Social Media Mining: Fundamental Issues and Challenges

Mohammad Ali Abbasi, *Huan Liu*, and *Reza Zafarani*

Data Mining and Machine Learning Lab
Arizona State University

December 10, 2013

Social Media Mining

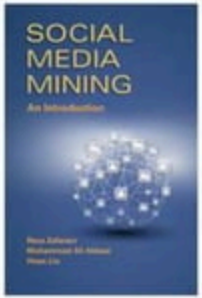
Home | Download | Errata | Slides | Table of Contents | Tutorials

Social Media Mining

An Introduction

A Textbook by Cambridge University Press

Reza Zafarani — Arizona State University
Mohammad Ali Abbasi — Arizona State University
Huan Liu — Arizona State University



CAMBRIDGE UNIVERSITY PRESS
amazon.com
BARNES & NOBLE BOOKSELLERS
eBooks.com

Download

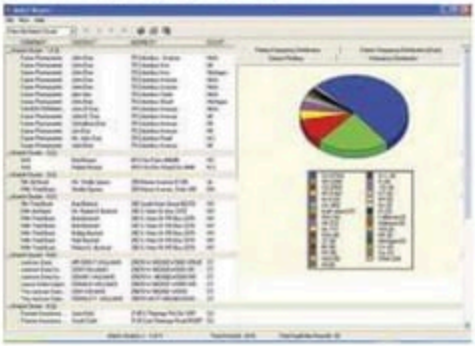
http://dmml.asu.edu/smm/

The growth of social media over the last decade has revolutionized the way individuals interact and industries conduct business. Individuals produce data at an unprecedented rate by interacting, sharing, and consuming content through social media. Understanding and processing this new type of data to glean actionable patterns presents challenges and opportunities for interdisciplinary research, novel algorithms, and tool development. Social Media Mining integrates social media, social network analysis, and data mining to provide a convenient and coherent platform for students, practitioners, researchers, and project managers to understand the basics and potentials of social media mining. It introduces the unique problems arising from social media data and presents fundamental concepts, emerging issues, and effective algorithms for network analysis and data mining. Suitable for use in advanced undergraduate and beginning graduate courses as well as professional short courses, the text contains exercises of different degrees of difficulty that improve understanding and help apply concepts, principles, and methods in various scenarios of social media mining.

Traditional Media and Data

Broadcast Media
One-to-Many

Communication Media
One-to-One



Traditional Data

Social Media: Many-to-Many

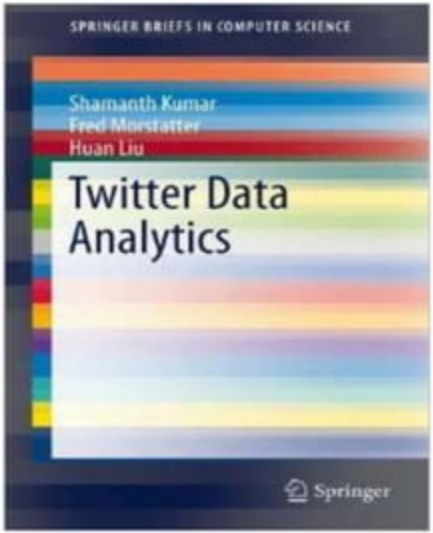
- Everyone can be a media outlet or producer
- Disappearing communication barrier
- Distinct characteristics
  - User generated content: Massive, dynamic, extensive, instant, and noisy
  - Rich user interactions
  - Collaborative environment, and wisdom of the crowd
  - Many small groups (the long tail phenomenon)
  - Attention is expensive

A Big Variety of Social Media

- Networking
- Sharing
- Publishing
- Gaming
- Discussing
- Location
- Marketing

Two DMML Books of SMM

Twitter Data Analytics
Nov. 2013



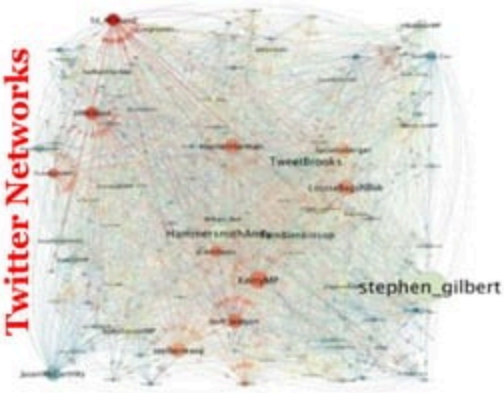
Social Media Mining
Feb. 2014

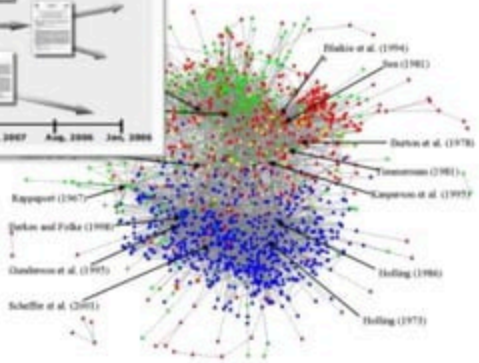
Networks are Pervasive

- A network is a graph.
  - Elements of the network have meanings
- Network problems can usually be represented in terms of graph theory

Internet

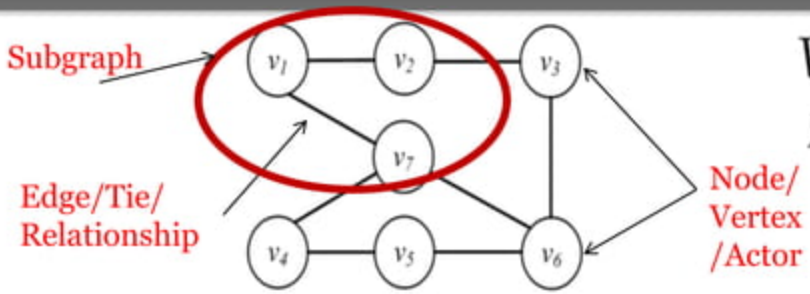
Twitter Networks





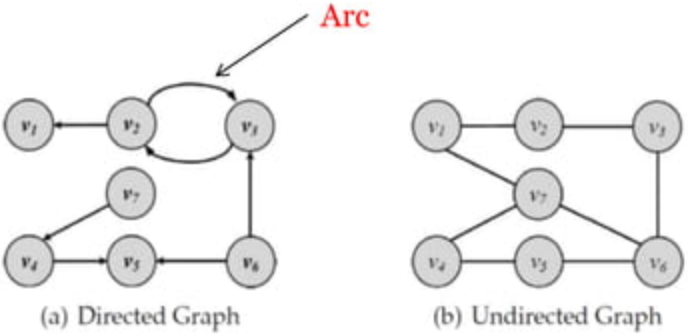
Citation Networks

Nodes/Edges/Degree/Neighborhood



$$V = \{v_1, v_2, \ldots, v_n\}$$

$$E = \{e_1, e_2, \ldots, e_m\}$$



For any node v, the set of nodes it is connected to via an edge is called its neighborhood and is represented as $N(v)$
$N(v1) = \{v2, v7\}$

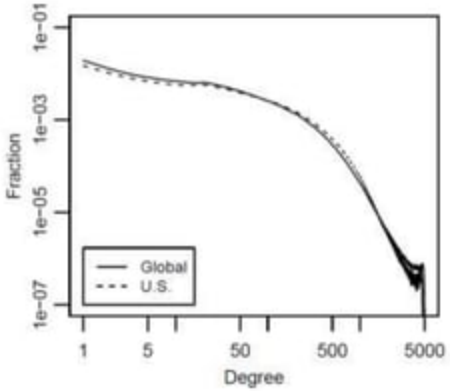
The size of the neighborhood of a node is its degree
$d(v1)=2$

Degree Distribution

When dealing with very large graphs, how nodes' degrees are distributed is an important concept to analyze and is called ***Degree Distribution***

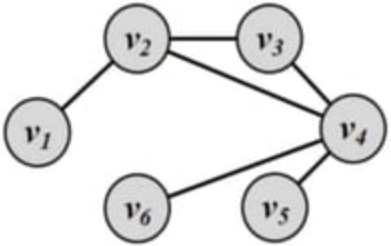
Degree distribution histogram

– The x-axis represents the degree and the y-axis represents the fraction of nodes having that degree



Graph Representation: Adjacency Matrix

$$A_{ij} = \begin{cases} 1, \text{ if there is an edge between nodes } vi \text{ and } vj \\ 0, \text{ otherwise} \end{cases}$$



	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	0	0	0	0
v_2	1	0	1	1	0	0
v_3	0	1	0	1	0	0
v_4	0	1	1	0	1	1
v_5	0	0	0	1	0	0
v_6	0	0	0	1	0	0

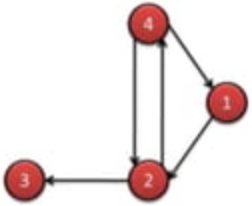
- Diagonal Entries are self-links or loops

Social media networks have very sparse Adjacency matrices

Types of Graphs

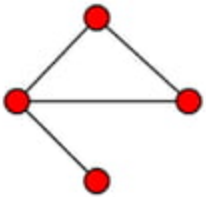
Undirected Graph

$A = A^T$

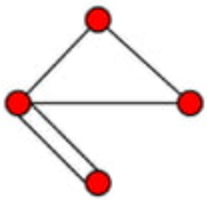


Directed graph

$A \neq A^T$



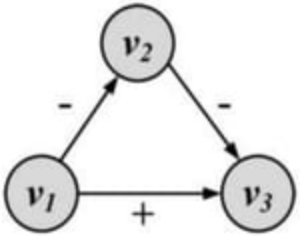
Simple graph



Multigraph

Weighted graph

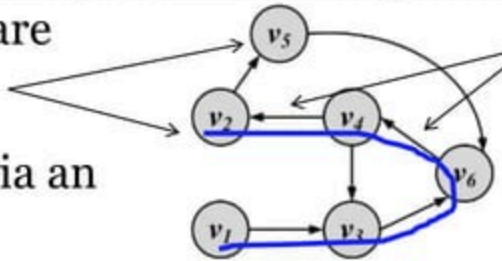
$$A_{ij} = \begin{cases} w, w \in R \\ 0, \text{There is no edge between i and j} \end{cases} \quad \mathbf{G(V, E, W)}$$



Signed graph

Adjacent nodes/Incident Edges/Paths/ Shortest Paths/Connectivity

Two nodes are adjacent if they are connected via an edge.

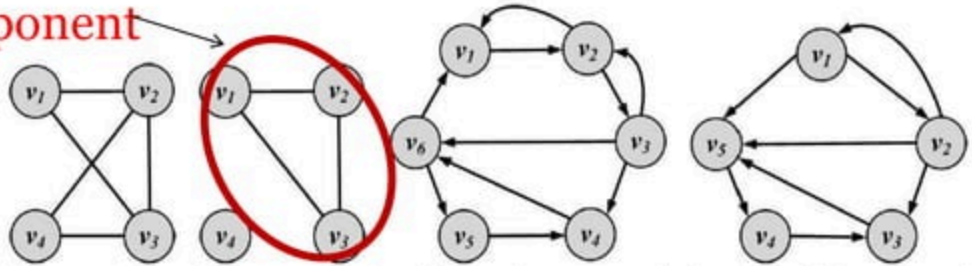


Two edges are incident, if they share one end-point

Length of path= 4

- A walk where **nodes and edges are distinct** is called a **path** and a closed path is called a **cycle**
- The length of a path or cycle is the number of edges visited in the path or cycle

Component



Connected Disconnected Strongly connected Weakly connected

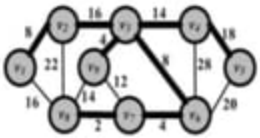
Shortest Path

- **Shortest Path** is the path between two nodes that has the shortest length.
  - We denote the length of the shortest path between nodes v_i and v_j as $l_{i,j}$.
- The concept of the neighborhood of a node can be generalized using shortest paths.
  - An **n-hop neighborhood** of a node is the set of nodes that are within n hops distance from the node.
- The diameter of a graph is the length of the longest shortest path between any pairs of nodes in the graph

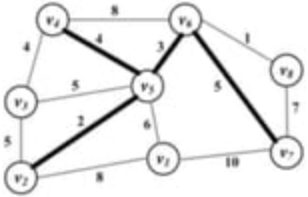
$$diameter_G = \max_{(v_i,v_j)\in V\times V} l_{i,j}.$$

Special Graphs/Properties

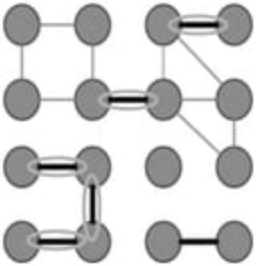
A forest with 3 trees



Spanning Tree

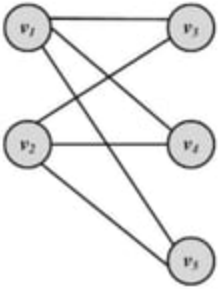


Steiner Tree

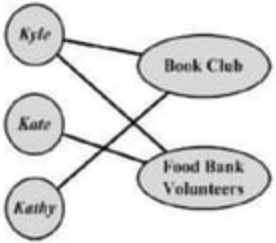


Bridges

Complete Graph

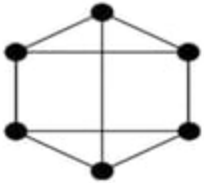


Bipartite Graph



Affiliation Network

Social-Affiliation Network



Regular Graph

Graph Algorithms

- Graph/Tree Traversal Algorithms
  - Depth-First Search (DFS)
  - Breadth-First Search (BFS)
- Shortest Path Algorithms
  - Dijktra's Algorithm
  - Bellman-Ford Algorithm
  - Floyd-Warshall Algorithm
- Minimum Spanning Tree Algorithms
  - Prim's Algorithm
  - Kruskal's Algorithm
- Maximum Flow Algorithms
  - Ford-Fulkerson Algorithm
- Matching Algorithms
  - Bipartite Matching
  - Weighted Matching

Network Measures

Why do we need measures?

