MODULE 1

INTRODUCTION TO SOCIAL NETWORK AND ANALYTICS

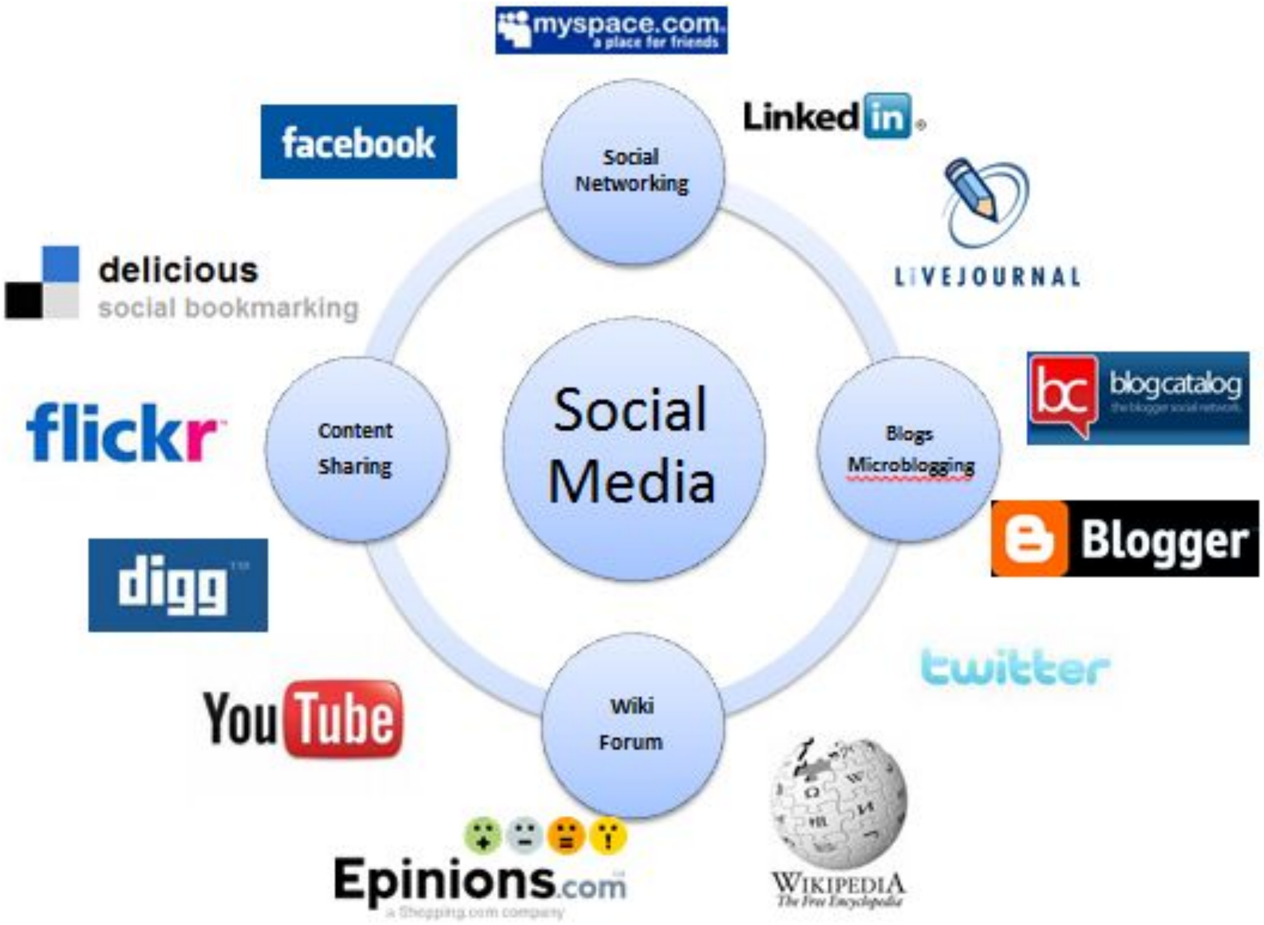
What is a social network?

- The term *social network* refers to the articulation of a social relationship, ascribed or achieved, among individuals, families, households, villages, communities, regions, and so on.

- Each of them can play dual roles, acting both as a unit or node of a social network as well as a social actor
- Actors –are the smallest unit of a network
  - Persons
  - Organizations
  - Countries
  - Companies
  - Animals
  - Words
  - Web pages
  - Families

- Social network analysis (SNA) is a collection of techniques, tools, and methods to map and measure the relationships among people and organizations
- SNA is multidisciplinary and deals with
  - Sociology
  - Graph theory
  - Computer science
  - Mathematics
  - Economics
  - Women Studies
  - Development Studies

Social Media: Many-to-Many


Social Media

Social Networking

Blogs Microblogging

Wiki Forum

Content Sharing

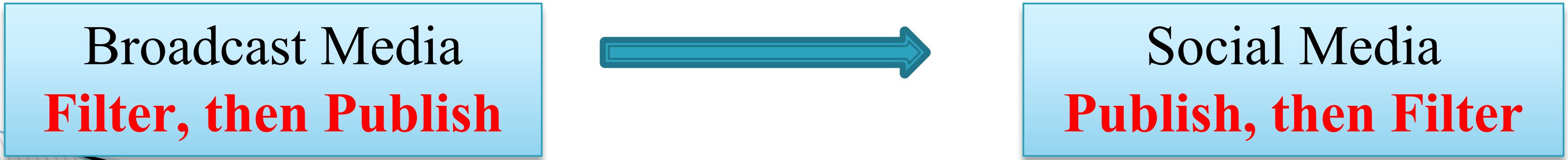
myspace.com a place for friends
LinkedIn
LIVEJOURNAL
blogcatalog
Blogger
twitter
WIKIPEDIA The Free Encyclopedia
Epinions.com a Shopping.com company
YouTube
digg
flickr
delicious social bookmarking
facebook

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, Insta
- 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

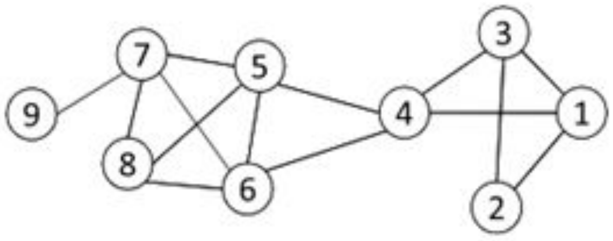

Broadcast Media
Filter, then Publish

Social Media
Publish, then Filter

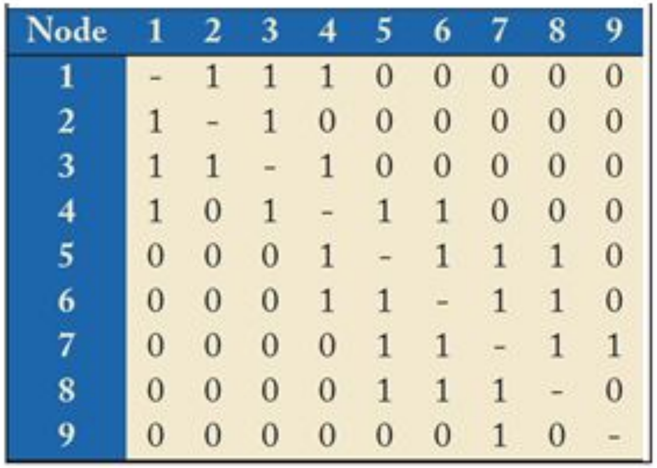
Networks and Representation

- 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	-

Network Applications

- Citation network
- Co-authorship network
- Terrorist networks
- Economic networks
- Family Networks
- Organization networks
- Sports Networks

Organizational Network Analysis

- ONA is a method for studying communication and socio-technical networks within an organization.
- Organizational network analysis (ONA) often refers to the use of SNA methods in the context of organization dynamics and development
- It is a quantitative descriptive technique for creating statistical and graphical models of the people, tasks, groups, knowledge and resources of organizational systems

To Find Subject Matter Experts

- Each node indicates people working in particular domain area .
- X --> Y means X seeks knowledge from Y.
- Two people are connected if one goes to the other for expertise in this domain .
- Potential of each node is shown in different colors based on their experience.

