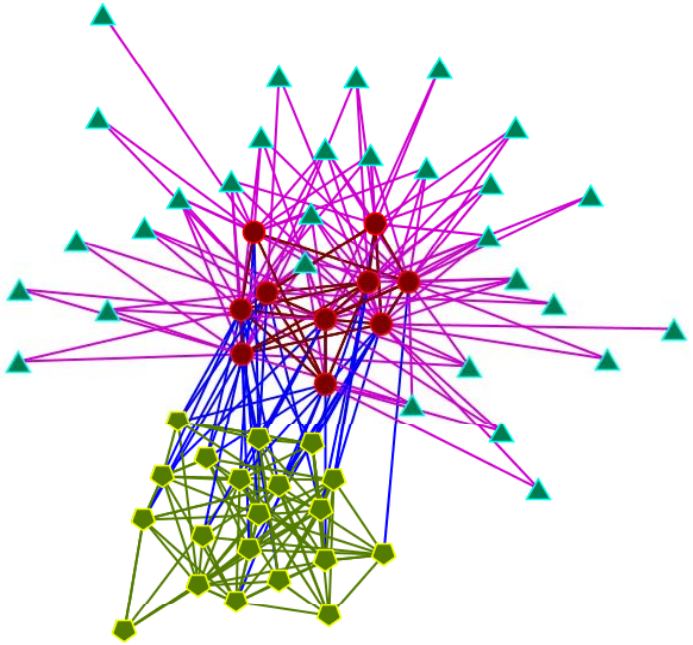
- Heterogeneous kinds of objects in social media
  - YouTube
    - Users, tags, videos, ads
  - Del.icio.us
    - Users, tags, bookmarks
- Heterogeneous types of interactions between actors
  - Facebook
    - Send email, leave a message
    - write a comment, tag photos
  - Same users interacting at different sites
    - Facebook, YouTube, Twitter

Multi-Mode Network

- Networks consists of multiple modes of nodes
- a.k.a. meta network



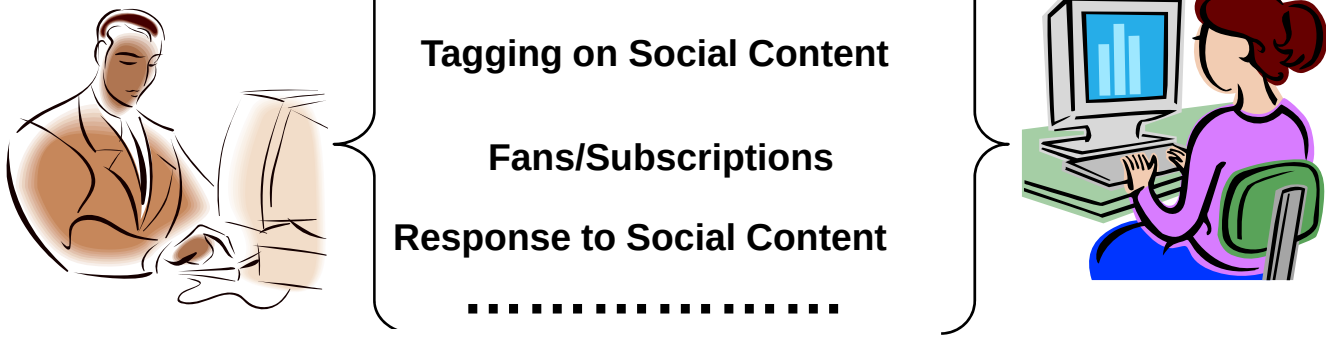
3-Mode Network
in YouTube

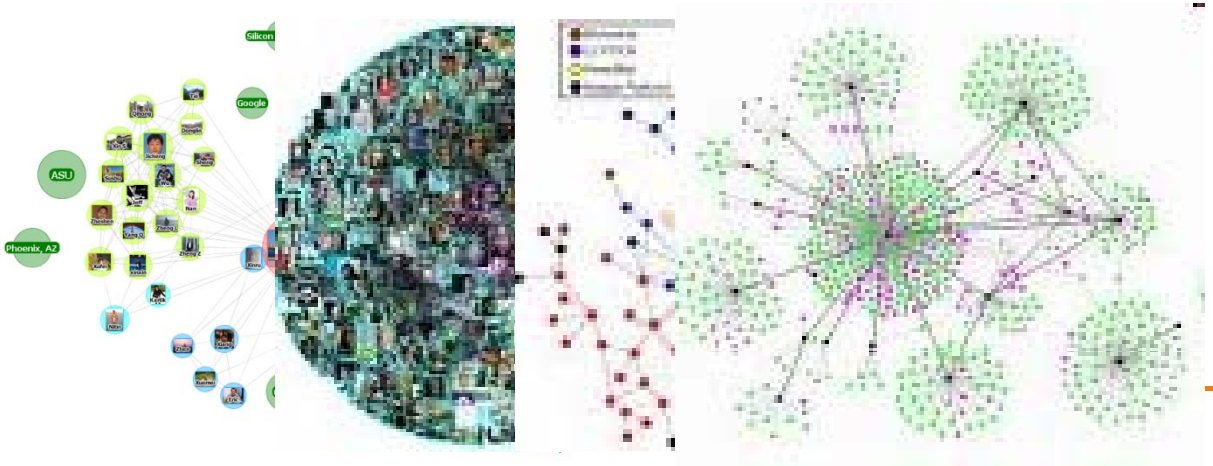


Visualization of a
3-mode network

Multi-Dimensional Network

- Networks consists of heterogeneous links between nodes
- a.k.a. multi-relational networks, multi-link networks





Network of Multiple Dimensions

Does Heterogeneity Matter?

- Social Media presents heterogeneity in networks
- Can we simply ignore the heterogeneity?

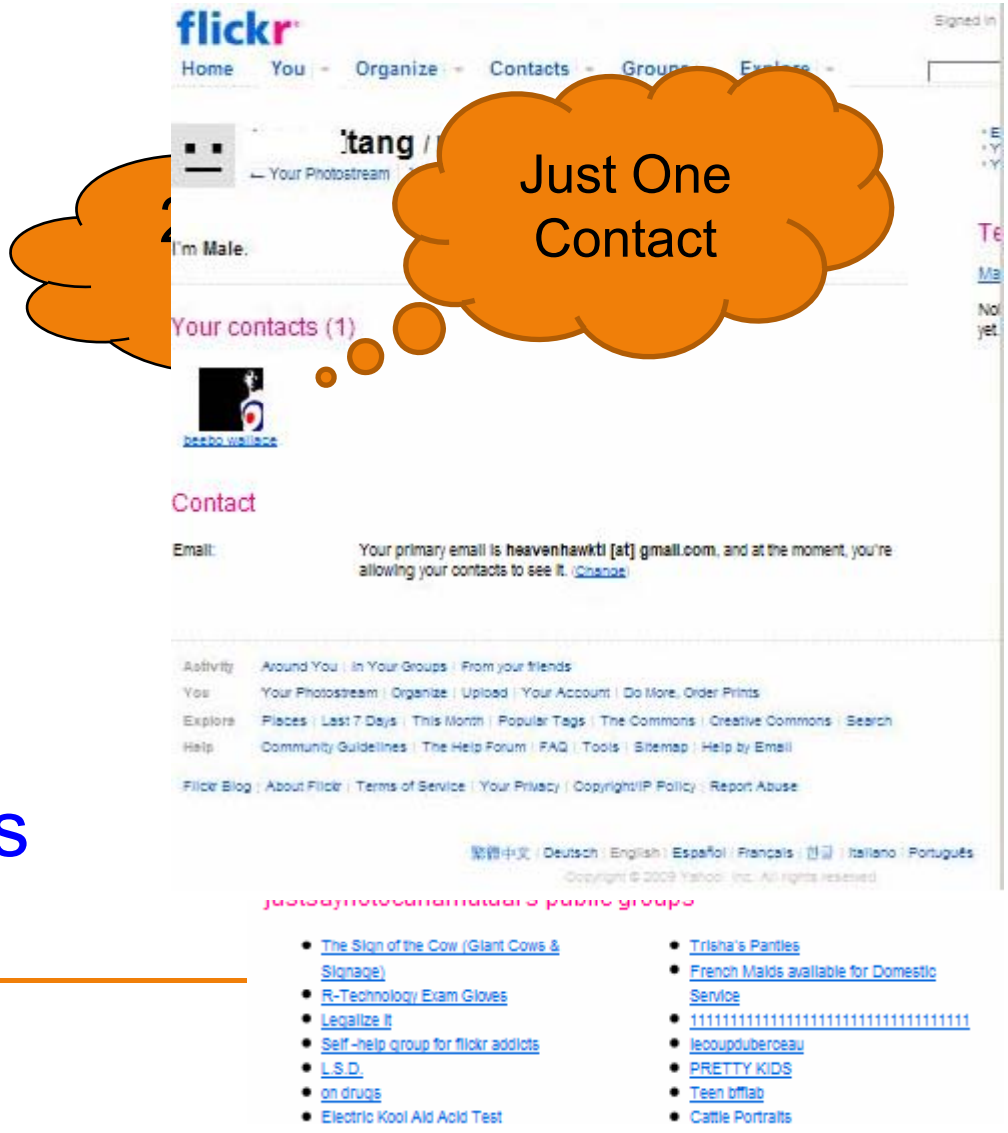
NO

Networks in Social Media are Noisy

Example of noisy friends network

- Too many friends?
- Too few friends?

