Social Media and Social Computing

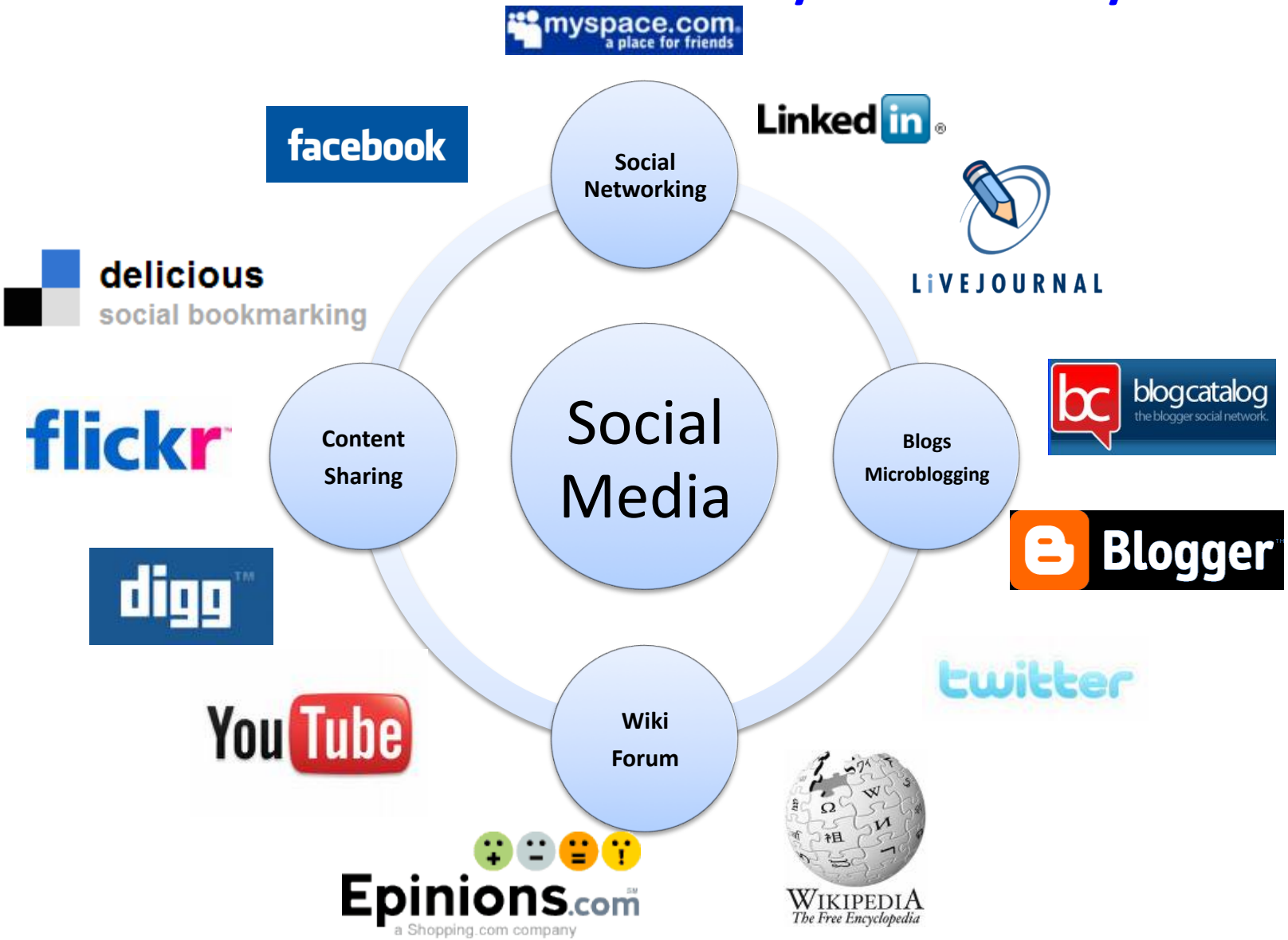
Chapter 1

Traditional Media

Broadcast Media: One-to-Many

Communication Media: One-to-One

Social Media: Many-to-Many

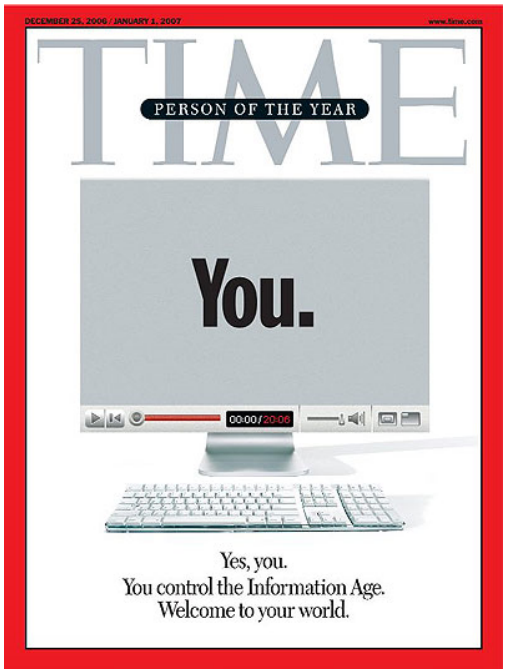


Various forms of Social Media

- Blog: Wordpress, blogspot, LiveJournal
- Forum: Yahoo! Answers, Epinions
- Media Sharing: Flickr, YouTube, Scribd
- Microblogging: Twitter, FourSquare
- Social Networking: Facebook, LinkedIn, Orkut
- Social Bookmarking: Del.icio.us, Diigo
- Wikis: Wikipedia, scholarpedia, AskDrWiki

Characteristics of Social Media

- "Consumers" become "Producers"
- Rich User Interaction
- User-Generated Contents
- Collaborative environment
- Collective Wisdom
- Long Tail