1. Who are the central figures (influential individuals) in the network?
2. What interaction patterns are common in friends?
3. Who are the like-minded users and how can we find these similar individuals?

To answer these and similar questions, one first needs to define measures for quantifying centrality (**centrality measures**), level of interactions, and similarity (**similarity measures**), among others.

Centrality

Centrality defines how important a node is within a network.

Centrality: Degree Centrality

- The degree centrality measure ranks nodes with more connections higher in terms of centrality

$$C_d(v_i) = d_i$$

- d_i is the degree (number of adjacent edges) for vertex v_i

In this graph degree centrality for vertex v_1 is $d_1 = 8$ and for all others is $d_j = 1, j \neq 1$

In directed graphs, we can either use the in-degree, the out-degree, or the combination as the degree centrality value:

$$C_d(v_i) = d_i^{in} \quad \textit{(prestige)},$$
$$C_d(v_i) = d_i^{out} \quad \textit{(gregariousness)},$$
$$C_d(v_i) = d_i^{in} + d_i^{out}.$$

Eigenvector Centrality

- Having more friends does not by itself guarantee that someone is more important, but having more **important friends** provides a stronger signal
- Eigenvector centrality tries to generalize degree centrality by incorporating the importance of the neighbors (undirected)
- For directed graphs, we can use incoming or at times, outgoing edges

$$c_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{j,i} c_e(v_j),$$

Let $\mathbf{C_e} = (C_e(v_1), C_e(v_2), \ldots, C_e(v_n))^T$

$$\lambda \mathbf{C}_e = A^T \mathbf{C}_e.$$

This means that C_e is an eigenvector of adjacency matrix A and λ is the corresponding eigenvalue

Katz Centrality

- A major problem with eigenvector centrality arises when it deals with directed graphs
- Centrality only passes over *outgoing* edges and in special cases such as when a node is in a directed acyclic graph centrality becomes zero even though the node can have many edge connected to it
- To resolve this problem, we add bias term β to the centrality values for all nodes
- We select $\alpha < 1/\lambda$, where λ is the largest eigenvalue of A^T For the matrix $(I - \alpha A^T)$ to be invertible

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{Katz}(v_j) + \beta.$$

Rewriting equation in a vector form

$$\mathbf{C}_{Katz} = \alpha A^T \mathbf{C}_{Katz} + \beta \mathbf{1}$$

Katz centrality:

$$\mathbf{C}_{Katz} = \beta(\mathbf{I} - \alpha A^T)^{-1} \cdot \mathbf{1}.$$

vector of all 1's

PageRank

- Problem with Katz Centrality: in directed graphs, once a node becomes an authority (high centrality), it passes **all** its centrality along **all** of its out-links
- This is less desirable since not everyone known by a well-known person is well-known
- To mitigate this problem we can divide the value of passed centrality by the number of outgoing links, i.e., out-degree of that node such that each connected neighbor gets a fraction of the source node's centrality

$$C_p(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} \frac{C_p(v_j)}{d_j^{out}} + \beta.$$

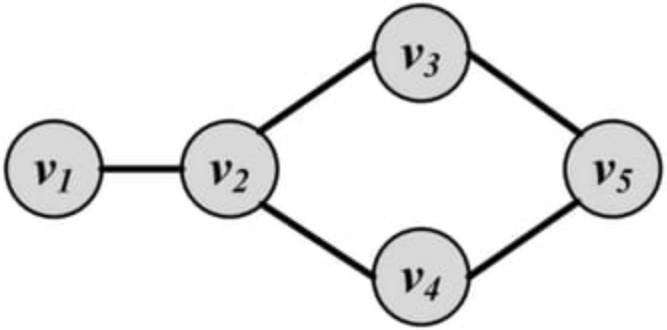
$$D = diag(d_1, d_2, \ldots, d_n)$$

$$\mathbf{C}_p = \alpha A^T D^{-1} \mathbf{C}_p + \beta \mathbf{1},$$

$$\mathbf{C}_p = \beta(\mathbf{I} - \alpha A^T D^{-1})^{-1} \cdot \mathbf{1},$$

Betweenness Centrality

Another way of looking at centrality is by considering how important nodes are in connecting other nodes



$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

$$C_b(v_2) = (\underbrace{(1/1)}_{s=v_1,t=v_3} + \underbrace{(1/1)}_{s=v_1,t=v_4} + \underbrace{(2/2)}_{s=v_1,t=v_5} + \underbrace{(1/2)}_{s=v_3,t=v_4} + \underbrace{0}_{s=v_3,t=v_5} + \underbrace{0}_{s=v_4,t=v_5}) = 3.5,$$

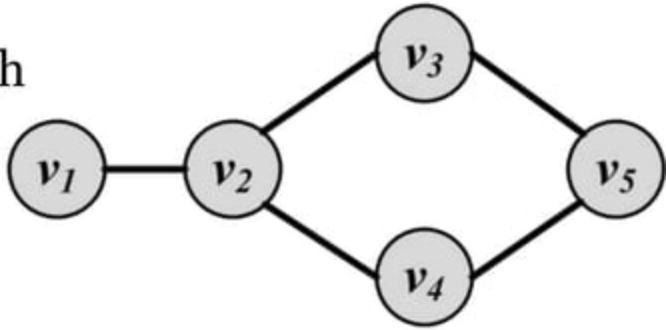
σ_{st} the number of shortest paths from vertex s to t – a.k.a. information pathways

$\sigma_{st}(v_i)$ the number of shortest paths from s to t that pass through vi

In undirected graphs we can assume s<t

Closeness Centrality

- The intuition is that influential and central nodes can quickly reach other nodes
- These nodes should have a smaller average shortest path length to other nodes



$$C_c(v_i) = \frac{1}{\bar{l}_{v_i}}, \qquad \bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

$C_c(v_3) =$

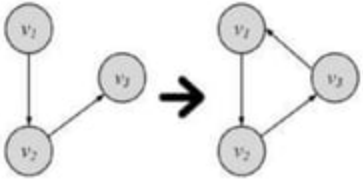
$C_c(v_1) = 1/((1+2+2+3)/4) = 0.5,$
$C_c(v_2) = 1/((1+1+1+2)/4) = 0.8,$
$C_b(v_4) = 1/((1+1+2+2)/4) = 0.66,$
$C_c(v_5) = 1/((1+1+2+3)/4) = 0.57$

Transitivity and Reciprocity

Transitivity

- Mathematic representation:
  - For a transitive relation R:

$$aRb \wedge bRc \rightarrow aRc$$



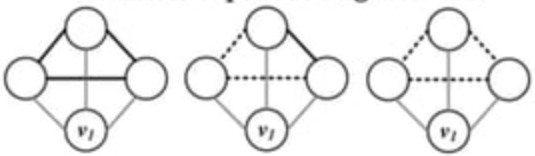
- In a social network:
  - ***Transitivity is when a friend of my friend is my friend***
  - Transitivity in a social network leads to a denser graph, which in turn is closer to a complete graph
  - We can determine how close graphs are to the complete graph by measuring transitivity

[Global] Clustering Coefficient

$$C = \frac{\text{|Paths of Length 2 that have the third edge|}}{\text{|Paths of Length 2|}}$$

Local clustering coefficient measures transitivity at the node level

$$C(v_i) = \frac{\text{number of pairs of neighbors of } v_i \text{ that are connected}}{\text{number of pairs of neighbors of } v_i}.$$

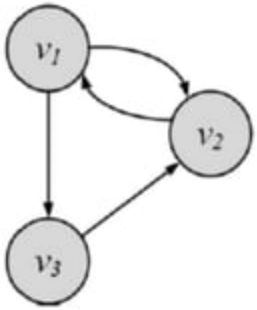


$C(v_1)=1$ $C(v_1)=1/3$ $C(v_1)=0$

Reciprocity

If you become my friend, I'll be yours

- Reciprocity is a more simplified version of transitivity as it considers closed loops of length 2
- If node v is connected to node u, u by connecting to v, exhibits reciprocity



$$R = \frac{\sum_{i,j,j<j} A_{i,j}A_{j,i}}{|E|/2} = \frac{2}{|E|}\sum_{i,j,j<j} A_{i,j}A_{j,i},$$
$$= \frac{2}{|E|} \times \frac{1}{2} Trace(A^2),$$
$$= \frac{1}{|E|} Trace(A^2),$$
$$= \frac{1}{m} Trace(A^2),$$

$$Trace(A) = A_{1,1} + A_{2,2} + \ldots + A_{n,n} = \sum_{i=1}^{n} A_{i,i}$$

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

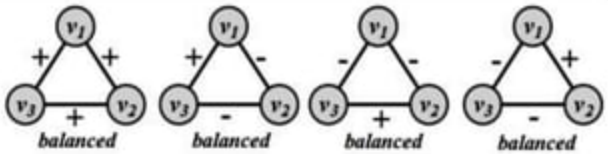
$$R = \frac{1}{m} Trace(A^2) = \frac{2}{4} = \frac{1}{2}$$

Balance and Status

- Measuring stability for an observed network

Social Balance/Social Status

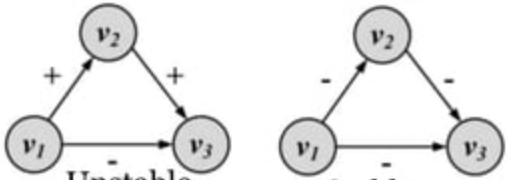
Social balance theory discusses consistency in friend/foe relationships among individuals.



The friend of my friend is my friend,
The friend of my enemy is my enemy,
The enemy of my enemy is my friend,
The enemy of my friend is my enemy.

$$w_{ij}w_{jk}w_{ki} \geq 0$$

- *Status* defines how prestigious an individual is ranked within a societ
- Social status theory measures how consistent individuals are in assigning status to their neighbors



If X has a higher status than Y and Y has a higher status than Z, then X should have a higher status than Z.

Similarity

- How similar two nodes are in a network?

Structural Equivalence

- In structural equivalence, we look at the neighborhood shared by two nodes; the size of this neighborhood defines how similar two nodes are.

For instance, two brothers have in common sisters, mother, father, grandparents, etc. This shows that they are similar, whereas two random male or female individuals do not have much in common and are not similar.

Structural Equivalence: Definitions

- Vertex similarity

$$\sigma(v_i, v_j) = |N(v_i) \cap N(v_j)|.$$

Jaccard Similarity: $$\sigma_{Jaccard}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|},$$

Cosine Similarity: $$\sigma_{Cosine}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}.$$

- In general, the definition of neighborhood N(v) excludes the node itself v.
  - Nodes that are connected and do not share a neighbor will be assigned zero similarity
  - This can be rectified by assuming nodes to be included in their neighborhoods

Regular Equivalence

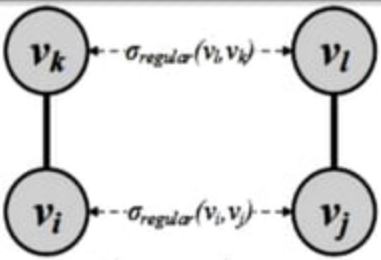
- In regular equivalence, we do not look at neighborhoods shared between individuals, but how neighborhoods themselves are similar

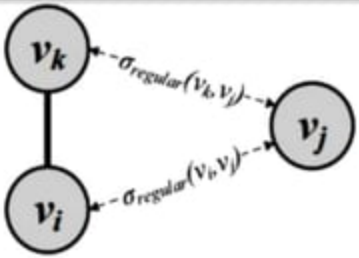
For instance, athletes are similar not because they know each other in person, but since they know similar individuals, such as coaches, trainers, other players, etc.

v_i, v_j are similar when their neighbors v_k and v_l are similar

$$\sigma_{regular}(v_i, v_j) = \alpha \sum_{k,l} A_{i,k} A_{j,l} \sigma_{Regular}(v_k, v_l).$$

Regular Equivalence





v_i, v_j are similar when v_j is similar to v_i's neighbors v_k

$$\sigma_{regular}(v_i, v_j) = \alpha \sum_{k,l} A_{i,k} A_{j,l} \sigma_{Regular}(v_k, v_l).$$

$$\sigma_{regular}(v_i, v_j) = \alpha \sum_{k} A_{i,k} \sigma_{Regular}(v_k, v_j)$$

A vertex is highly similar to itself, we guarantee this by adding an identity matrix to the equation

$$\sigma_{regular} = \alpha A \sigma_{Regular}$$

$$\sigma_{regular} = \alpha A \sigma_{Regular} + \mathbf{I}$$

$$\sigma_{regular} = (\mathbf{I} - \alpha A)^{-1}$$

Network Models

Why should I use network models?