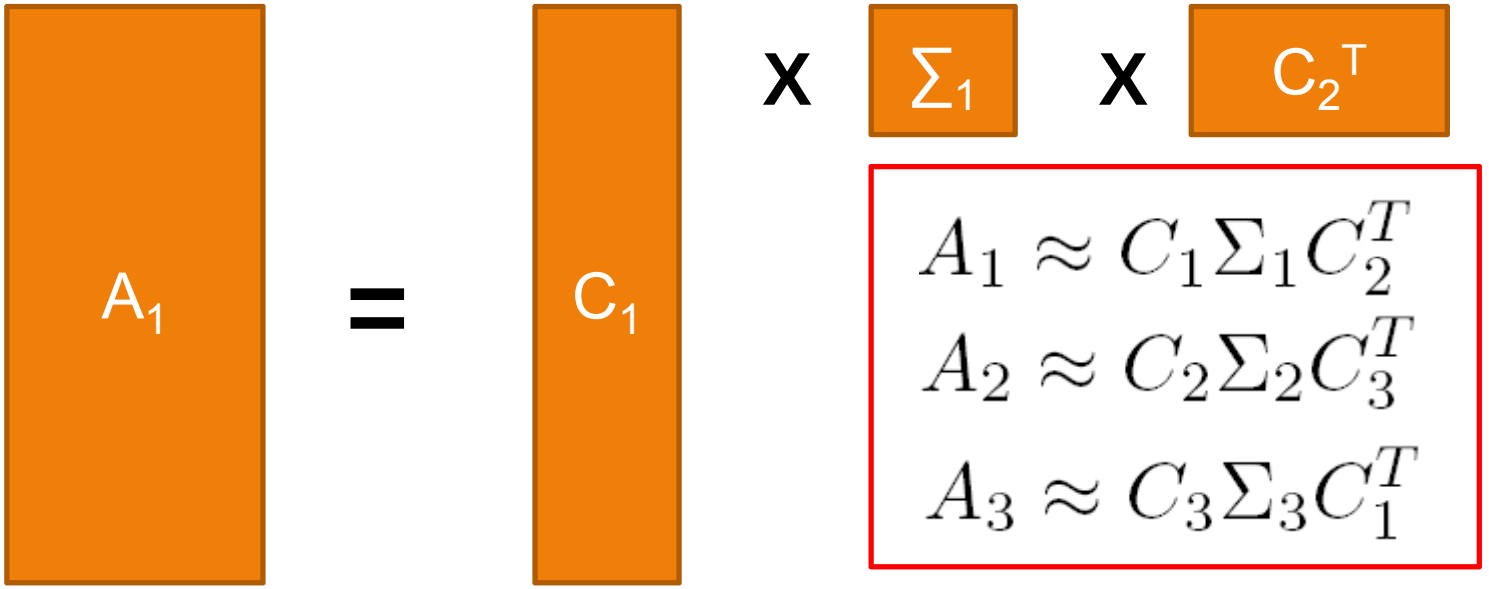
- Friends network tells limited info for some users
- Interaction at other modes or dimensions might help



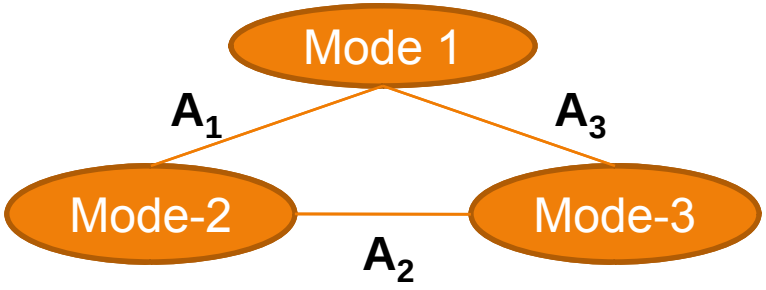
Reducing the Noise

- A multi-mode network presents correlations between different kinds of objects
  - e.g., Users of similar interests are likely to have similar tags
- Multi-dimensional networks can present complementary information at different dimensions
  - e.g., Some users seldom send email to each other, but might comment on each other's photos
- Taking into account of heterogeneity helps reduce the noise

Block Model for Multi-Mode Network



$$A_1 \approx C_1 \Sigma_1 C_2^T$$
$$A_2 \approx C_2 \Sigma_2 C_3^T$$
$$A_3 \approx C_3 \Sigma_3 C_1^T$$

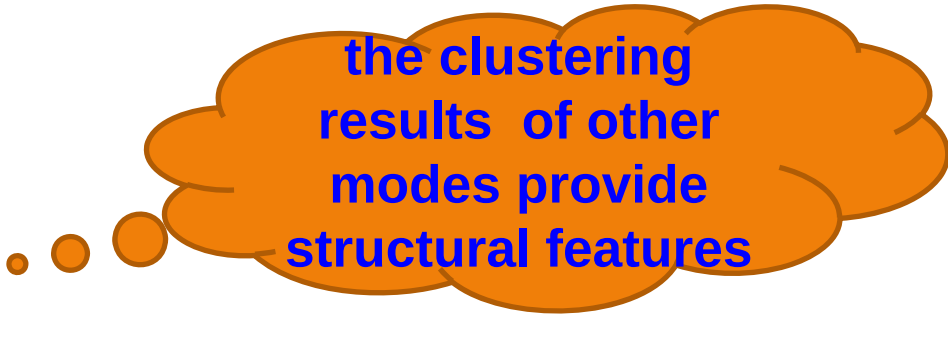


$$\sum_i \| A_i - C_i \Sigma_i C_{i+1}^T \|_F^2$$

Alternating Optimization

- No analytical solution
- Iteratively compute the optimal clustering in one mode while fixing the clustering of other modes
- C_j corresponds to the top left-singular vectors of P_j which is concatenated by the following matrix in column-wise:

$$P1 = [A_1 C_2, A_3^T C_3]$$
$$P2 = [A_2 C_3, A_1^T C_1]$$
$$P3 = [A_3 C_1, A_2^T C_2]$$



- Essentially apply PCA to data of the above format

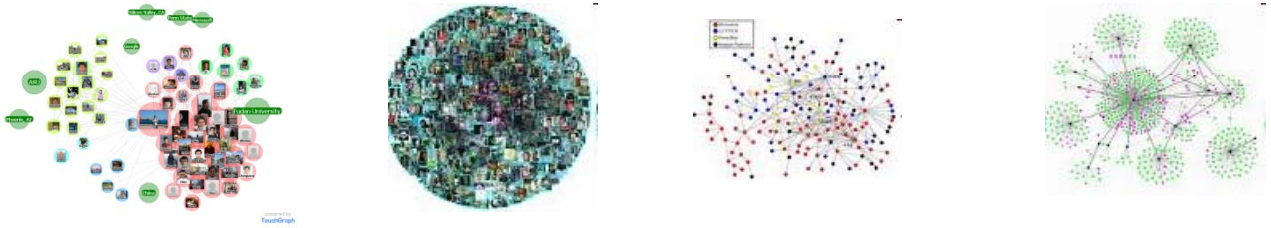
Shared Community Structure in Multi-Dimensional Networks

- A latent community structure is shared in a multi-dimensional network
  - a group sharing similar interests
  - users interacted at different social media sites

- Goal: Find out the shared community structure by integrating the network information of different dimensions

Communities in Multi-Dimensional Networks

Multi-Dimensional Networks



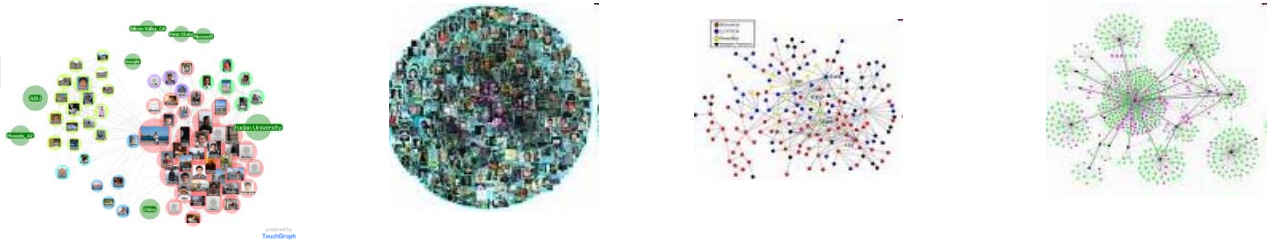
Extract Structural Features via Community Detection

Denoise the interaction at each dimension

- These structural features are not necessarily similar, but are highly correlated.
- Transform these features into a shared space such that their correlation is maximized.
- Solution: Generalized Canonical Correlation Analysis (CCA)

Communities in Multi-Dimensional Networks

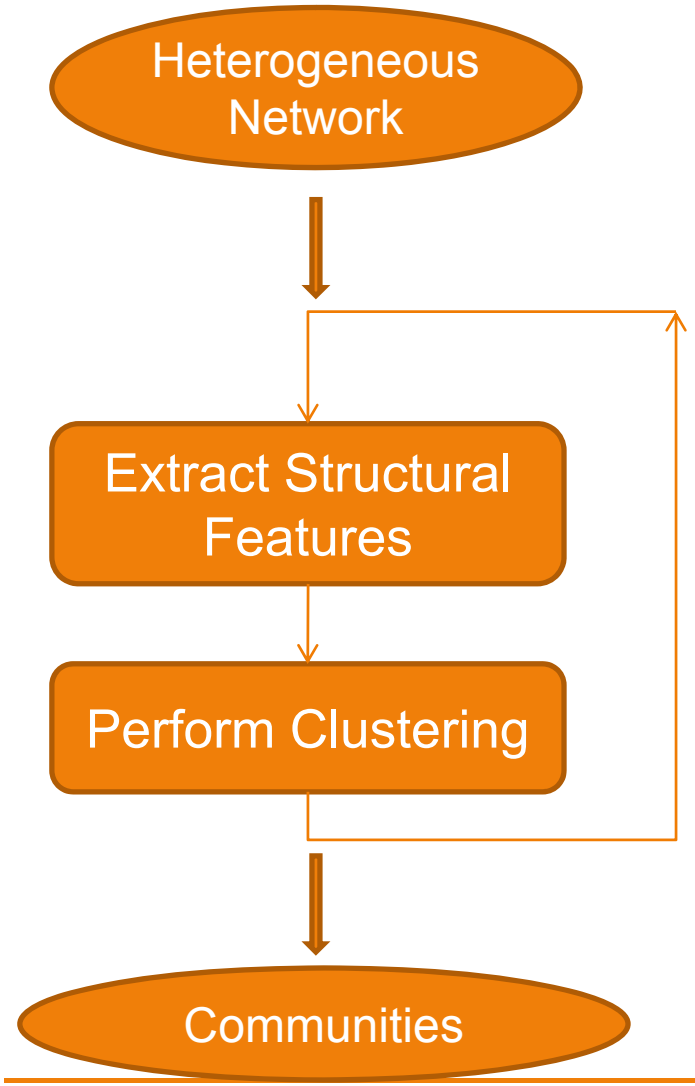
Multi-Dimensional Networks



Extract Structural Features via Community Detection

Combine all the structural features and perform **Principal Component Analysis**)

A Unified View



- Clustering at different modes or dimensions provides structural features
- Apply PCA or other community detection methods to find out the clustering

EVALUATION STRATEGY FOR COMMUNITY DETECTION

Next Section

Challenge of Evaluation

- Many methods of community detection
- Optimal methods depend on the data, tasks, and computational resources
- More often than not, no ground truth in reality!
- How to evaluate?
  - Whether the extracted communities are reasonable?
  - Which method works best under what conditions?