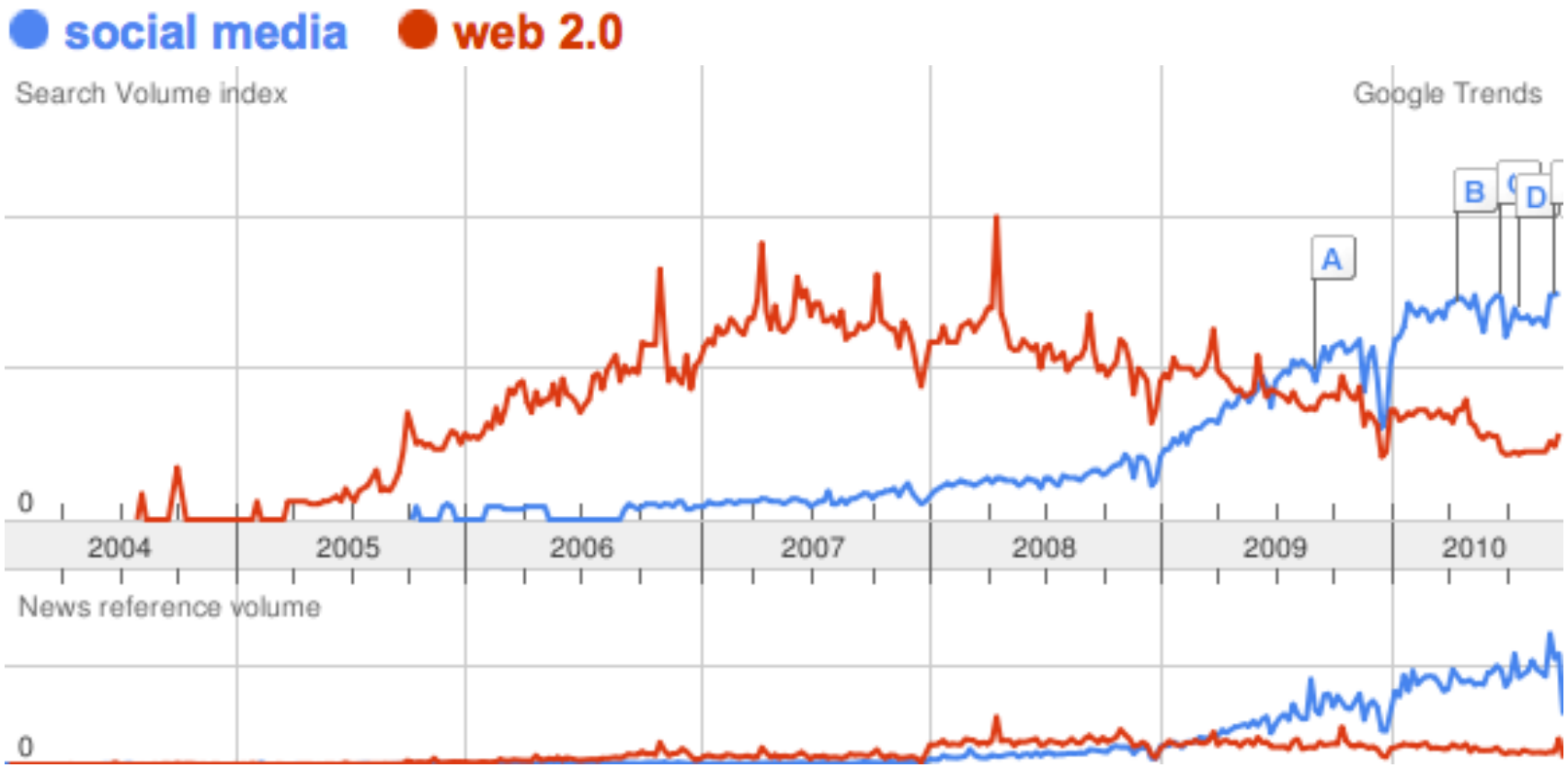
Broadcast Media
Filter, then Publish

Social Media
Publish, then Filter

Top 20 Websites at USA

1	Google.com	11	Blogger.com
2	Facebook.com	12	msn.com
3	Yahoo.com	13	Myspace.com
4	YouTube.com	14	Go.com
5	Amazon.com	15	Bing.com
6	Wikipedia.org	16	AOL.com
7	Craigslist.org	17	LinkedIn.com
8	Twitter.com	18	CNN.com
9	Ebay.com	19	Espn.go.com
10	Live.com	20	Wordpress.com

40% of websites are social media sites


social media
web 2.0
Search Volume index
Google Trends
A
B
D
0
2004
2005
2006
2007
2008
2009
2010
News reference volume
0

How Tweet It Is!: Library Acquires Entire Twitter Archive

April 14th, 2010 by Matt Raymond

[**UPDATE**: We posted an FAQ on April 28]

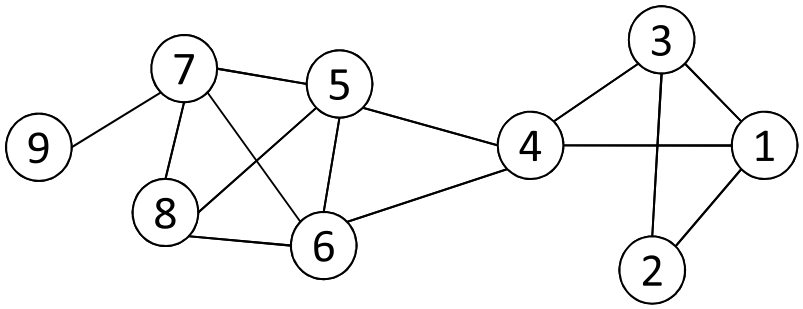
Have you ever sent out a "tweet" on the popular Twitter social media service? Congratulations: Your 140 characters or less will now be housed in the Library of Congress.

That's right. Every public tweet, ever, since Twitter's inception in March 2006, will be archived digitally at the Library of Congress. That's a LOT of tweets, by the way: Twitter processes more than 50 million tweets every day, with the total numbering in the billions.

Networks and Representation

Social Network: A social structure made of nodes (individuals or organizations) and edges that connect nodes in various relationships like friendship, kinship etc.