- In may 2011, Facebook had 721 millions users. A Facebook user at the time had an average of 190 users -> a total of 68.5 billion friendships
  - What are the principal underlying processes that help initiate these friendships
  - How can these seemingly independent friendships form this complex friendship network?
- In social media there are many networks with millions of nodes and billions of edges.
  - They are complex and it is difficult to analyze them

So, what do we do?

- We design models that generate, on a smaller scale, graphs similar to real-world networks.
- Hoping that these models simulate properties observed in real-world networks well, the analysis of real-world networks boils down to a cost-efficient measuring of different properties of simulated networks
  - Allow for a better understanding of phenomena observed in real-world networks by providing concrete mathematical explanations; and
  - Allow for controlled experiments on synthetic networks when real-world networks are not available.
- These models are designed to accurately model properties observed in real-world networks

Properties of Real-World Networks

Power-law Distribution,
High Clustering Coefficient, and
Small Average Path Length

Power-law Degree Distribution

- Many sites are visited less than a 1,000 times a month whereas a few are visited more than a million times daily.
- Social media users are often active on a few sites whereas some individuals are active on hundreds of sites.
- There are exponentially more modestly priced products for sale compared to expensive ones.
- There exist many individuals with a few friends and a handful of users with thousands of friends

(Degree Distribution)

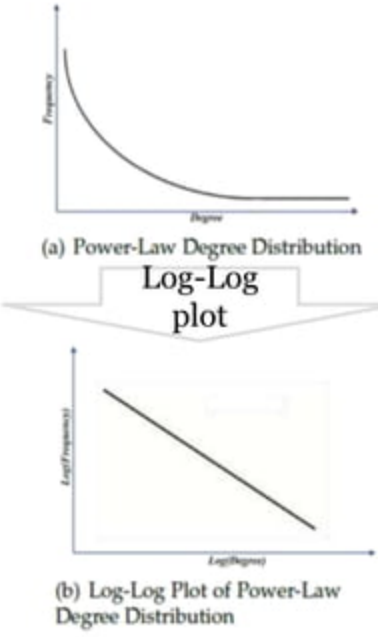
Power-Law Distribution

- When the frequency of an event changes as a power of an attribute:
  - the frequency follows a **power-law**
- Let *f(k)* denote the number individuals having degree k.

$$f(k) = ak^{-b}$$

$$\ln f(k) = -b \ln k + \ln a$$

b: the power-law exponent and its value is typically in the range of [2, 3]
a: power-law intercept



Networks with power-law degree distribution are often called **scale free** networks

High Clustering Coefficient

- In real-world networks, friendships are highly transitive, i.e., friends of an individual are often friends with one another
  - These friendships form triads -> **high average [local] clustering coefficient**

- In May 2011, Facebook had an average clustering coefficient of 0.5 for individuals who had 2 friends.

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
0.081	0.14 (with 100 friends)	0.31	0.33	0.17	0.13

Small Average [Shortest] Path Length

- In real-world networks, any two members of the network are usually connected via short paths. In other words, the average path length is small
  - Six degrees of separation:
    - **Stanley Milgram,** in the well-known small-world experiment conducted in the 1960's, conjectured that people around the world are connected to one another via a path of at most 6 individuals
  - Four degrees of separation:
    - **Lars Backstrom et al.** in May 2011, the average path length between individuals in the Facebook graph was 4.7. (4.3 for individuals in the US)

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
16.12	4.7	5.67	5.88	4.25	5.10

Random Graphs

Random Graphs

- We start with the most basic assumption on how friendships are formed.

Random Graph's main assumption:

Edges (i.e., friendships) between nodes (i.e., individuals) are formed randomly.

Formally, we can assume that for a graph with a fixed number of nodes n, any of the $\binom{n}{2}$ edges can be formed independently, with probability p. This graph is called a random graph and we denote it as G(n, p) model

This model was first proposed independently by Edgar Gilbert and Solomonoff and Rapoport.

Expected Degree

The expected number of edges connected to a node (expected degree) in G(n, p) is $c=(n-1)p$

- **Proof:**
  - A node can be connected to at most n-1 nodes (or n-1 edges)
  - All edges are selected independently with probability p
  - Therefore, on average, (n - 1)p edges are selected

- c=(n-1)p or equivalently,

$$p = \frac{c}{n-1}.$$

Properties of Random Graphs

- **Degree Distribution** $P(d_v = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d},$
  - This is a binomial degree distribution. In the limit this will become the Poisson degree distribution
- **Global Clustering coefficient**
  - The global clustering coefficient of any graph defines the probability of two neighbors of the same node that are connected. This probability is the same for any two nodes and is p
- **Average Path length**
  - (sketch) When moving average path length number of steps away from a node, almost all nodes are visited.

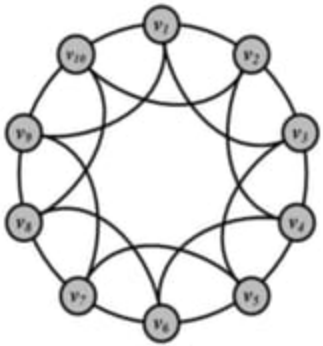
$$c^D \approx c^l \approx |V|.$$

$$l \approx \frac{\ln |V|}{\ln c},$$

Small-World Model

Small-world Model

- In real-world interactions, many individuals have a limited and often at least, a fixed number of connections
- In graph theory terms, this assumption is equivalent to embedding individuals in a regular network.
- A regular (ring) lattice is a special case of regular networks where there exists a certain pattern on how ordered nodes are connected to one another.
- In particular, in a regular lattice of degree c, nodes are connected to their previous c/2 and following c/2 neighbors. Formally, for node set V an edge exists between node i and j if and only if



$$0 < |i - j| \leq c/2.$$

Small-World Model Properties

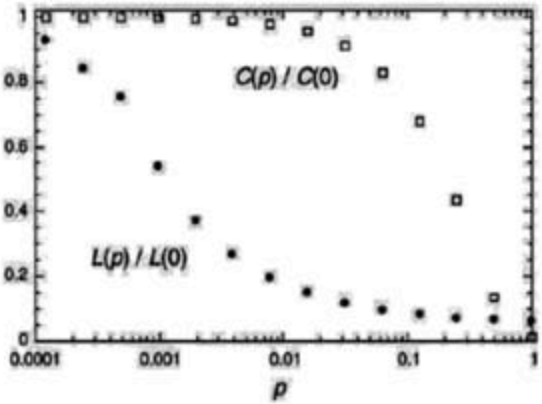
- **Degree distribution:**

$$P(d_v = d) = \sum_{n=0}^{\min(d-c/2,c/2)} \binom{c/2}{n} (1-\beta)^n \beta^{c/2-n} \frac{(\beta c/2)^{d-c/2-n}}{(d-c/2-n)} e^{-\beta c/2},$$

  - In the graph generated by the small world model, most nodes have similar degrees due to the underlying lattice.

- **Clustering Coefficient and Average Path Length:**

C(o) and L(o) are the clustering Coefficient and average path Length of the lattice. We change the value of p such that C(p)/C(o) and L(p)/L(o) Are desirable



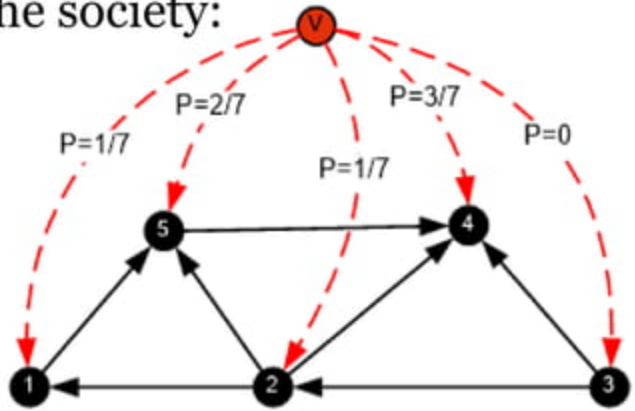
Preferential Attachment Model

Preferential Attachment Model

- Networks:
  - When a new user joins the network, the probability of connecting to an existing node v_i is proportional to the degree of v_i

$$P(v_i) = \frac{d_i}{\sum_j d_j}$$

- Distribution of wealth in the society:
  - The rich get richer
  - Node v arrives
    - P(1) = 1/7
    - P(2) = 1/7
    - P(3) = 0
    - P(4) = 3/7
    - P(5) = 2/7



Preferential Attachment Model Properties

- **Degree Distribution:**

$$P(d) = \frac{2m^2}{d^3},$$

- **Clustering Coefficient:**

$$C = \frac{m_0 - 1}{8} \frac{(\ln t)^2}{t},$$

- **Average Path Length:**

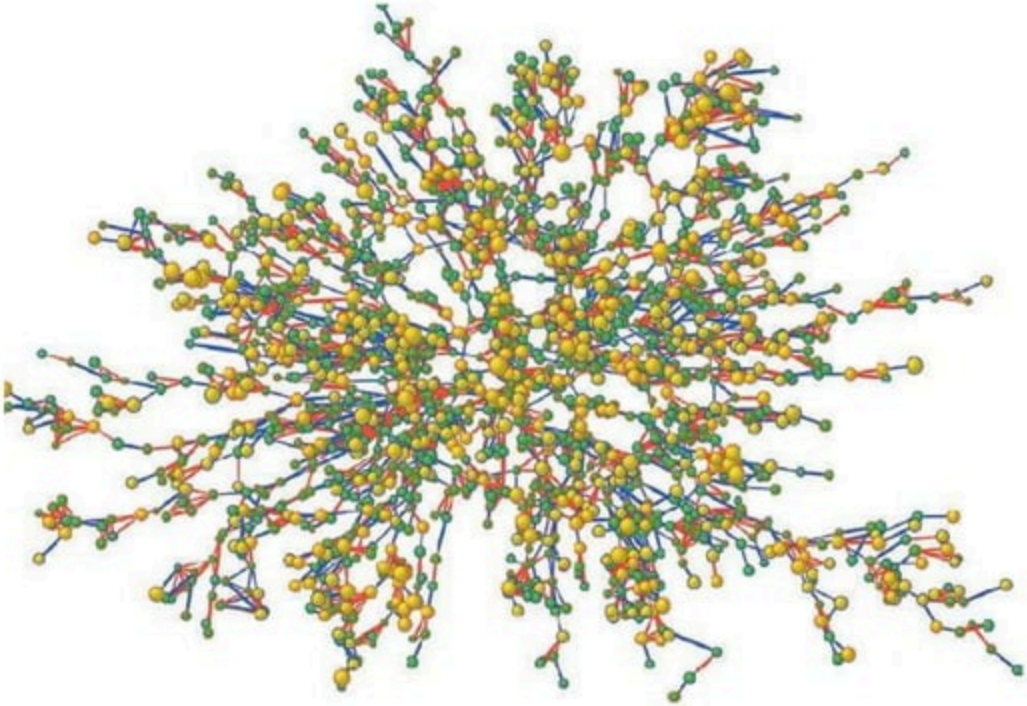
$$l \sim \frac{\ln |V|}{\ln(\ln |V|)}.$$