Self-consistent Community Definition

- To find a community with desired properties
  - e.g., Clique, k-clan, k-plex, etc.
  - Can be examined immediately
- To compare community size
  - e.g. clique or quasi-clique
- To enumerate as many communities as possible
  - The method returning maximum number of communities is the winner

Networks with Ground Truth

- Community Membership of each actor is known
- Commonly used in small networks or synthetic networks
- Measure: normalized mutual information in[0,1]

$$NMI(\pi^a, \pi^b) = \frac{\sum_{h=1}^{k^{(a)}} \sum_{\ell=1}^{k^{(b)}} \log\left(\frac{n \cdot n_{h,l}}{n_h^{(a)} \cdot n_\ell^{(b)}}\right)}{\sqrt{\left(\sum_{h=1}^{k^{(a)}} n_h^{(a)} \log \frac{n_h^a}{n}\right)\left(\sum_{\ell=1}^{k^{(b)}} n_\ell^{(b)} \log \frac{n_\ell^b}{n}\right)}}$$

Networks with Semantic Information

- Some networks come with attribute information
  - Blog, web with content information
  - Co-authorship with research interests information
- Check whether the extracted communities based on networks connectivity are consistent with semantics or shared attributes
- Pros
  - Help understand the community
- Cons
  - Requiring human subjects in evaluation
  - Applicable only to small numbers of communities
  - Only a qualitative evaluation

Networks without Ground Truth or Semantic Information

- Only network structure information is available
- More common in the real world
- Evaluation follows a cross-validation style
- Randomly sample some links to find communities
  - Approximate the remaining ones using the community structure
  - Adopt certain quantitative measure to calibrate the matching
    - Modularity
    - Network difference

Outline

- Social Media
- Data Mining Tasks and Evaluation

- Principles of Community Detection
- Communities in Heterogeneous Networks
- Evaluation Methodology for Community Detection

- Behavior Prediction via Social Dimensions

- Identifying Influential Bloggers in a Community

BEHAVIOR STUDY IN SOCIAL MEDIA

Basic Questions

- Q1: How do communities influence human behavior? Can we predict user behavior given partial observations?
- Q2: How do people interact in a community? Who is the leader in a group?

Social Computing Application I:

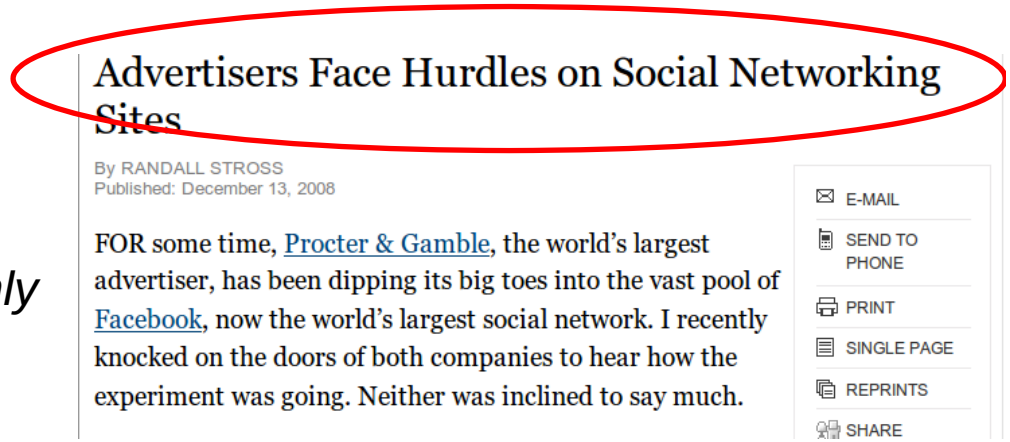
BEHAVIOR PREDICTION VIA SOCIAL DIMENSIONS

Motivation from Advertizing

Recent Boom of Social Media

vs.

"In 2008, 57% of all users of social networks clicked on an ad and only 11% of those clicks lead to a purchase"



Advertisers Face Hurdles on Social Networking Sites

By RANDALL STROSS
Published: December 13, 2008

FOR some time, Procter & Gamble, the world's largest advertiser, has been dipping its big toes into the vast pool of Facebook, now the world's largest social network. I recently knocked on the doors of both companies to hear how the experiment was going. Neither was inclined to say much.

E-MAIL
SEND TO PHONE
PRINT
SINGLE PAGE
REPRINTS
SHARE

Reality:
Limited user profile information
Readily available Social Network

Core Problem:
How to utilize Social Network information to help predict user preference or potential behavior?

Behavior Prediction

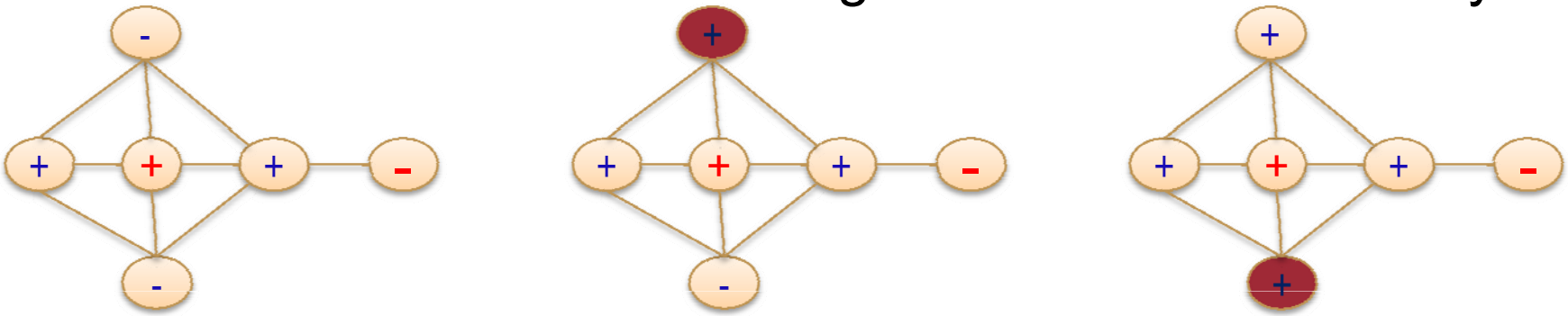
- User Preference or Behavior can be represented by labels (+/-)
  - Whether or not clicking on an ad
  - Whether or not interested in certain topics
  - Subscribed to certain political views
  - Like/Dislike a product

- **Given:**
  - A social network (i.e., connectivity information)
  - Some actors with identified labels

- **Output:**
  - Labels of other actors within the same network

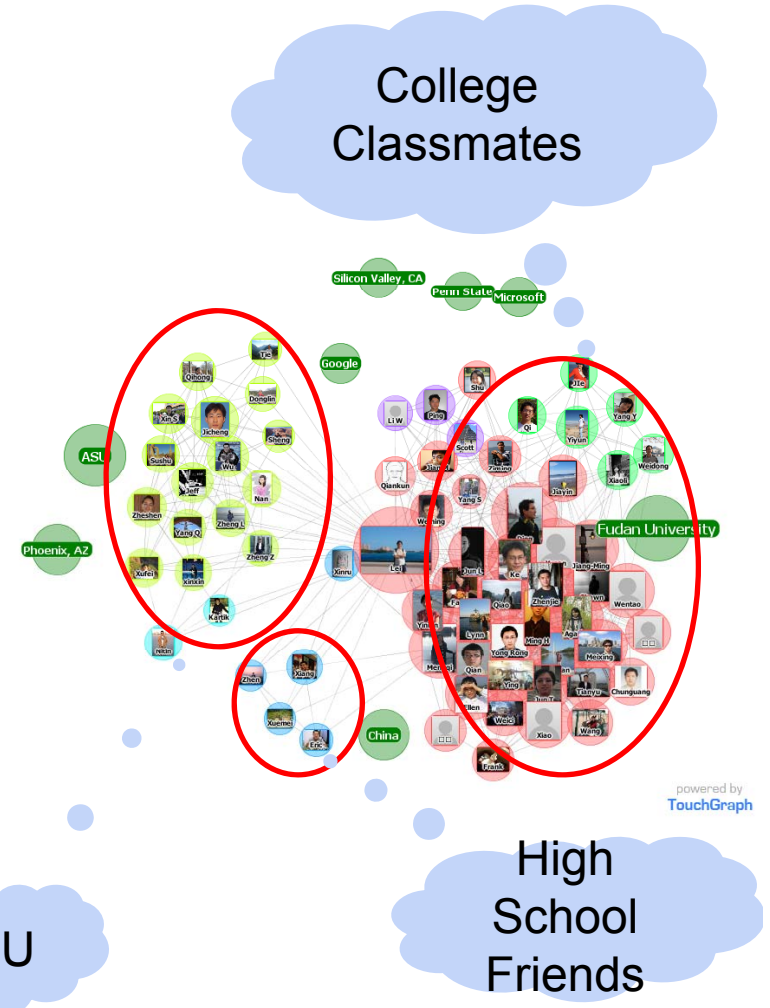
Approach I: Collective Inference

- Markov Assumption
  - The label of one node depends on that of its neighbors
- Training
  - Build a relational model based on labels of neighbors
- Prediction --- **Collective inference**
  - Predict the label of one node while fixing labels of its neighbors
  - Iterate until convergence
- Same as classical thresholding model in behavior study