Maximizing Organizational Productivity

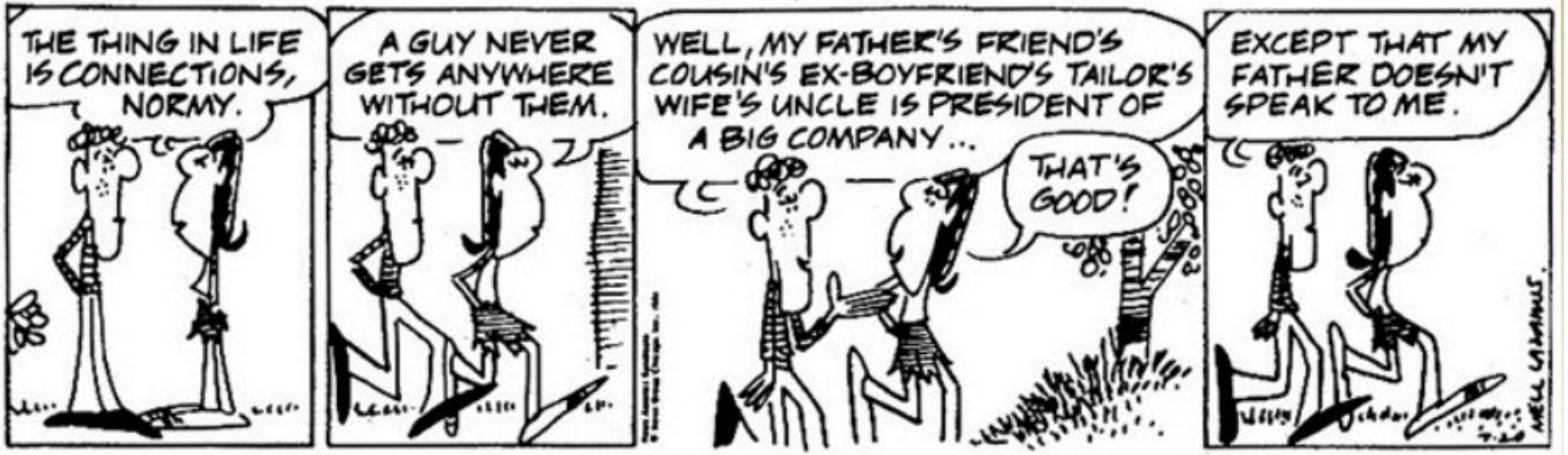
- How valuable is the information I receive from this person?
- How well does this person collaborate with me to solve problems and make decisions?
- How aware is this person of my skills?
- How accessible is this person to me?
- How “engaged” is this person with me?
- How safe is it to communicate with this person?
- What is the level of quality of conversation with this person?

- To what degree is my productivity improved by this person?
- How much power and influence does this person have?
- How much do I like this person?
- To what degree does this person support the achievement of my career goals?
- To what degree does this person support the achievement of my personal goals?
- To what degree does this person energize (or exhaust) me?
- To what degree do I trust this person?

Broader Applications of SNA

- Accelerate diffusion by identifying opinion leaders
- Reveal how infections spread among patients and staff in a hospital
- Map executive's personal network based on email flows
- Map interactions amongst blogs on various topics
- Map communities of expertise in various fields
- Discover emergent communities of interest amongst faculty at various universities
- Discover useful patterns in click streams on the WWW
- Viral spread: disease, fads and fashions, ideas, YouTube videos
- To Find Subject Matter Experts in Particular Area

MOMMA By Mell Lazarus


THE THING IN LIFE IS CONNECTIONS, NORMY.
A GUY NEVER GETS ANYWHERE WITHOUT THEM.
WELL, MY FATHER'S FRIEND'S COUSIN'S EX-BOYFRIEND'S TAILOR'S WIFE'S UNCLE IS PRESIDENT OF A BIG COMPANY...
THAT'S GOOD!
EXCEPT THAT MY FATHER DOESN'T SPEAK TO ME.
MELL LAZARUS
7-20

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

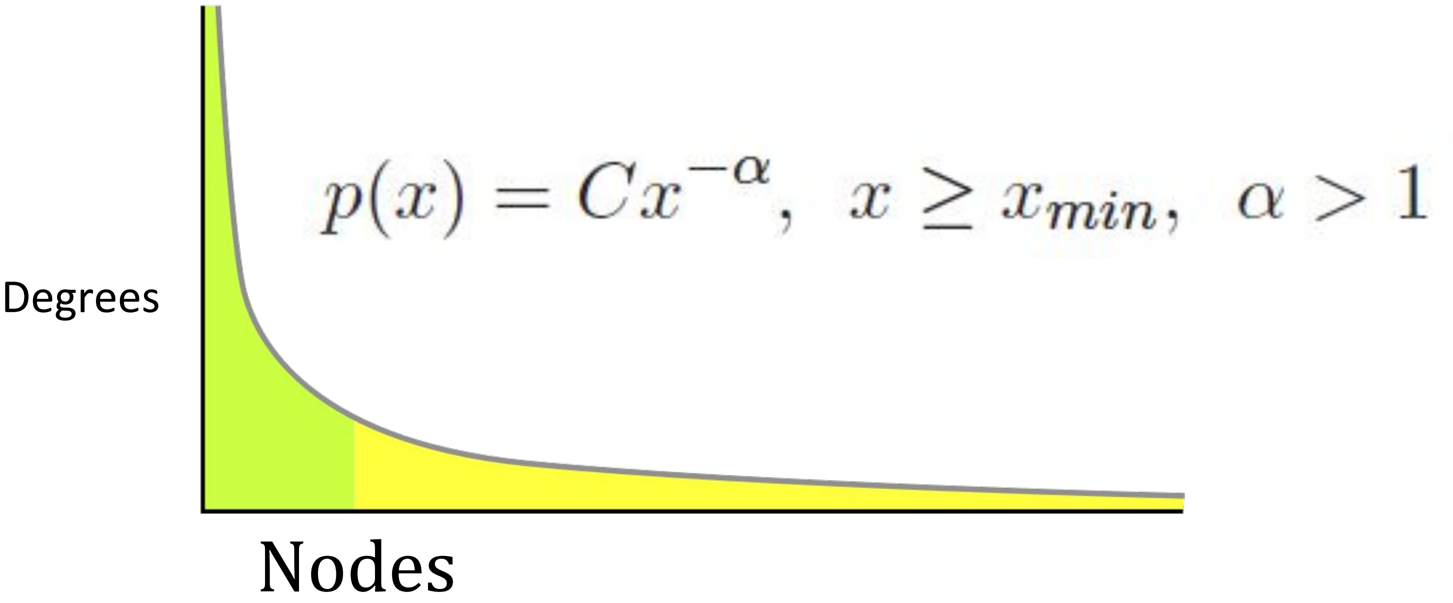
Properties of Large-Scale Networks

- 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.