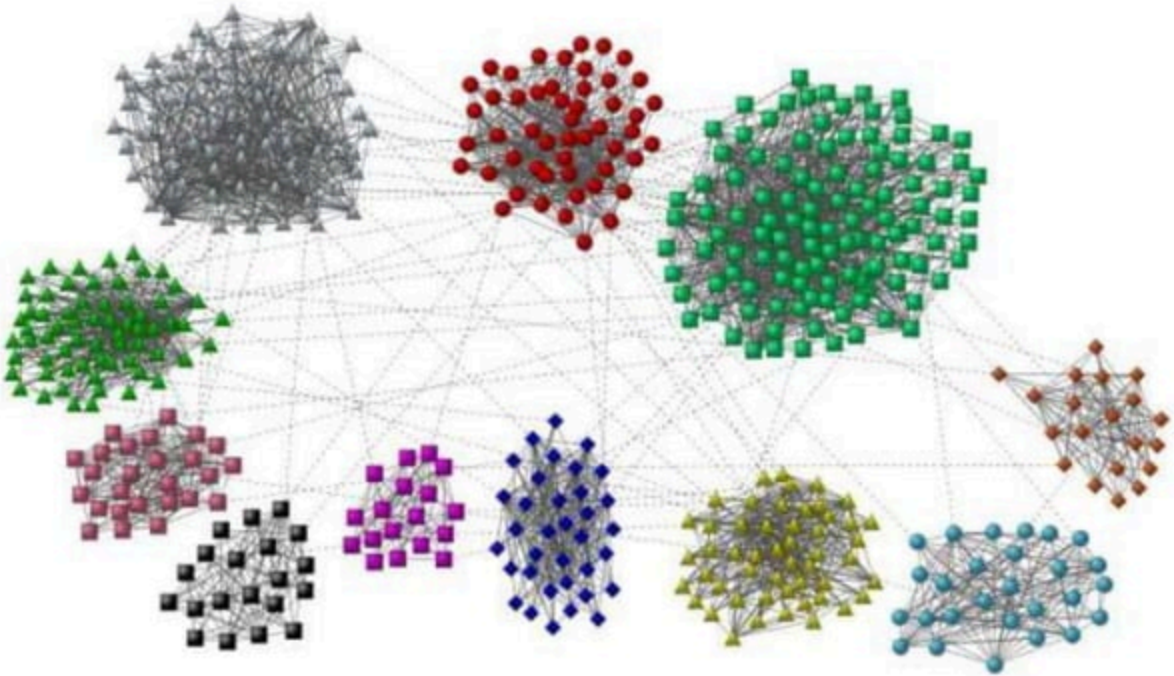
Assortativity in Social Networks

- Why connected people are similar?

Birds of a feather flock together

Social influences on obesity

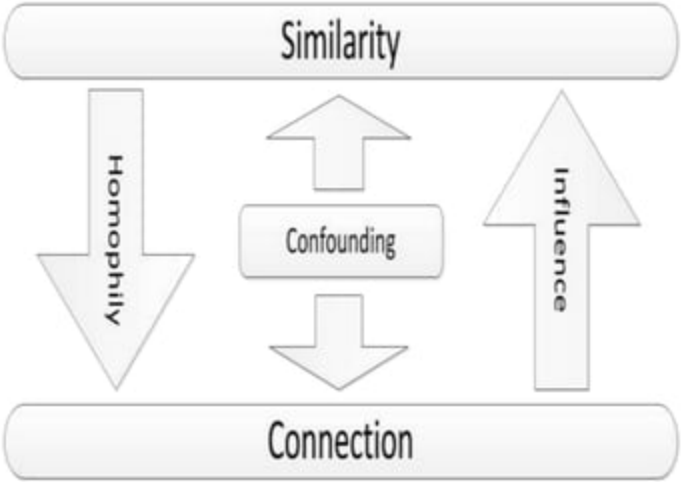


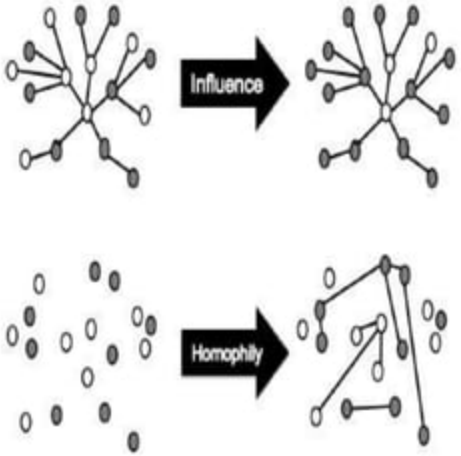


Why connected people are similar?

- **Influence**
  - Influence is the process by which an individual (the influential) affects another individual such that the influenced individual becomes more similar to the influential figure.
    - If most of one's friends switch to a mobile company, he might be influenced by his friends and switch to the company as well.
- **Homophily**
  - It is realized when similar individuals become friends due to their high similarity.
    - Two musicians are more likely to become friends.
- **Confounding**
  - Confounding is environment's effect on making individuals similar
    - Two individuals living in the same city are more likely to become friends than two random individuals

Influence, Homophily, and Confounding





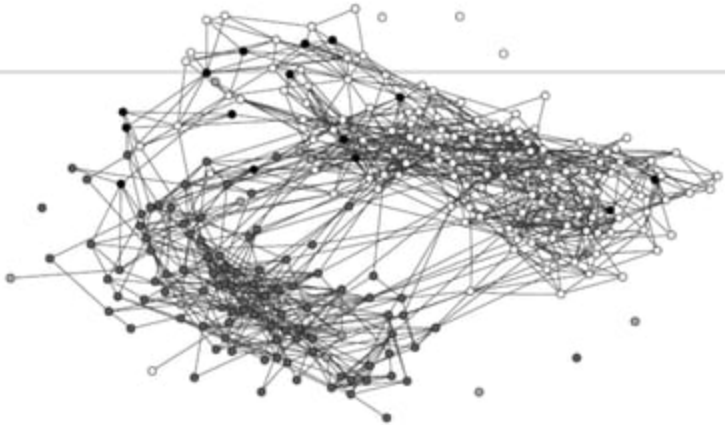
Similarity of Connected Nodes in Social Networks

- Race
- Religion
- Education
- Income level
- Job and skills
- Language
- Interests and preferences

Topics

- Measuring Assortativity
- Measuring Influence and Homophily
- Distinguishing Influence and Homophily

Measuring Assortativity



Measuring Assortativity for Nominal Attributes

- Where nominal attributes are assigned to nodes (race), we can use edges that are between nodes of the same type (i.e., attribute value) to measure assortativity of the network
  - Node attributes could be nationality, race, sex, etc.

$$\frac{1}{m}\sum_{(v_i,v_j)\in E}\delta(t(v_i),t(v_j)) = \frac{1}{2m}\sum_{ij}A_{ij}\delta(t(v_i),t(v_j))$$

$t(v_i)$ denotes type of vertex v_i

$$\delta(x,y) = \begin{cases} 0, & \text{if } x \neq y \\ 1, & \text{if } x = y \end{cases}$$

Kronecker delta function

Assortativity Significance

- Assortativity significance measures the difference between the measured assortativity and its expected assortativity
  - The higher this value, the more significant the assortativity observed
- **Example**
  - Consider a school where half the population is white and half the population is Hispanic. It is expected for 50% of the connections to be between members of different races. If all connections in this school were between members of different races, then we have a significant finding

Assortativity Significance: Measuring

Assortativity

The expected assortativity in the whole graph

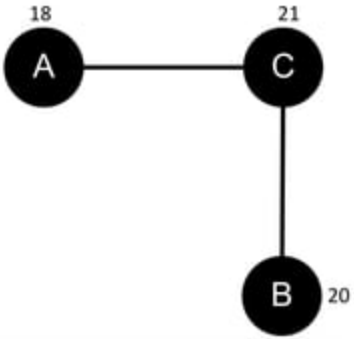
$$Q = \frac{1}{2m}\sum_{ij} A_{ij}\delta(t(v_i), t(v_j)) - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}\delta(t(v_i), t(v_j))$$

$$= \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{d_i d_j}{2m})\delta(t(v_i), t(v_j)).$$

This measure is called modularity

Measuring Assortativity for Ordinal Attributes

- A common measure for analyzing the relationship between ordinal values is covariance.
- It describes how two variables change together.
- In our case we are interested in how values of nodes that are connected via edges are correlated.

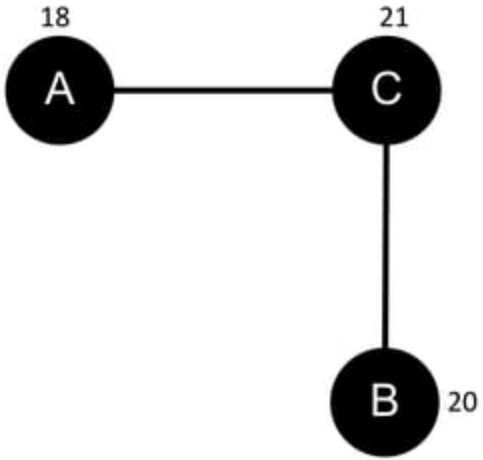


Covariance Variables

- We construct two variables X_L and X_R, where for any edge (v_i, v_j) we assume that x_i is observed from variable X_L and x_j is observed from variable X_R.
- In other words, X_L represents the ordinal values associated with the left node of the edges and X_R represents the values associated with the right node of the edges
- Our problem is therefore reduced to computing the covariance between variables X_L and X_R

Covariance Variables: Example

List of edges:
((A, C),
(C, A),
(C, B),
(B, C))



- X_L : (18, 21, 21, 20)
- X_R : (21, 18, 20, 21)

→

$$E(X_L) = E(X_R),$$
$$\sigma(X_L) = \sigma(X_R).$$

Social Influence

- Measuring Influence
- Modeling Influence

Social Influence: Definition

- The act or power of producing an effect without apparent exertion of force or direct exercise of command

Measuring Influence

- Measuring influence is assigning a number to each node that represents the influential power of that node
- We assume that an individual's attribute or the way she is situated in the network predicts how influential she will be.

- For instance, we can assume that the gregariousness (e.g., number of friends) of an individual is correlated with how influential she will be. Therefore, it is natural to use any of the centrality measures to calculate influence
- An example:
  - On Twitter, in-degree (number of followers) is a benchmark for measuring influence commonly used

Measuring Social Influence on Twitter

Measuring Social Influence on Twitter

- In Twitter, users have an option of following individuals, which allows users to receive tweets from the person being followed
- Intuitively, one can think of the number of followers as a measure of influence (in-degree centrality)

Measuring Social Influence on Twitter: Measures

- **Number of Followers**
  - The number of users following a person on Twitter
  - Indegree denotes the "audience size" of an individual.
- **Number of Mentions**
  - The number of times an individual is mentioned in a tweet, by including @username in a tweet.
  - The number of mentions suggests the "ability in engaging others in conversation"
- **Number of Retweets**:
  - Tweeter users have the opportunity to forward tweets to a broader audience via the retweet capability.
  - The number of retweets indicates individual's ability in generating content that is worth being passed on.
- **Number of Tweets**
- **PageRank (TwitterRank)**

Influence Modeling

- Linear threshold model

Influence Modeling

- At time stamp t_1, node v is activated and node u is not activated
- Node u becomes activated at time stamp t_2, as the effect of the influence
- Each node is started as active or inactive;
- A node, once activated, will activate its neighboring nodes
- Once a node is activated, this node cannot be deactivated

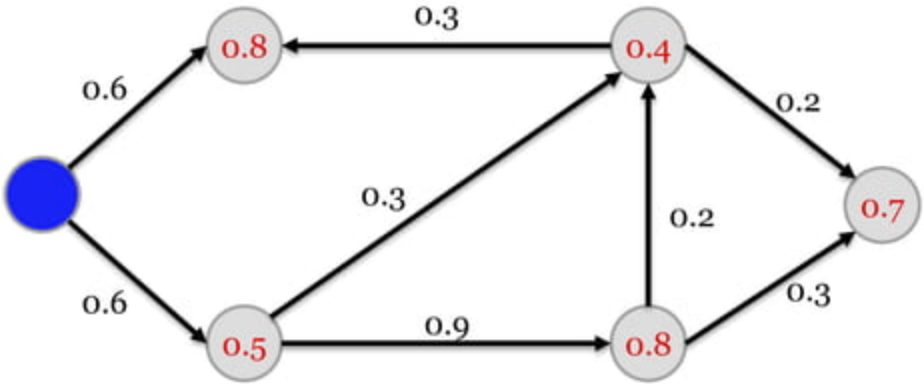
Linear Threshold Model (LTM)

An actor would take an action if the number of his friends who have taken the action exceeds (reach) a certain threshold

- Each node *i chooses a threshold* θ_i *randomly from a uniform* distribution in an interval between 0 and 1.
- In each discrete step, all nodes that were active in the previous step remain active
- The nodes satisfying the following condition will be activated

$$\sum_{j \in N(i), j \in \mathcal{A}} w_{j,i} \geq \theta_i,$$

Linear Threshold Model, an Example



Homophily

"Birds of a feather flock together"

Homophily- Definition

- Homophily (i.e., "love of the same") is the tendency of individuals to associate and bond with similar others
- People interact more often with people who are "like them" than with people who are dissimilar

- What leads to Homophily?
  - Race and ethnicity, Sex and Gender, Age, Religion, Education, Occupation and social class, Network positions, Behavior, Attitudes, Abilities, Believes, and Aspirations

Measuring Homophily: Idea

- To measure homophily, one can measure how the assortativity of the network changes over time
  - Consider two snapshots of a network $G_t(V, E)$ and $G_{t'}(V, E')$ at times t and t', respectively, where $t' > t$
  - Assume that the number of nodes stay fixed and edges connecting them are added or removed over time.