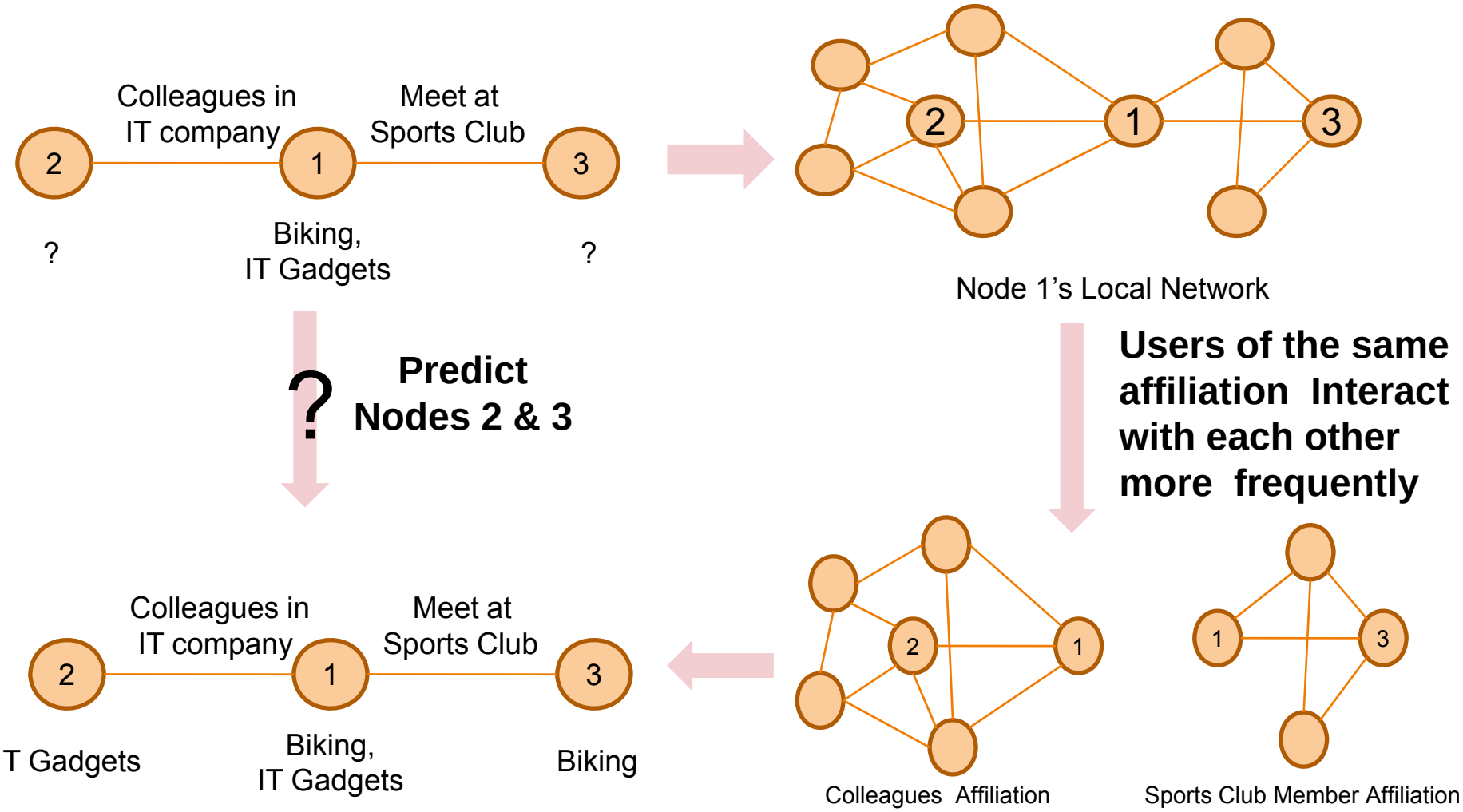


Heterogeneous Relations

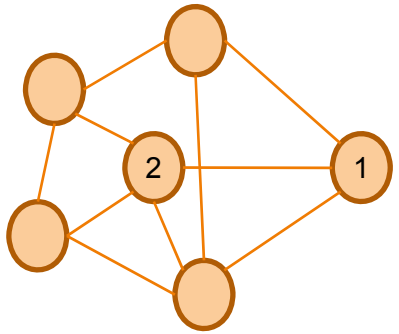
- Connections in a social network are heterogeneous
- Relation type information in social media is not always available
- Direct application of collective inference to social media treats all connections equivalently



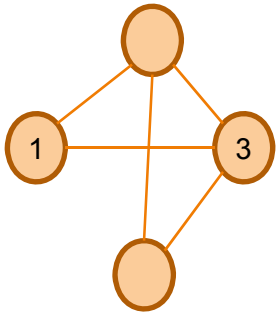
Extracting Actor Affiliations



Social Dimensions



Affiliation 1

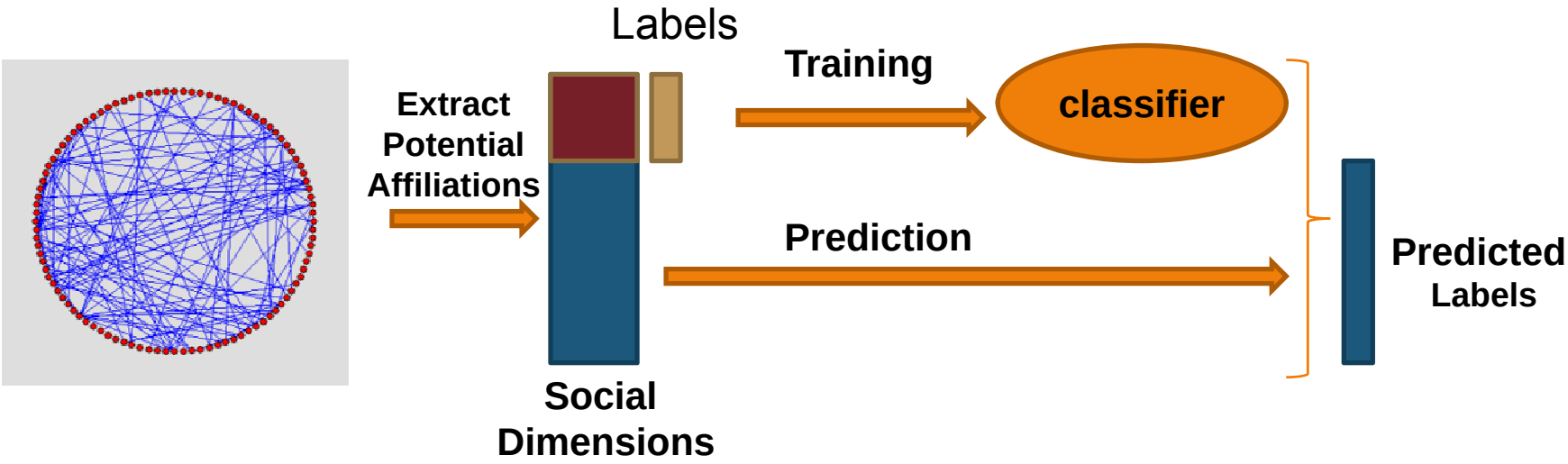


Affiliation 2

Actor	Affiliation 1	Affiliation 2
1	1	1
2	1	0
3	0	1
…	……	……

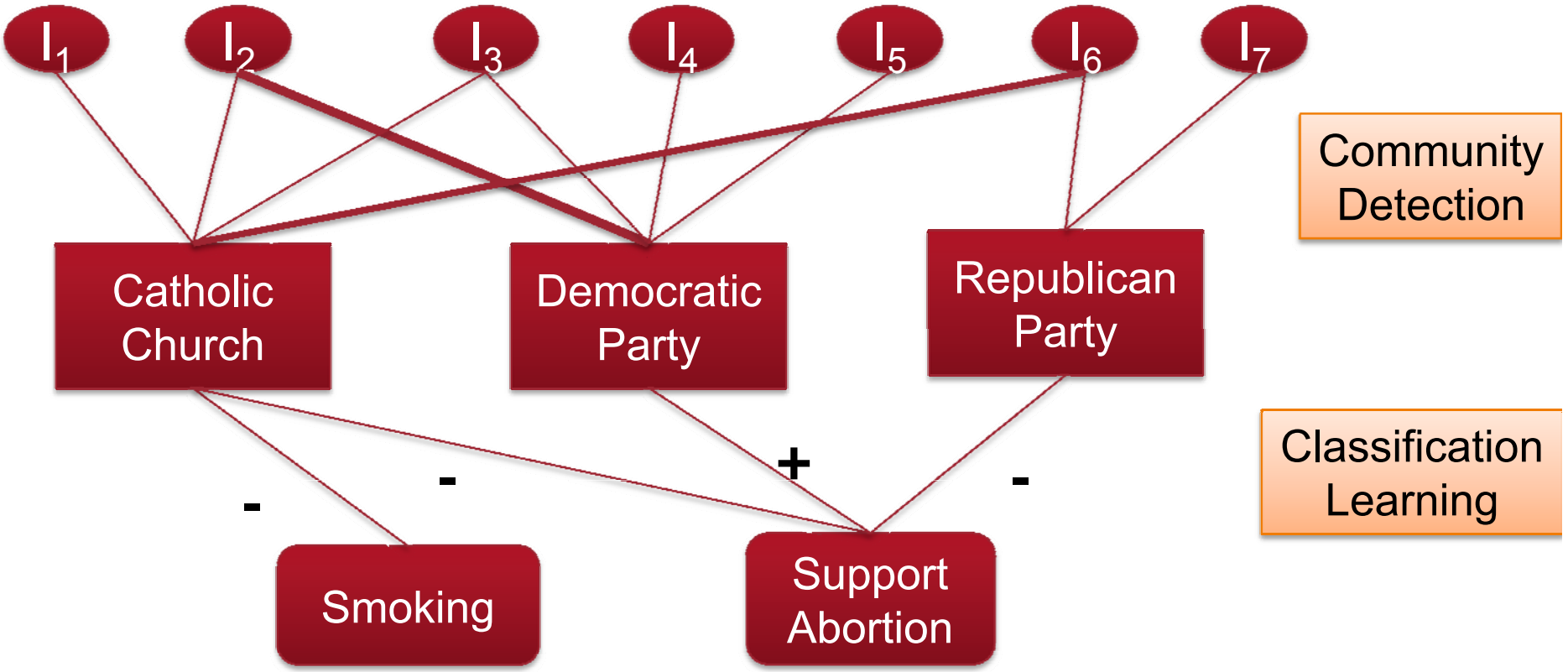
- Affiliations of actors are represented as social dimensions
- Each Dimension represents one potential affiliation
- Social dimensions capture prominent interaction patterns presented in the network