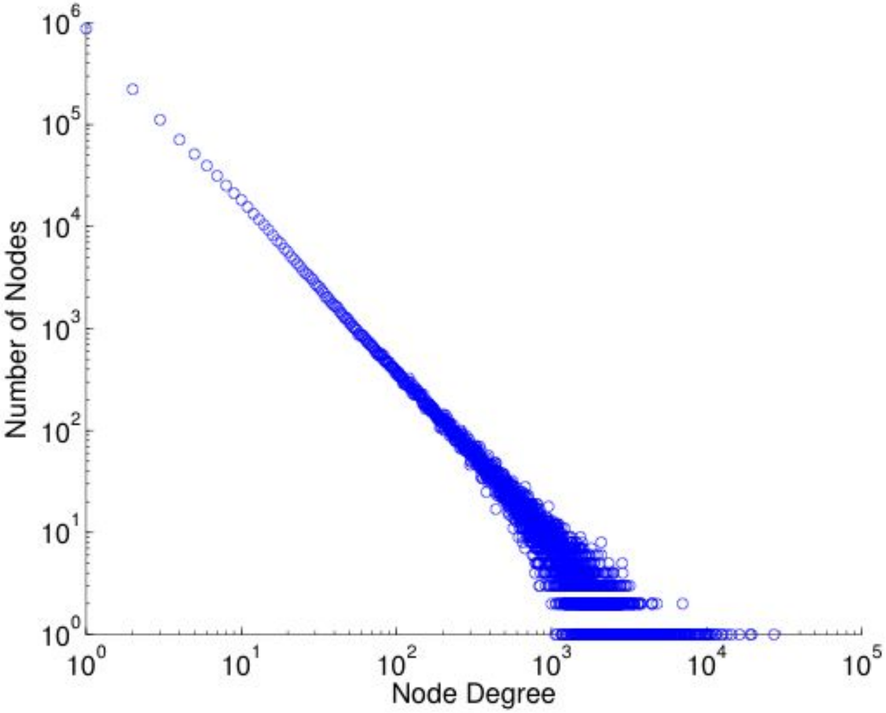


$$p(x) = Cx^{-\alpha}, \;\; x \geq x_{min}, \;\; \alpha > 1$$

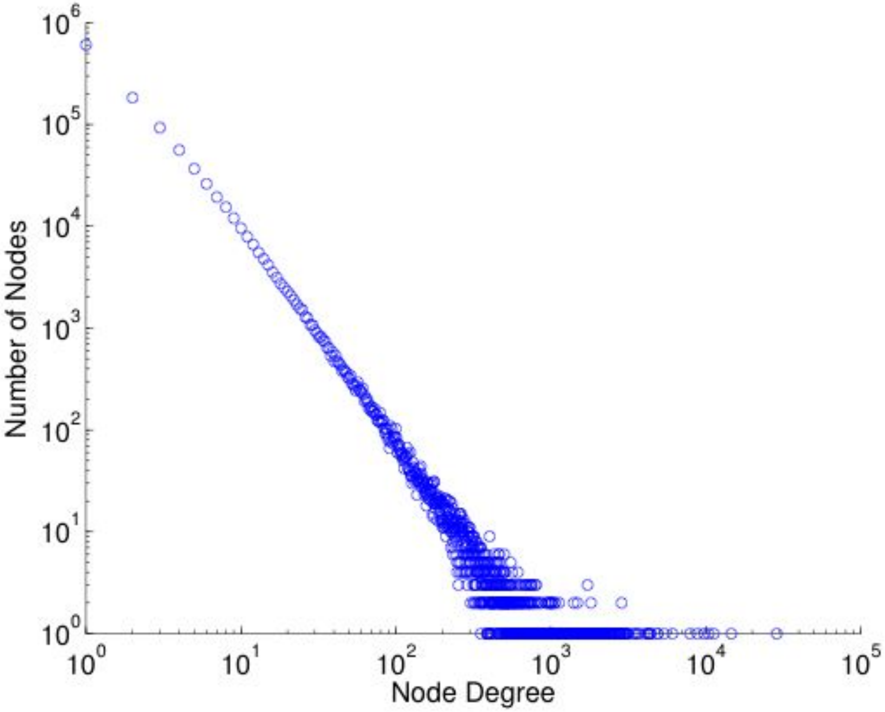
- A.k.a. long tail distribution, scale-free distribution

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

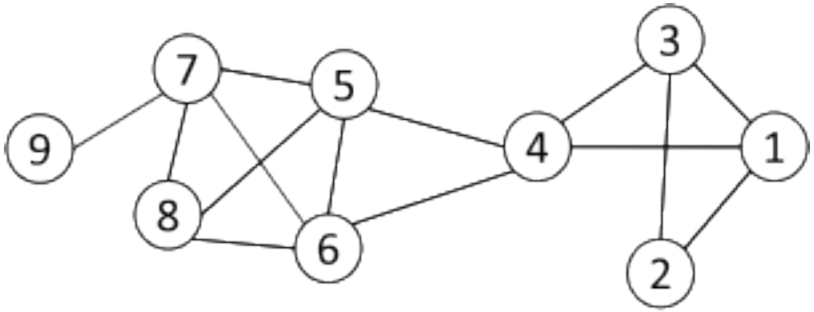
- "Six Degrees of Separation"

- 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.

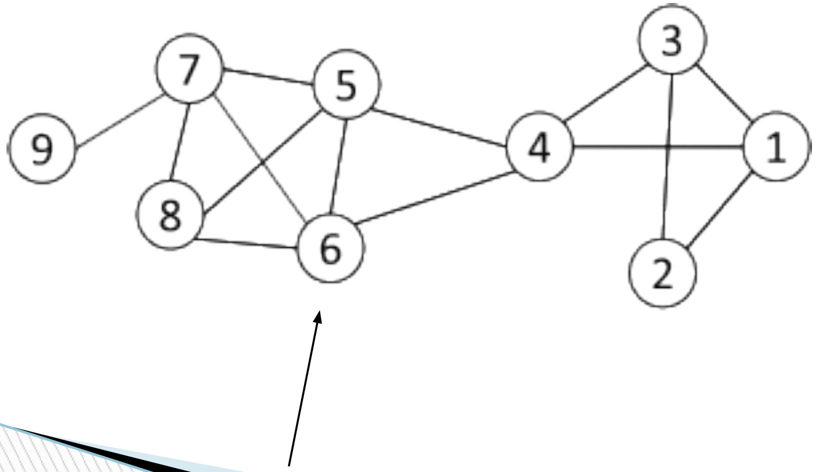
Community Structure

- Community: People in a group interact with each other more frequently than those outside the group
- *ki* = number of edges among node Ni’s neighbors
- Friends of a friend are likely to be friends as well
- Measured by clustering coefficient:
  - density of connections among one’s friends

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

Clustering Coefficient

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



- 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.