- Graph Representation



- Matrix Representation

Node	1	2	3	4	5	6	7	8	9
1	-	1	1	1	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0
3	1	1	-	1	0	0	0	0	0
4	1	0	1	-	1	1	0	0	0
5	0	0	0	1	-	1	1	1	0
6	0	0	0	1	1	-	1	1	0
7	0	0	0	0	1	1	-	1	1
8	0	0	0	0	1	1	1	-	0
9	0	0	0	0	0	0	1	0	-

Basic Concepts

- A: the adjacency matrix
- V: the set of nodes
- E: the set of edges
- v_i: a node v_i
- $e(v_i, v_j)$: an edge between node v_i and v_j
- N_i: the neighborhood of node v_i
- d_i: the degree of node v_i
- geodesic: a shortest path between two nodes
  - geodesic distance

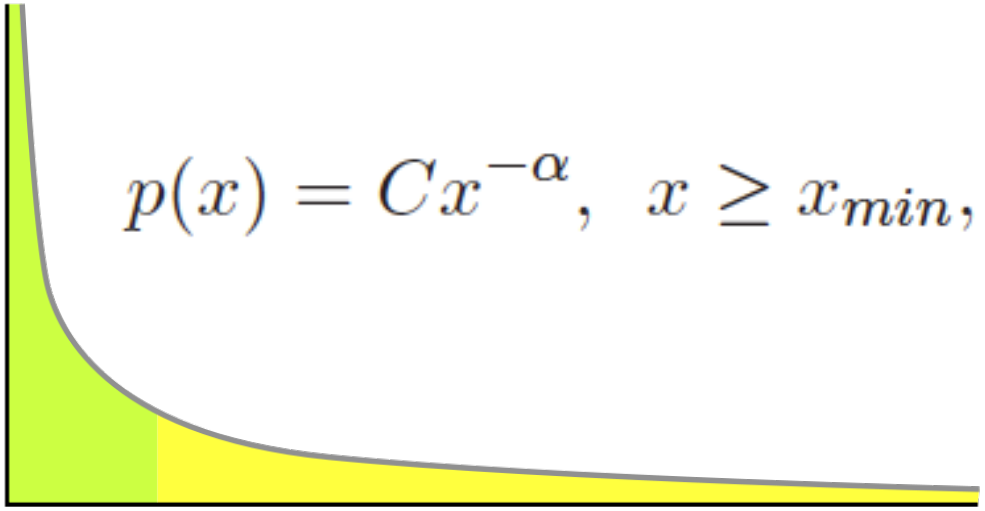
Properties of Large-Scale Networks

- Networks in social media are typically huge, involving millions of actors and connections.

- Large-scale networks in real world demonstrate similar patterns
  - Scale-free distributions
  - Small-world effect
  - Strong Community Structure

Scale-free Distributions

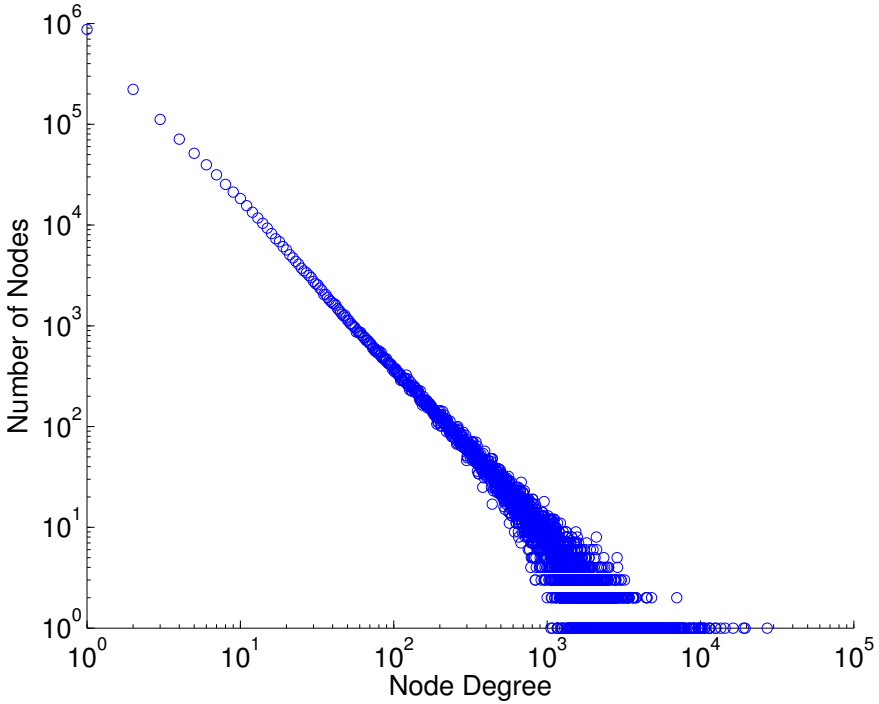
- Degree distribution in large-scale networks often follows a power law.