Measuring Homophily

- For nominal attributes, the homophily index is defined as

$$H = Q^{t'}_{normalized} - Q^{t}_{normalized}$$

- For ordinal attributes, the homophily index can be defined as the change in Pearson correlation

$$H = \rho^{t'} - \rho^{t}$$

Distinguishing influence and Homophily

Distinguishing Influence and Homophily

- We are often interested in understanding which social force (influence or homophily) resulted in an assortative network.
- To distinguish between an influence-based assortativity or homophily-based one, statistical tests can be used
- Note that in all these tests, we assume that several temporal snapshots of the dataset are available where we know exactly, when each node is activated, when edges are formed, or when attributes are changed

Shuffle Test

IDEA:

- The basic idea behind the shuffle test comes from the fact that influence is temporal but homophily is not!
- When u influences v, then v should have been activated after u.
  - Define a temporal assortativity measure.
  - Assume that if there is no influence, then a shuffling of the activation timestamps should not affect the temporal assortativity measurement.

Shuffle Test

The key idea of the shuffle test is that if influence does not play a role, the timing of activations should be independent of users. Thus, even if we randomly shuffle the timestamps of user activities, we should obtain a similar *assortativity*

User	A	B	C
Time	1	2	3

User	A	B	C
Time	2	3	1

Test of Influence:
After we shuffle the timestamps of user activities, if the new estimate of social correlation is significantly different from the estimate based on the user's activity log, **there is evidence of influence**.

The Edge-reversal Test

If influence resulted in activation, then the direction of edges should be important (who influenced whom).

- Reverse directions of all the edges
- Run the same logistic regression on the data using the new graph
- If correlation is not due to influence, then α should not change

Community Detection

Social Media Communities

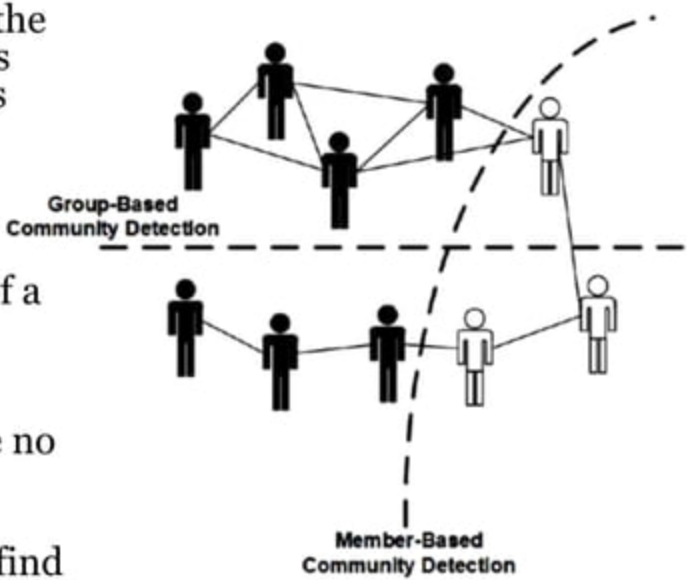
- Any formation of a community requires
  - 1) a set of at least two nodes sharing some interest and
  - 2) interactions with respect to that interest.

- Two types of groups in social media
  - Explicit Groups: formed by user subscriptions
  - Implicit Groups: implicitly formed by social interactions
    - (individuals calling Canada from the United States) -> phone operator considers them one community for marketing purposes
- We may see *group*, *cluster*, *cohesive subgroup*, or *module* in different contexts instead of "community"

- Community detection
  - Discovering implicit communities

Definition

- Community Detection is the process of finding clusters ("communities") of nodes with strong internal connections and weak connections between different clusters
- An ideal decomposition of a large graph is into completely disjoint communities (groups of particles) where there are no interactions between different communities.
- In practice, the task is to find a partition into communities which are maximally decoupled.



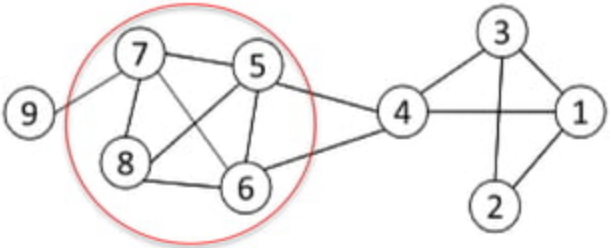
Member-Based Community Detection

Methods that concentrate on properties of nodes and in most cases assume that nodes with similar characteristics represent a community

- Node Characteristics:
  - **Degree:** node with same (or similar) degree are in one community
    - cliques
  - **Reachability:** nodes that are close (small shortest path) are in the same community
    - k-clique, k-club, and k-clan
  - **Similarity:** similar nodes are in the same community

Node-Degree

- **Clique**: a maximum complete subgraph in which all nodes are adjacent to each other



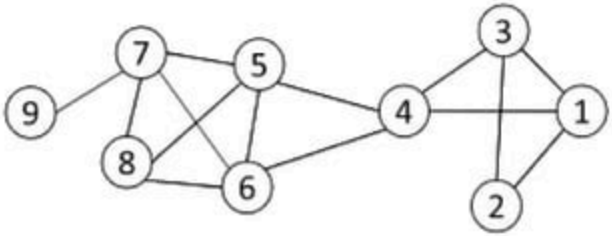
Nodes 5, 6, 7 and 8 form a clique

- NP-hard to find the maximum clique in a network
- Finding cliques is computationally expensive
- The definition of clique is very strict; often, cliques are
  - Relaxed: k-plex; or
  - used as cores or seeds to find larger communities
    - CPM is a method to find communities with overlap

Using cliques as seeds: Clique Percolation Method (CPM): Algorithm

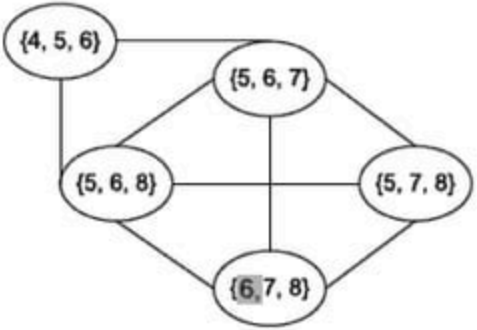
- Input
  - A parameter k, and a network
- Procedure
  - Find out all cliques of size k in the given network
  - Construct a clique graph.
    - Two cliques are adjacent if they share k-1 nodes
  - Each connected components in the clique graph form a community

Clique Percolation Method: Example



Cliques of size 3:

{1, 2, 3}, {1, 3, 4}, {4, 5, 6}, {5, 6, 7}, {5, 6, 8}, {5, 7, 8}, {6, 7, 8}



Communities:

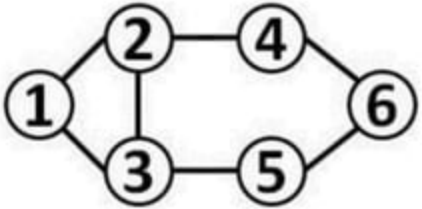
{1, 2, 3, 4}

{4, 5, 6, 7, 8}

Node-Reachability

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
- **k-Club**: it follows the same definition as k-clique with an additional constraint that nodes on the shortest paths should be part of the subgraph
- **k-Clans:** it is a k-clique where for all shortest paths within the subgraph the distance is equal or less than k. All k-clans are k-cliques, but not vice versa.

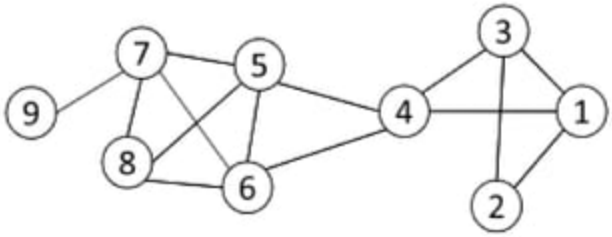


Cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}
2-clubs: {1,2,3,4}, {1, 2, 3, 5}, {2, 3, 4, 5, 6}
2-clans: {2, 3, 4, 5, 6}

Node Similarity

- Apply k-means or similarity-based clustering to nodes
- Vertex similarity is defined in terms of the similarity of their neighborhood
- **Structural equivalence**: two nodes are structurally equivalent iff they are connecting to the same set of actors

Nodes 1 and 3 are structurally equivalent,
So are nodes 5 and 7.



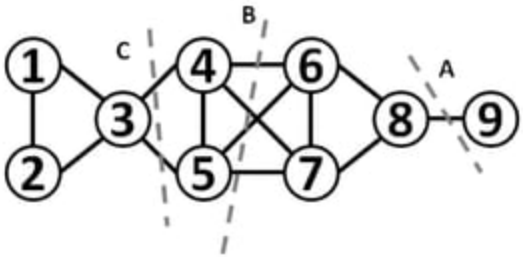
- Structural equivalence is too restrict for practical use.

Group-Based Community Detection

- In group-based community detection, the global network information and topology is considered to determine communities
- We search for communities that are:
  - **Balanced** -> *spectral clustering*
  - **Modular** -> *modularity maximization*
  - **Hierarchical** -> *hierarchical clustering / Girvan-Newman algorithm*
  - **Dense** -> *Quasi-cliques*
  - **Robust** -> *k-connected graphs*

Balanced Communities: Spectral Clustering: Cut

- Most interactions are within group whereas interactions between groups are few
- Cut: A partition of vertices of a graph into two disjoint sets
- Minimum cut problem: find a graph partition such that the number of edges between the two sets is minimized
- Community detection → minimum cut problem



Ratio Cut & Normalized Cut

- Minimum cut often returns an imbalanced partition, with one set being a singleton
- Change the objective function to consider community size

$$\text{Ratio Cut}(P) = \frac{1}{k}\sum_{i=1}^{k}\frac{\text{cut}(P_i, \bar{P}_i)}{|P_i|}$$

$$\text{Normalized Cut}(P) = \frac{1}{k}\sum_{i=1}^{k}\frac{\text{cut}(P_i, \bar{P}_i)}{\text{vol}(P_i)}$$

$\bar{P}_i = V - P_i$ is the complement cut set, $cut(P_i, \bar{P}_i)$ is the size of the cut, and $vol(P_i) = \sum_{v \in P_i} d_v$.

Spectral Clustering

- Both ratio cut and normalized cut can be reformulated as

$$\min_{X \in \{0,1\}^{|V| \times k}} \mathrm{Tr}(X^R L X)$$

$$L = \begin{cases} D - A & \text{Ratio Cut Laplacian, i.e., Unnormalized Laplacian} \\ I - D^{-1/2} A D^{-1/2} & \text{Normalized Laplacian for Normalized Cut.} \end{cases}$$

$D = diag(d_1, d_2, \cdots, d_n)$ Represnts diagonal degree matrix

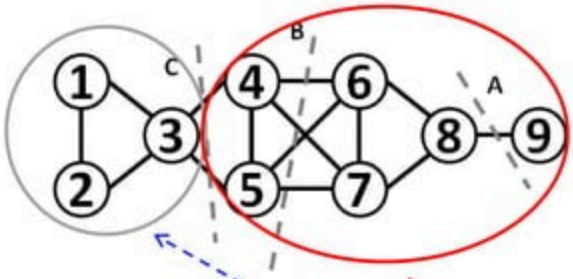
- Spectral relaxation:

$$\min_X \mathrm{Tr}(X^R L X), \quad s.t. \;\; X^T X = I_k$$

Optimal solution:
Top eigenvectors with the smallest eigenvalues

Spectral Clustering: Example

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$



D = diag(2, 2, 4, 4, 4, 4, 4, 3, 1)

$$L = D - A = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 4 & -1 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Eigenvectors

$$\begin{bmatrix} 0.33 & -0.46 \\ 0.33 & -0.46 \\ 0.33 & -0.26 \\ 0.33 & 1.16 \times 10^{-16} \\ 0.33 & 1.16 \times 10^{-16} \\ 0.33 & 0.13 \\ 0.33 & 0.13 \\ 0.33 & 0.33 \\ 0.33 & 0.59 \end{bmatrix}$$

