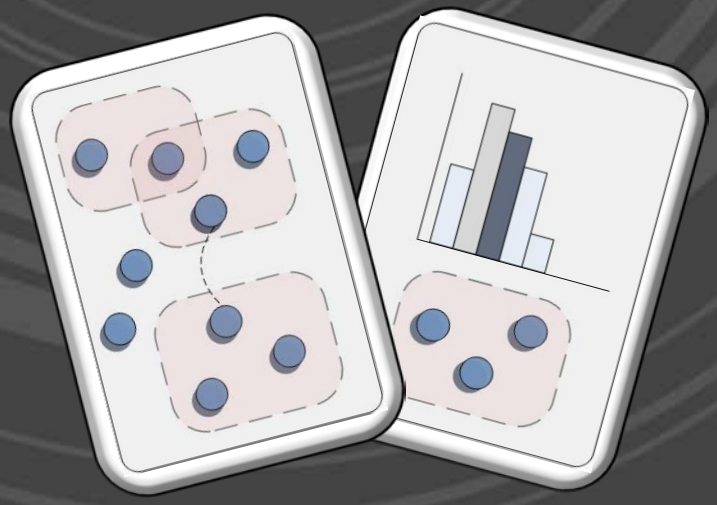
Network Analysis Basics and applications to online data

Katherine Ognyanova
University of Southern California

Prepared for the Annenberg Program for Online Communities, 2010.

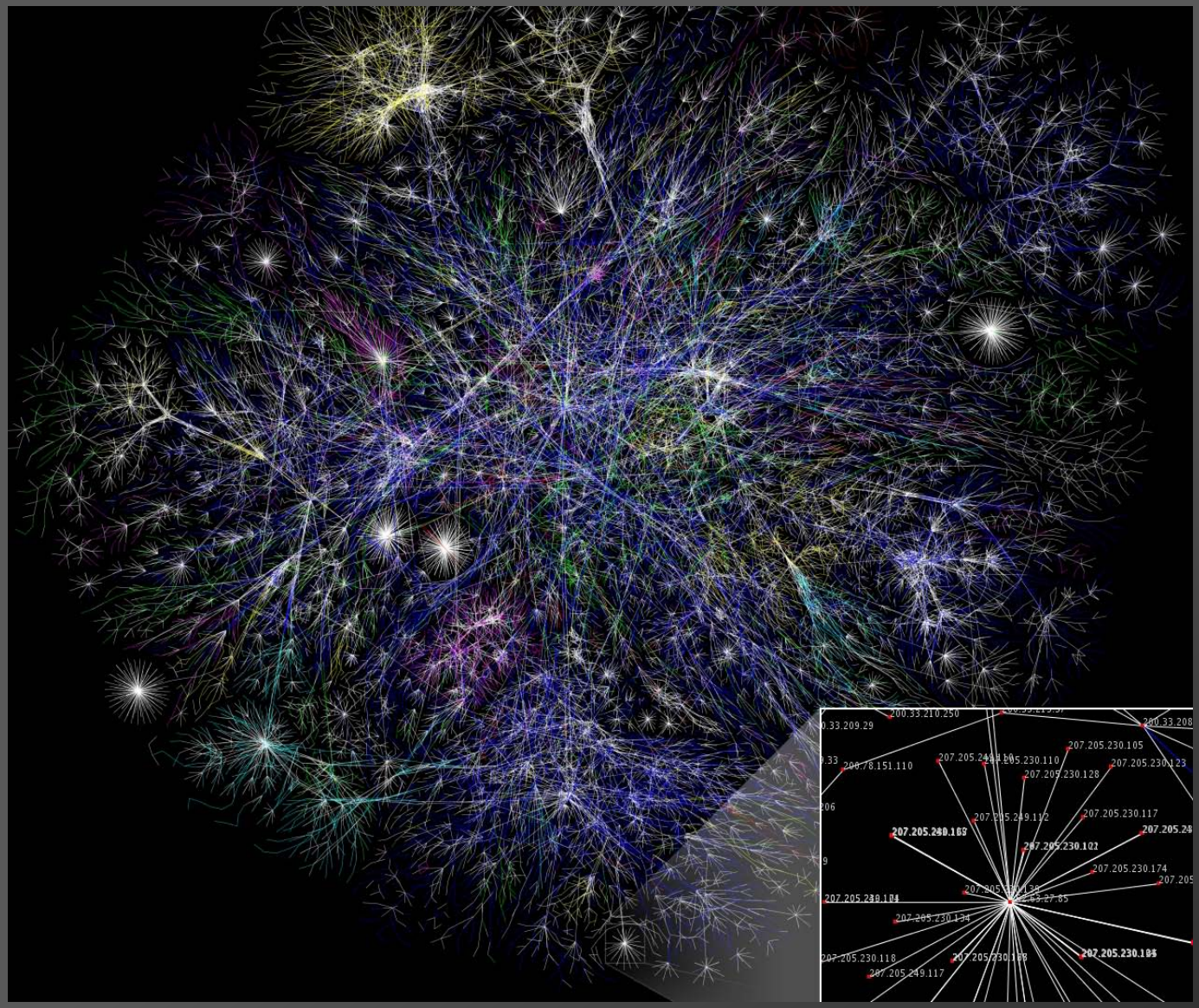


Relational data

Node (actor, vertex, etc.)

Tie (link, relation, edge, etc.)

Internet



Internet Backbone: Each line is drawn between two nodes, representing two IP addresses. Source: Wikipedia

Global Media Networks

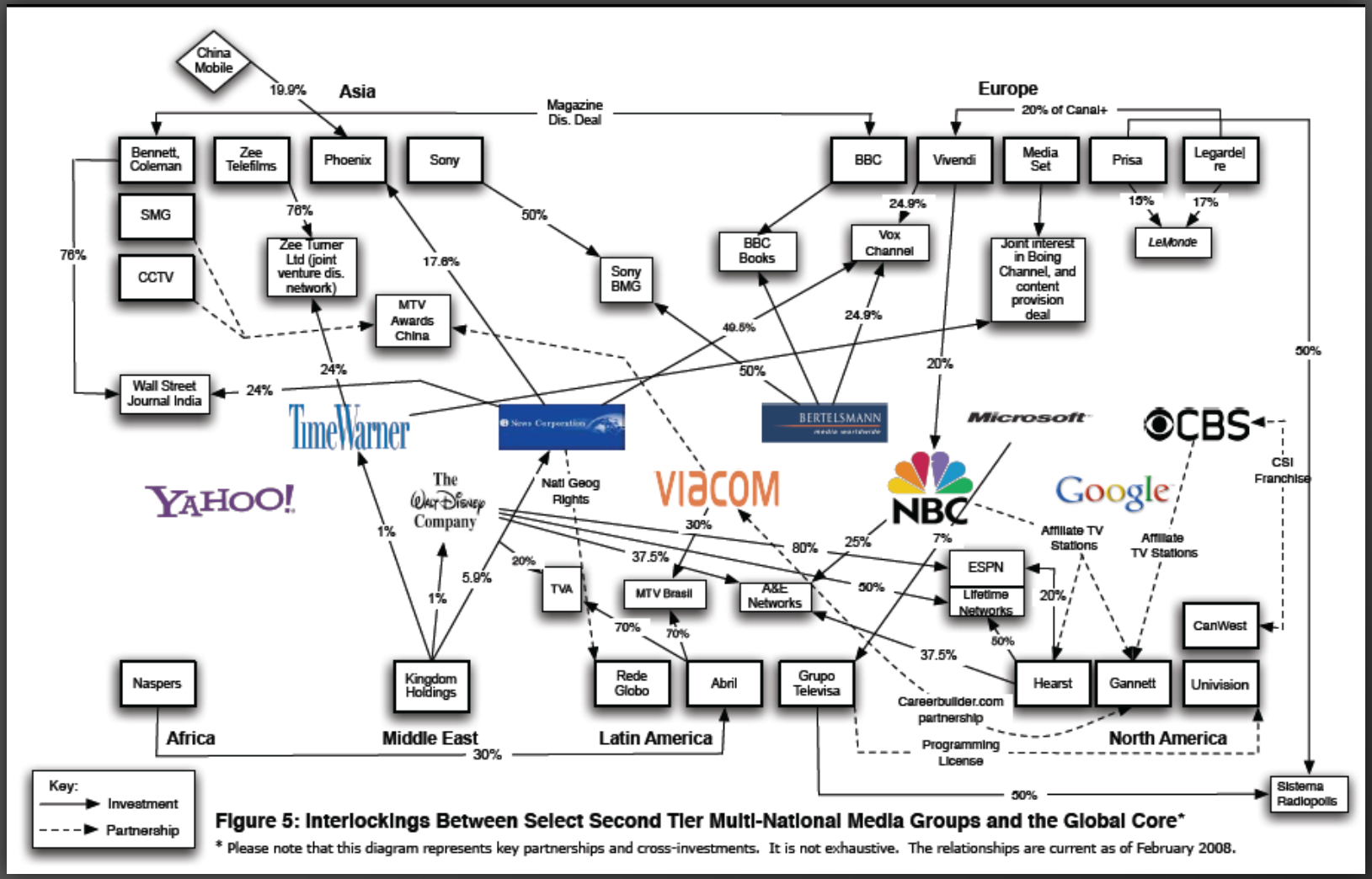


Figure 5: Interlockings Between Select Second Tier Multi-National Media Groups and the Global Core*

* Please note that this diagram represents key partnerships and cross-investments. It is not exhaustive. The relationships are current as of February 2008.