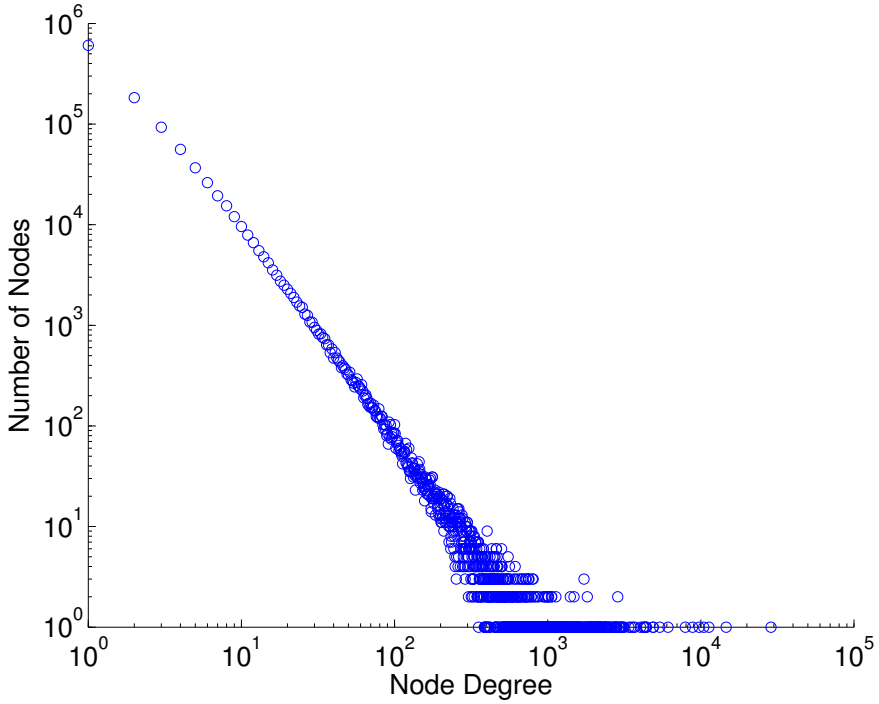
- A.k.a. long tail distribution, scale-free distribution

log-log plot

- Power law distribution becomes a straight line if plot in a log-log scale



Friendship Network in Flickr



Friendship Network in YouTube

Small-World Effect

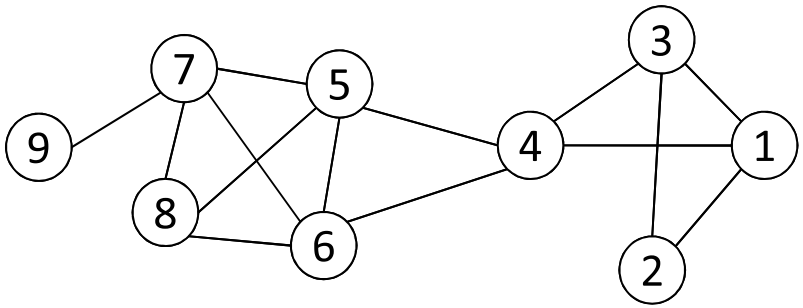
- "Six Degrees of Separation"

- A famous experiment conducted by Travers and Milgram (1969)
  - Subjects were asked to send a chain letter to his acquaintance in order to reach a target person
  - The average path length is around 5.5

- Verified on a planetary-scale IM network of 180 million users (Leskovec and Horvitz 2008)
  - The average path length is 6.6

Diameter

- Measures used to calibrate the small world effect
  - Diameter: the longest shortest path in a network
  - Average shortest path length



- The shortest path between two nodes is called geodesic.
- The number of hops in the geodesic is the geodesic distance.

- The geodesic distance between node 1 and node 9 is 4.
- The diameter of the network is 5, corresponding to the geodesic distance between nodes 2 and 9.