Modular Communities: Modularity Maxmization

- Consider a graph G(V, E), |E| = m where the degrees are known beforehand however edges are not
  - Consider two vertices v_i and v_j with degrees d_i and d_j.
- Now what is an expected number of edges between these two nodes?
- For any edge going out of v_i randomly the probability of this edge getting connected to vertex v_j is

$$\frac{d_j}{\sum_i d_i} = \frac{d_j}{2m}$$

Modularity Maximization: Main Idea

- Given a degree distribution, we know the expected number of edges between any pairs of vertices
- We assume that real-world networks should be far from random. Therefore, the more distant they are from this randomly generated network, the more structural they are.
- Modularity defines this distance and modularity maximization tries to maximize this distance

Normalized Modularity

Consider a partitioning of the data, $P = (P_1, P_2, P_3, ..., P_k)$

For partition P_x, this distance can be defined as

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

This distance can be generalized for a partitioning P

$$\sum_{x=1}^{k} \sum_{i,j \in P_x} A_{ij} - \frac{d_i d_j}{2m}$$

The normalized version of this distance is defined as Modularity

$$Q = \frac{1}{2m} \sum_{x=1}^{k} \sum_{i,j \in P_x} A_{ij} - \frac{d_i d_j}{2m}$$

Modularity Maximization

Modularity matrix

$$B = A - dd^T/2m$$

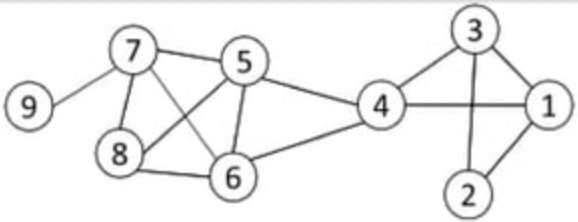
$d \in \mathbb{R}^{n*1}$ is the degree vector for all nodes

Reformulation of the modularity

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

Where, $X \in \mathbb{R}^{n \times k}$ is the indicator (partition membership) function, i.e., $X_{ij} = 1 \Leftrightarrow v_i \in P_j$. This objective can be maximized such that the best membership function is extracted with respect to modularity. Relaxing X to have an orthogonal structure, the optimal X can be computed using the top k eigenvectors of B.

Modularity Maximization: Example



Two Communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

$$B = A - \mathbf{d}\mathbf{d}^T/2m \quad (B_{ij} = A_{ij} - d_i d_j/2m)$$

k-means

$$B = \begin{bmatrix}
-0.32 & 0.79 & 0.68 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\
0.79 & -0.14 & 0.79 & -0.29 & -0.29 & -0.29 & -0.29 & -0.21 & -0.07 \\
0.68 & 0.79 & -0.32 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\
0.57 & -0.29 & 0.57 & -0.57 & 0.43 & 0.43 & -0.57 & -0.43 & -0.14 \\
-0.43 & -0.29 & -0.43 & 0.43 & -0.57 & 0.43 & 0.43 & 0.57 & -0.14 \\
-0.43 & -0.29 & -0.43 & 0.43 & 0.43 & -0.57 & 0.43 & 0.57 & -0.14 \\
-0.43 & -0.29 & -0.43 & -0.57 & 0.43 & 0.43 & -0.57 & 0.57 & 0.86 \\
-0.32 & -0.21 & -0.32 & -0.43 & 0.57 & 0.57 & 0.57 & -0.32 & -0.11 \\
-0.11 & -0.07 & -0.11 & -0.14 & -0.14 & -0.14 & 0.86 & -0.11 & -0.04
\end{bmatrix}$$

Modularity Matrix

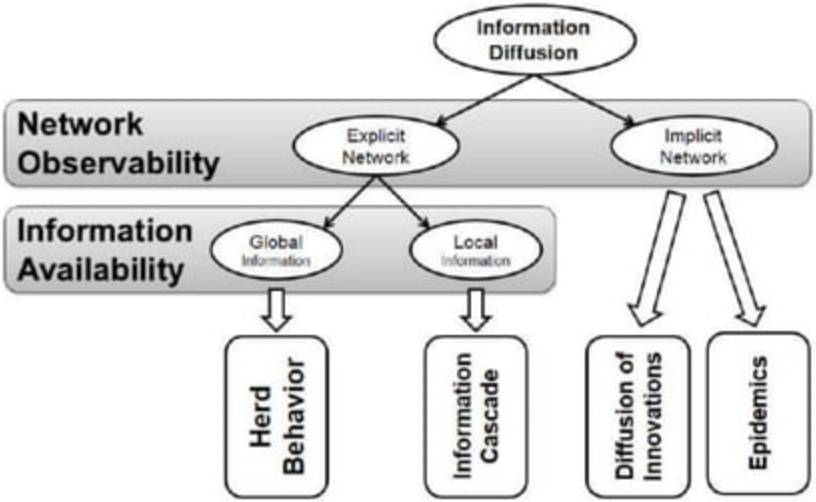
$$S = \begin{bmatrix}
0.44 & -0.00 \\
0.38 & 0.23 \\
0.44 & -0.00 \\
0.17 & -0.48 \\
-0.29 & -0.32 \\
-0.29 & -0.32 \\
-0.38 & 0.34 \\
-0.34 & -0.08 \\
-0.14 & 0.63
\end{bmatrix}$$

Information Diffusion

Information Diffusion

- **Sender(s).** A sender or a small set of senders that initiate the information diffusion process;
- **Receiver(s).** A receiver or a set of receivers that receive diffused information. Commonly, the set of receivers is much larger than the set of senders and can overlap with the set of senders; and
- **Medium.** This is the medium through which the diffusion takes place. For example, when a rumor is spreading, the medium can be the personal communication between individuals

Information Diffusion Types



We define the process of interfering with information diffusion by expediting, delaying, or even stopping diffusion as Intervention

Herd Behavior

- Network is observable
- Only public information is available

Herd Behavior Example

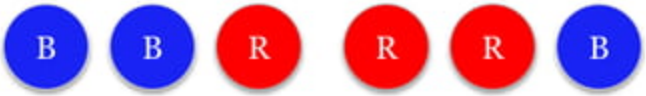
- Consider people participating in an online auction.
- In this case, individuals can observe the behavior of others by monitoring the bids that are being placed on different items.
- Individuals are connected via the auction's site where they can not only observe the bidding behaviors of others, but can also often view profiles of others to get a feel for their reputation and expertise.
- In these online auctions, it is common to observe individuals participating actively in auctions, where the item being sold might otherwise be considered unpopular.
- This is due to individuals trusting others and assuming that the high number of bids that the item has received is a strong signal of its value. In this case, Herd Behavior has taken place.

Herding: Elevator Example

http://www.youtube.com/watch?v=zNNzoyzHcw

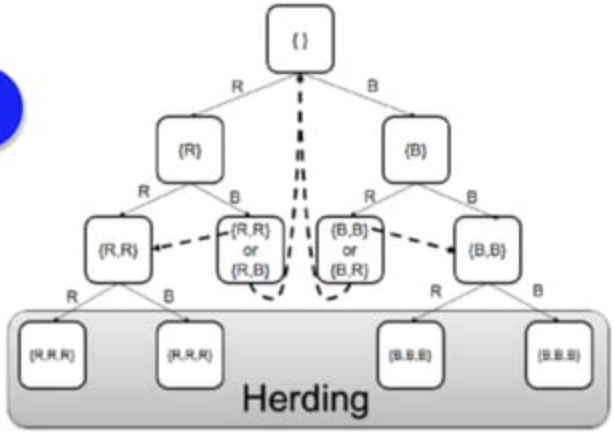
Herding: Urn Experiment

- There is an urn in a large class with three marbles in it



50% 50%

- During the experiment, each student comes to the urn, picks one marble, and checks its color in private.
- The student predicts majority blue or red, writes her prediction on the blackboard, and puts the marble back in the urn.
- Students can't see the color of the marble taken out and can only see the predictions made by different students regarding the majority color on the board



Designing a Herd Behavior Experiment

1. There needs to be a decision made.
   - In our example, it is going to a restaurant

2. Decisions need to be in sequential order;

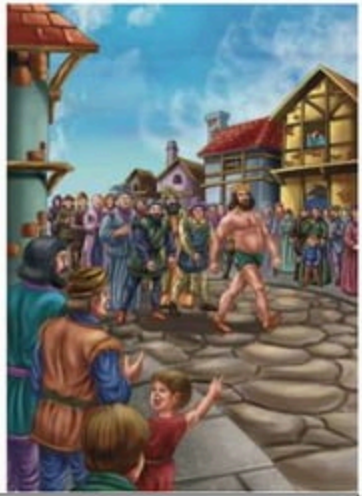
3. Decisions are not mindless and people have private information that helps them decide; and

4. No message passing is possible. Individuals don't know the private information of others, but can infer what others know from what is observed from their behavior.

Herding Intervention

In herding, the society only has access to public information.
Herding may be intervened by releasing private information which was not accessible before

The little boy in "The Emperor's New Clothes" story *intervenes* the herd by shouting "There is no clothe"



Information Cascade

- In the presence of a network
- Only local information is available

Information Cascade

- In social media, individuals commonly repost content posted by others in the network. This content is often received via immediate neighbors (friends).
- An Information Cascade occurs as information propagates through friends

- An information cascade is defined as a piece of information or decision being cascaded among a set of individuals, where
  - 1) individuals are connected by a network and
  - 2) individuals are only observing decisions of their immediate neighbors (friends).

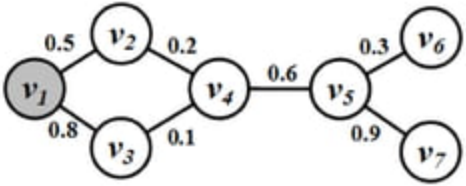
- Therefore, cascade users have less information available to them compared to herding users, where almost all information about decisions are available.

In cascading, local information is available to the users, but in herding the information about the population is available.

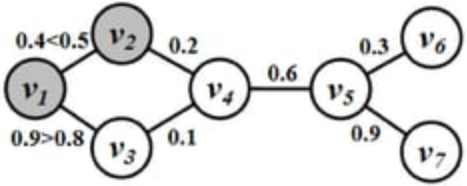
Underlying Assumptions for Cascade Models

- The network is represented using a directed graph. Nodes are actors and edges depict the communication channels between them. A node can only influence nodes that it is connected to;
- Decisions are binary - nodes can be either active or inactive. An active nodes means that the node decided to adopt the behavior, innovation, or decision;
- A node, once activated, can activate its neighboring nodes; and
- Activation is a progressive process, where nodes change from inactive to active, but not vice versa 1.

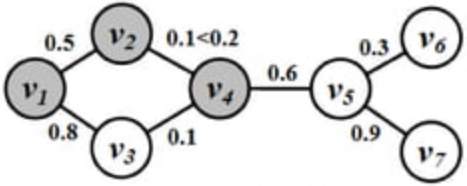
Independent Cascade Model: An Example



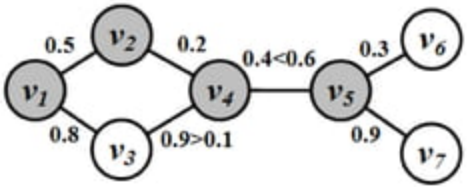
Step 1



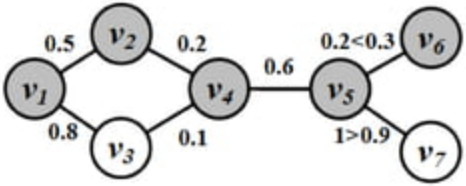
Step 2



Step 3



Step 4



Step 5

Diffusion of Innovations

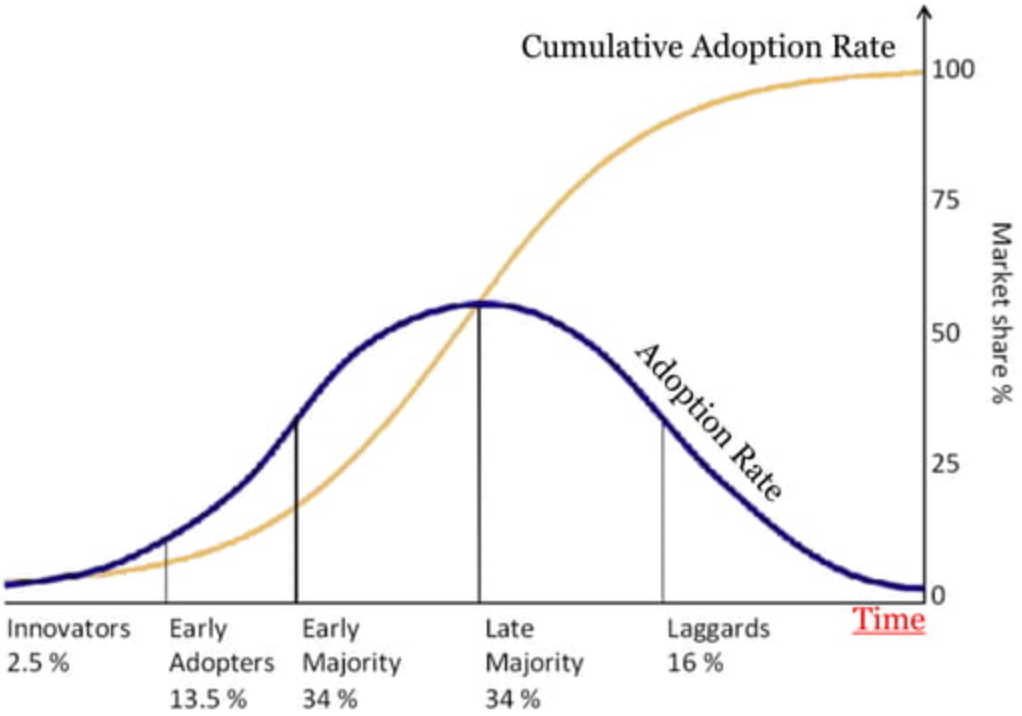
- The network is not observable
- Only public information is observable

Diffusion of Innovation

- an innovation is "an idea, practice, or object that is perceived as new by an individual or other unit of adoption"

- The theory of diffusion of innovations aims to answer why and how these innovations spread. It also describes the reasons behind the diffusion process, individuals involved, as well as the rate at which ideas spread.

