Networks Out of Control: Real-World Networks 4

Announcement: Project Selection

Project

Details:

- Pair project
- 4 page (2-column) report due Wednesday 23rd May.
- 15 min presentation + 5 min questions on Monday 28th May to be scheduled between 2pm-7pm.

Contact Farnood with the following by Friday 20 April:

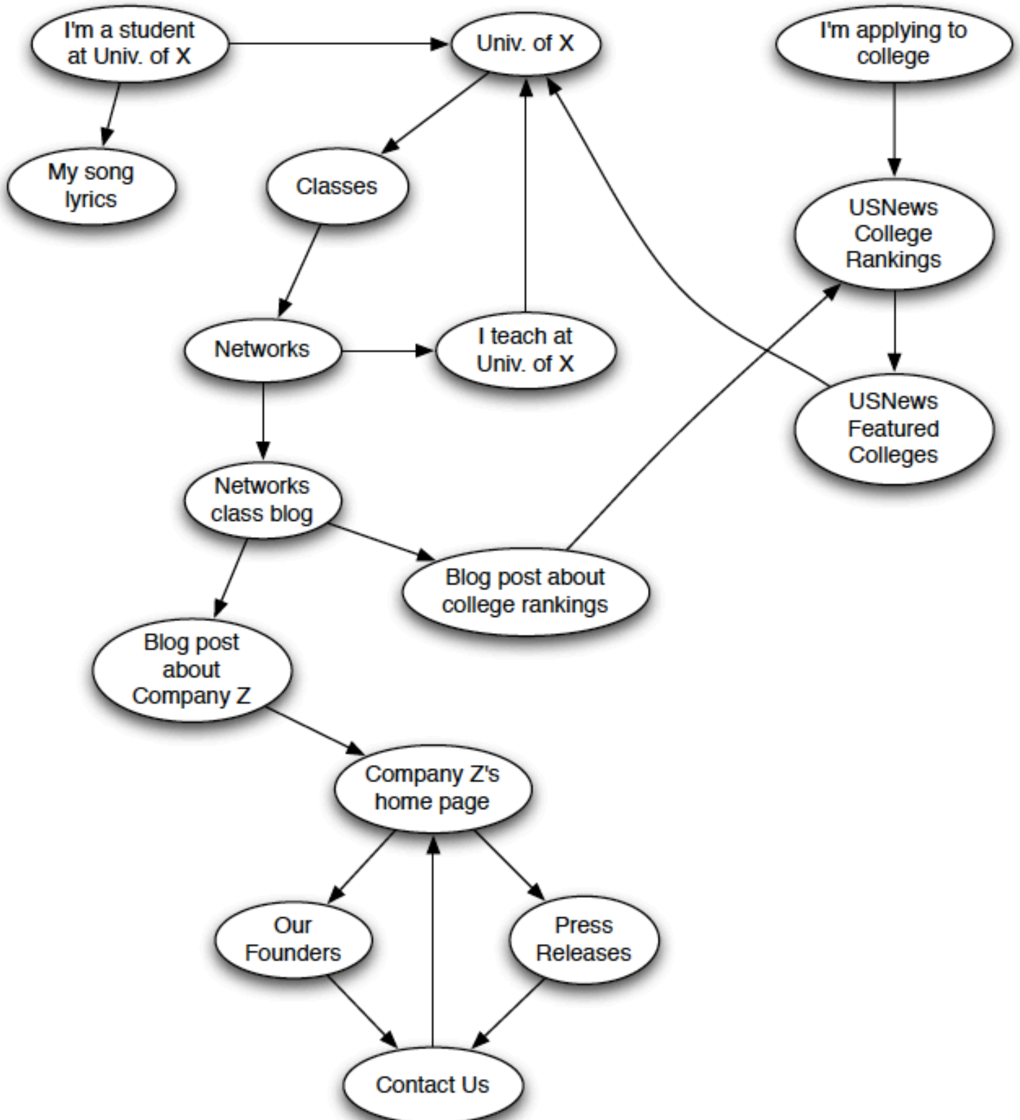
- Who your team is (pairs)
  - In case you do not find a team, respond to the other two points and we will match you with someone
- List of 3 preferred papers, in order, on which you would like to do your project (ties get broken by when the email was sent).
- Availability to present in the 2pm-4pm time slot on Monday.

Real-World Example 1: The Internet Graph

Internet Graph (Virtual Layer)

- Nodes are webpages. (Directed) edges are hyperlinks from one webpage to another.
- Analyze properties of this network as a first step in order to understand efficiency of routing algorithms, vulnerability to attacks on vertices or edges, predicting user behavior, etc.