Media corporations network. Source: Arsenault & Castells, 2008

US Senate

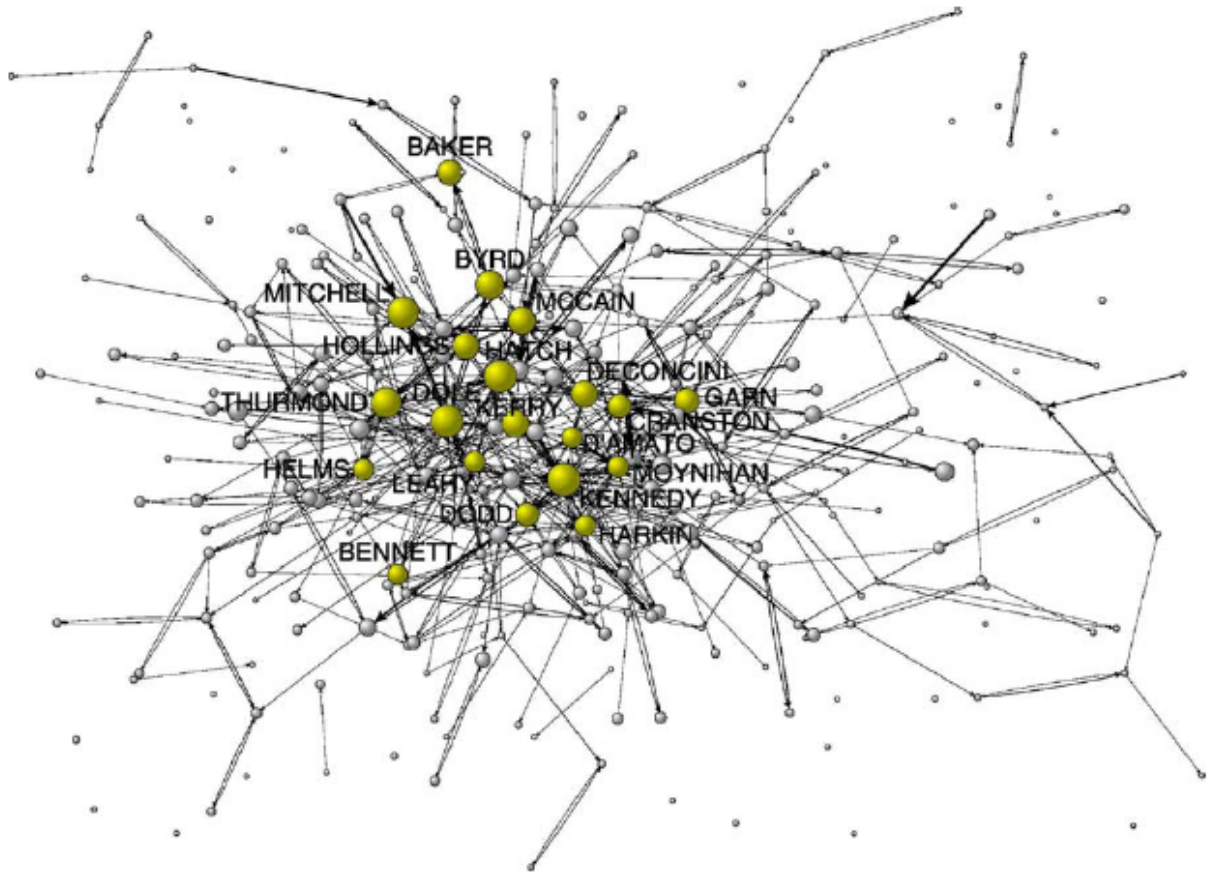


Fig. 4. Strongest weighted cosponsorship ties in the full Senate network, 1973–2004. Note: Size of each vertex is proportional to the Senator's connectedness score, the width of each arrow is proportional to the weighted quantity of bills cosponsored w_{ij} (values of $w_{ij} < 10$ not shown) and vertices that represent the top 20 Senators are identified by name. Figure drawn using Kamada–Kawai algorithm in Pajek (de Nooy et al., 2005).