Social Computing Tasks

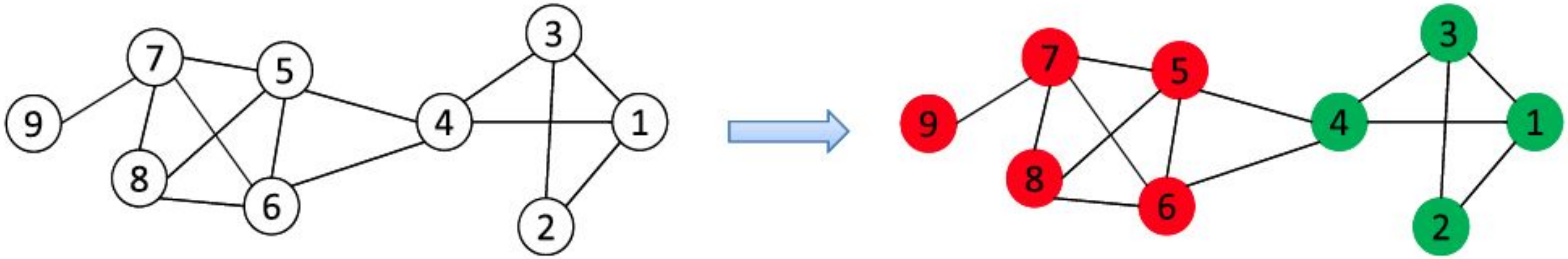
- Social Computing: a young and vibrant field
- Conferences: KDD, WSDM, WWW, ICML, AAAI/IJCAI, etc.
- Tasks
  - Network Modeling
  - Centrality Analysis and Influence Modeling
  - Community Detection
  - Classification and Recommendation
  - Privacy, Spam and Security

Centrality Analysis and Influence Modeling

- Centrality Analysis:
  - Identify the most important actors or edges
    - E.g. PageRank in Google
  - Various other 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

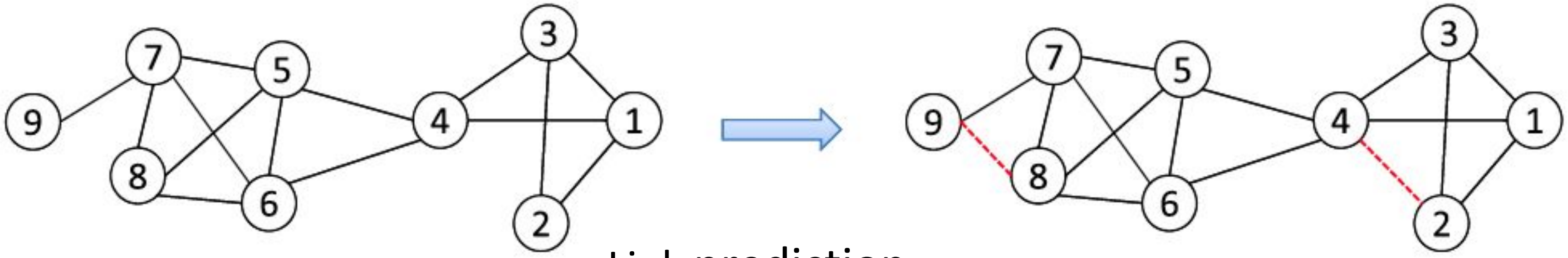
- A community is a set of nodes between which the interactions are (relatively) frequent
  - A.k.a., *group, cluster, cohesive subgroups, modules*



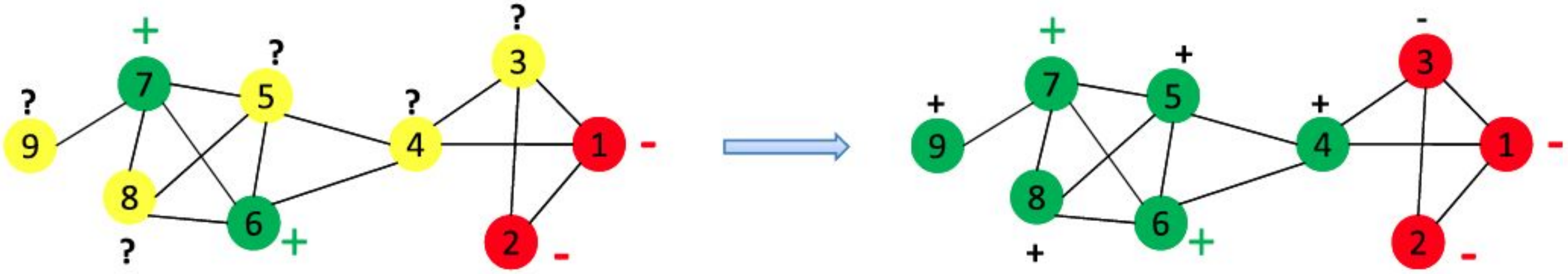
- *Network Compression, Visualization of a huge network*
- New lines of research in social media
  - Community Detection in Heterogeneous Networks
  - Community Evolution in Dynamic Networks
  - Scalable Community Detection in Large-Scale Networks

Classification and Recommendation

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



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 anonymization 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

Development of social network analysis

- In its first incarnation, modern social network analysis was introduced by a psychiatrist, Jacob L. Moreno, and a psychologist, Helen Jennings. They conducted elaborate research, first among the inmates of a prison and later among the residents in a reform school for girls (Moreno, 1934).
- Moreno and Jennings named their approach *sociometry.* At first, sociometry generated a great deal of interest, particularly among American psychologists and sociologists. But that interest turned out to be short lived; by the 1940s most American social scientists had returned to their traditional focus on the characteristics of individuals.

- The third version of social network analysis emerged when a German psychologist, Kurt Lewin, took a job at the University of Iowa in 1936. There, Lewin worked with a large number of graduate students and post-docs.