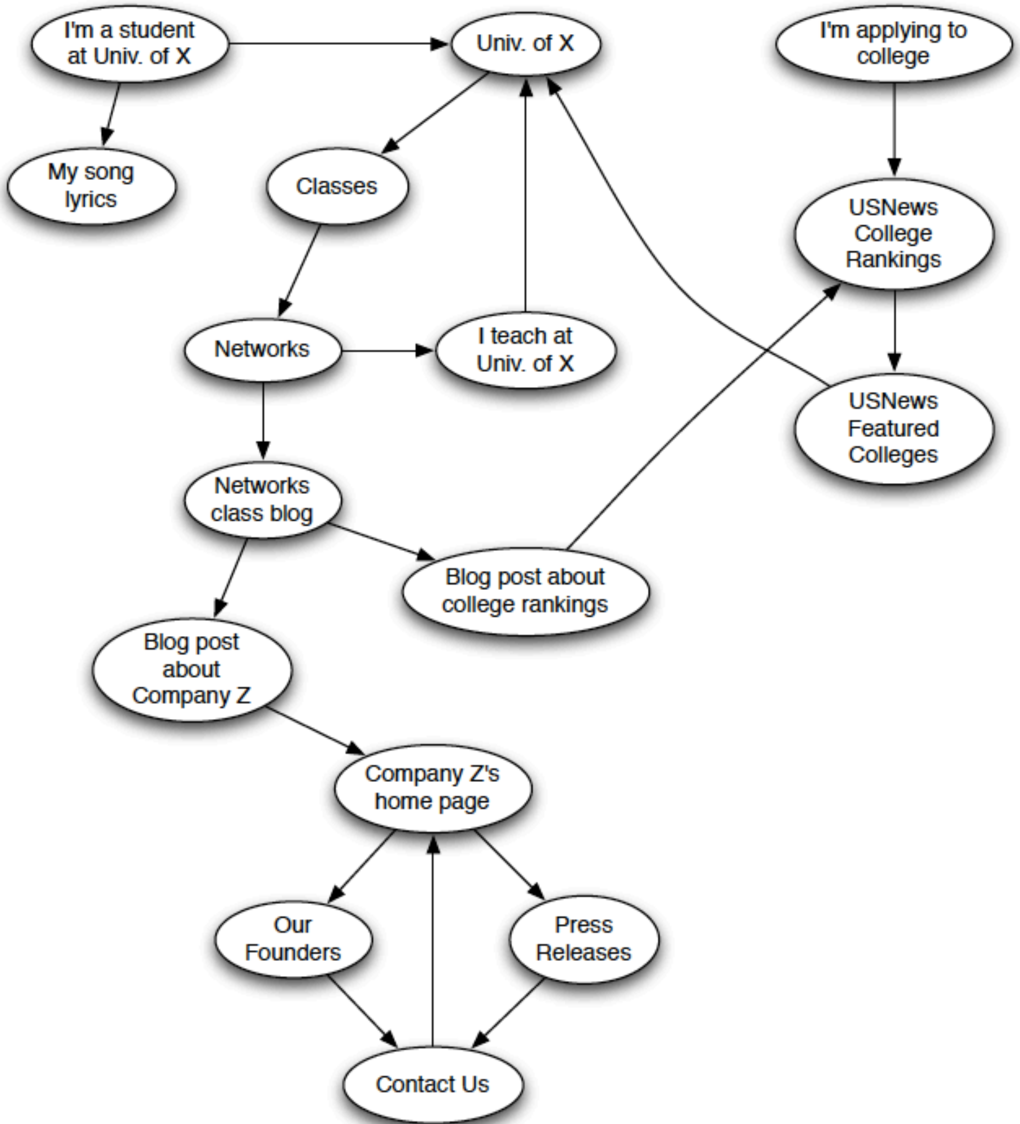
Internet Graph Structure



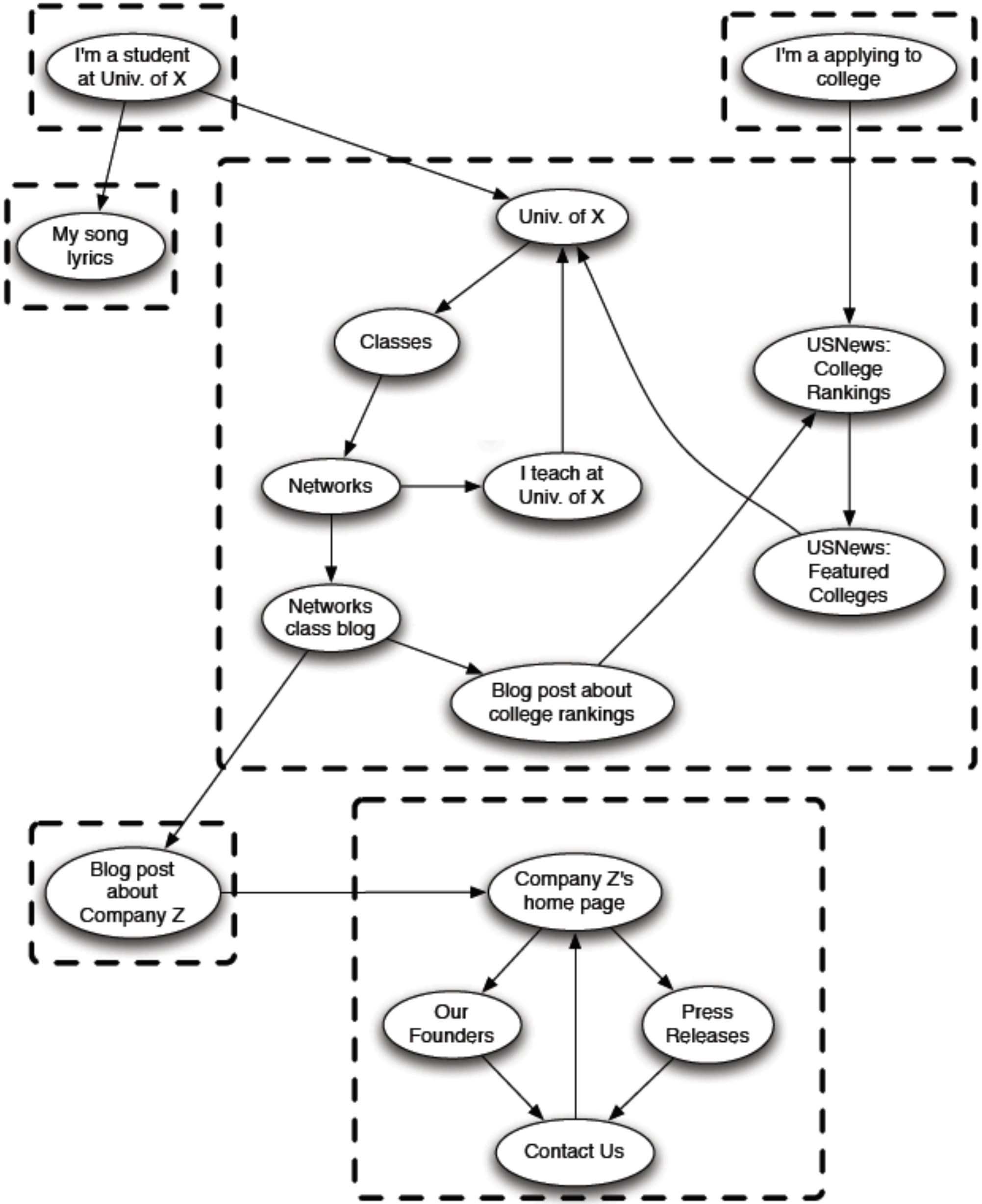
+ Internet Graph Structure

- The graph is *directed, so* $uv \in E \not\Rightarrow vu \in E$
- We say nodes *u* and *v* are *strongly connected* if there is a path from *u* to *v* and *also* a path from *v* to *u*.
- A *strongly connected component* $S \subseteq V$ is a *maximal* set such that all pairs of vertices in *S* are strongly connected.

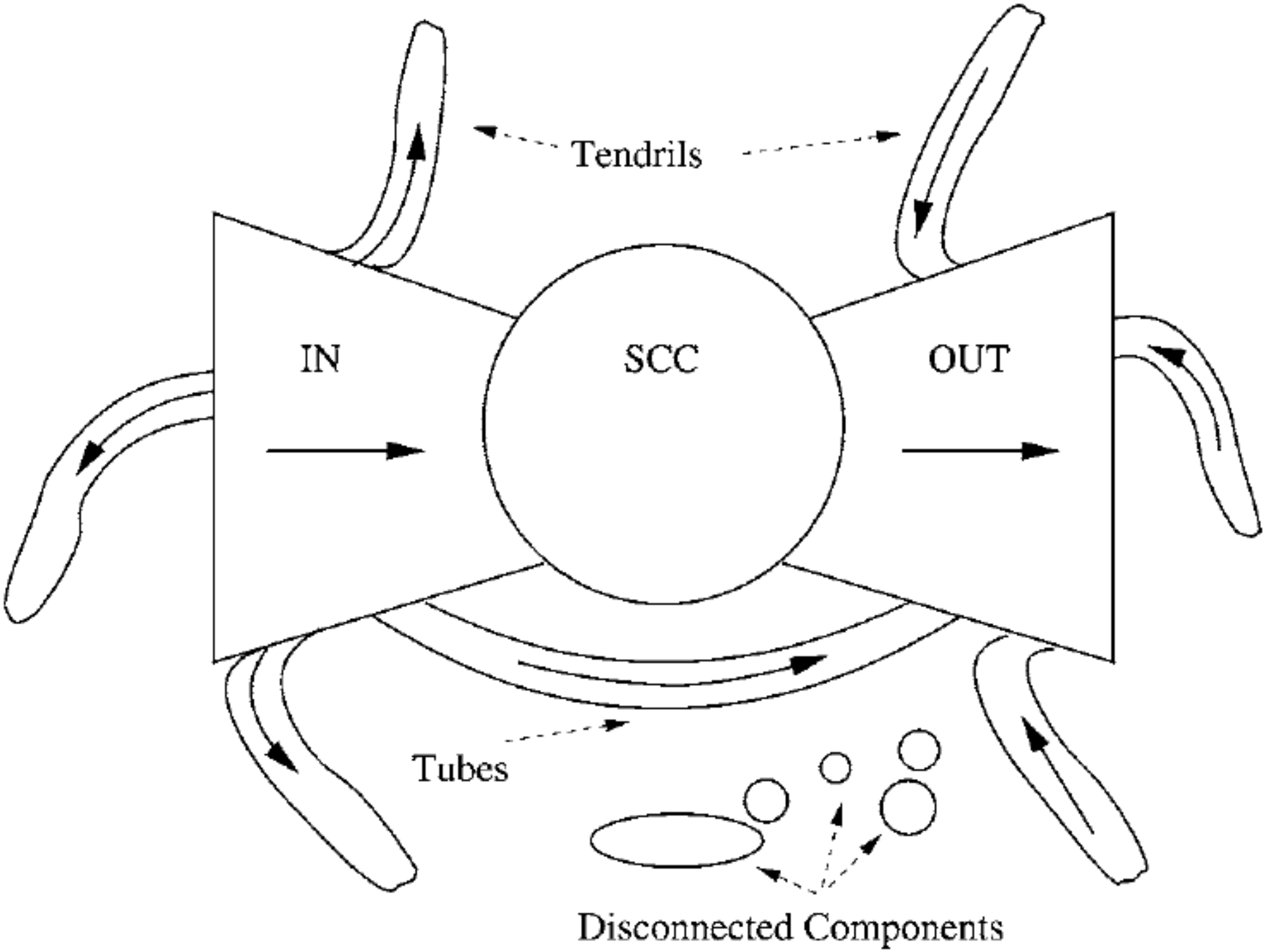
+ Internet Graph Structure



Internet Graph Structure



The Bowtie Structure of the Internet



+ Degree Distribution

- Let *D* be the degree of a randomly selected node.
- The degree distribution is: $\mathbb{P}[D = d]$
- G(n,p): $\mathbb{P}[D = d] = \binom{n-1}{d} p^d (1-p)^{n-1-k}$
- G(n,k): $\mathbb{P}[D = d] = \begin{cases} 1 & if\ d = k \\ 0 & otherwise \end{cases}$
- Using the same matching method used for G(n,k), can construct networks with arbitrary degree distributions!

Power Law Degree Distributions

- Degree distributions in the internet graph resemble a power law.

- Let *D be the degree of a randomly selected node.*
  - Power law: $\mathbb{P}[D > d] \sim d^{-\gamma}$
  - Most distributions we have been working with have "light tails":
    - Exponential, Geometric, Gaussian, Poisson, … $\mathbb{P}[D > d] \sim e^{-\alpha d}$

- Networks with power-law degree distributions are called *Scale-Free Networks.*
  - In effect, such networks have a large number of "hubs", i.e., vertices of very large degree.