US Senate cosponsorship network 1973-2004. Source: Fowler, 2006

High-School Romance

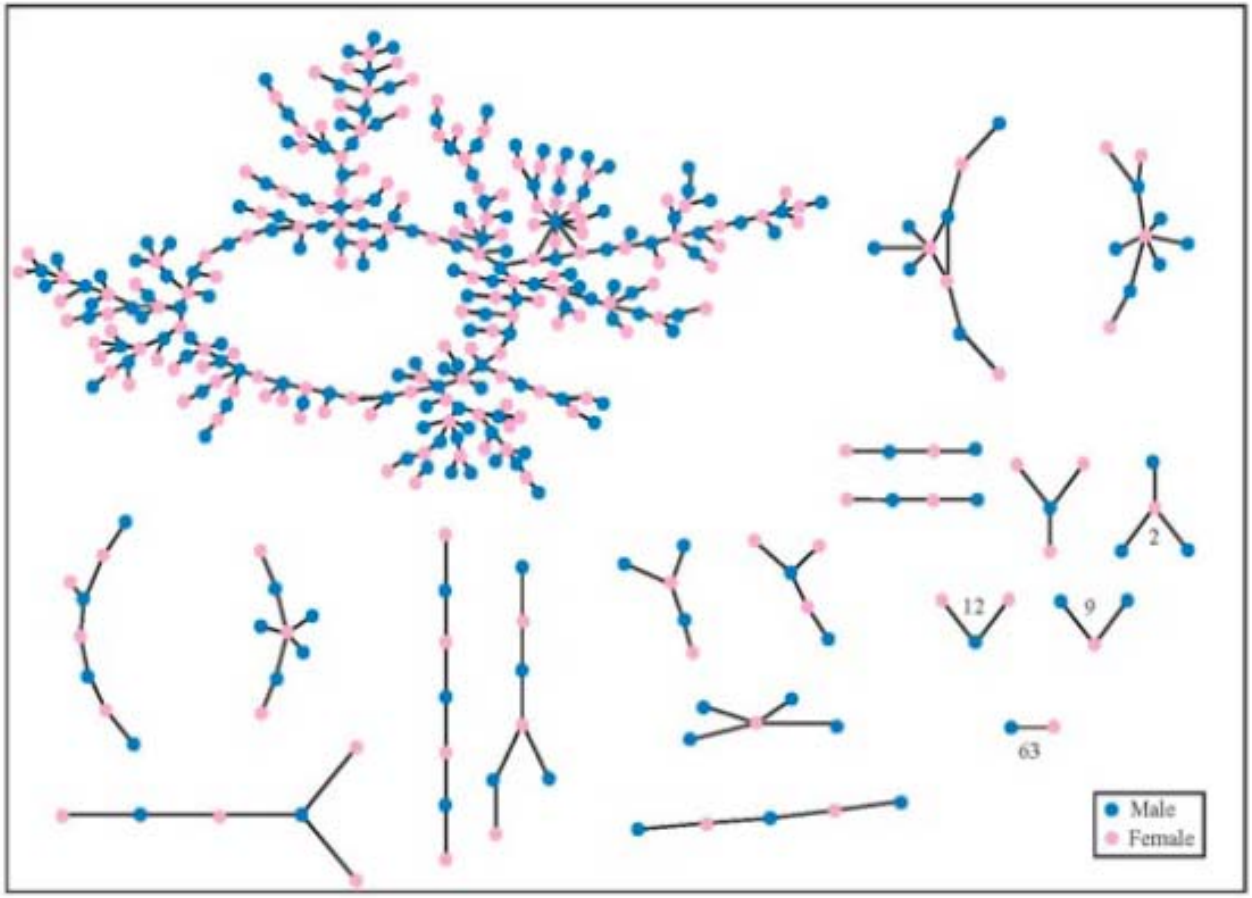
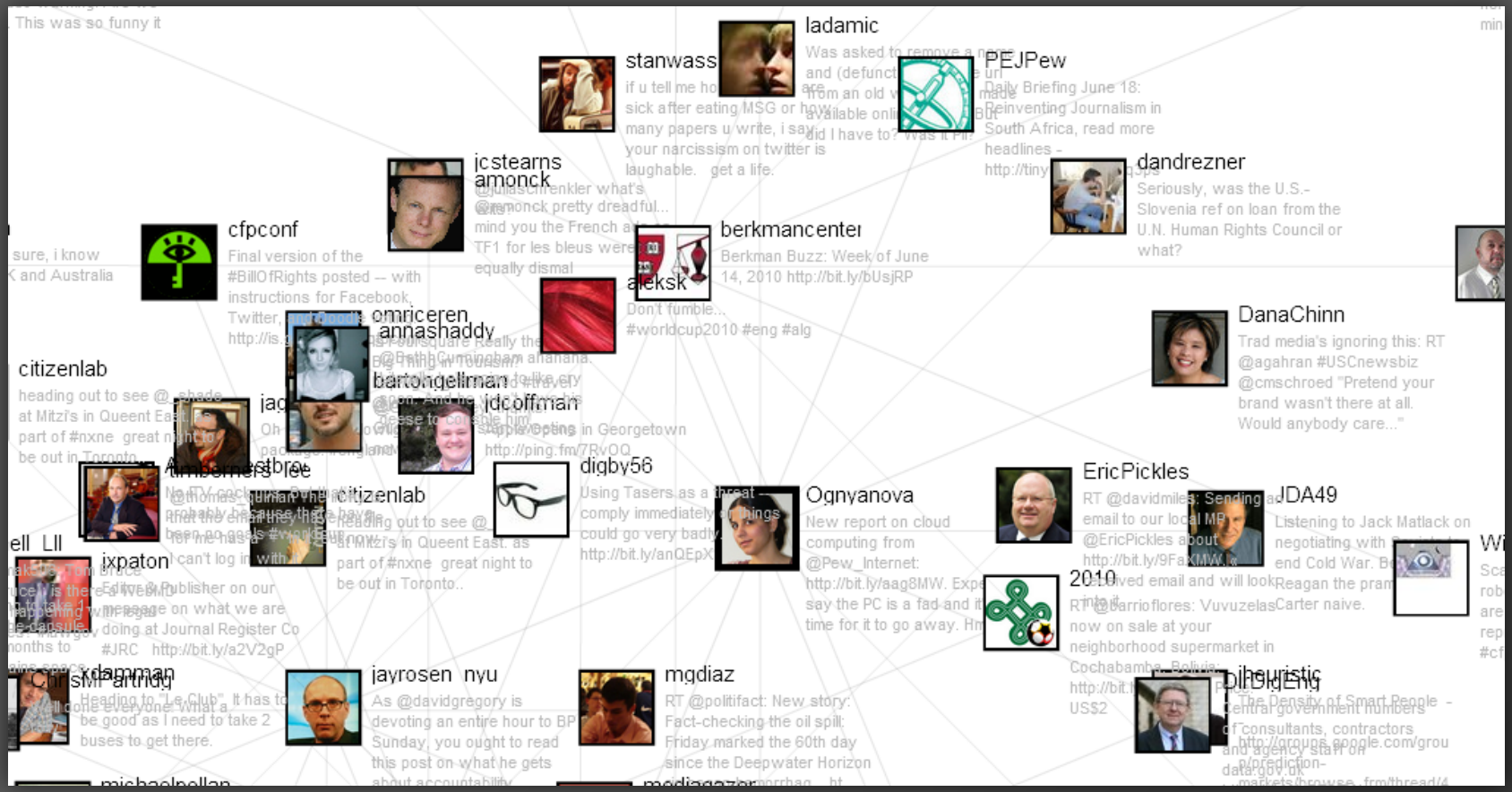


Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [50].

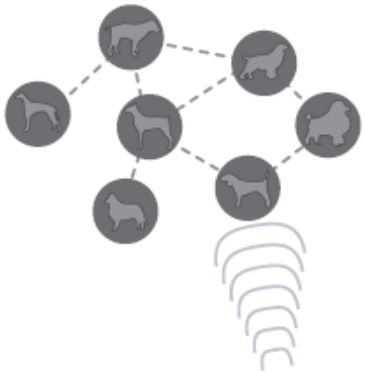
Source: Easley, Kleinberg (2010) Networks, Crowds and Markets

Online Social Networks



Twitter, 2010

SNiF



SNIF: Social Networking in Fur

Group: Noah Fields, Jonathan Gips, Philip Liang, Arnaud Pilpré

What

We present a system that allows pet owners to interact through their pets' social networks. Inexpensive, unobtrusive hardware can be affixed to pet collars and paraphernalia in order to augment pet-to-pet, pet-to-owner, and owner-to-owner interactions. SNIF devices aggregate pertinent environmental, social, and individual information that can be broadcast or addressed to other participating community members.

Why

Pets already function as social devices. Walking a dog in the park can lead to conversations that one might not otherwise have. Pets function as active icebreakers that will go up to anyone without any notion of social inhibition. Furthermore, pet-owners love buying products for their pets: sweaters, leashes, collars, toys, dishes, and beds. These items provide a set of rich interactions that can be brought into the digital world.

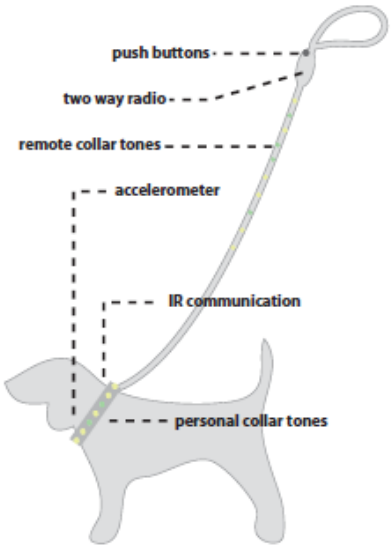
How

The SNIF starter kit includes a leash and collar as well as membership in the online community.

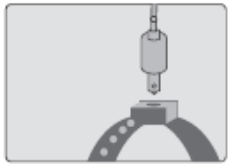
SNIF collars contain an LED display, an IR transceiver, and various sensors such as accelerometers and digital thermometers. They function as output devices that display personalized "collar tones" when the pet comes in proximity to another pet. They serve as input devices that sense activity levels, microclimate conditions, and other pets' presence.

The SNIF leash contains a two-way RF device, such as the Ambient Devices platform, and serves multiple purposes in the SNIF system. When attached to a pet's collar, it can upload information from the collar to the SNIF servers. When disconnected, the leash functions as an ambient device that displays real-time information, which is streamed from the SNIF servers, relevant to the pet and pet owner. For example, the leash displays the "collar tones" of frequently encountered pets that are going out for a walk. It may also give an indication of the general pet-walking index.

The online community portion of SNIF allows pet-owners to set privacy preferences, communicate with other pet owners, arrange pet outings, and customize the ambient information that their SNIF leashes display.

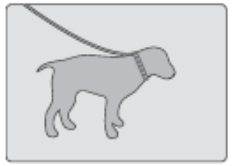


1. leash up



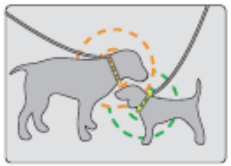
By connecting the leash to the collar, you signal the network that you are about to head out to play.

2. walk



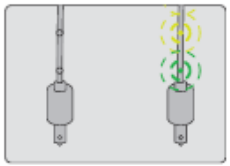
While you are on your walk, your collar keeps an eye out for your pals.

3. sniff



When you discover another dog, your collar displays a unique sequence of flashing lights, these are your collar tones. Your friend's collar tones flash on his collar.

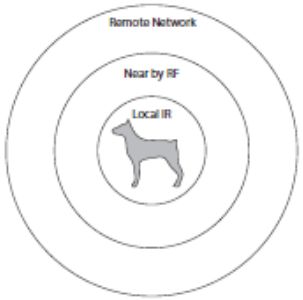
4. friend!