Community Structure

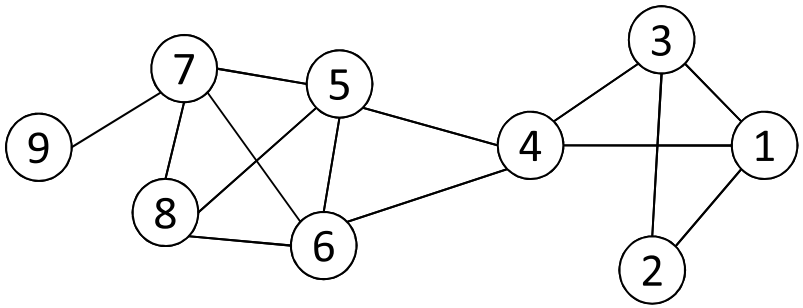
- Community: People in a group interact with each other more frequently than those outside the group

- Friends of a friend are likely to be friends as well
- Measured by clustering coefficient:
  - density of connections among one's friends

$$C_i = \begin{cases} \frac{k_i}{d_i \times (d_i - 1)/2} & d_i > 1 \\ 0 & d_i = 0 \text{ or } 1 \end{cases}$$

Clustering Coefficient

$$C_i = \begin{cases} \dfrac{k_i}{d_i \times (d_i - 1)/2} & d_i > 1 \\ 0 & d_i = 0 \text{ or } 1 \end{cases}$$



- d_6=4, N_6= {4, 5, 7,8}
- k_6=4 as e(4,5), e(5,7), e(5,8), e(7,8)
- C_6 = 4/(4*3/2) = 2/3
- Average clustering coefficient

  C = (C_1 + C_2 + … + C_n)/n

- C = 0.61 for the left network
- In a random graph, the expected coefficient is 14/(9*8/2) = 0.19.

Challenges

- Scalability
  - Social networks are often in a scale of millions of nodes and connections
  - Traditional Network Analysis often deals with at most hundreds of subjects
- Heterogeneity
  - Various types of entities and interactions are involved
- Evolution
  - Timeliness is emphasized in social media
- Collective Intelligence
  - How to utilize wisdom of crowds in forms of tags, wikis, reviews
- Evaluation
  - Lack of ground truth, and complete information due to privacy