+ Pareto Distribution

- The Pareto Distribution:

$$\mathbb{P}[D > d] = \begin{cases} \dfrac{d}{\beta}^{-\gamma} & \textit{if } d \geq \beta \\ 1 & \textit{otherwise}. \end{cases}$$

- The moments of a Pareto distribution only exist up to γ

$$\mathbb{E}[D^k] = \begin{cases} \left(\dfrac{\gamma \cdot \beta}{\gamma - k}\right)^k & \textit{if } k \leq \gamma \\ \infty & \textit{otherwise}. \end{cases}$$

Examples of Observed Power Laws

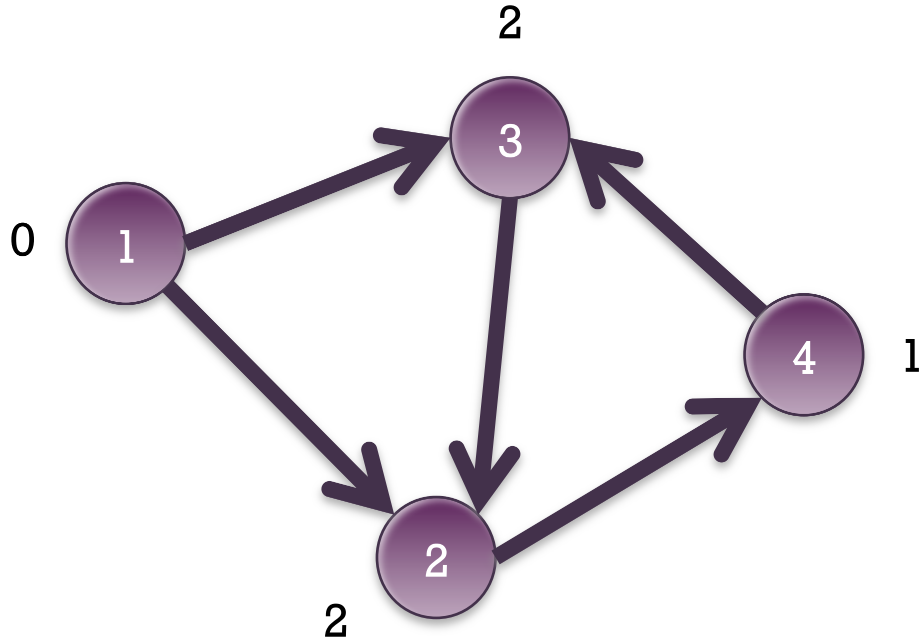
- Sizes of cities
- Phone call length
- Wealth & income distribution
- Word frequencies in prose
- Internet graph degree distribution
- …

Why Power Laws?

- If you were new to Switzerland, would you take an apartment in Prevange or Lausanne?
  - More likely Lausanne, because **more people are already there**
- The "rich-get-richer" phenomena:
  - It is easier to make $1 when you have $1,000,000 than when you have $10.
- *Dynamic* model.

+ Preferential Attachment

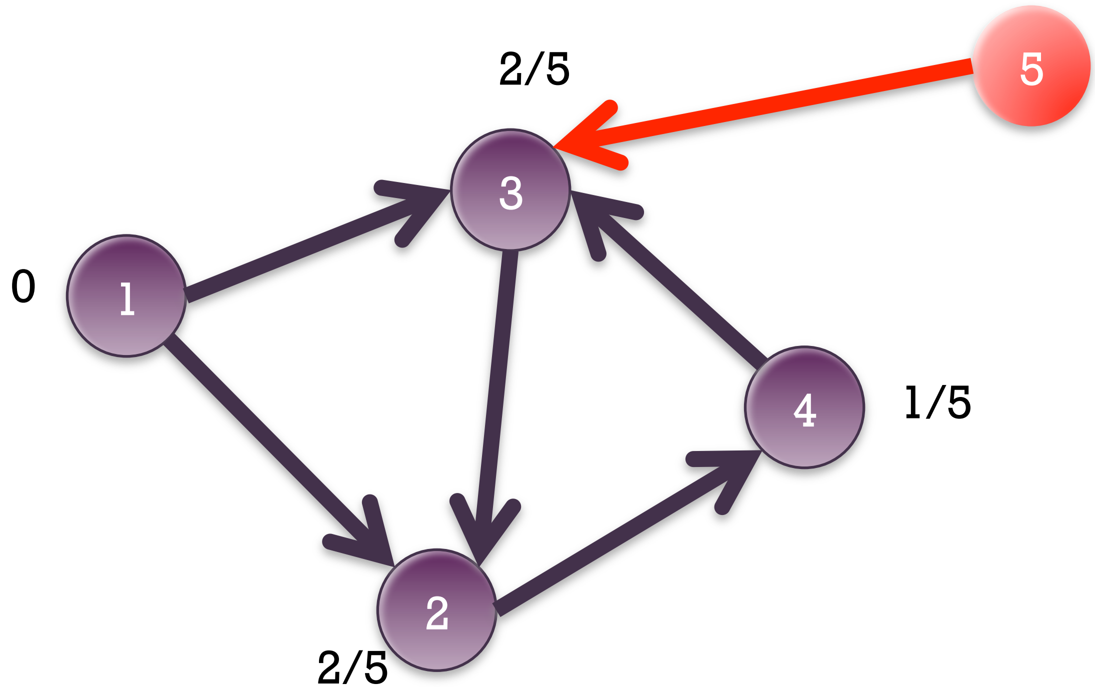
- Directed graph (in the homework will see an undirected version).
- In-degree, denoted by $d_{in}(v)$, measures "popularity" of a node



Preferential Attachment

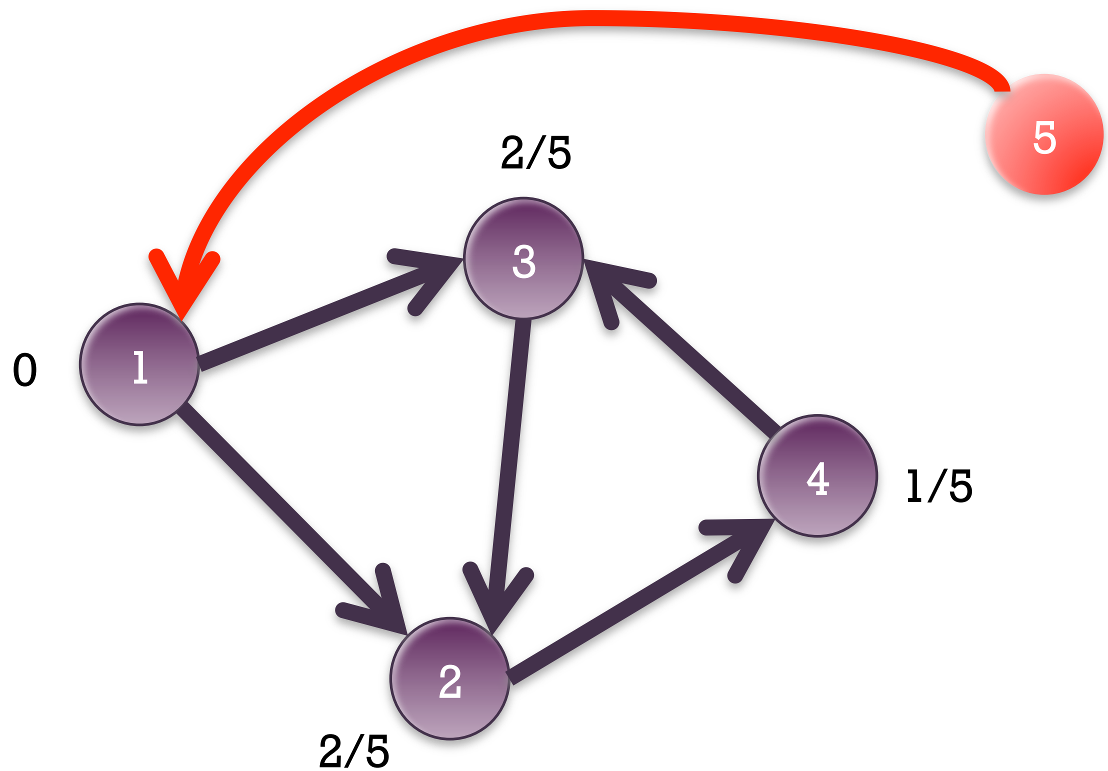
- Nodes arrive one-by-one and connect one edge to an existing node.

- Probability of connecting to a node is proportional to it's indegree.



+ Preferential Attachment

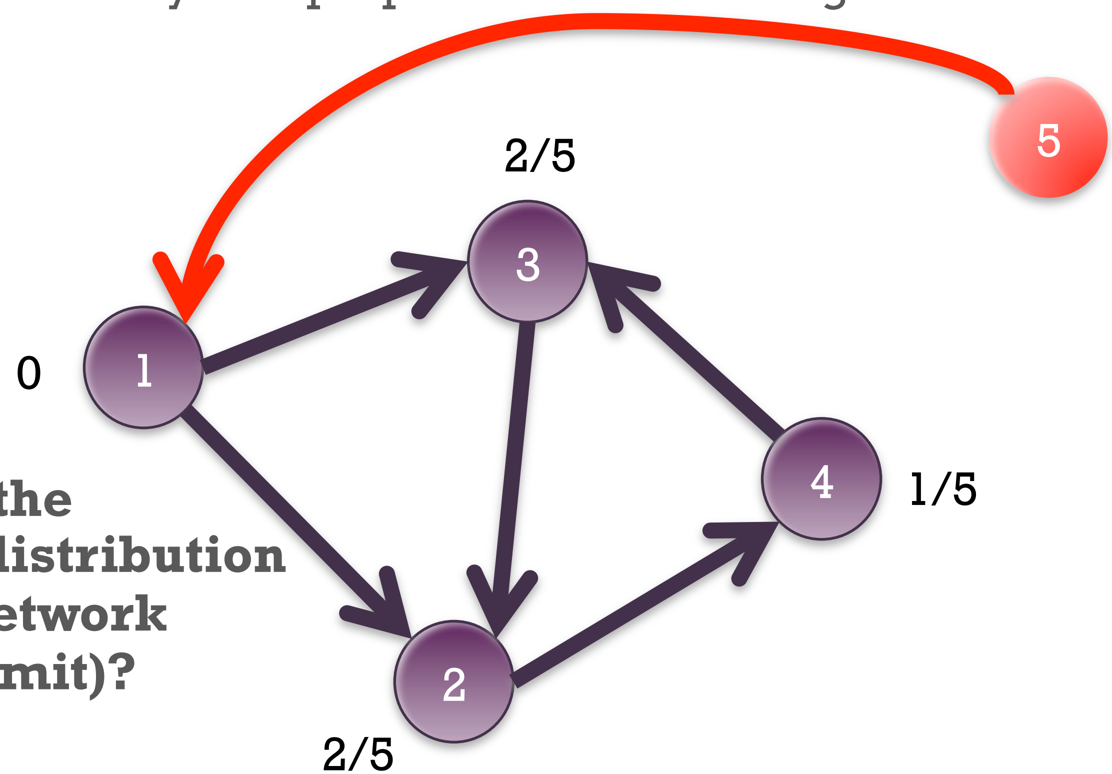
- Nodes arrive one-by-one and connect one edge to an existing node.
- Probability of connecting to a node is proportional to it's indegree.