When you are back at your house, you can keep an eye on your companions. When one of your pals goes out to play, their collar tones are displayed on your leash!

Extensions

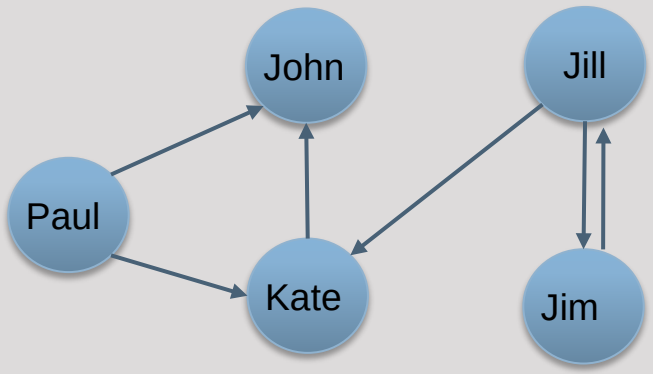
Pet toys that serve as tangible interfaces for the pet.
Degrees of separation between pets that changes as they interact.
Remote monitoring of pet's activity.
Local RF detection to display degrees of separation from the other pets in the vicinity.



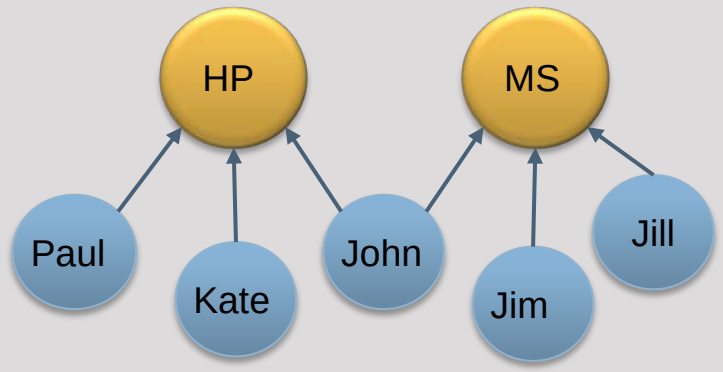
Source: noah.cx, 2010

Basic Types of Networks

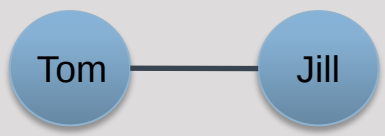
Adjacency
(e.g. friendship nets)



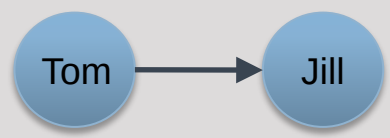
Affiliation
(e.g. employer-employee nets)



Undirected
(e.g. Facebook friends)



Directed
(e.g. Twitter followers)

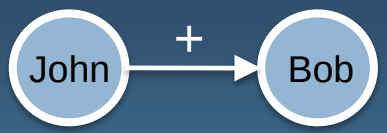


A closer look at links

Binary

- **Link or no link?** (1 or 0)
  Typical friendship networks – links exist or they don't, Bob and John are friends or they aren't.

Signed

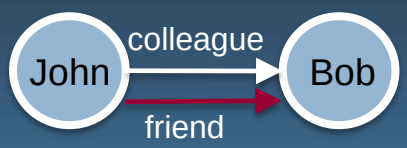


- **Positive or negative** (+, - or 0)
  You can get signed links by asking John "Do you like Bob, dislike him or are neutral about him?"

Valued

- **Weighted links** (each link is assigned a number)
  Weights can represent the strength/duration of a relationship. "How many times a week do you call Bob?"

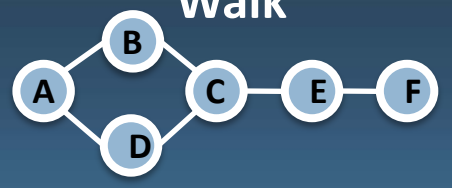
Multiplex



- **Multiplex** (more than one type of link)
  Add different relationships between the same set of nodes.

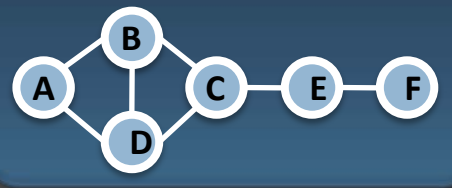
Distance in Networks

Walk



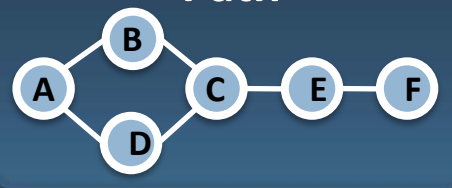
- A route through the network from one node to another. No restraints on repeated nodes or links: A-B-C-D-A-B-C-E is a walk.

Trail



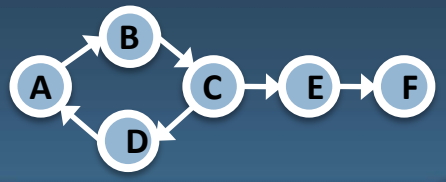
- A route between two nodes that does not use the same link (edge) twice. A-B-C-D-B is a trail, A-B-C-D-A-B is not (the link from A to B is used twice)

Path



- A route between two nodes that does not pass through the same node twice. A-B-C-D is a path, A-B-C-D-A-B is not. Only exception: closed paths – A-B-C-D-A

Directed Versions



- Walks, trails and paths have the same definition, except now direction matters. If a route includes a link in the wrong direction: semi-path, semi-trail, semi-walk.

A Closer Look at Nodes

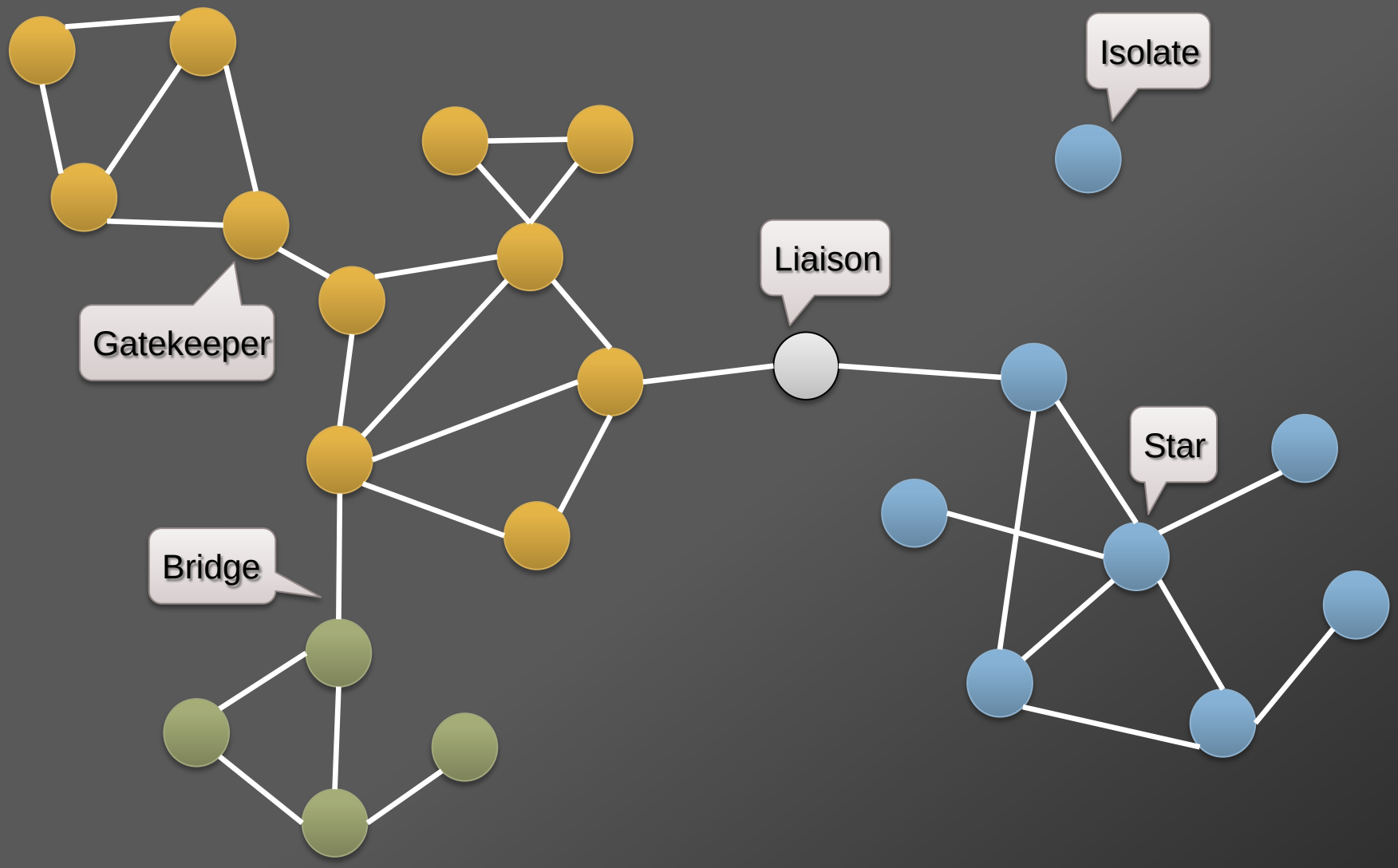
Attributes

- **Every node in a network can be assigned a range of attributes**:
  - Binary (e.g. 1 if the person has Internet access, 0 for those who are not online)
  - Categorical (e.g. 1 for Democrats, 2 for Republicans, 3 for undecided, etc.)
  - Continuous (e.g. age, income, etc.)