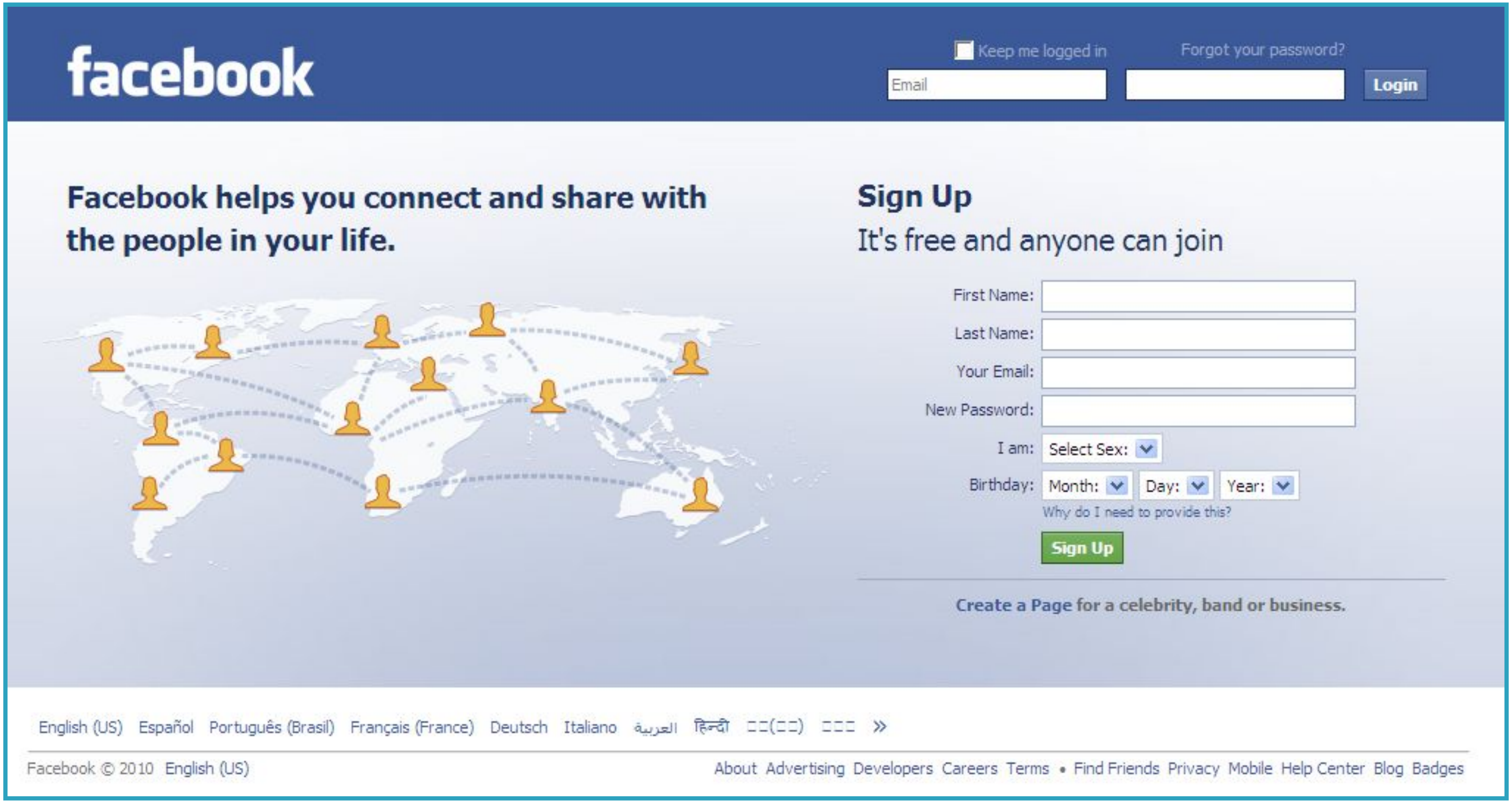
- Together, they developed a structural perspective and conducted social network research in the field of social psychology

- One of Lewin's students, Alex Bavelas, remained at MIT where he spearheaded a famous study of the impact of group structure on productivity and morale. This work was influential in the field of organizational behavior, but most of its influence was limited to that field.

- By 1970, then, sixteen centers of social network research had appeared. With the development of each, knowledge and acceptance of the structural approach grew.
- Still, however, none of these centers succeeded in providing a generally recognized paradigm for the social network approach to social science research.

Online Social Networking (OSN)

- Online Web services enabling people to connect with each other, share information
  - Common friends, interests, personal info, …
  - Post photos, videos, etc. for others to see
  - Communicate via email, instant message, etc.
- Major OSN services: Facebook, Twitter, MySpace, LinkedIn, etc.



facebook
☐ Keep me logged in
Forgot your password?
Email
Login

Facebook helps you connect and share with the people in your life.

Sign Up

It's free and anyone can join

First Name:
Last Name:
Your Email:
New Password:
I am: Select Sex:
Birthday: Month: Day: Year:

Why do I need to provide this?

Create a Page for a celebrity, band or business.

English (US) Español Português (Brasil) Français (France) Deutsch Italiano العربية हिन्दी □□(□□) □□□ »
Facebook © 2010 English (US)
About Advertising Developers Careers Terms • Find Friends Privacy Mobile Help Center Blog Badges

"Giving people the power to share and make the world more open and connected."

OSN Popularity

- Over 2.7-2.9 Billion Facebook users worldwide
  - Over 225 million in U.S.
  - Over 450 million access via mobile
  - 300 million pictures uploaded to Facebook daily
- Over 330 million Twitter users; over 500 million Tweets sent daily
- Over 810 million LinkedIn members in over 200 countries

Benefits of OSN Communication

- Vast majority of college students use OSNs
  - Organizations want to market products, services, etc. to this demographic
  - OSNs can help them reach these potential buyers
- OSNs provide communal forum for expression (self, group, mass), collaboration, etc.
  - Connect with old friends, find new friends and connect
  - Play games with friends, e.g., Mafia Wars, Scrabulous
  - Commerce in “virtual items”
- But using OSNs poses security issues for orgs as well as individuals

OSN Security Threats/Attacks

- Malware distribution
- Cyber harassment, stalking, etc.
- Information “shelf life” in cyberspace
- Privacy issues:
  - Information about person posted by him/herself, others
  - Information about people collected by OSNs
- Information posted on OSNs impacts unemployment, insurance, etc.
- Organizations’ concerns: brand, laws, regulations

MSN Security Threat/Attacks

- Personal information leakage
  - Particularly dangerous because of physical proximity
- Malware distribution

“Common Sense” Measures (1)

- Use strong, unique passwords
- Provide minimal personal information: avoid entering birthdate, address, etc.
- Review privacy settings, set them to “maximum privacy”
  - “Friends of friends” includes far more people than “friends only”
- Exercise discretion about posted material:
  - Pictures, videos, etc.
  - Opinions on controversial issues
  - Anything involving coworkers, bosses, classmates, professors
  - Anything related to employer (unless authorized to do so)
- Be wary of 3rd party apps, ads, etc.
- Supervise children’s OSN activity

“Common Sense” Measures (2)

- “If it sounds too good to be true, it probably is”
- Use browser security tools for protection:
  - Anti-phishing filters (IE, Firefox)
  - Web of Trust (crowdsourced website trust)
  - AdBlock/NoScript/Do Not Track Plus
- Personal reputation management:
  - Search for yourself online, look at the results…
  - Google Alerts: emails sent daily to you about results for any search query (free), e.g., your name
- Extreme cases:
  - Cease using OSNs, delete accounts
  - Contact law enforcement relentless online harassment

Final Remarks on OSNs

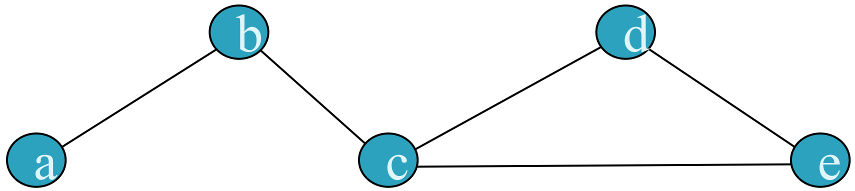
- On-line social networking systems are very popular and mobile social networking systems are emerging
- Malware distribution and personal information leakage are two most prominent threats and attacks
- Personal countermeasures are most effective

Social Network Data

- The unit of interest in a network are the combined sets of actors and their relations.