Diffusion of innovation: Adopter Categories



Rogers: Diffusion of Innovations: The Process

- **Awareness**
  - The individual becomes aware of the innovation, but her information regarding the product is limited
- **Interest**
  - The individual shows interest in the product and seeks more information
- **Evaluation**
  - The individual tries the product in his mind and decides whether or not to adopt it
- **Trial**
  - The individual performs a trial use of the product
- **Adoption**
  - The individual decides to continue the trial and adopts the product for full use

Social Media Mining Challenges

Some New Challenges in Mining Social Media

- Evaluation Dilemma
  - Evaluation without conventional test data, but how?
- Big-Data Paradox
  - Often we get a small sample of (still big) data. How can we ensure if the data can offer credible findings?
- Noise-Removal Fallacy
  - How do we remove noise without losing too much?

Challenge 1: Evaluation Dilemma

- In conventional data mining, training and test datasets are used to validate findings and compare performance.
- Without training-test data and with the need to evaluate, how can we do it?
  - User study, Amazon Mechanical Turk, …
  - Are they scalable, reproducible, or applicable?
- We need to explore new ways of evaluation.

Understanding User Migration Patterns in Social Media

Joint work with Shamanth Kumar and Reza Zafarani

AAAI 2011, San Francisco, CA

Migration in Social Media

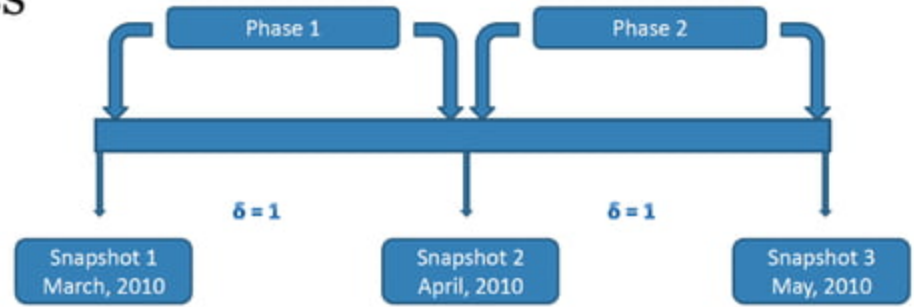
- What is migration?
  - Migration can be described as the movement of users away from one location toward another, either due to necessity, or attraction to the new environment.
- Migration in social media can be of two types
  - Site migration
  - Attention migration

Why is Migration Important?

- Users are a primary source of revenue
  - Ads, Recommendations, Brand loyalty
- New social media sites need to attract new users to expand their user base
- Existing sites need to retain their users by migration prevention
- Competition for attention entails the understanding of migration patterns

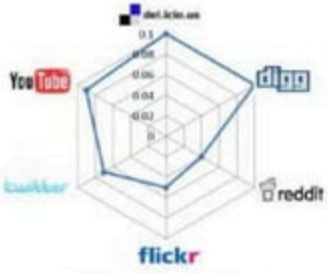
Obtaining User Migration Patterns

- Goal: Identifying trends of attention migration of users across the two phases of the collected data.
- Process

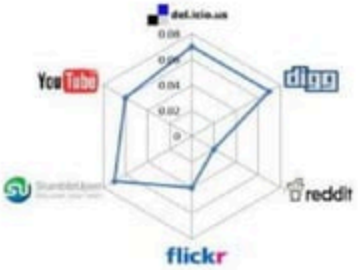


Patterns from Observation

(a) Delicious



(e) StumbleUpon



(f) Twitter

(b) Digg

(d) Reddit

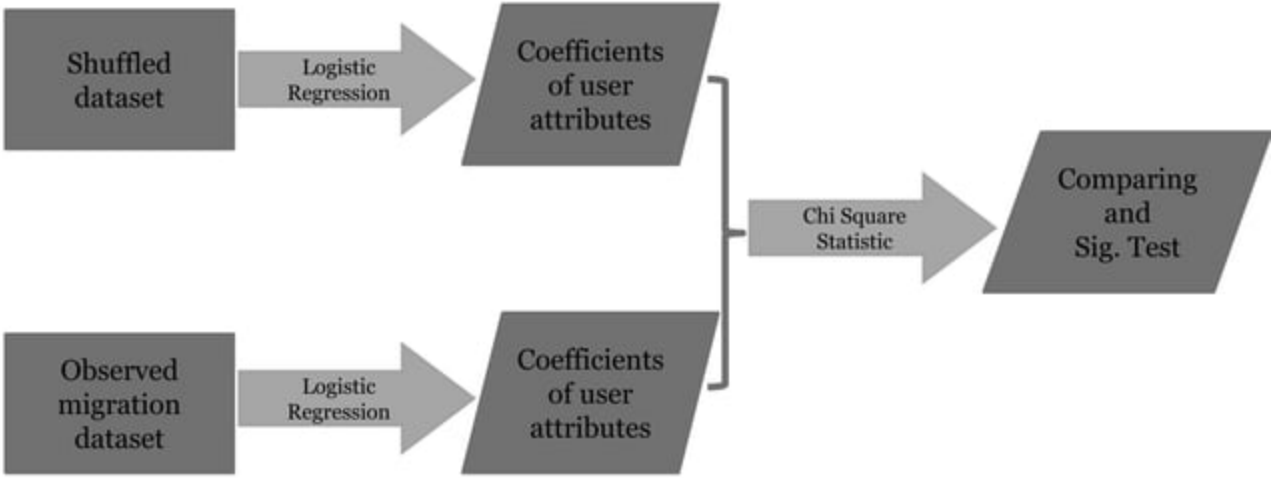
Facing an Evaluation Dilemma

- Important to know if they are valid or not
  - If yes, we investigate further how we use the patterns to: prevention or promotion.
  - If not, why not? And what can we do?
- We would like to evaluate migration patterns, but without ground truth
- How?
  - User study or AMT?

Evaluating Patterns' Validity

- One way is to verify if these patterns are fortuitous
- Null Hypothesis: *Migration of individuals is a **random** process*
  - Generating another similar dataset for comparison
    - Potential migrating population includes overlapping users from Phase 1 and Phase 2
    - Shuffled datasets are generated by picking random active users from the potential migrating population
    - The number of random users selected for each dataset is the same as the real migrating population

A Significance Test



Evaluation Results

- Significant differences observed in StumbleUpon, Twitter, and YouTube
- Patterns from other sites are not statistically significant. Potential cause:
  - Insufficient Data?

Table 2: χ^2 test results on the observed and shuffled data

Site	Observed Coefficients			Shuffled Coefficients			p-value	Statistical Significance
	N	A	R	N	A	R		
Delicious	0.2858	0.4585	-	0.6029	0.5921	-	0.65	Not significant
Digg	0.4796	0.8066	-	0.52	0.5340	-	0.70	Not significant
Flickr	1	1	0.9797	0.2922	0.2759	0.4982	0.13	Not significant
Reddit	0.5385	0.6065	-	0.4846	0.6410	-	0.92	Not significant
StumbleUpon	1	1	-	0.4191	0.2059	-	0.0492	Significant
Twitter	0.5215	1	0.5335	0.2811	0.0365	0.4009	0.0001	Extremely significant
YouTube	0	1	0.1644	0.7219	0.0040	0.4835	0.0001	Extremely significant

Summary

- Mitigating or promoting migration by targeting high net-worth individuals
  - Identifying users with high value to the network, e.g., high network activity, user activity, and external exposure
- Social media migration is first studied in this work
- Migration patterns can be evaluated without test data

Challenge 2: Big-Data Paradox

- What is Big Data?
  - 3Vs, 4Vs, or 5Vs ...
- Is Social Media Data big?
  - Yes, it is obviously so (e.g., FB and Twitter)
  - But, we are often limited by source APIs (e.g., Twitter Streaming API offers 1% data)
- What can we do facing the cold reality?
  - It depends on if you're rich, famous, lazy, or curious

Is the Sample Good Enough? Comparing Data from Twitter's Streaming API and Data from Twitter's Firehose

Joint Work with Fred Morstatter, Jürgen Pfeffer, and Kathleen Carley

AAAI ICWSM2013, Boston, MA

Big-Data Problems

- Twitter provides two main outlets for researchers to access tweets in real time:
  - Streaming API (~1% of all public tweets, free)
  - Firehose (100% of all public tweets, costly)
- Streaming API data is often used to by researchers to validate hypotheses.
- How ***well*** does the sampled Streaming API data measure the true activity on Twitter?

Facets of Twitter Data

- Compare the data along different facets
- Selected facets commonly used in social media mining:
  - Top Hashtags
  - Topic Extraction
  - Network Measures
  - Geographic Distributions

Preliminary Results

Top Hashtags	Topic Extraction
• No clear correlation between Streaming and Firehose data.	• Topics are close to those found in the Firehose.

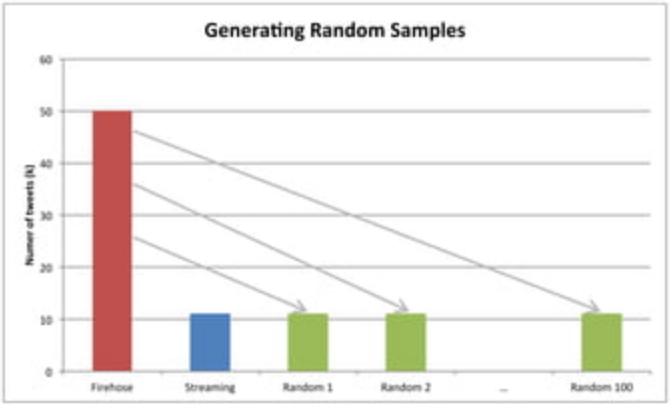
Network Measures	Geographic Distributions
• Found ~50% of the top tweeters by different centrality measures. • Graph-level measures give similar results between the two datasets.	• Streaming data gets >90% of the geotagged tweets. • Consequently, the distribution of tweets by continent is very similar.

How Good are These Results?

- Accuracy of Streaming API varies with analysis the researcher wants to perform.
- These results are about single cases of streaming API.
- Are these findings significant, or just an artifact of random sampling?
- How do we verify that our results indicate sampling bias or not?

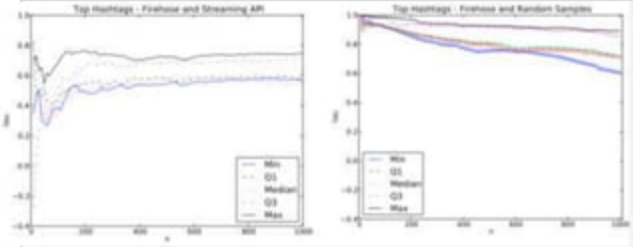
Probing Further

- Aggregate data by day
- Select 5 days with different levels of coverage
- Create random samples from the Firehose data to compare against the Streaming API



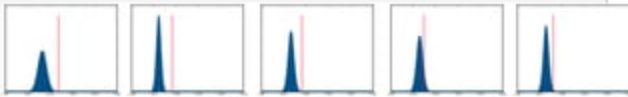
Comparative Results

Top Hashtags



- No correlation between Streaming and Firehose data.
- **Not so in random samples**

Topic Extraction



- Topics are close to those found in the Firehose.
- **Topics extracted with the random data are significantly better.**

Summary

- Streaming API data can be biased in some facets.
- Our results were obtained with the help of Firehose.
- Without Firehose data, challenges are to figure out which facets have biases, and how to compensate them in search of credible mining results

Challenge 3: Noise-Removal Fallacy

- A common complaint: "99% Twitter data is useless".
  - "Had eggs, sunny-side-up, this morning"
  - Can we remove the noise as we usually do in DM?
- What is left after noise removal?
  - Twitter data can be rendered useless after conventional noise removal
- As we are certain there is noise in data, how can we remove it?