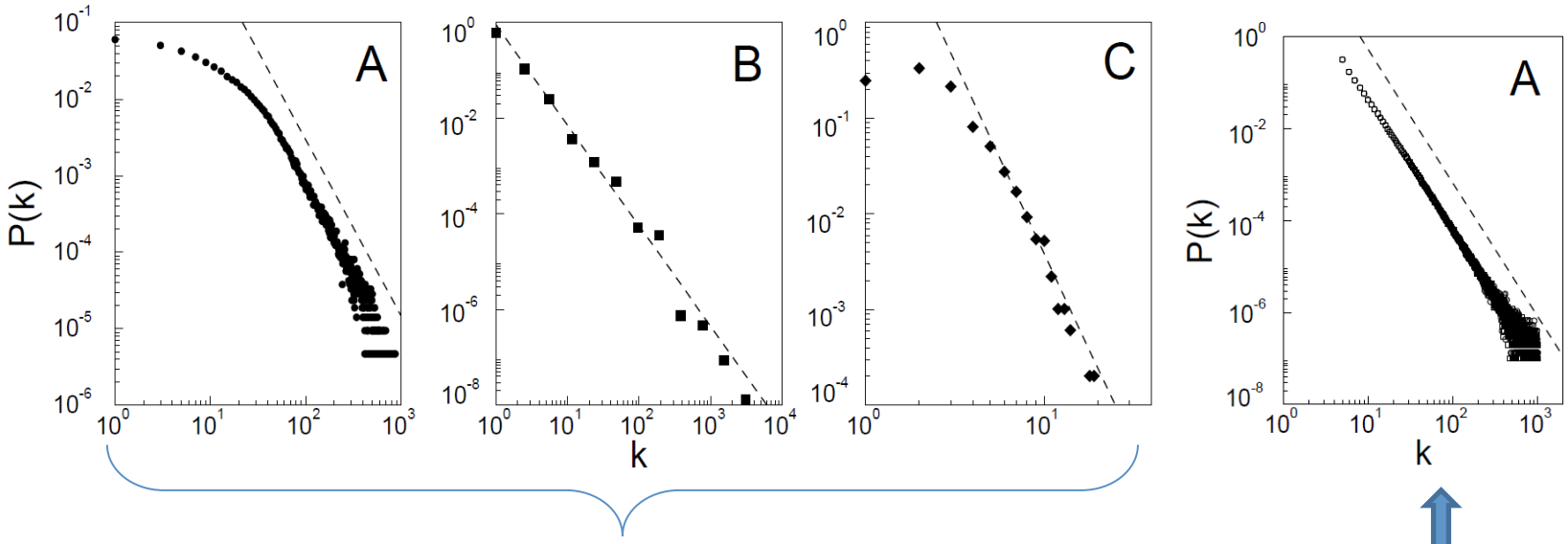
Social Computing Tasks

- Social Computing: a young and vibrant field
- Many new challenges
- Tasks
  - Network Modeling
  - Centrality Analysis and Influence Modeling
  - Community Detection
  - Classification and Recommendation
  - Privacy, Spam and Security

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, Watts and Strogatz model
- 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

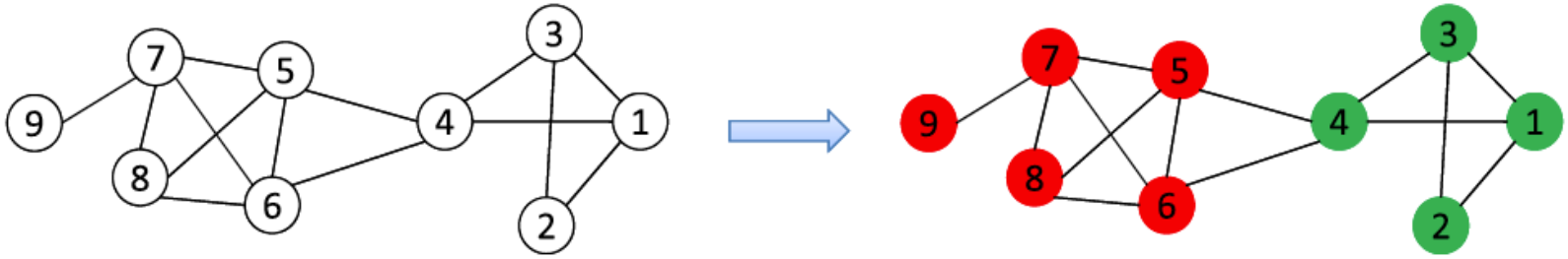
(Figures borrowed from "*Emergence of Scaling in Random Networks*")

Centrality Analysis and Influence Modeling

- Centrality Analysis:
  - Identify the most important actors or edges
  - Various criteria