Roles

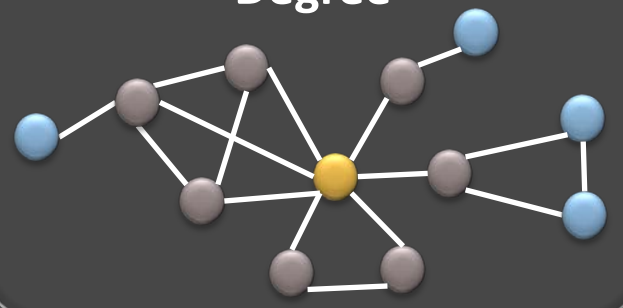
- **Star** - An actor who is highly central (has many connections)
- **Liaison** - An actor who has links to two or more groups that would otherwise not be linked, but is not a member of either group.
- **Bridge** - An actor who links/belongs to two or more groups (strict definitions may require that no other link between the two groups exists)
- **Gatekeeper** - An actor who mediates or controls the flow (is the single link) between one part of the network and another
- **Isolate** - An actor who has no links to other actors

Node Types



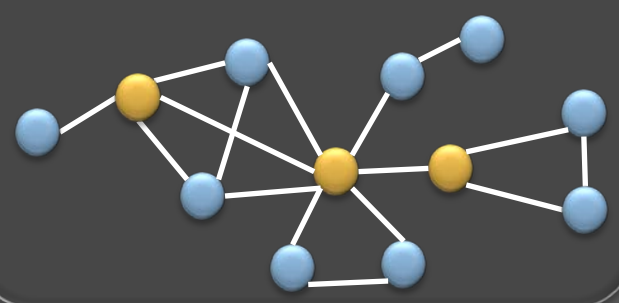
Node Centrality

Degree



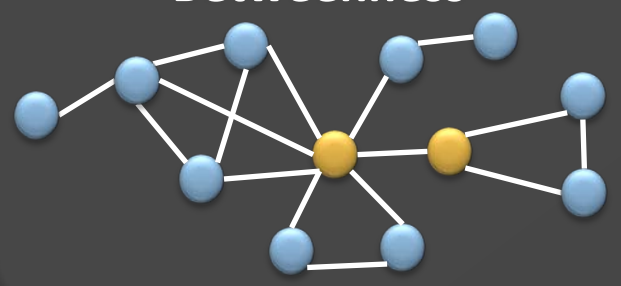
- **Degree:** Number of ties a node has (also called nominations)
  In a directed network: in-degree and out-degree
  In-degree (number incoming ties) also called *prestige.*
  Normalized: divide by total possible number of ties a single node can have (N-1, where N is the number of nodes in the net)
  Shows how well connected the node is to the network.

Closeness



- **Closeness:** based on the average distance from a node to every other reachable node in the network. Calculated as the inverted sum of the shortest paths between the node and every other node. The smallest possible sum(star linked to every node): N-1
  Indicates how quickly an actor can reach everyone in the network – to spread information, etc.

Betweenness



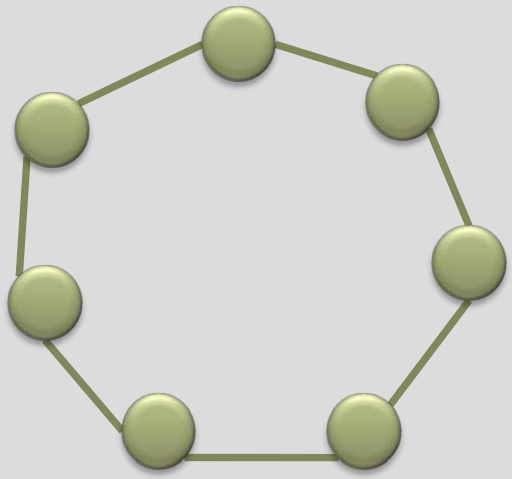
- **Betweenness:** based on the number of cases in which a node lies on the shortest path between two other nodes in the network. Adjusted by total number of shortest paths.
  Tells you whether a node is in a strategic position (gatekeeper, liaison) controlling the spread of information/resources.

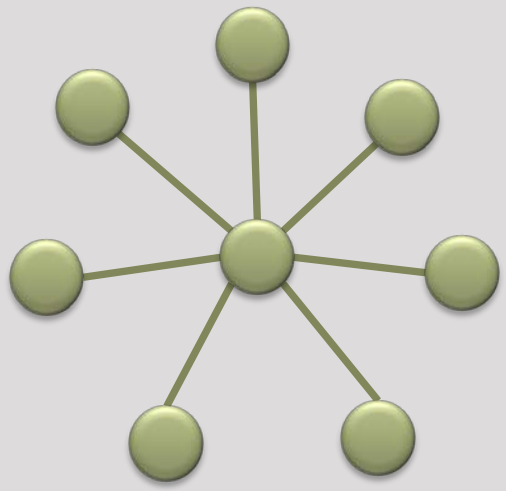
Network Centralization & Density

Degree, **closeness** and **betweenness** are **centrality** measures for individual nodes.

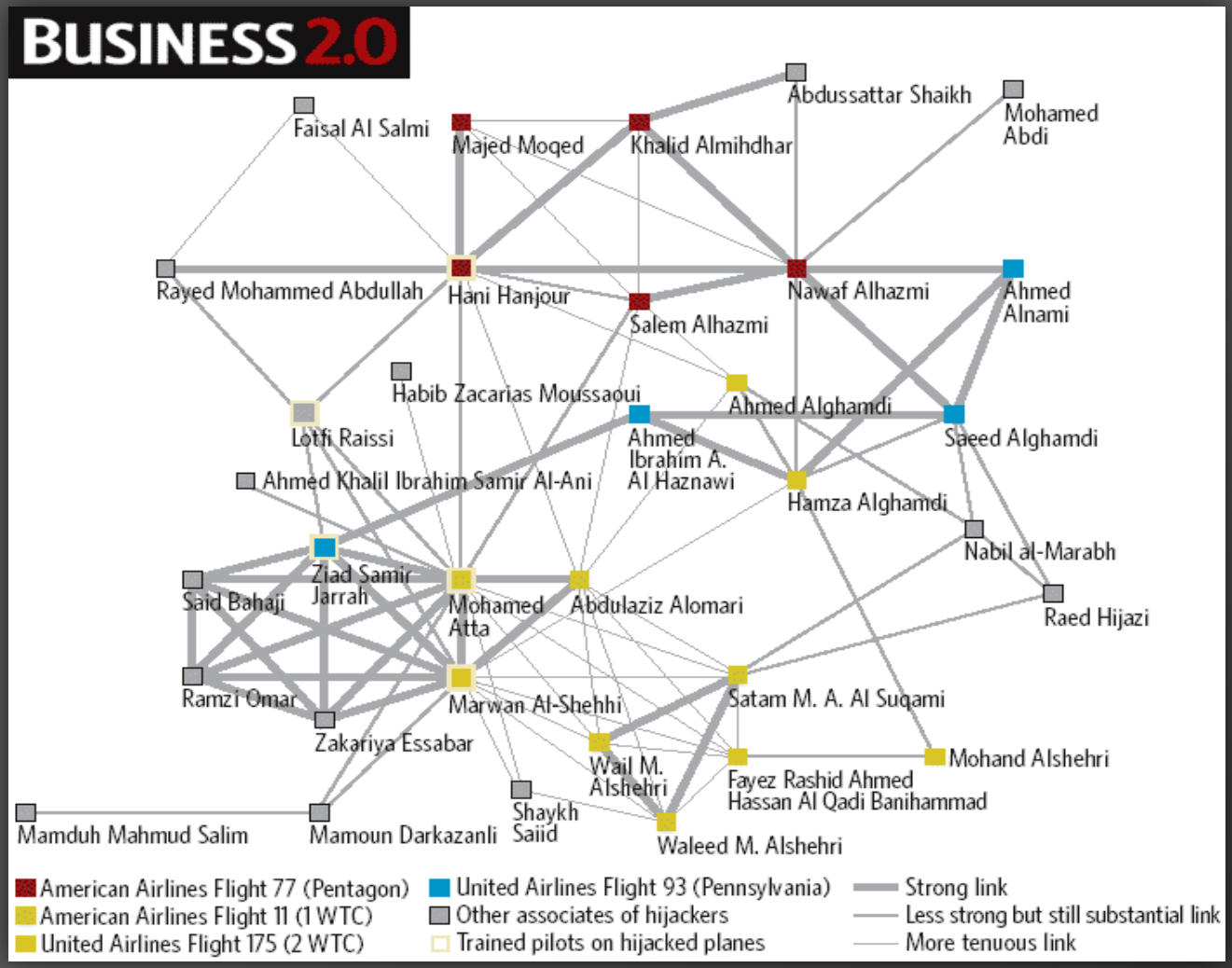
Centralization is a network-level measure. It measures the degree to which an entire network is focused around a few central nodes. In a decentralized network, the links are more or less evenly distributed among nodes.
Centralization is calculated based on the differences in degree centrality between nodes divided by the maximum possible sum of differences.