- We represent *actors* with points and *relations* with lines.
- Actors are referred to variously as:
- Nodes, vertices or points
- Relations are referred to variously as:
- Edges, Arcs, Lines, Ties

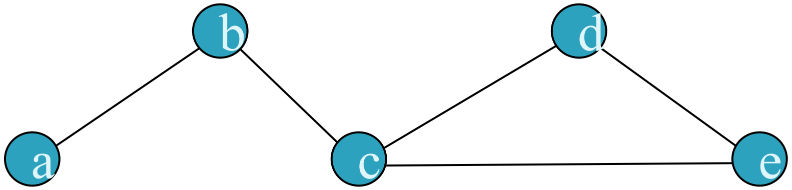
Example:



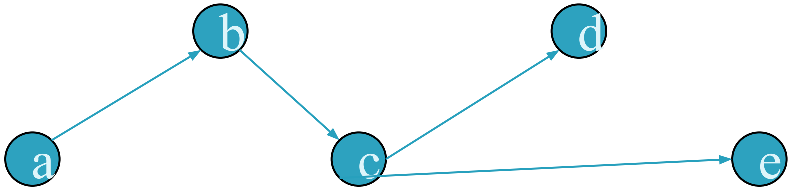
Social Network Data

In general, a relation can be:

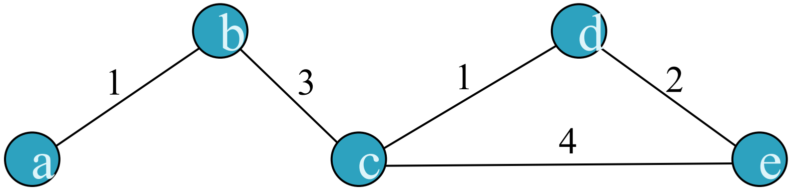
- Binary or Valued
- Directed or Undirected



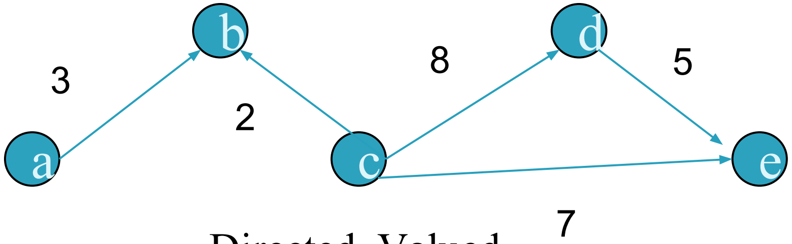
Undirected, binary



Directed, binary



Undirected, Valued



Directed, Valued

Social Network Data

- Social network data are substantively divided by the number of *modes* in the data.
- 1-mode data represents edges based on *direct* contact between actors in the network. All the nodes are of the same type (people, organization, ideas, etc).
- Examples: Communication, friendship, giving orders, sending email.
- 1-mode data are usually singly reported (each person reports on their friends), but you can use multiple-informant data, which is more common in child development research

Social Network Data

- 2-mode data represents nodes from two separate classes, where all ties are across classes. Examples:
  - *People* as members of *groups*
  - *People* as authors on *papers*
  - *Words* used often by *people*
  - *Events* in the life history of *people*
- The two modes of the data represent a duality: you can project the data as people connected to people through joint membership in a group, or groups to each other through common membership
- *There may be multiple relations of multiple types connecting your nodes.*

Social Network Data

We can examine networks across multiple levels:

1) Ego-network

- Have data on a respondent (ego) and the people they are connected to (alters). Example: 1985 GSS module
- May include estimates of connections among alters

2) Partial network

- Ego networks plus some amount of tracing to reach contacts of contacts
- Something less than full account of connections among all pairs of actors in the relevant population
- Example: CDC Contact tracing data for STDs

Social Network Data

3) Complete or "Global" data

- Data on *all* actors within a particular (relevant) boundary
- Never exactly complete (due to missing data), but boundaries are set
- Example: Co-authorship data among all writers in the social sciences, friendships among all students in a classroom

Social Network Data

Collecting Network Data

Data capture any connection between the nodes. Sources include surveys, published accounts, special informants, etc.

In general, you can only make conclusions about relations among the set of nodes you have collected, *so it is important to observe as much of the network as possible.*

Social Network Data

Collecting Network Data

If you use surveys to collect data, some general rules of thumb:

a) Network data collection can be time consuming. It is better to have *breadth* over *depth.* Having detailed information on <50% of the sample will make it very difficult to draw conclusions about the general network structure.

b) Question format:
- If you ask people to *recall* names (an open list format), fatigue will result in under-reporting
- If you ask people to check off names from a full list, you can often get over-reporting

c) It is common to limit people to ~5 nominations. This will bias network stats for stars, but is sometimes the best choice to avoid fatigue.

d) Concrete relational indicators are best (who did you talk to?) over attitudes that are harder to define (who do you like?)

Social Network Data

Collecting Network Data

Existing Sources of Social Network Data

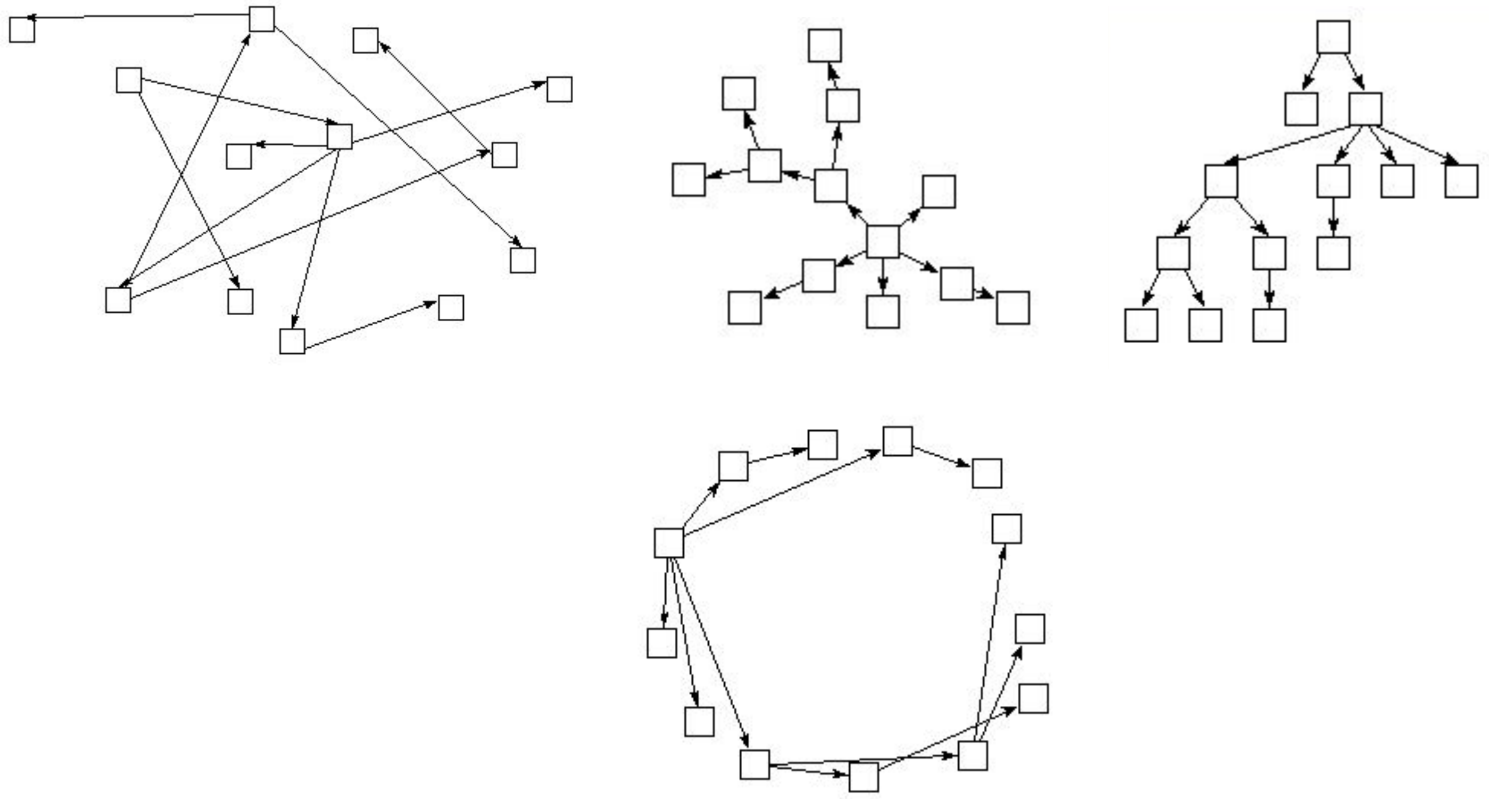
1) Check INSNA: The International Network of Social Network Analysis
2) Many secondary sources (particularly for 2-mode data)
3) National Longitudinal Survey of Adolescent Health (Add Health)