- Influence modeling:
  - How is information diffused?
  - How does one influence each other?
- Related Problems
  - Viral marketing: word-of-mouth effect
  - Influence maximization

Community Detection

- A community is a set of nodes between which the interactions are (relatively) frequent
  - A.k.a., *group, cluster, cohesive subgroups, modules*

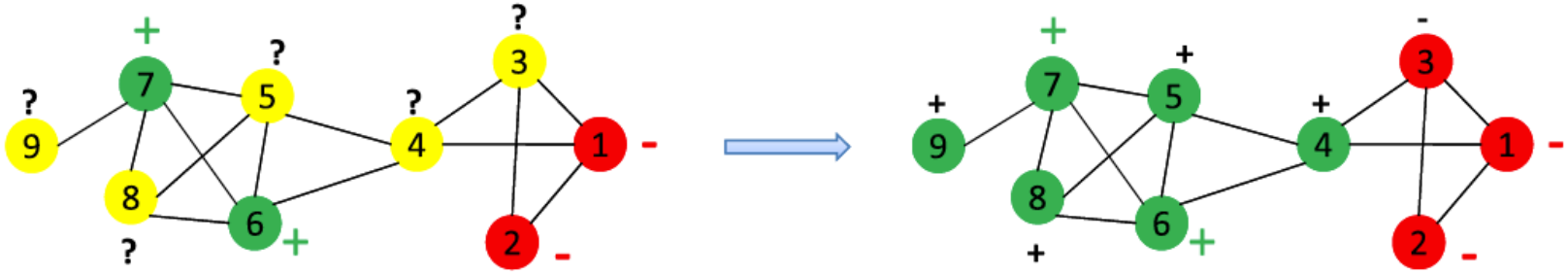


- Applications: *Recommendation based communities, Network Compression, Visualization of a huge network*
- New lines of research in social media
  - Community Detection in Heterogeneous Networks
  - Community Evolution in Dynamic Networks
  - Scalable Community Detection in Large-Scale Networks