Density : Ratio of the number of links to the number of possible links in the network
Size: The number of nodes in a network





Example: Terrorist Networks



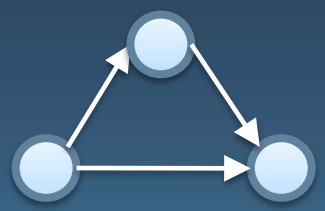
Source: Business 2.0 December 2001. *Six Degrees of Mohamed Atta*

Social Network Measures & Mechanisms

Reciprocity

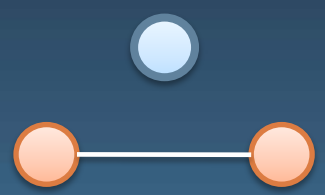
- **Reciprocity/mutuality**: tie A to B and B to A
  Can be calculated as a network measure: proportion reciprocal ties from the total number of ties.
  Reciprocity is calculated for directed networks - in undirected ones all links are by definition reciprocated.

Transitivity



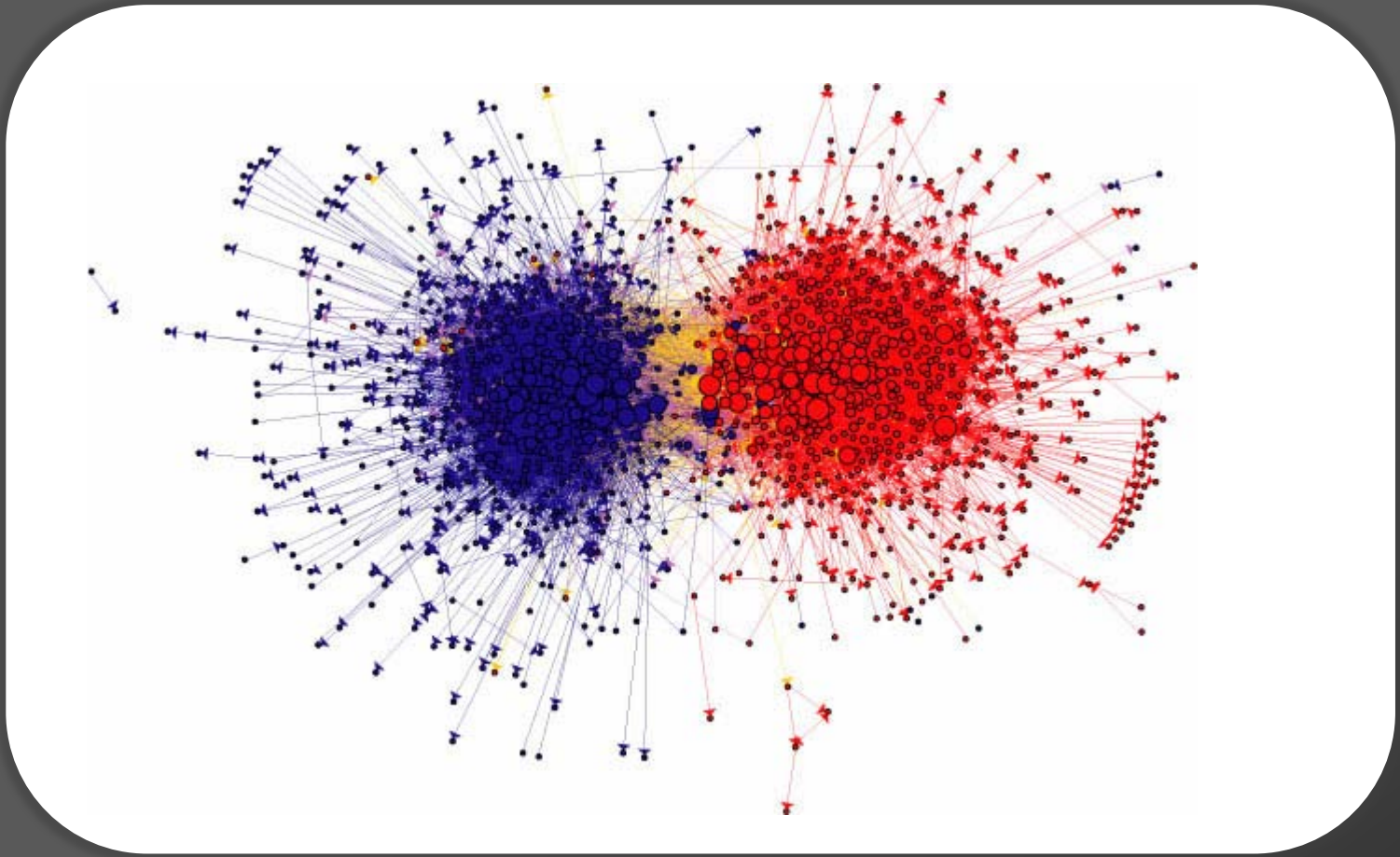
- **Transitivity:** "the friends of my friends are my friends" principle. People look for a balanced environment and try to reduce intransitive triads to avoid dissonance.
- **Clustering:** a measure linked to the number of closed triangles in a network. Higher clustering means more "cliquishness"

Homophily



- **Homophily** is not a network measure – rather it is a mechanism that may motivate link creation and dissolution. The idea behind homophily is that people who are similar in some way (gender, political leanings, etc.) are more likely to be connected. This may be because we create links to people who are like us – or because of social influence (our friends become similar to us with time).

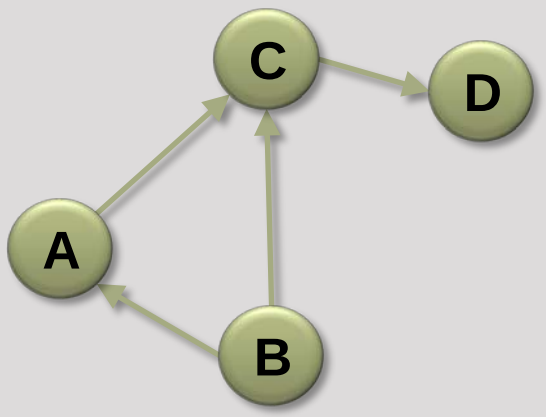
US Political Blogosphere



Linking patterns in the US political blogosphere. Source: Adamic & Glance (2005)

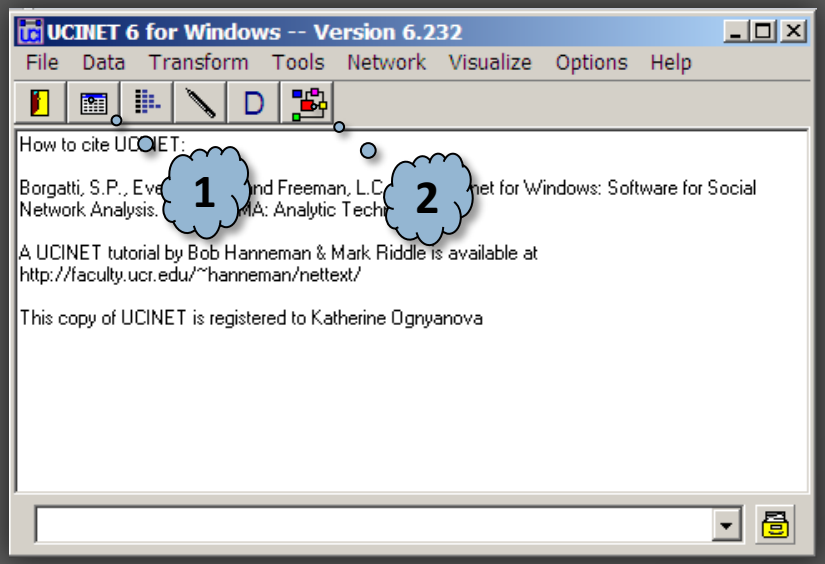
What does network data look like?