Feature Selection with Linked Data in Social Media

Joint Work with Jiliang Tang

SDM2012 and KDD2012

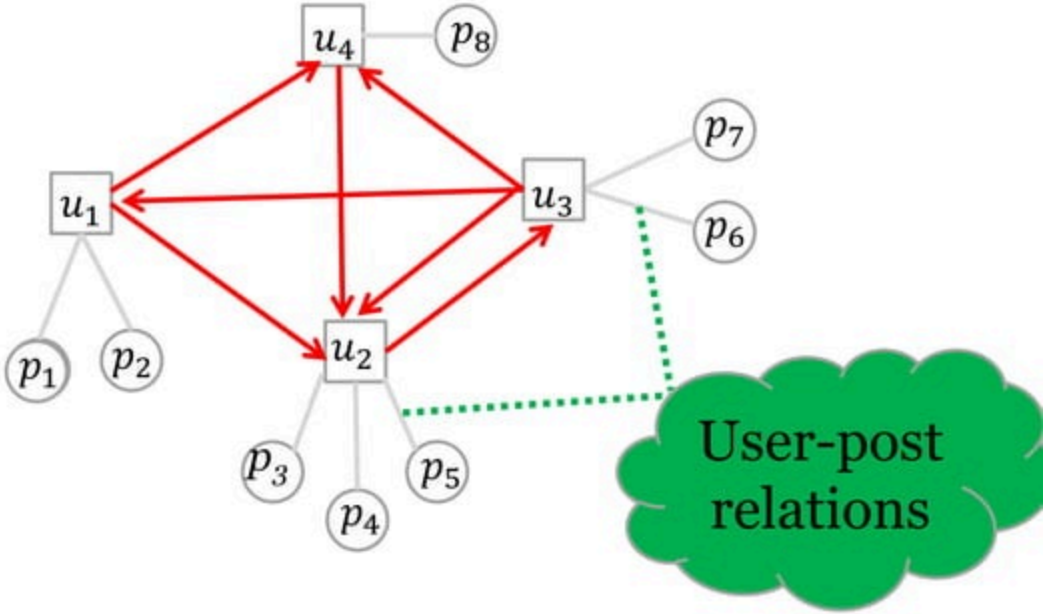
Social Media Data

- Massive and high-dimensional social media data poses unique challenges to data mining tasks
  - Scalability
  - Curse of dimensionality
- Social media data is inherently linked
  - A key difference between social media data and attribute-value data

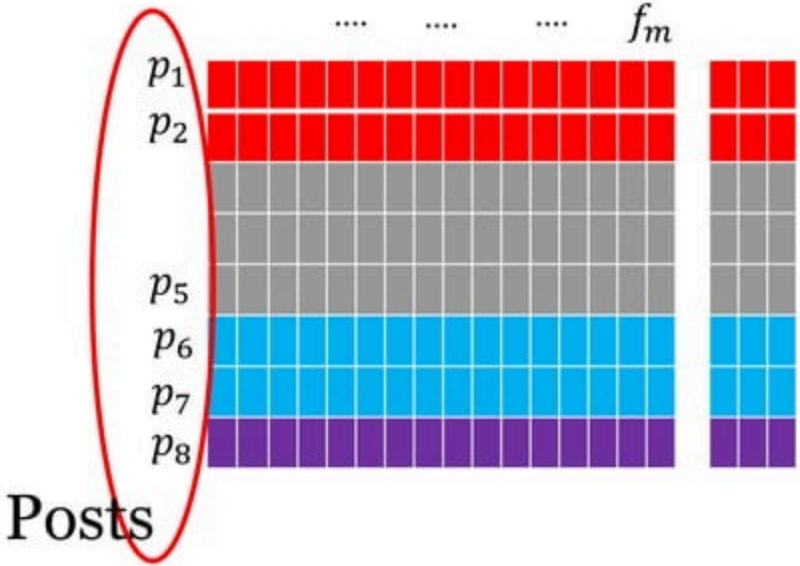
Feature Selection of Social Data

- Feature selection has been widely used to prepare large-scale, high-dimensional data for effective data mining
- Traditional feature selection algorithms deal with only "flat" data (*attribute-value data*).
  - Independent and Identically Distributed (i.i.d.)
- We need to take advantage of linked data for feature selection

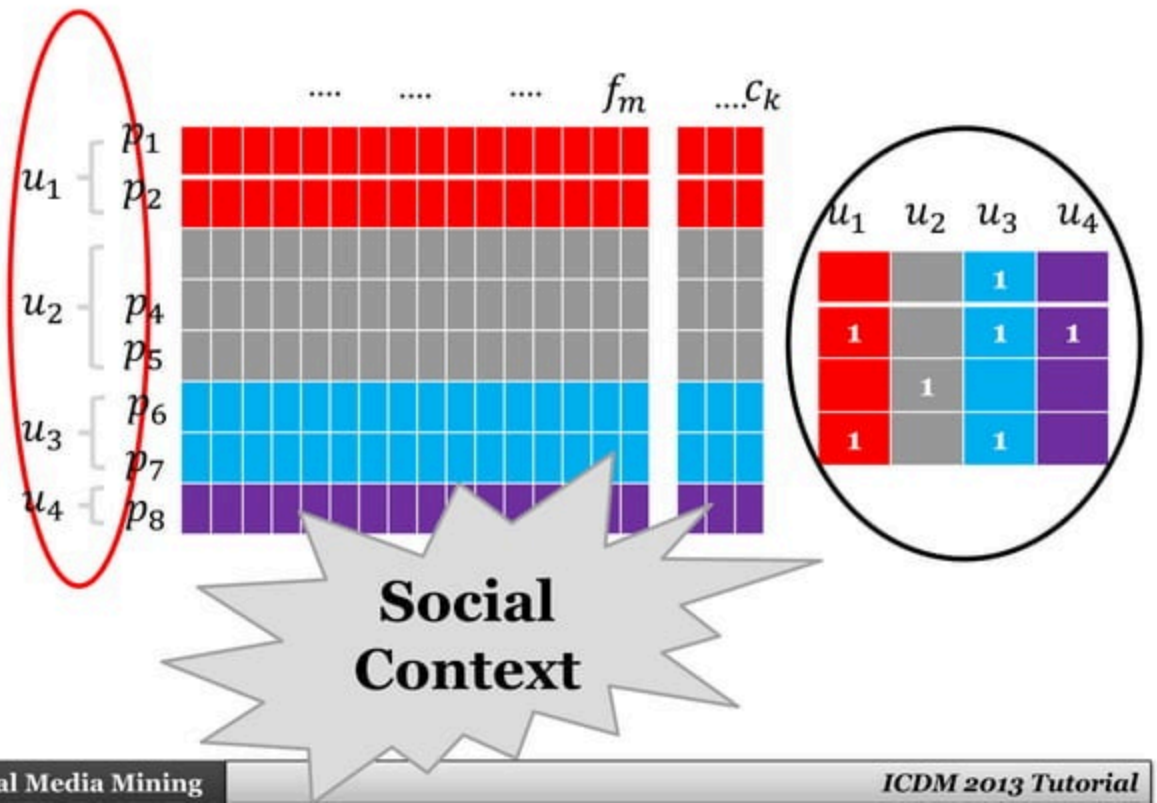
An Example of Social Media Data



Representation for Attribute-Value Data



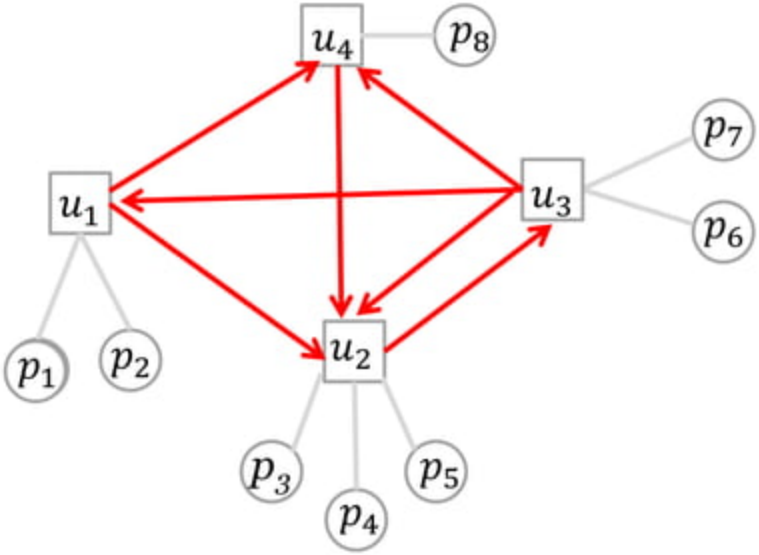
Representation for Social Media Data



How to Use Link Information

- The new question is how to proceed with additional information for feature selection
- Two basic technical problems
  - Relation extraction: What are distinctive relations that can be extracted from linked data
  - Mathematical representation: How to use these relations in feature selection formulation
- Do we have theories to guide us?

Relation Extraction



1. CoPost
2. CoFollowing
3. CoFollowed
4. Following

Relations, Social Theories,

- Social correlation theories suggest that the four relations may affect the relationships between posts
- Social correlation theories
  - Homophily: People with similar interests are more likely to be linked
  - Influence: People who are linked are more likely to have similar interests
- Thus, four relations lead to four hypotheses

Modeling CoFollowing

- Users' topic interests

$$\hat{T}(u_k) = \frac{\sum_{f_i \in F_k} T(f_i)}{|F_k|} = \frac{\sum_{f_i \in F_k} W^T f_i}{|F_k|}$$

- Two co-following users have similar interested topics

$$\min_{W} \| X^T W - Y \|_F^2 + \alpha \| W \|_{2,1} + \beta \sum_{u} \sum_{u_i, u_j \in N_u} \| \hat{T}(u_i) - \hat{T}(u_j) \|_2^2$$

Evaluation Results on Digg

Table 3: Classification Accuracy of Different Feature Selection Algorithms in Digg

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
$\mathcal{T}_5$	50	45.45	44.50	46.33	45.27	**58.82**	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	**59.43**	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	**63.32**
	300	54.04	55.24	56.54	59.27	65.30	60.40	59.93	**66.19**
$\mathcal{T}_{25}$	50	49.91	50.08	51.54	56.02	**58.90**	57.76	57.01	**58.90**
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	**65.02**
	200	59.97	57.37	60.07	64.36	**67.33**	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	**69.52**	65.46	65.01	67.95
$\mathcal{T}_{50}$	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	**60.77**
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	**65.74**
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	**71.32**
	300	61.47	62.35	64.77	69.58	77.86	71.40	70.50	**78.65**
$\mathcal{T}_{100}$	50	51.74	56.06	55.94	58.08	**61.51**	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	**65.65**
	200	60.49	62.78	65.18	66.87	**69.75**	67.40	67.00	67.31
	300	62.97	66.35	67.12	69.27	**73.01**	70.99	69.50	72.64

Summary

- LinkedFS is evaluated under varied circumstances to understand how it works.
  - Link information can help *feature selection for social media data.*
- Unlabeled data is more often in social media, unsupervised learning is more sensible, but also more challenging.
- An unsupervised method is showcased in our KDD12 paper following social correlation theories

Thank You All

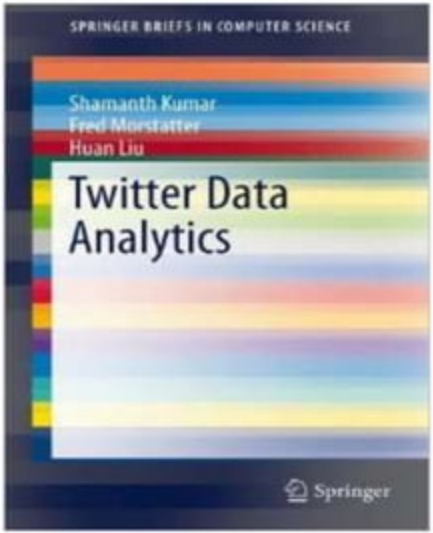
Acknowledgments: Projects are, in part, sponsored by ARO, NSF, and ONR; thanks to passionate and creative DMML members and our collaborators.

Thanks for Your Interests and Participation

Two DMML Books of SMM

Twitter Data Analytics

Nov. 2013



Social Media Mining

Feb. 2014