Preferential Attachment

- Nodes arrive one-by-one and connect one edge to an existing node.

- Probability of connecting to a node is
  - With probability α uniformly at random.
  - With probability 1-α proportional to it's indegree.



- **What is the degree distribution of this network (in the limit)?**

Preferential Attachment: Degree Distribution

- Let $X_j(t)$ be the number of nodes with in-degree j at time t.
  - Total number of nodes and edges at time t is exactly t
- How does $X_j(t)$ change with time?

$$\mathbb{P}[X_j(t+1) = X_j(t)+1] =$$

$$\mathbb{P}[X_j(t+1) = X_j(t)-1] =$$

- In general, such stochastic processes are difficult to analyze. Instead of doing it directly, we will use a *continuous mean-field approximation.*
  - ***Assume the mean (i.e., expected) number of nodes at each step***

$$\mathbb{E}[X_j(t+1) - X_j(t)] = \alpha \frac{X_{j-1}(t) - X_j(t)}{t} + (1-\alpha)\frac{(j-1)X_{j-1}(t) - j\,X_j(t)}{t}.$$

Preferential Attachment: Degree Distribution

- Let $X_j(t)$ be the number of nodes with in-degree j at time t.
  - Total number of nodes and edges at time t is exactly t
- How does $X_j(t)$ change with time?

$$\mathbb{P}[X_j(t+1) = X_j(t)+1] = \alpha \frac{X_{j-1}(t)}{t} + (1-\alpha)$$

$$\mathbb{P}[X_j(t+1) = X_j(t)-1] = \alpha \frac{X_j(t)}{t} + (1-\alpha)\frac{j \cdot X_j(t)}{t}.$$

- In general, such stochastic processes are difficult to analyze. Instead of doing it directly, we will use a *continuous mean-field approximation.*
  - ***Take the continuous limit***

$$\frac{dX_j(t)}{dt} = \alpha \frac{X_{j-1}(t) - X_j(t)}{t} + (1-\alpha)\frac{(j-1)X_{j-1}(t) - j\,X_j(t)}{t}.$$

+ Preferential Attachment: Degree Distribution

$$\frac{dX_j(t)}{dt} = \alpha \frac{X_{j-1}(t) - X_j(t)}{t} + (1-\alpha)\frac{(j-1)X_{j-1}(t) - j\,X_j(t)}{t}.$$

- We can now evaluate this dynamical system, which we can analyze using our standard tools.
- Note that the degree distribution in this case is given by

$$\mathbb{P}[D = d] = \frac{X_d}{t}$$

- To sketch the remainder of the proof, assume that the degree distribution converges (this requires proof!), i.e.,

$$\frac{X_d}{t} \to c_d \text{ as } t \to \infty$$

- We now just have to solve for c_d

Preferential Attachment: Degree Distribution

$$\frac{dX_j(t)}{dt} = \alpha \frac{X_{j-1}(t) - X_j(t)}{t} + (1-\alpha)\frac{(j-1)X_{j-1}(t) - j\,X_j(t)}{t}.$$

- Rearranging the above equation using $c_d = \dfrac{X_d}{t}$ we get that

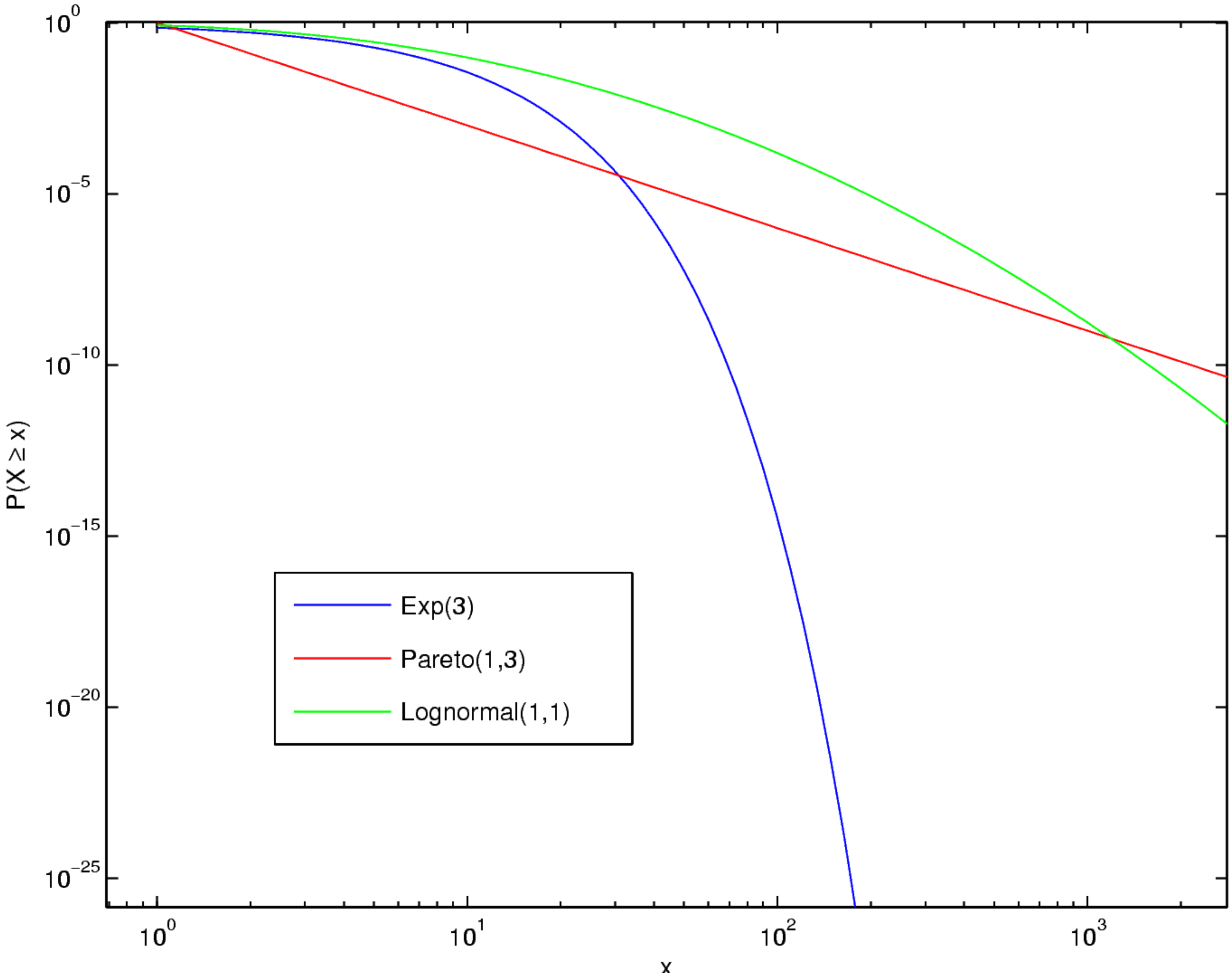
$$\frac{c_j}{c_{j-1}} = \frac{\alpha + (1-\alpha)(j-1)}{1+\alpha+(1-\alpha)j} = 1 - \frac{2-\alpha}{1+\alpha+(1-\alpha)j}$$

- Using standard approximations for large enough *j*, we get

$$\frac{c_j}{c_{j-1}} \approx 1 - \frac{2-\alpha}{1-\alpha}\cdot j^{-1} \approx \left(1-\frac{1}{j}\right)^{\frac{2-\alpha}{1-\alpha}} \approx \left(\frac{j}{j-1}\right)^{-\frac{2-\alpha}{1-\alpha}}$$

- So $c_j \approx j^{-\frac{2-\alpha}{1-\alpha}}$

Power Law vs Lognormal



Power Law vs Lognormal

- Power Law:
  - Asymptotically like a Pareto Distribution
  - Generative Mechanism: "rich-get-richer"
- Log-normal Law:
  - All moments exist (not heavy-tailed), but "looks like" a power law.
  - Generative mechanism: product of many independent random variables (log satisfies the central-limit theorem).
- Long-running controversy.
- More generally – preferential attachment is not the only way to get a heavy-tailed distribution.

+

Complex Networks

Complex Networks

- So far, all the networks we have studied are composed solely of nodes and edges.
  - What if there is additional *structure* to the network?