Approach II: Social-Dimension Approach (SocDim)

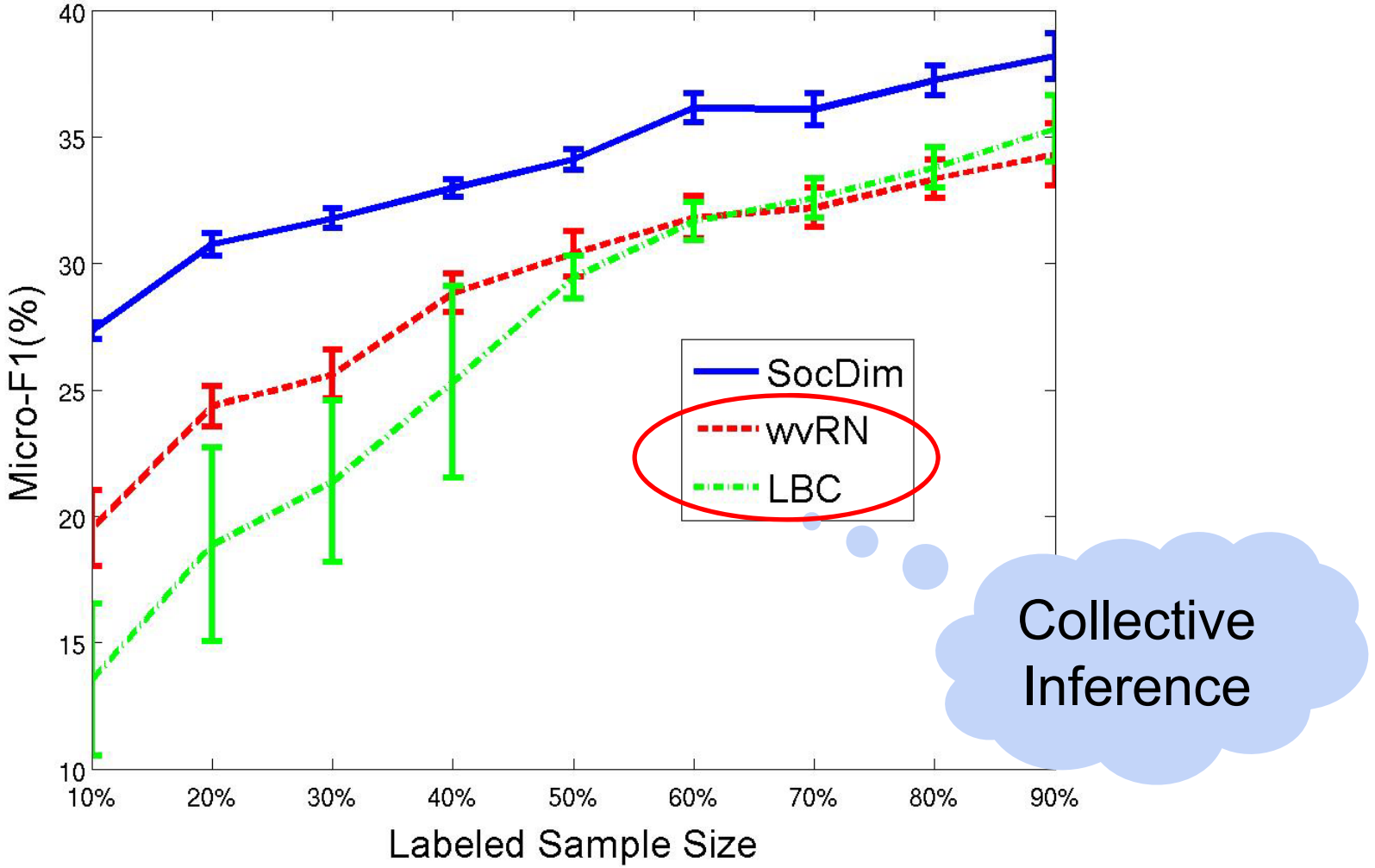


- Training:
  - Extract social dimensions to represent potential affiliations of actors
    - Any community detection methods is applicable (block model, spectral clustering)
  - Build a classifier to select those discriminative dimensions
    - Any discriminative classifier is acceptable (SVM, Logistic Regression)
- Prediction:
  - Predict labels based on one actor's latent social dimensions
  - No collective inference is necessary

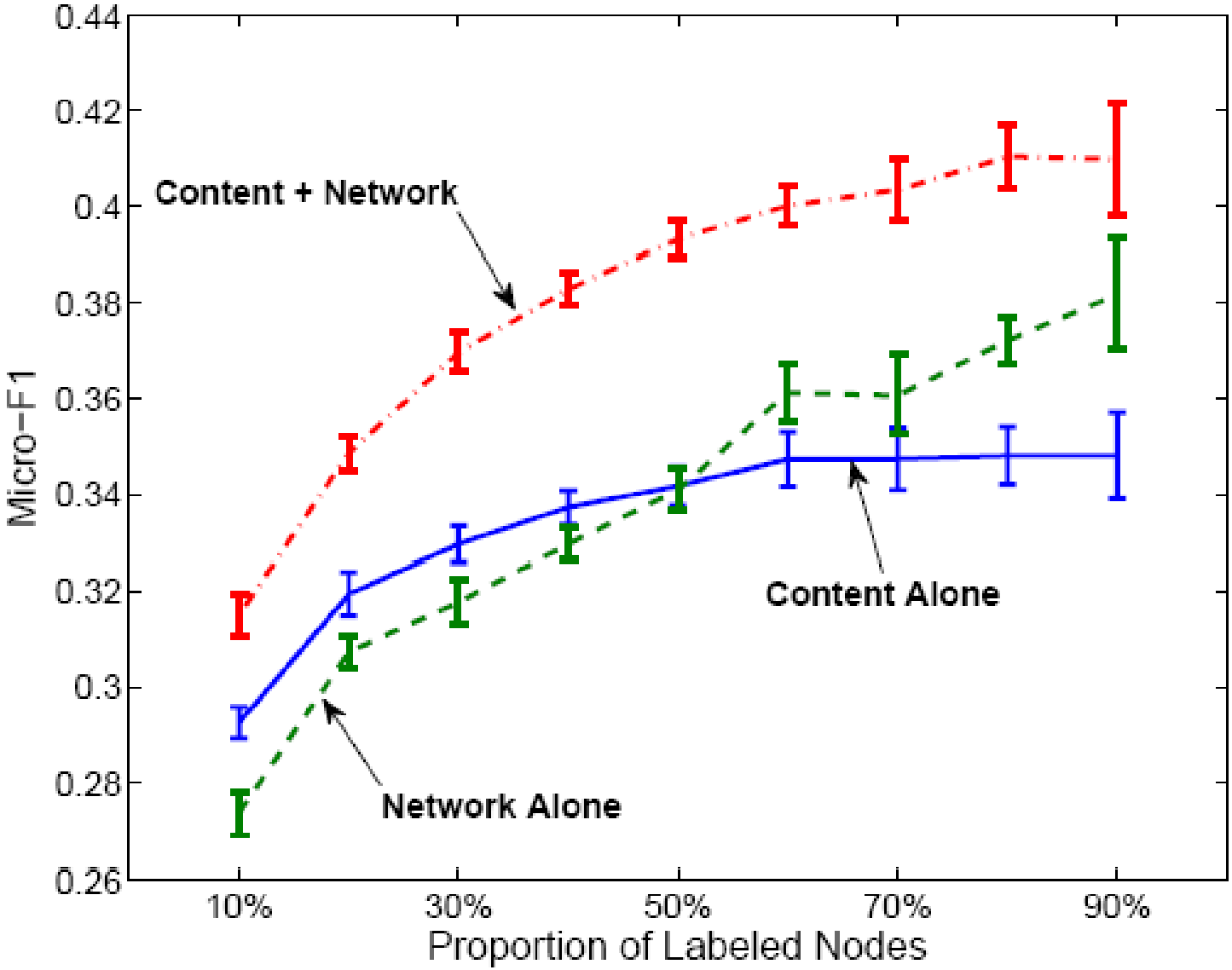
An Example of SocDim Model



SocDim vs. Collective Inference

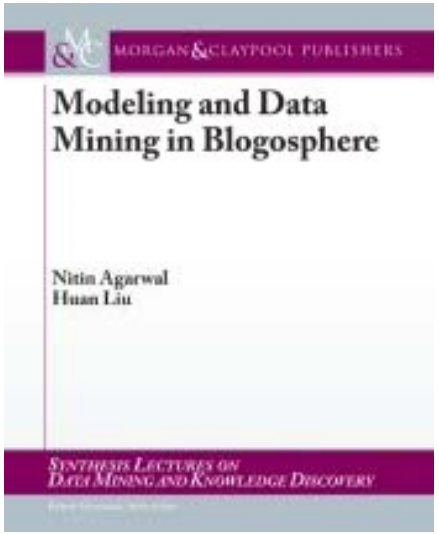


SocDim with Actor Features



Summary

- Networks in social media are *noisy and heterogenous*
- SocDim proposes to extract social dimensions to capture potential affiliations of actors
- Community Detection can be used to extract social dimensions from networks
- Social dimensions can be combined with other content and/or profile features
- SocDim outperforms other representative collective inference methods
- Recent advancement of SocDim can handle networks of 1 million nodes in 10 mins.