Social Network Data

Basic Data Structures

Working with pictures.

No standard way to draw a sociogram: each of these are equal:



Social Network Data

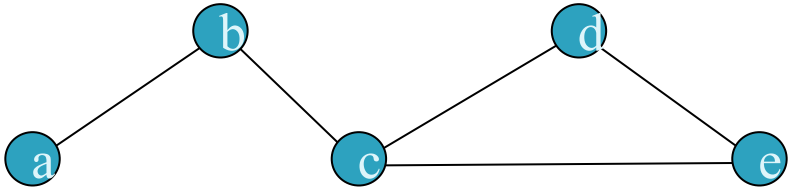
Basic Data Structures

In general, graphs are cumbersome to work with analytically, though there is a great deal of good work to be done on using visualization to build network intuition.

Social Network Data

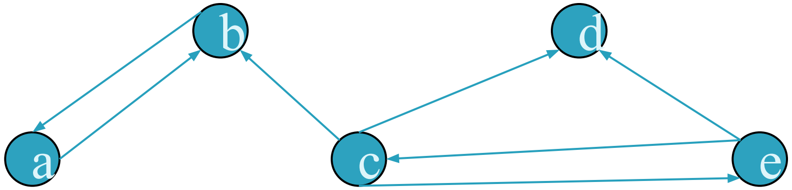
Basic Data Structures

From pictures to matrices



Undirected, binary

	a	b	c	d	e
a		1			
b	1		1		
c		1		1	1
d			1		1
e			1	1	



Directed, binary

	a	b	c	d	e
a		1			
b	1				
c		1		1	1
d					
e			1	1	

Social Network Data

Basic Data Structures

From matrices to lists

	a	b	c	d	e
a		1			
b	1		1		
c		1		1	1
d			1		1
e			1	1	

Adjacency List

a	b
b	a c
c	b d e
d	c e
e	c d

Arc List

a	b
b	a
b	c
c	b
c	d
c	e
d	c
d	e
e	c
e	d

Measuring Networks: Flow

In addition to the simple probability that one actor passes information on to another (p_{ij}), two factors affect flow through a network:

Topology

-the shape, or form, of the network

- Example: one actor cannot pass information to another unless they are either directly or indirectly connected

Time

- the timing of contact matters
- Example: an actor cannot pass information he has not received yet

Measuring Networks: Flow

Two features of the network's topology are known to be important: *connectivity* and *centrality*

Connectivity refers to how actors in one part of the network are connected to actors in another part of the network.

- *Reachability*: Is it possible for actor *i* to reach actor *j*? This can only be true if there is a chain of contact from one actor to another.

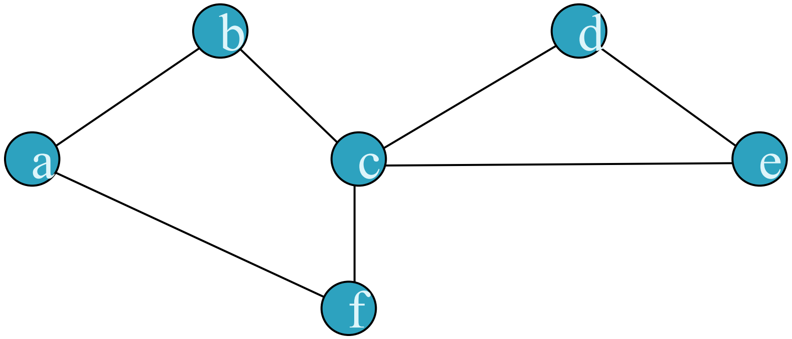
- *Distance*: Given they can be reached, how many steps are they from each other?

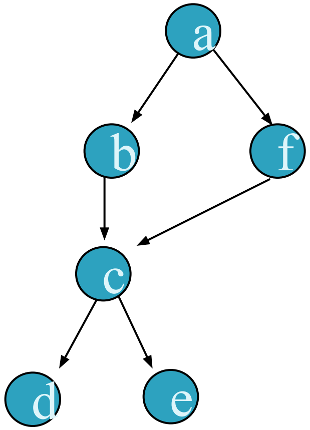
- *Number of paths*: How many different paths connect each pair?

Measuring Networks: Flow

Reachability

Indirect connections are what make networks systems. One actor can *reach* another if there is a *path* in the graph connecting them.





Paths can be directed, leading to a distinction between "strong" and "weak" components

Measuring Networks: Flow

Reachability

Basic elements in connectivity

- A *path* is a sequence of nodes and edges starting with one node and ending with another, tracing the indirect connection between the two. On a path, you never go backwards or revisit the same node twice.
  Example: a → b → c → d

- A *walk* is any sequence of nodes and edges, and may go backwards.
  Example: a → b → c → b → c → d

- A *cycle* is a path that starts and ends with the same node. Example:
  c → d → e → c