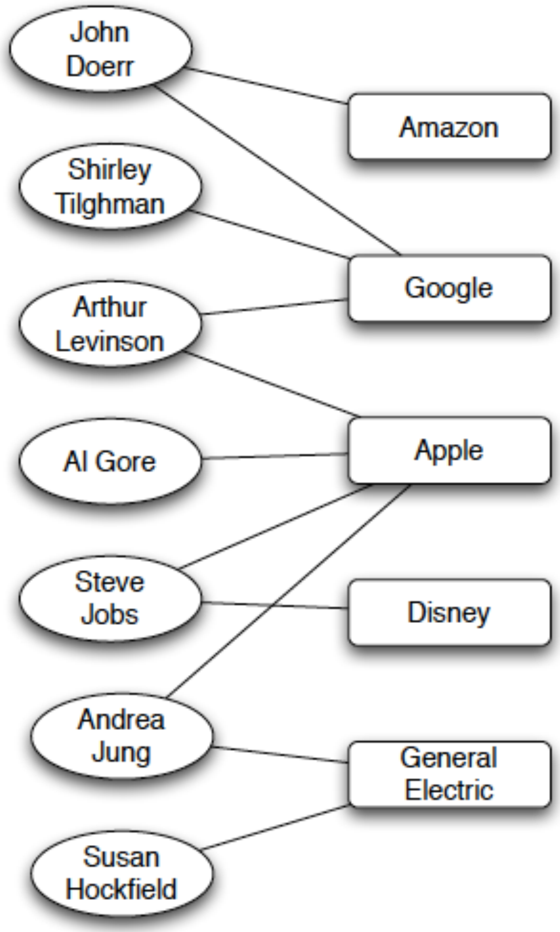
- Affiliation Networks: more than one type of *node*.
  - People and affiliations (e.g., clubs/employers)

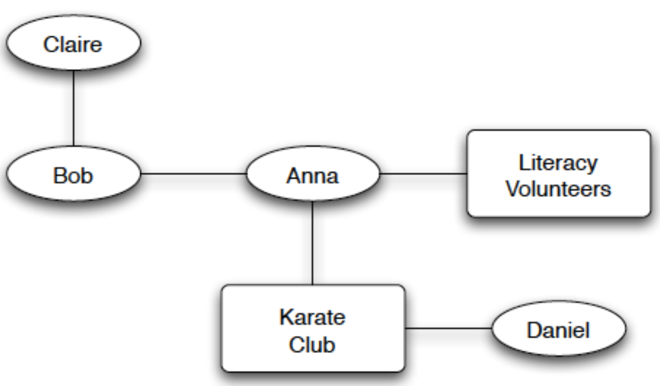
- Signed Networks: more than one type of *edge.*
  - Positive and negative edges (e.g., friends and enemies)

Affiliations

Note Types

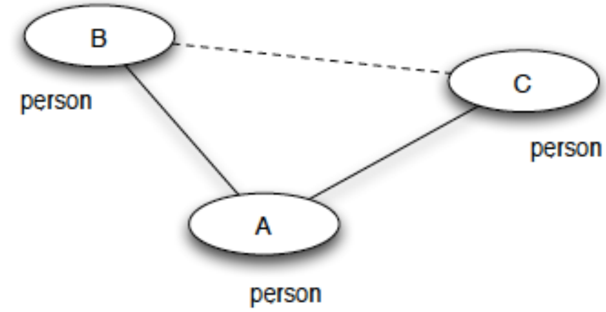
- Nodes are either "individuals" or "clubs" (also called *foci*).
- Affiliation Networks: The graph is *bipartite*
  - Edges are only between individuals and clubs.
  - Individuals are (implicitly) linked if they are affiliated with the same club, and clubs are (implicitly) linked if they have the common individuals affiliated with them.
- Social-Affiliation Networks: A social network superimposed with an affiliation network.
  - Edges from individual to club, or from individual to individual (not from club to club)



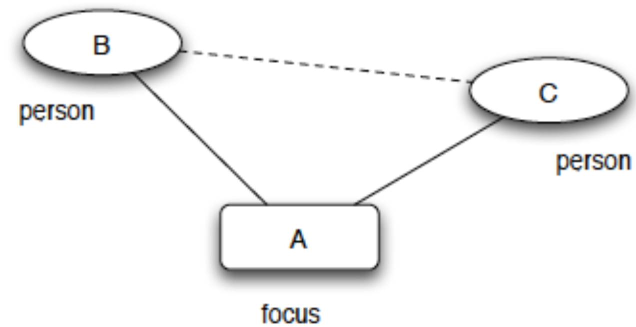


+ Closure

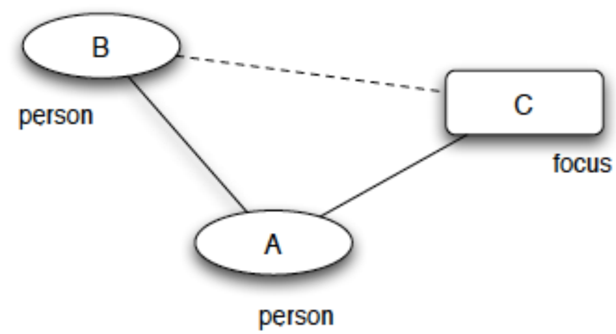
- Recall from earlier lectures:
  - Clustering coefficient (e.g., fraction of possible triangles) tends to be high in social networks.
  - *Triadic closure*: "friends of friends" are likely to meet and form an edge.
- Social-Affiliation Networks: Two more kinds of closure can form.
  - *Focal closure*: Edges between individuals who have a focus in common are likely to form.
  - *Membership closure*: Edges between an individual and a focus that a friend participates in are likely to form.