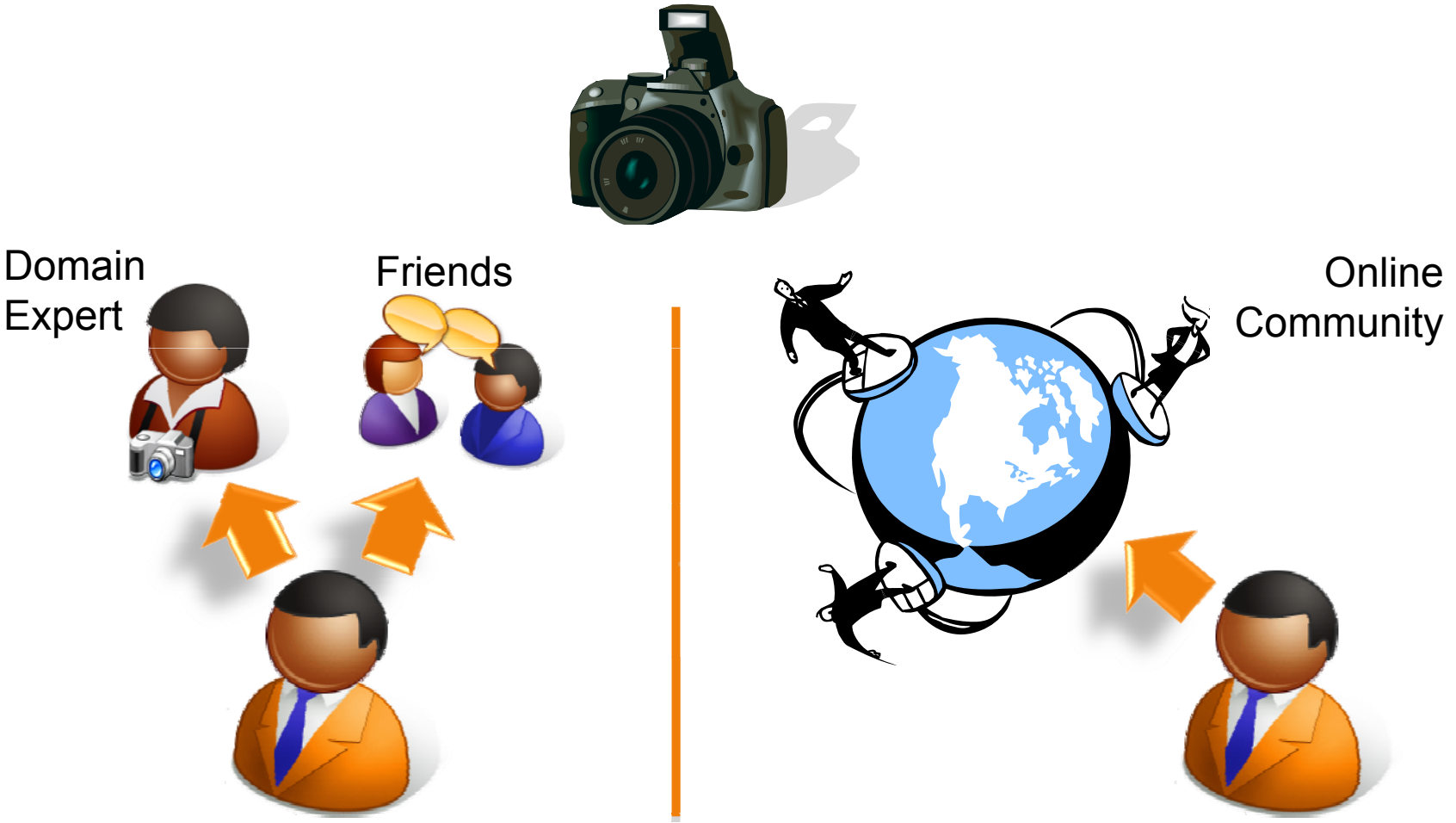


Social Computing Applications II:

IFINDER: IDENTIFYING INFLUENTIAL BLOGGERS IN A COMMUNITY (VIDEO)

Go to the End

Physical and Virtual World



Introduction

- Inspired by the analogy between real-world and blog communities, we answer:

Who are the influentials in Blogosphere?

Can we find *them*?

?

Active Bloggers = Influential Bloggers

- Active bloggers may not be influential
- Influential bloggers may not be active

Searching The Influentials

- Active bloggers
  - Easy to define
  - Often listed at a blog site
  - Are they necessarily influential
- How to define an influential blogger?
  - Influential bloggers have influential posts
  - Subjective
  - Collectable statistics
  - How to use these statistics

Intuitive Properties

- Social Gestures (*statistics*)
  - Recognition: Citations (incoming links)
    - An influential blog post is recognized by many. The more influential the referring posts are, the more influential the referred post becomes.
  - Activity Generation: Volume of discussion (comments)
    - Amount of discussion initiated by a blog post can be measured by the comments it receives. Large number of comments indicates that the blog post affects many such that they care to write comments, hence influential.
  - Novelty: Referring to (outgoing links)
    - Novel ideas exert more influence. Large number of outlinks suggests that the blog post refers to several other blog posts, hence less novel.
  - Eloquence: "goodness" of a blog post (length)
    - An influential is often eloquent. Given the informal nature of Blogosphere, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so.
- *Influence Score = f(Social Gestures)*

A Preliminary Model

- Additive models are good to determine the combined value of each alternative [Fensterer, 2007]. It also supports preferential independence of all the parameters involved in the final decision. A weighted additive function can be used to evaluate trade-offs between different objectives [Keeney and Raiffa, 1993].

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n)$$

$$I(p) \propto w_{comm} \gamma_p + InfluenceFlow(p)$$

$$I(p) = w(\lambda) \times (w_{comm} \gamma_p + InfluenceFlow(p))$$

$$iIndex(B) = \max(I(p_l))$$

Understanding the Influentials

- Are influential bloggers simply active bloggers?
- If not, in what ways are they different?
  - Can the model differentiate them?

- Are there different types of influential bloggers?

- What other parameters can we include to evolve the model?
- Are there temporal patterns of the influential bloggers?

How to Evaluate the Model

- Where to find the ground truth?
  - Lack of Training and Test data
  - Any alternative?
- About the parameters
  - How can they be determined
  - Are they all necessary?
    - Are any of these correlated?
- Data collection
  - A real-world blog site
  - “The Unofficial Apple Weblog”

Active & Influential Bloggers

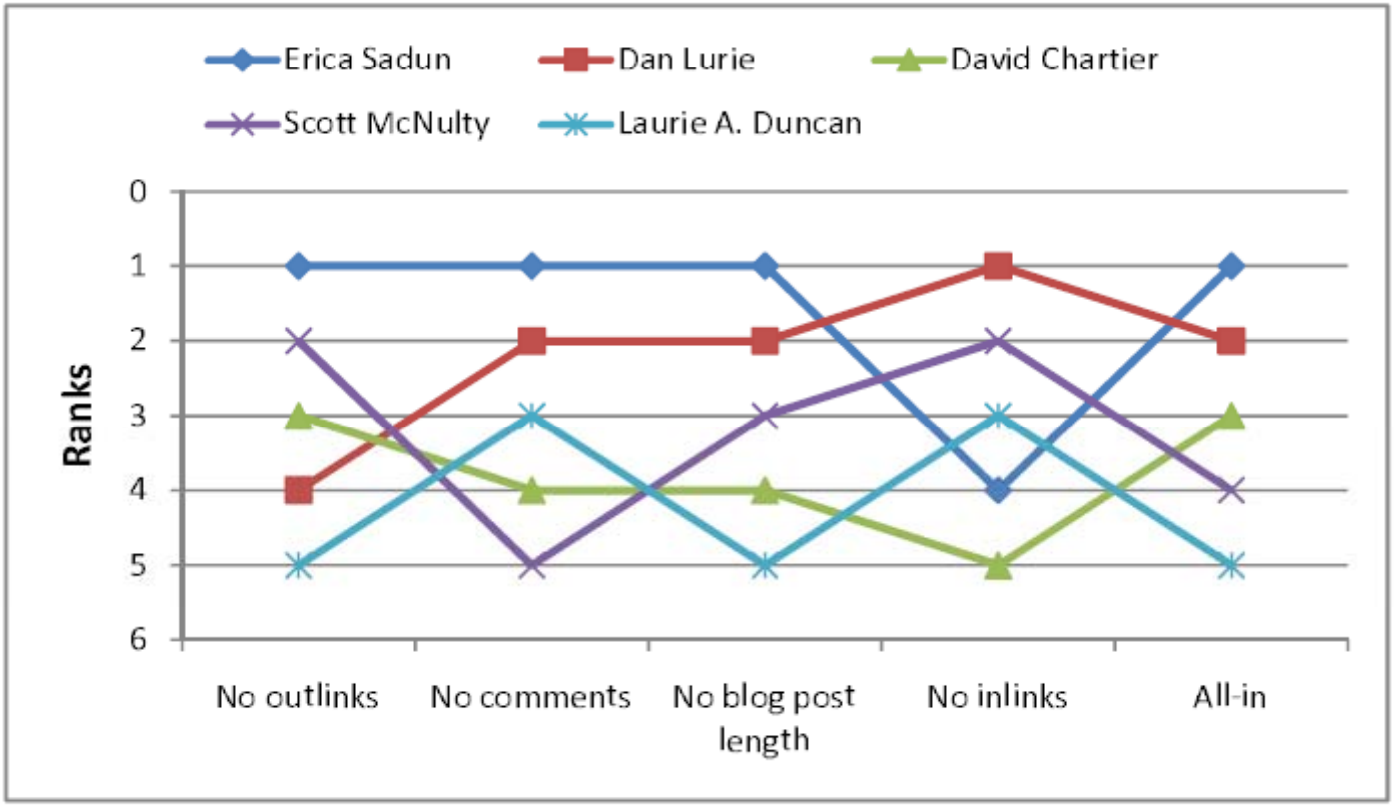
Top 5 TUAW Bloggers	Top 5 Influential Bloggers
Erica Sadun	*Erica Sadun*
Scott McNulty	Dan Lurie
Mat Lu	*David Chartier*
David Chartier	*Scott McNulty*
Michael Rose	Laurie A. Duncan

- Active and Influential Bloggers
- Inactive but Influential Bloggers
- Active but Non-influential Bloggers

- We don't consider "Inactive and Non-influential Bloggers", because they seldom submit blog posts. Moreover, they do not influence others.