Although there are other options (edge lists, node lists, etc.) network data is typically used in the form of a matrix .
Row X column Y is 1 if there is a link from X to Y - and 0 if there is no link.
The diagonal represents self-loops: links from X back to X.



	A	B	C	D
A	0	0	1	0
B	1	0	1	0
C	0	0	0	1
D	0	0	0	0

Using UCINET for Network Analysis



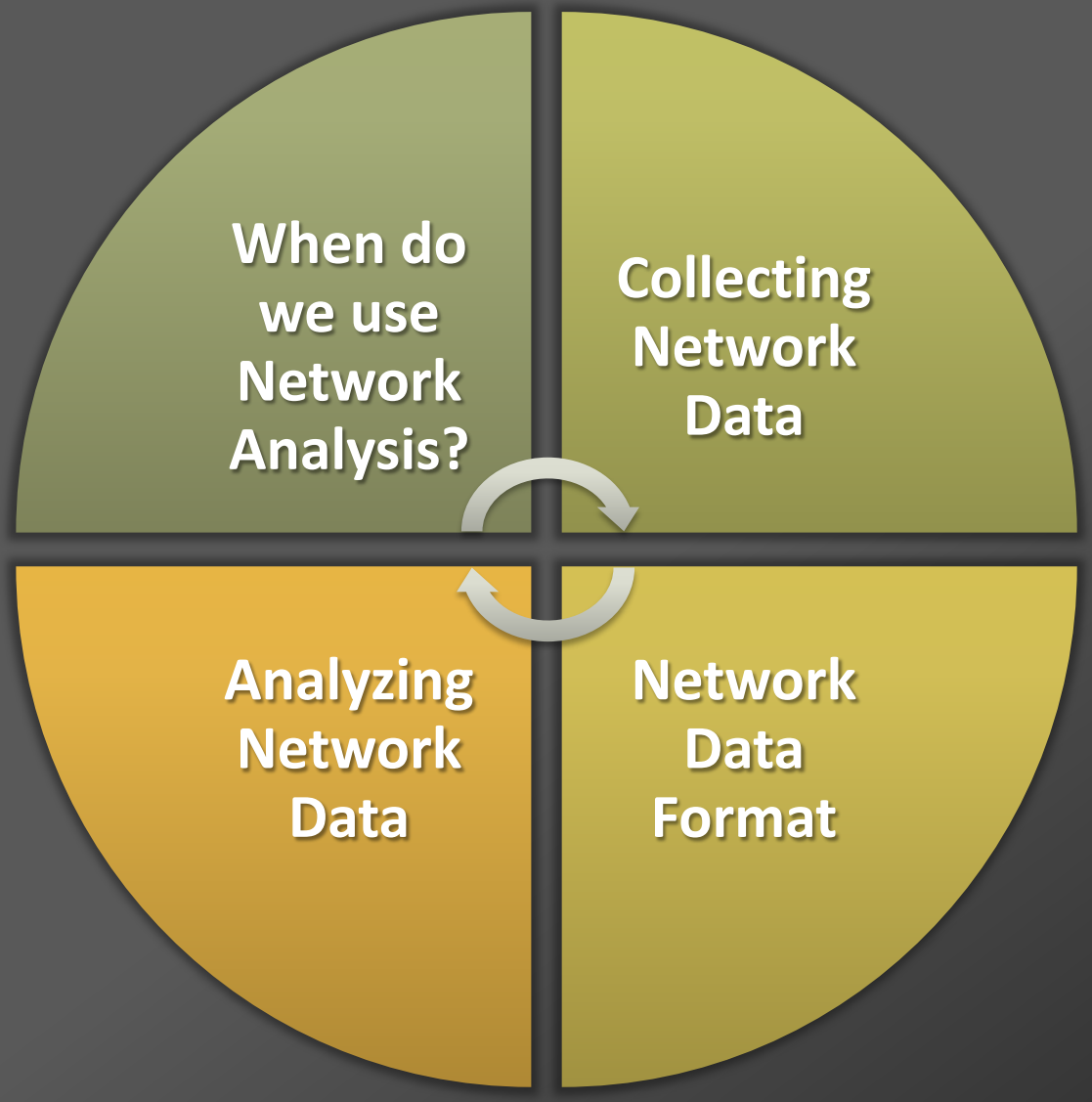
- **UCINET**: download from this website. You get two months of free trial.

- **Data Input**:
  - Direct: Click on (1) – copy and paste from Excel or enter manually a matrix.
  - Import: Data menu -> Import via Spreadsheet - > Full matrix w/ multiple sheets.

For every network Ucinet will create two files: .##h and .##d. If you move them, move both.

- **Density** : Network menu -> Cohesion -> Density
- **Reciprocity / Transitivity / Clustering :** Network menu -> Cohesion -> Reciprocity / Transitivity / Clustering
- **Node centrality :** Network -> Centrality -> Multiple measures (old)
- **Visualization**: Press button (2) for NetDraw.

Studying Social Structures Through Network Analysis



Collecting Network Data

Complete Networks

- Administer a network questionnaire to the full population. Ask everyone about their own relationships. Typical questions involve naming close friends, people one goes to for advice, etc.
- Good idea to have a roster that people can select from.

Cognitive Social Structure

- Ask not only for a person's relationships, but also for perceived relationships between other people in your population.
  Different ways to compile a network from CSS based on your interests.

Snowball Sampling

- Often used in preventive medicine. Contact a high-risk group, ask each person in your sample to bring in a number of their contacts (friends, sexual partners, etc.)

Secondary Data

- Online social networks are typically studied through secondary data. Instead of asking individuals, a researchers can compile a network based on hyperlinks, social network site friends/followers, etc.