(a) *Triadic closure*

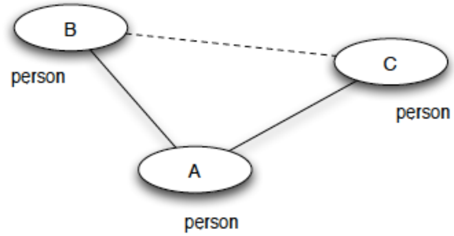


(b) *Focal closure*

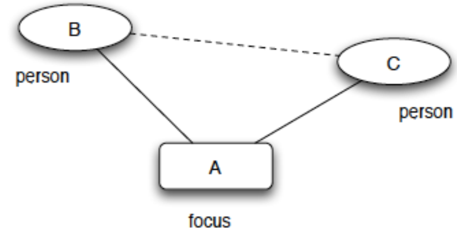


(c) *Membership closure*

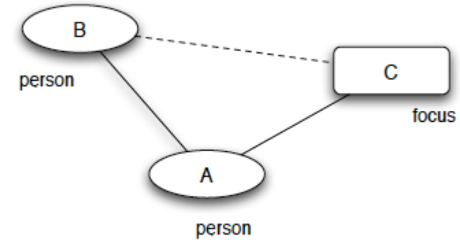
+ Closure



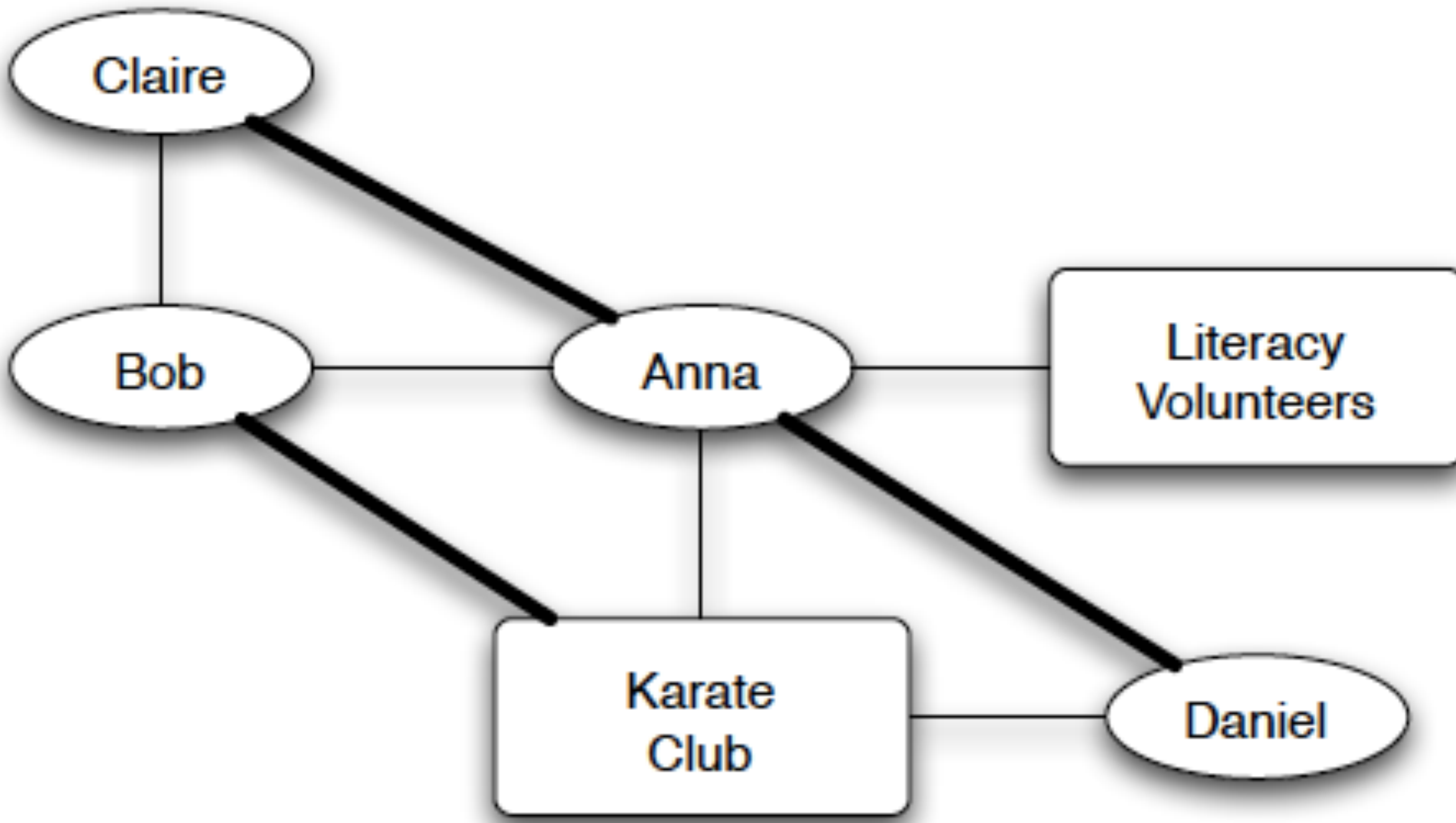
(a) *Triadic closure*



(b) *Focal closure*



(c) *Membership closure*

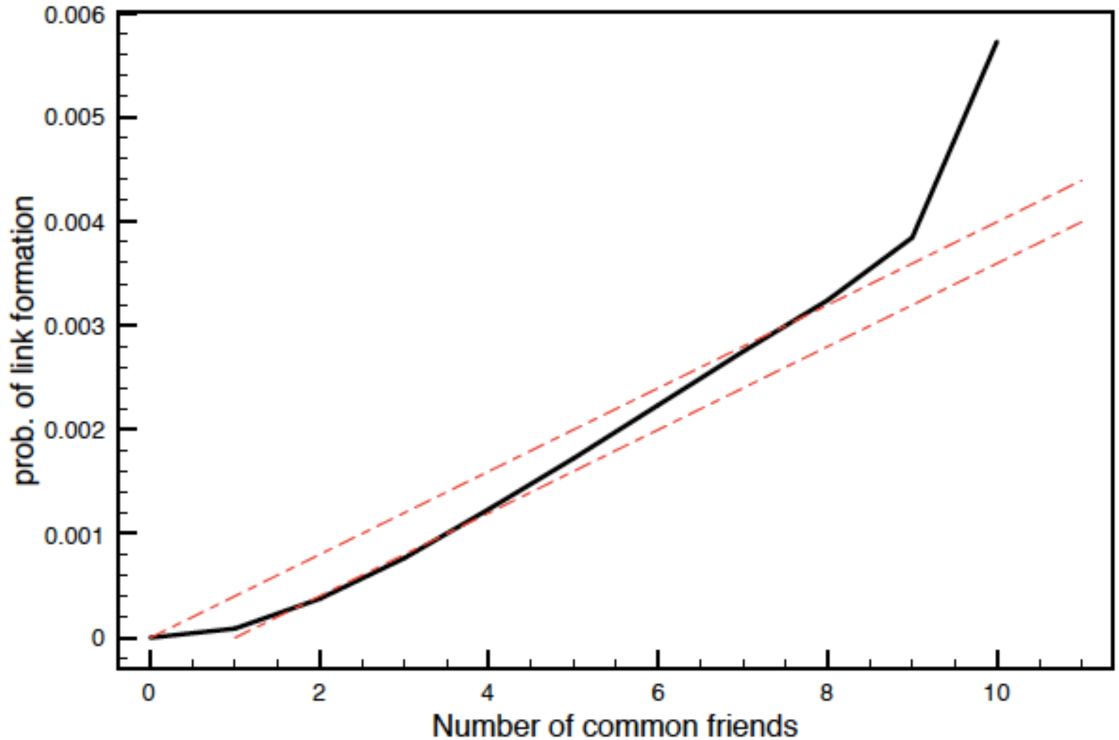


Tracking Link Formation in Data

- *How much more likely is a link to form between two people in a social network if they have a friend in common?*
- *How much more likely is a link to form between two people in a social network if have **multiple** friends in common?*

- Empirical Study:
  - Take two snapshots of the network at different points in time.
  - For each k, identify all pairs of nodes that have exactly k friends in common at the first snapshot but are not connected by an edge.
  - Let T(k) be the fraction of pairs that have formed an edge in the second snapshot.

Tracking Link Formation in Data

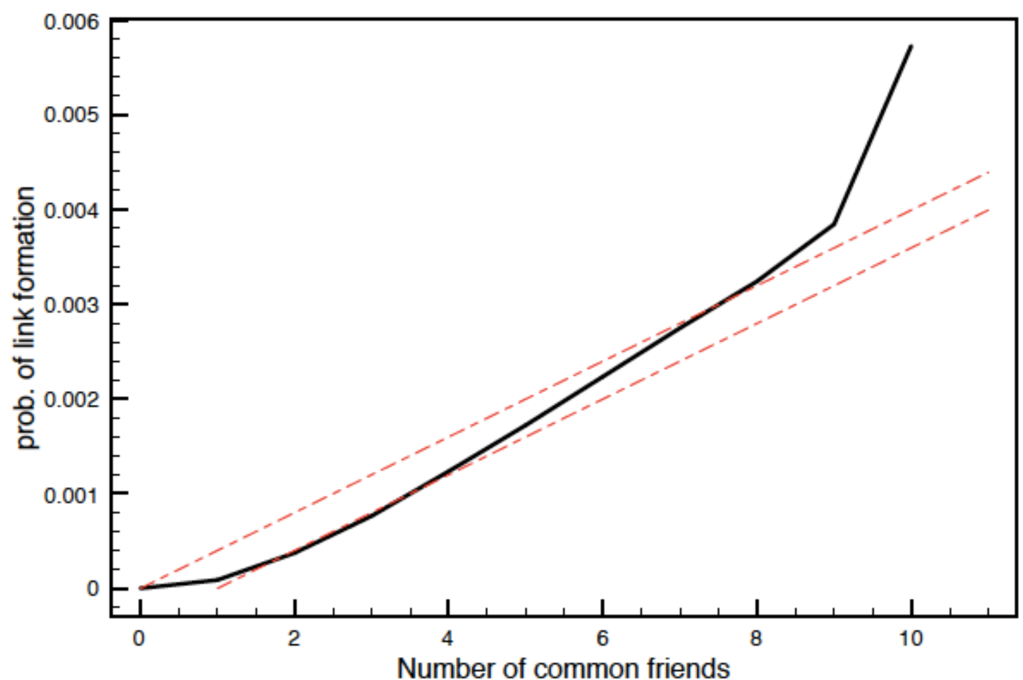


- Empirical Study:
  - Take two snapshots of the network at different points in time.
  - For each k, identify all pairs of nodes that have exactly k friends in common at the first snapshot but are not connected by an edge.
  - Let T(k) be the fraction of pairs that have formed an edge in the second snapshot.

Tracking Link Formation in Data

- Mathematical baseline:
  - Suppose that each common friend ads an (independent) probability p of forming a link on a given day.
  - If two people have k friends in common, then the probability that they do not form a link on a given day is $(1-p)^k$
  - Thus, our baseline probability that it does form is:

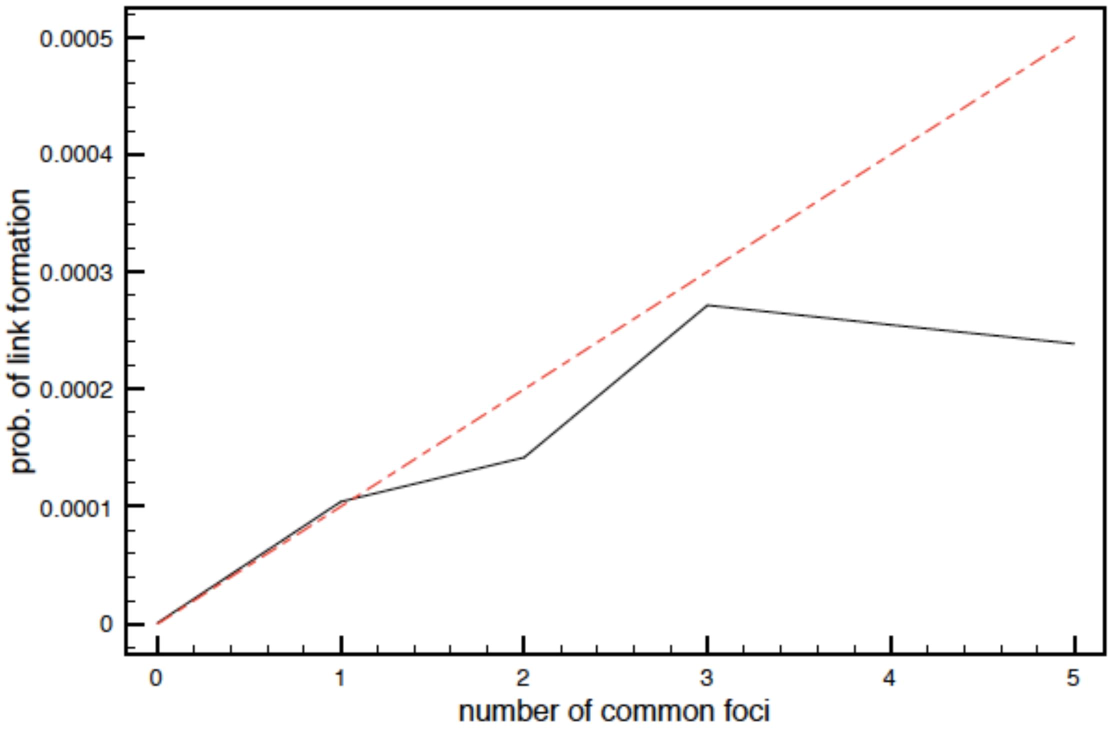
$$T_b(k) = 1 - (1-p)^k$$



- Compare data against $T_b(k)$ and $T_b(k-1)$.