Measuring Networks: Flow

Reachability

Reachability

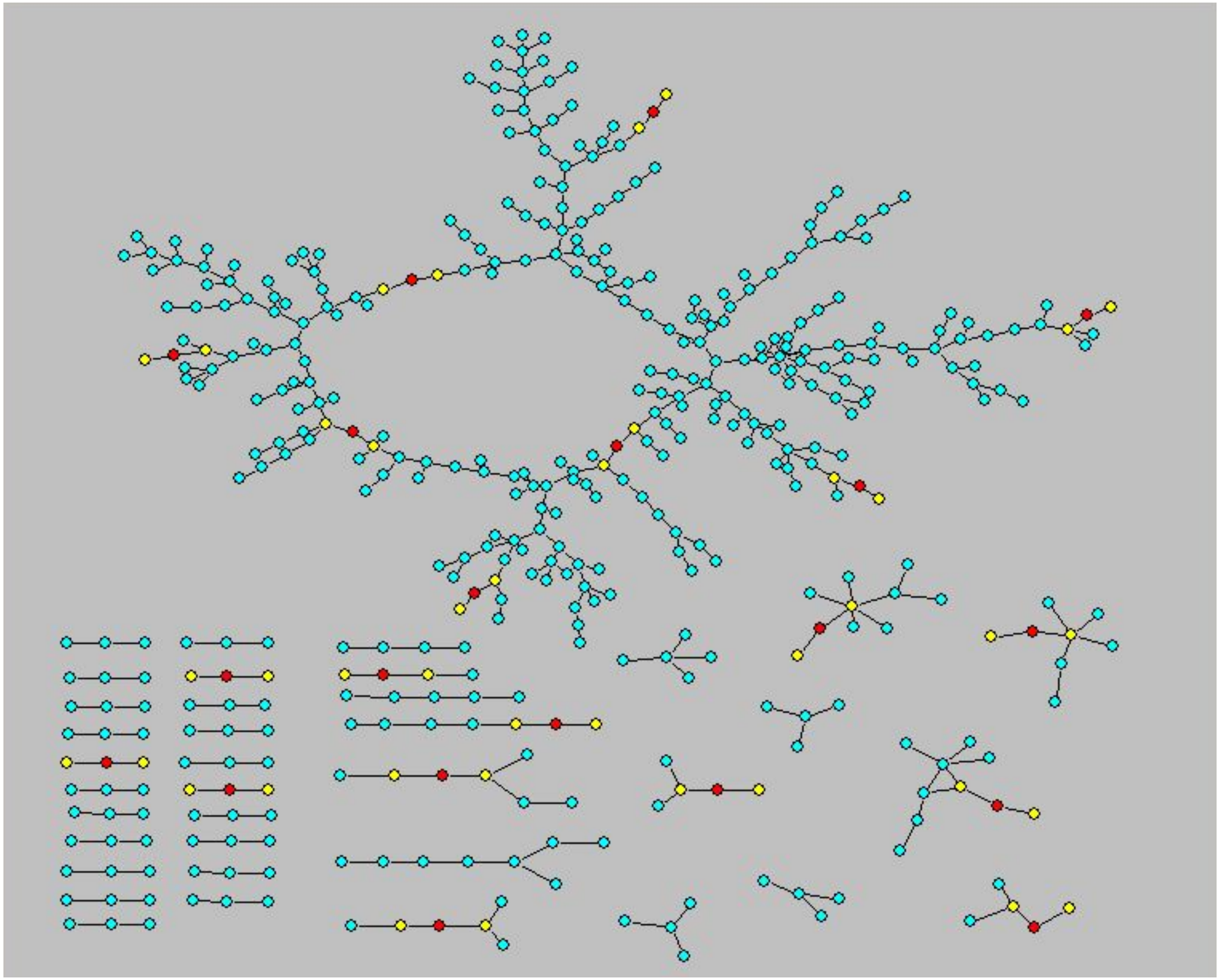
If you can trace a sequence of relations from one actor to another, then the two are reachable. If there is at least one path connecting every pair of actors in the graph, the graph is *connected* and is called a *component*.

Intuitively, a component is the set of people who are all connected by a chain of relations.

Measuring Networks: Flow

Reachability

This example contains many components.



Measuring Networks: Flow

Reachability

In general, components can be directed or undirected.

For a graph with any directed edges, there are two types of components:

Strong components consist of the set(s) of all nodes that are *mutually reachable*

Weak components consist of the set(s) of all nodes where at least one node can reach the other.

Measuring Networks: Flow

Centrality

Centrality refers to (one dimension of) *location,* identifying where an actor resides in a network.

- For example, we can compare actors at the edge of the network to actors at the center.

- In general, this is a way to formalize intuitive notions about the distinction between insiders and outsiders.

Measuring Networks: Flow

Centrality

At the individual level, one dimension of *position* in the network can be captured through centrality.

Conceptually, centrality is fairly straight forward: we want to identify which nodes are in the 'center' of the network. In practice, identifying exactly what we mean by 'center' is somewhat complicated, but substantively we often have reason to believe that people at the center are very important.

Three standard centrality measures capture a wide range of "importance" in a network:

- Degree
- Closeness
- Betweenness

Measuring Networks: Flow

Centrality

A second measure of centrality is closeness centrality. An actor is considered important if he/she is relatively close to all other actors.

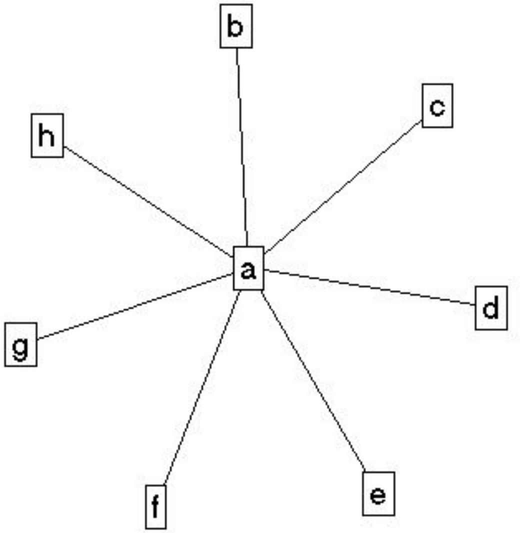
Closeness is based on the inverse of the distance of each actor to every other actor in the network.

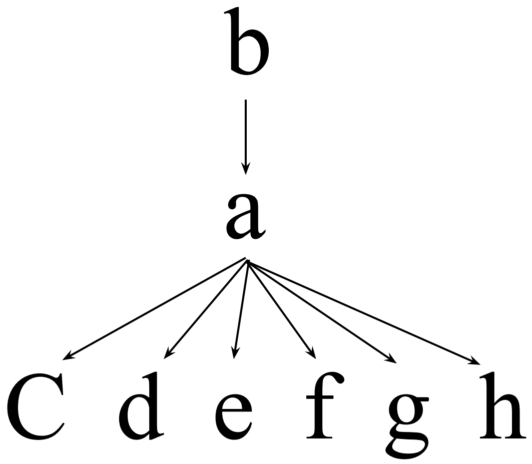
Measuring Networks: Flow

Centrality

Betweenness Centrality:

Model based on communication flow: A person who lies on communication paths can control communication flow, and is thus important. Betweenness centrality counts the number of <u>shortest</u> paths between *i* and *k* that actor *j* resides on.





b

a

C d e f g h

Measuring Networks: Flow

Centrality

Comparing across centrality values

- Generally, the 3 centrality types will be positively correlated
- When they are not correlated, it probably tells you something interesting about the network.

	Low Degree	Low Closeness	Low Betweenness
High Degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant - communication bypasses him/her
High Closeness	Key player tied to important important/active alters		Probably multiple paths in the network, ego is near many people, but so are many others
High Betweenness	Ego's few ties are crucial for network flow	Very rare cell. Would mean that ego monopolizes the ties from a small number of people to many others.	

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