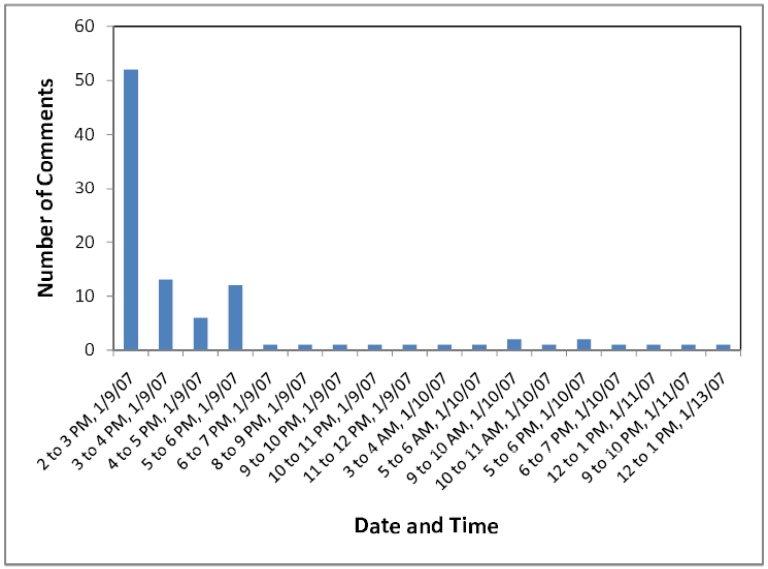
Lesion Study

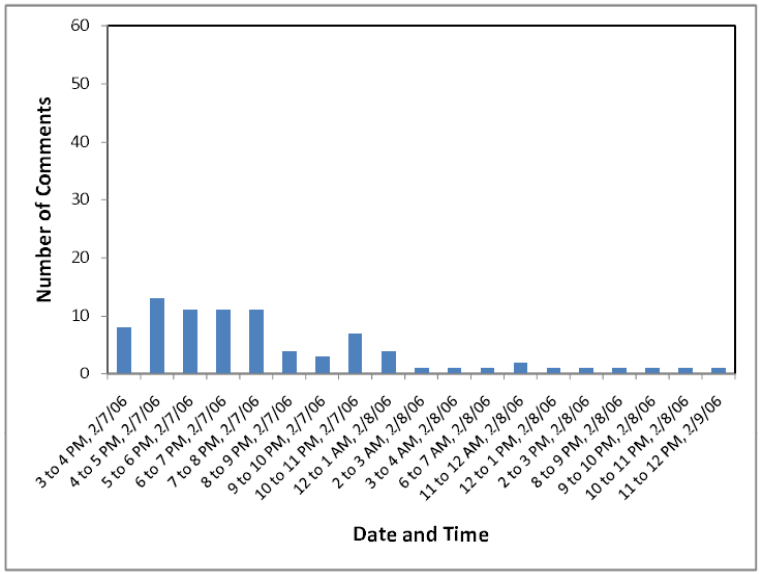
- To observe if any parameter is irrelevant.



Other Parameters

- Rate of Comments

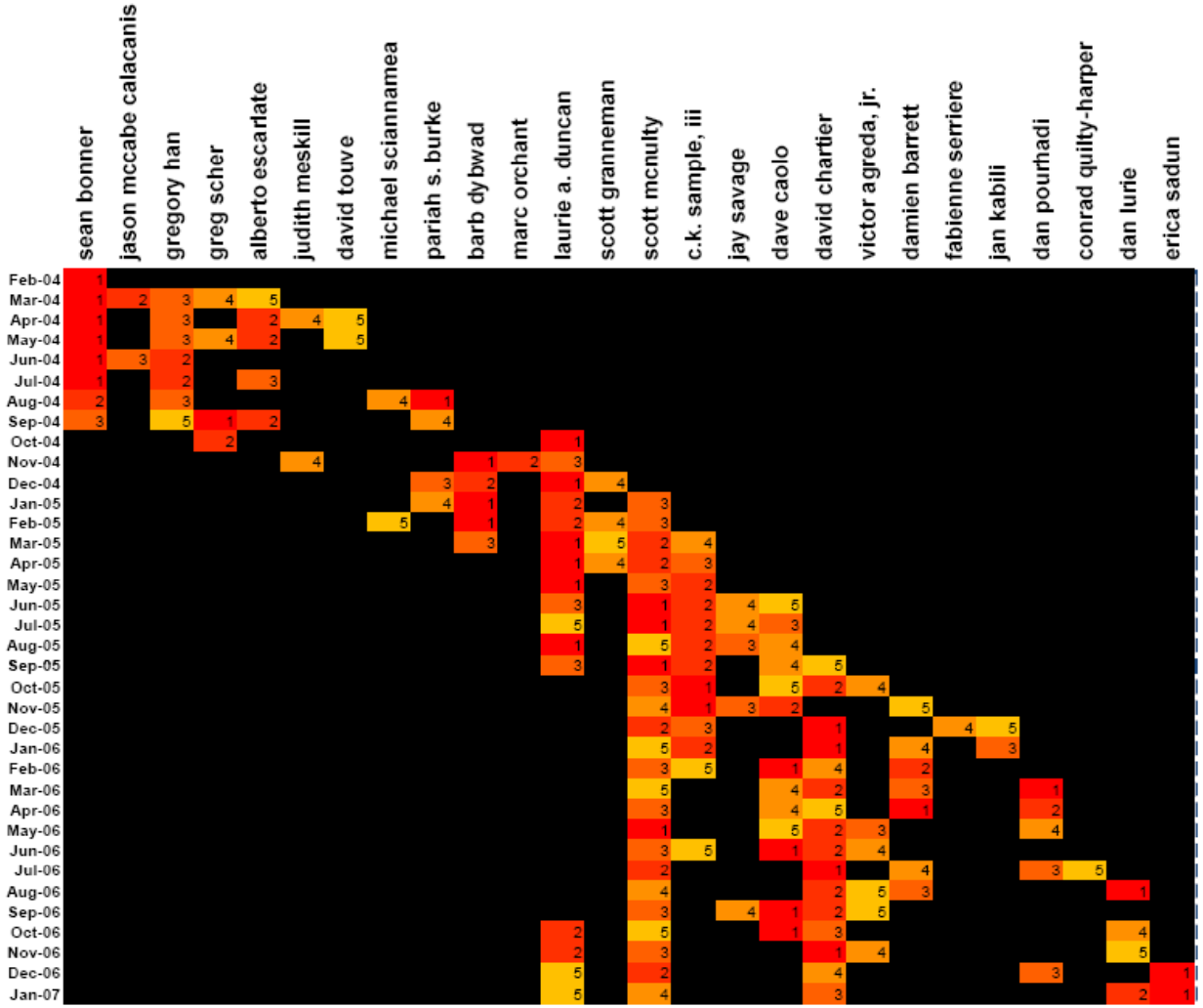




"Spiky" comments reaction

"Flat" comments reaction

Temporal Patterns of Influential Bloggers



- Long term Influentials
- Average term Influentials
- Transient Influentials
- Burgeoning Influentials

Verification of the Model

- Revisit the challenges
  - No training and testing data
  - Absence of ground truth
  - Subjectivity
- We use another Web 2.0 website, Digg as a reference point.
- “Digg is all about user powered content. Everything is submitted and voted on by the Digg community. Share, discover, bookmark, and promote stuff that‘s important to you!”
- The higher the digg score for a blog post is, the more it is liked.
- A not-liked blog post will not be submitted thus will not appear in Digg.

Verification of the Model

- Digg records top 100 blog posts.
- Top 5 influential and top 5 active bloggers were picked to construct 4 categories
- For each of the 4 categories of bloggers, we collect top 20 blog posts from our model and compare them with Digg top 100.

Bloggers	Active	Inactive
Influential	S1: 17	S2: 7
Non-influential	S3: 3	S4: 0/1

Bloggers	Active	Inactive
Influential	S1: 71	S2: 14
Non-influential	S3: 8	S4: 7

Bloggers	Active	Inactive
Influential	S1: 327	S2: 42
Non-influential	S3: 131	S4: 35

- Distribution of Digg top 100 and TUAW's 535 blog posts

Verification of the Model

- Observe how much our model aligns with Digg.
- Compare top 20 blog posts from our model and Digg.
- Considered last six months

	Jun 2007	May 2007	Apr 2007	Mar 2007	Feb 2007	Jan 2007
All-in	14	16	12	15	10	12
No Inlinks	3	4	3	3	1	0
No Comments	8	8	5	4	5	4
No Outlinks	11	8	5	4	4	7
No Blog post length	12	14	11	15	9	10

- Considered all configuration to study relative importance of each parameter.
- **Inlinks > Comments > Outlinks > Blog post length**

Outline

- Social Media
- Data Mining Tasks
- Evaluation

- Principles of Community Detection
- Communities in Heterogeneous Networks
- Evaluation Methodology for Community Detection

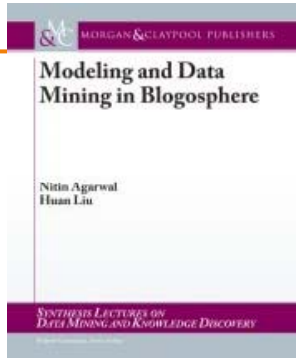
- Behavior Prediction via Social Dimensions
- Identifying Influential Bloggers in a Community
  - A related tutorial on Blogosphere

References

- General
- Social Computing
- Community Detection
- Heterogeneous Networks
- Behavior Prediction

- Related Tutorial and Talk
  - KDD'08 Tutorial
  - WSDM'08 Presentation

References: General



- Tang, L. & Liu, H. (Forthcoming), Graph Mining Applications to Social Network Analysis'Managing and Mining Graph Data'.
- Agarwal, N. & Liu, H. (2009), *Modeling and Data Mining in Blogosphere*, Morgan and Claypool.
- Shirky, C. (2008), *Here Comes Everybody: The Power of Organizing without Organizations*, The Penguin Press.
- (2008), 'What is Social Media? An eBook from iCrossing'.
- Chakrabarti, D. &Faloutsos, C. (2006), 'Graph mining: Laws, generators, and algorithms', *ACM Comput. Surv.***38**(1), 2.
- Wasserman, S. & Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press.