Example: Snowball Sampling

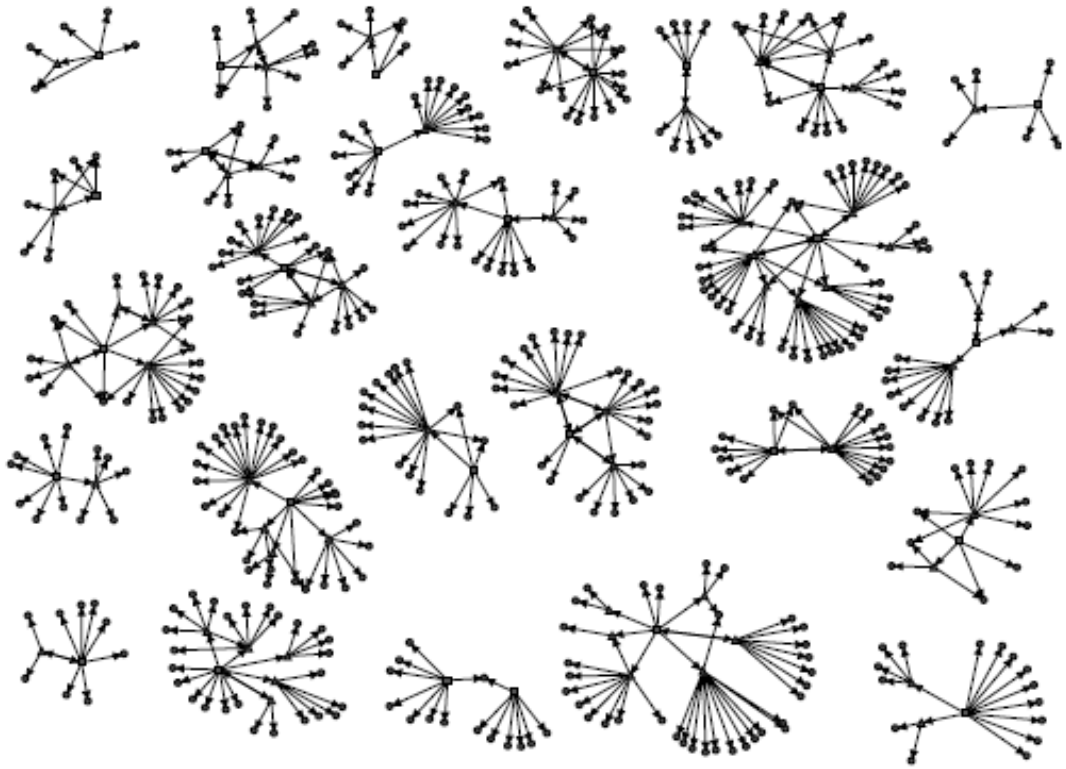


Figure 4–4. Networks of HIV-positive patients in which at least one alter was enrolled in an HIV vaccine preparedness study. Indexes are squares, enrolled alters are triangles, and nonenrolled alters are circles. Links are coded by willingness to invite the alter to participate in vaccine preparedness activities: solid arrows represent willingness, dashed arrows, not willing; dash-dotted arrows, absence of a response; and dash-dot-dot, enrolled but not named first-degree alters.

Source: Valente (2010) Social Networks and Health

The Web as a Network

Hyperlinks as Network Ties

- Hyperlinks are directed ties from one webpage to another.

  Two approaches to assessing the meaning of a link:
  - **Hyperlinks as affiliation**: linking occurs between sites which share some common ground, talk about similar things, etc.
  - **Hyperlinks as identifying value:** in this view links (not individual ones, but in the aggregate) are expected to point to content that is credible, relevant or authoritative. In this case links serve as reputation markers.

- PageRank – search engine algorithms are grounded in understanding of the hyperlink as a nomination/vote for the page it links to.
- Search engine ranking takes into account the affiliation aspect.

Power Law Distribution

- "The rich get richer" effect and preferential attachment: pages that already have many links are more likely to get linked to. (By the by: similar distribution appears for visitor numbers)
- How does this relate to concerns about fragmentation of online audiences?

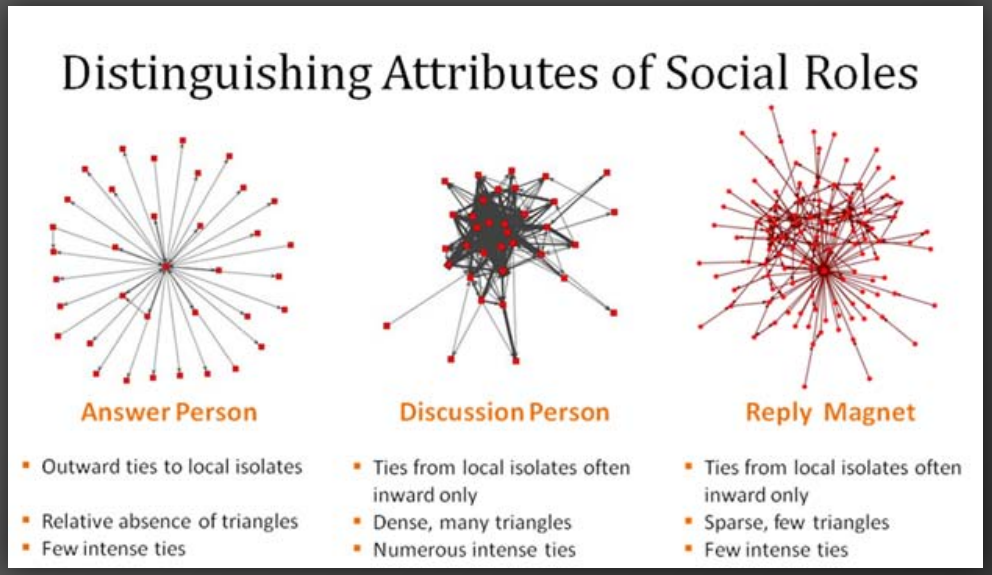
E-Mail and Discussion Forum Networks

E-Mail Networks

- Nodes: e-mail accounts/people
- Ties: directed, X sends an e-mail to Y (may be valued – number of e-mails). Example: the Enron e-mail network.

Discussion Forums

- Nodes: individual users
- Ties: directed, X replies to/comments on a post by Y. New, specific roles emerge for nodes in this network.



Source: Welser, Gleave, Fisher & Smith, 2007

Diffusion Through Online Social Networks

- **Relevant networks:** online social networks (Facebook, etc.), blogs, and microblogging services(Twitter, etc.)
- **Relevant Questions:**
  - How does information flow through the network? Who are the influential nodes? How can we get a message out to everyone?
  - How can we minimize the spread of undesirable things (e.g. computer viruses) by blocking a minimal number of links?

- **Epidemiology**: the **S-I-R model** (Susceptible-Infected-Recovered).
  In the online-diffusion version of SIR, users may become *susceptible* to a topic when it is suggested to them by a friend (either through a blog post or via a service like Twitter or Facebook). The person may then be *"Infected"* with the topic: they write a post about it or publish a link on a social platform. With this, the individual is considered to have *recovered* from the topic, although a relapse is possible when something new appears on the subject.

- **Threshold models of diffusion:** an actor's decision to disseminate a topic is based on the proportion of their connections who have already started discussing the subject.

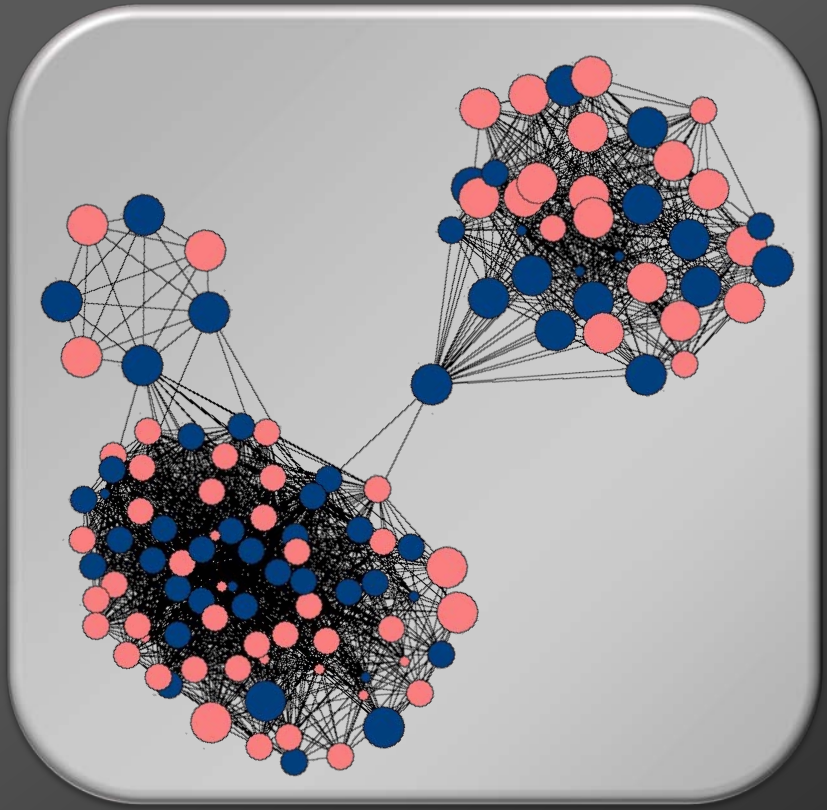
- **Cascade models:** each time an actor is "infected" with a new topic there is a certain probability that the infection will spread to neighboring nodes.

Online Resources for Social Networks

How can we collect data about online networks?

- **Facebook:** Name Gen Web by Bernie Hogan is a Facebook app that will collect data about your network and let you download it in a UCINET file format.

- NodeXL is an add-on for network analysis in Excel. It can collect and analyze Twitter, Flickr and e-mail nets.

- SNAP (Stanford Network Analysis Platform) has a data library with big network datasets from Google, Amazon, Wikipedia, social network sites, etc.

- IssueCrawler, LexiURL and SocSciBot are some of the available solutions for crawling web pages and collecting hyperlink networks.