Classification and Recommendation

- Common in social media applications
  - Tag suggestion, Friend/Group Recommendation, Targeting

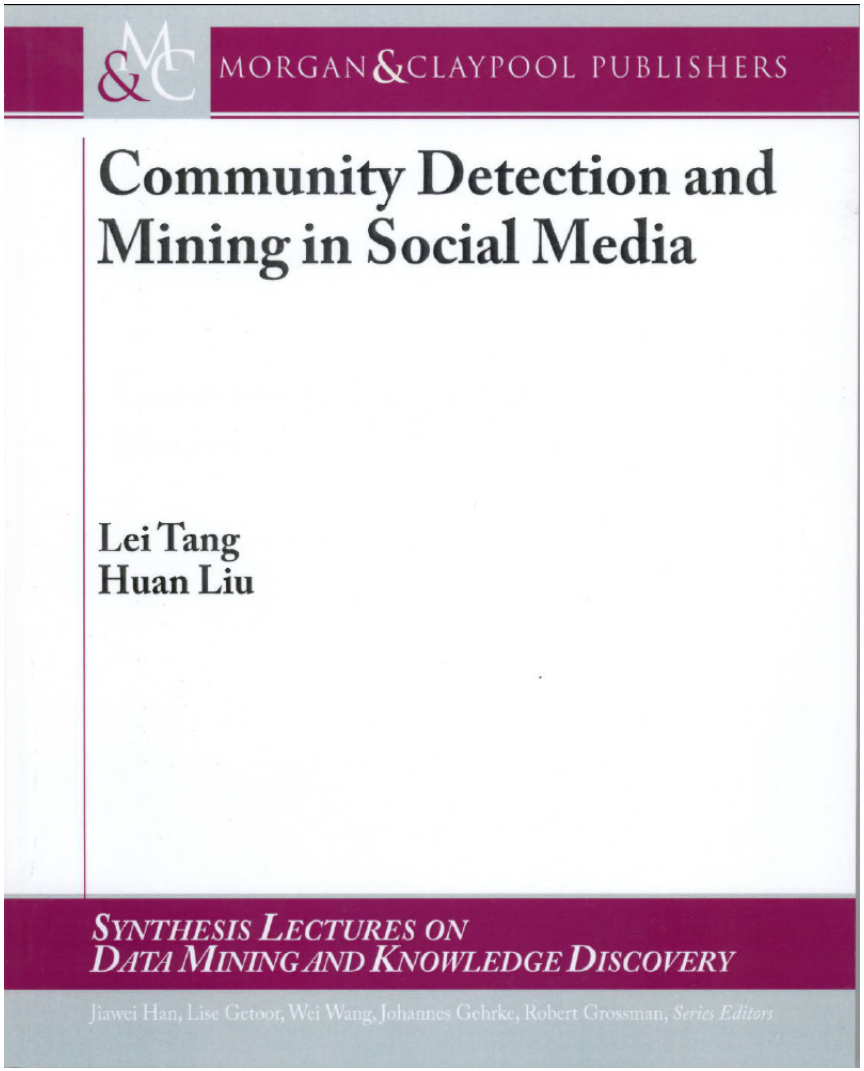
Link prediction



Network-Based Classification

Privacy, Spam and Security

- Privacy is a big concern in social media
  - Facebook, Google buzz often appear in debates about privacy
  - NetFlix Prize Sequel cancelled due to privacy concern
  - Simple annoymization does not necessarily protect privacy
- Spam blog (splog), spam comments, Fake identity, etc., all requires new techniques
- As private information is involved, a secure and trustable system is critical
- Need to achieve a balance between sharing and privacy



Book Available at

- Morgan & claypool Publishers
- Amazon

If you have any comments, please feel free to contact:

- **Lei Tang**, Yahoo! Labs, ltang@yahoo-inc.com
- **Huan Liu**, ASU huanliu@asu.edu