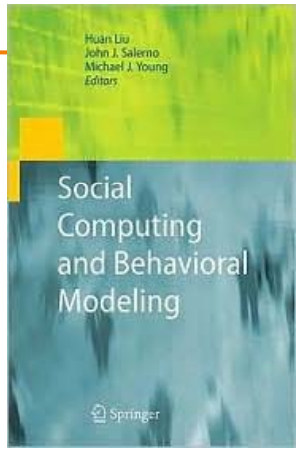
Return to Menu

References: Social Computing

- Tang, L. & Liu, H. (2009), Scalable Learning of Collective Behavior based on Sparse Social Dimensions, *in* 'The 18th ACM Conference on Information and Knowledge Management'.
- Tang, L. & Liu, H. (2009), Relational learning via latent social dimensions, *in* 'KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 817--826.
- Agarwal, N.; Galan, M.; Liu, H. &Subramanya., S. (2009), 'WisColl: Collective Wisdom based Blog Clustering', *Journal of Information Science: Special Issue on Collective Intelligence***http://dx.doi.org/10.1016/j.ins.2009.07.010**.
- Zafarani, R. & Liu, H. (2009), Connecting Corresponding Identities across Communities, *in* 'Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM)'.
- Agarwal, N.; Liu, H.; Tang, L. & Yu, P. S. (2008), Identifying the influential bloggers in a community, *in* 'WSDM '08: Proceedings of the international conference on Web search and web data mining', ACM, New York, NY, USA, pp. 207--218.
- Leskovec, J.; Lang, K. J.; Dasgupta, A. & Mahoney, M. W. (2008), Statistical properties of community structure in large social and information networks, *in* 'WWW '08: Proceeding of the 17th international conference on World Wide Web', ACM,.

Return to Menu

References: Social Computing



- Tang, L.; Liu, H.; Zhang, J. &Nazeri, Z. (2008), Community evolution in dynamic multi-mode networks, *in* 'KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 677--685.
- Tang, L.; Liu, H.; Zhang, J.; Agarwal, N. & Salerno, J. J. (2008), 'Topic taxonomy adaptation for group profiling', *ACM Trans. Knowl. Discov. Data***1**(4), 1--28.
- Liben-Nowell, D. & Kleinberg, J. (2007), 'The link-prediction problem for social networks', *J. Am. Soc. Inf. Sci. Technol.***58**(7), 1019--1031.
- Newman, M. (2005), 'Power laws, Pareto distributions and Zipf's law', *Contemporary physics***46**(5), 323--352.
- Richardson, M. &Domingos, P. (2002), Mining knowledge-sharing sites for viral marketing, *in* 'KDD', pp. 61-70.
- Barabási, A.-L. & Albert, R. (1999), 'Emergence of Scaling in Random Networks', *Science***286**(5439), 509-512.
- Travers, J. &Milgram, S. (1969), 'An Experimental Study of the Small World Problem', *Sociometry***32**(4), 425-443.

Return to Menu

References: Community Detection

- Tang, L. & Liu, H. (Forthcoming), Graph Mining Applications to Social Network Analysis'Managing and Mining Graph Data'.
- Abello, J.; Resende, M. G. C. &Sudarsky, S. (2002), Massive Quasi-Clique Detection, *in* 'LATIN', pp. 598-612.
- Agarwal, N.; Galan, M.; Liu, H. &Subramanya., S. (2009), 'WisColl: Collective Wisdom based Blog Clustering', *Journal of Information Science: Special Issue on Collective Intelligence*http://dx.doi.org/10.1016/j.ins.2009.07.010.
- Borg, I. &Groenen, P. (2005), *Modern Multidimensional Scaling: theory and applications*, Springer.
- Borgatti, S. P.; Everett, M. G. &Shirey, P. R. (1990), 'LS Sets, Lambda Sets and other cohesive subsets', *Social Networks***12**, 337-357.
- Brandes, U.; Delling, D.; Gaertler, M.; Gorke, R.; Hoefer, M.; Nikoloski, Z. & Wagner, D. (2006), 'Maximizing Modularity is hard', *Arxiv preprint physics/0608255*.
- Clauset, A.; Mewman, M. & Moore, C. (2004), 'Finding community structure in very large networks', *Arxiv preprint cond-mat/0408187*.
- Clauset, A.; Moore, C. & Newman, M. E. J. (2008), 'Hierarchical structure and the prediction of missing links in networks', *Nature***453**, 98-101.

Return to Menu

References: Community Detection

- Flake, G. W.; Lawrence, S. & Giles, C. L. (2000), Efficient identification of Web communities, *in* 'KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 150--160.
- Fortunato, S. &Barthelemy, M. (2007), 'Resolution limit in community detection', *PNAS***104**(1), 36--41.
- Gibson, D.; Kumar, R. & Tomkins, A. (2005), Discovering large dense subgraphs in massive graphs, *in* 'VLDB '05: Proceedings of the 31st international conference on Very large data bases', VLDB Endowment, , pp. 721--732.
- Handcock, M. S.; Raftery, A. E. & Tantrum, J. M. (2007), 'Model-based clustering for social networks', *Journal Of The Royal Statistical Society Series A***127**(2), 301-354.
- Hoff, P. D. & Adrian E. Raftery, M. S. H. (2002), 'Latent Space Approaches to Social Network Analysis', *Journal of the American Statistical Association***97**(460), 1090-1098
- von Luxburg, U. (2007), 'A tutorial on spectral clustering', *Statistics and Computing***17**(4), 395--416.

Return to Menu

References: Community Detection

- Newman, M. (2006), 'Modularity and community structure in networks', *PNAS***103**(23), 8577-8582.
- Newman, M. (2006), 'Finding community structure in networks using the eigenvectors of matrices', *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)***74**(3).
- Newman, M. & Girvan, M. (2004), 'Finding and evaluating community structure in networks', *Physical Review E***69**, 026113.
- Nowicki, K. &Snijders, T. A. B. (2001), 'Estimation and Prediction for Stochastic Blockstructures', *Journal of the American Statistical Association***96**(455), 1077-1087.
- Sarkar, P. & Moore, A. W. (2005), 'Dynamic social network analysis using latent space models', *SIGKDD Explor. Newsl.***7**(2), 31--40.
- Shi, J. &Malik, J. (1997), Normalized Cuts and Image Segmentation, *in* 'CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)', IEEE Computer Society, Washington, DC, USA, pp. 731.
- White, S. & Smyth, P. (2005), A spectral Clustering Approaches To Finding Communities in Graphs, *in* 'SDM'.

Return to Menu

References: Heterogeneous Networks

- Tang, L. & Liu, H. (Forthcoming), Understanding Group Structures and Properties in Social Media'Link Mining: Models, Algorithms and Applications', Springer, .
- Tang, L. & Liu, H. (2009), Uncovering Cross-Dimension Group Structures in Multi-Dimensional Networks, *in* 'SDM workshop on Analysis of Dynamic Networks'
- Zafarani, R. & Liu, H. (2009), Connecting Corresponding Identities across Communities, *in* 'Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM)'.
- Carley, K.; Reminga, J.; Storrick, J. &DeReno, M. (2009), 'ORA User's Guide', Technical report, Carnegie Mellon University.
- Tang, L.; Liu, H.; Zhang, J. &Nazeri, Z. (2008), Community evolution in dynamic multi-mode networks, *in* 'KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM,, pp. 677--685.
- Long, B.; Zhang, Z. (M.; Wú, X. & Yu, P. S. (2006), Spectral clustering for multi-type relational data, *in* 'ICML '06: Proceedings of the 23rd international conference on Machine learning', ACM, New York, NY, USA, pp. 585--592.
- Strehl, A. &Ghosh, J. (2003), 'Cluster ensembles --- a knowledge reuse framework for combining multiple partitions', *J. Mach. Learn. Res.***3**, 583--617.
- Kettenring, J. (1971), 'Canonical analysis of several sets of variables', *Biometrika***58**, 433-451.

Return to Menu

References: Behavior Prediction

- Tang, L. (2009), Collective Behavior Prediction in Social Media, *in* 'SIAM Data Mining Doctoral Student Forum (SDM)'.
- Tang, L. & Liu, H. (2009), Scalable Learning of Collective Behavior based on Sparse Social Dimensions, *in* 'The 18th ACM Conference on Information and Knowledge Management'.
- Tang, L. & Liu, H. (2009), Relational learning via latent social dimensions, *in* 'KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 817--826.
- Agarwal, N.; Liu, H.; Tang, L. & Yu, P. S. (2008), Identifying the influential bloggers in a community, *in* 'WSDM '08: Proceedings of the international conference on Web search and web data mining', ACM, New York, NY, USA, pp. 207--218.
- Macskassy, S. A. & Provost, F. (2007), 'Classification in Networked Data: A Toolkit and a Univariate Case Study', *J. Mach. Learn. Res.***8**, 935--983.
- Jensen, D.; Neville, J. & Gallagher, B. (2004), Why collective inference improves relational classification, *in* 'KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining',, pp. 593--598.
- McPherson, M.; Smith-Lovin, L. & Cook, J. M. (2001), 'BIRDS OF A FEATHER: Homophily in Social Networks', *Annual Review of Sociology***27**, 415-444.
- Granovetter, M. (1978), 'Threshold Models of Collective Behavior', *The American Journal of Sociology***83**(6), 1420-1443.
- Schelling, T. C. (1971), 'Dynamic models of segregation', *Journal of Mathematical Sociology***1**, 143—186.

Return to Menu

Thank You!

Please feel free to contact **Lei Tang** (L.Tang@asu.edu) if you have any questions!