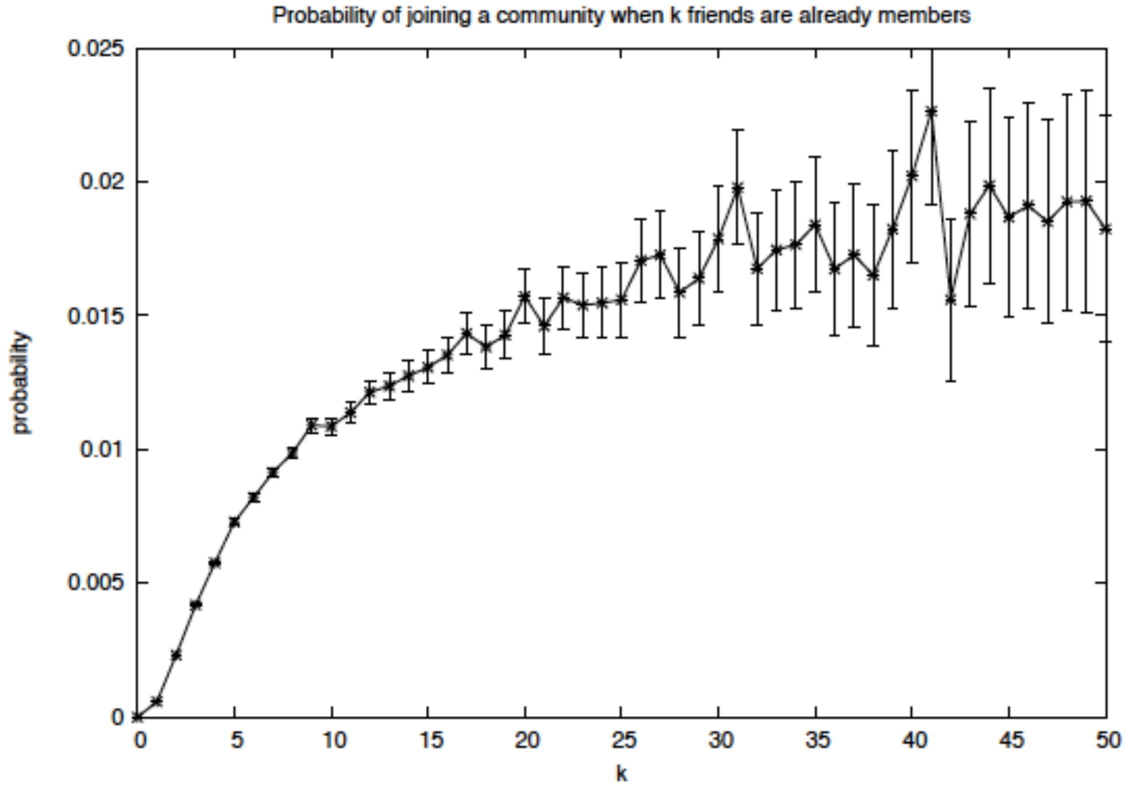
Tracking Link Formation in Data

- *How much more likely is a link to form between two people in a social-affiliation network if they have a focus in common?*
- *How much more likely is a link to form between two people in a social-affiliation network if have **multiple** foci in common?*



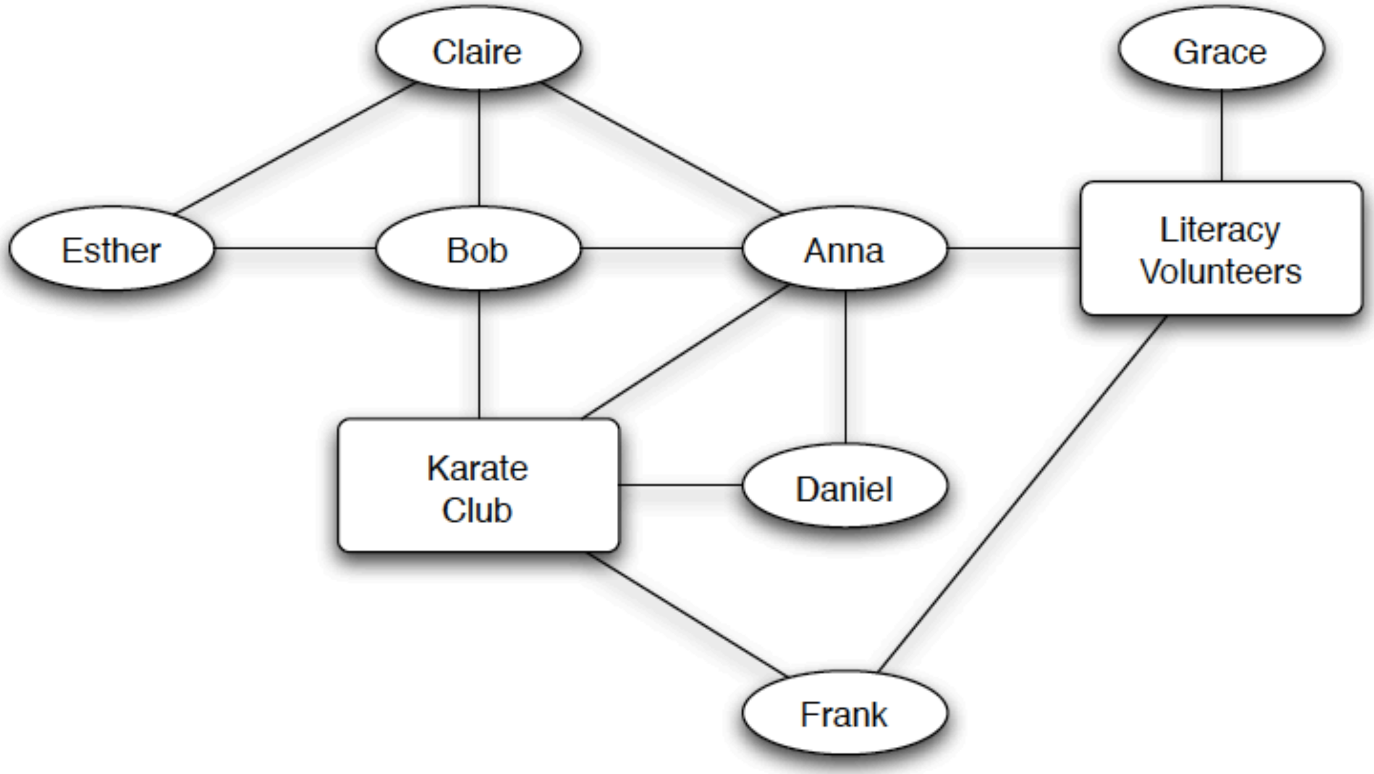
Tracking Link Formation in Data

- *How much more likely is a link to form between a person and a focus if they have a friend that has that focus?*
- *How much more likely is a link to form between a person and a focus if have **multiple** friends that have that focus?*



Tracking Link Formation in Data

- *Which effect is taking place if Bob and Daniel form an edge?*



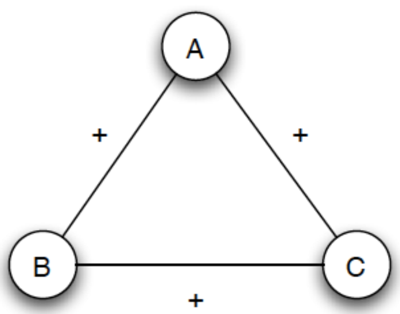
Mini-Exercise

- Consider an "inferred" social network created from a (bipartite) affiliation network where there is an edge between two nodes in the inferred social network if they are neighbors of the same focus in the affiliation network.
  - Can the same inferred social network arise from different affiliation networks?

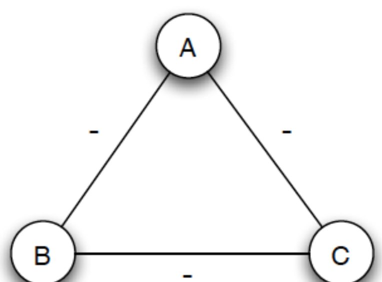
Signed Networks

Signed Networks

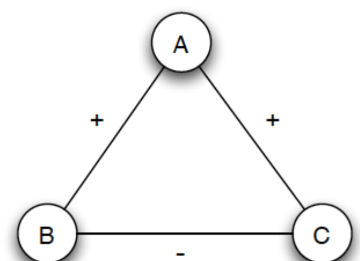
- Edges have sentiment:
  - In the simplest form, all edges are labeled either + or –
- A triangle in the network is either *balanced* or *unbalanced*.
  - A triangle is balanced if it has either 1 or 3 positive edges.



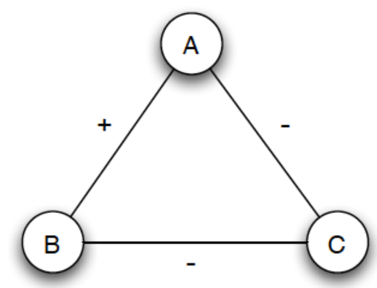
(a) *A, B, and C are mutual friends: balanced.*



(d) *A, B, and C are mutual enemies: not balanced.*



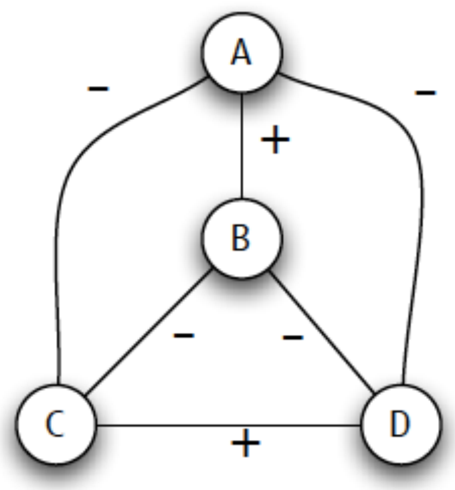
(b) *A is friends with B and C, but they don't get along with each other: not balanced.*

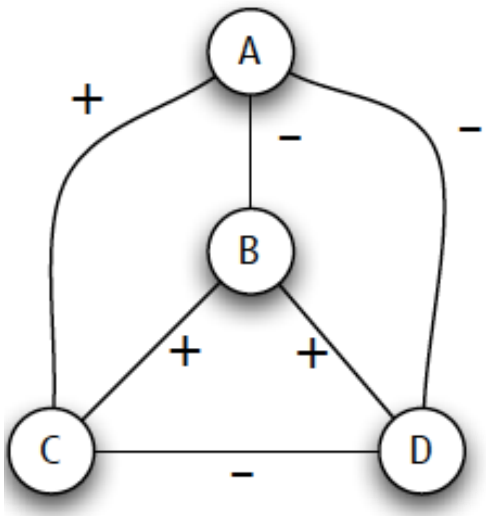


(c) *A and B are friends with C as a mutual enemy: balanced.*

Structural Balance

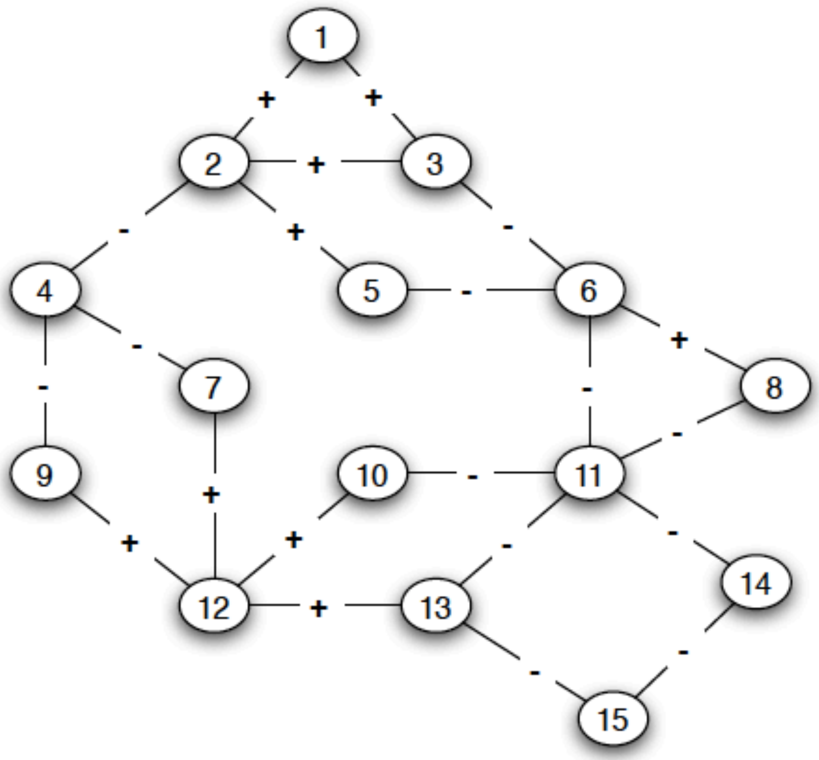
- A complete network is balanced if each of its complete triangles are balanced.





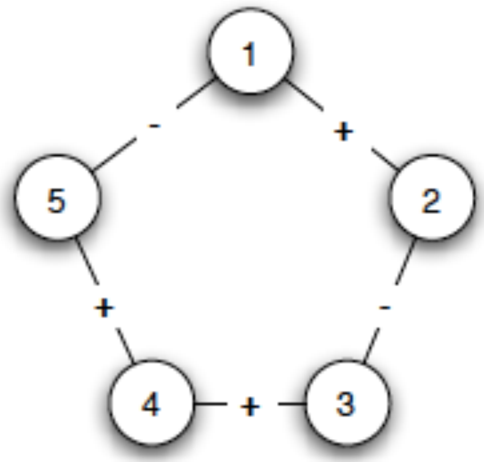
Structural Balance

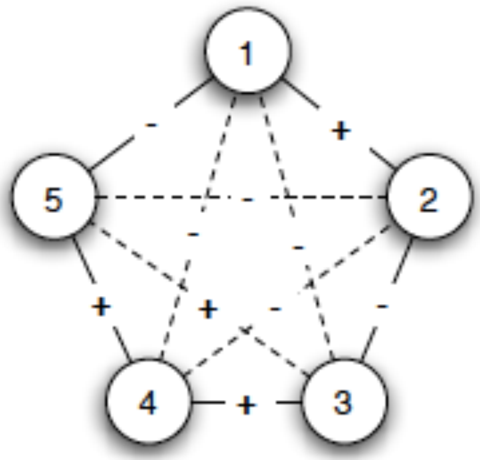
- What about non-complete graphs?



Structural Balance

- A non-complete graph is balanced if *all missing edges can be added* in a way such that the resulting network is balanced.



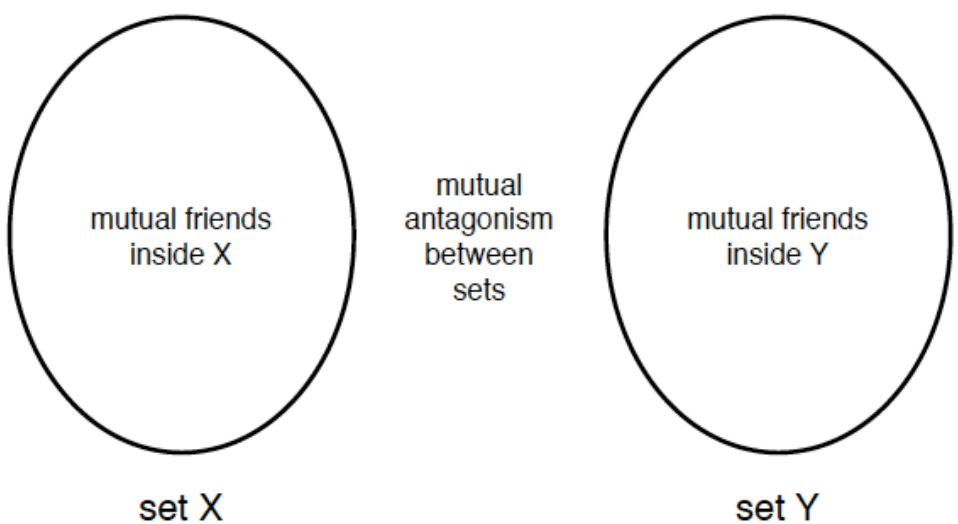


Structural Balance

- Given an arbitrary (unsigned) network, can one always label the edges such as the resulting signed network is balanced?
  - Yes! (e.g., all positive edges).
- Given a signed network, can we give a global characterization of balanced vs unbalanced networks?

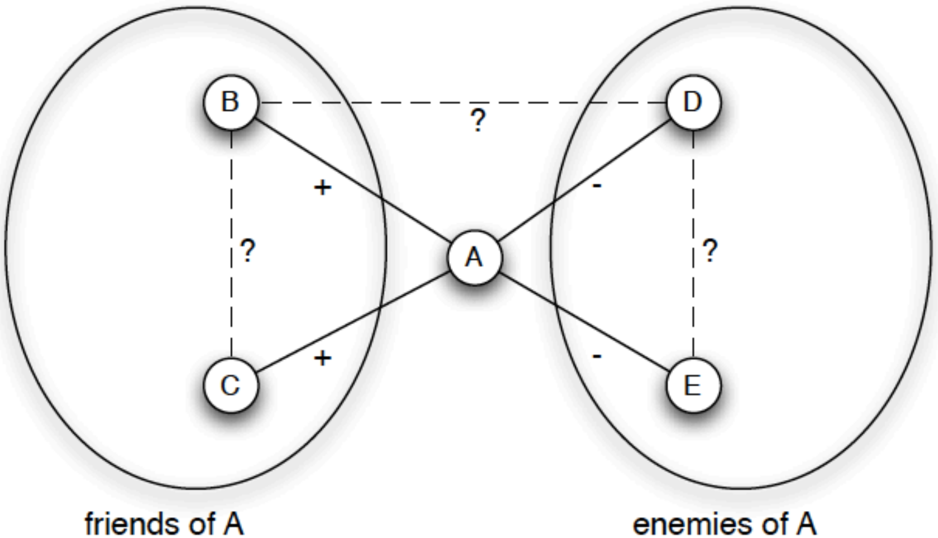
Structural Balance

- Theorem: A complete graph is balanced if and only if it can be decomposed into sets X and Y as below.
  - Assume we have a complete graph with this structure.
    - Check that any triangle has either 3 positive edges or exactly 1 positive edge.
  - In fact true even without completeness assumption!

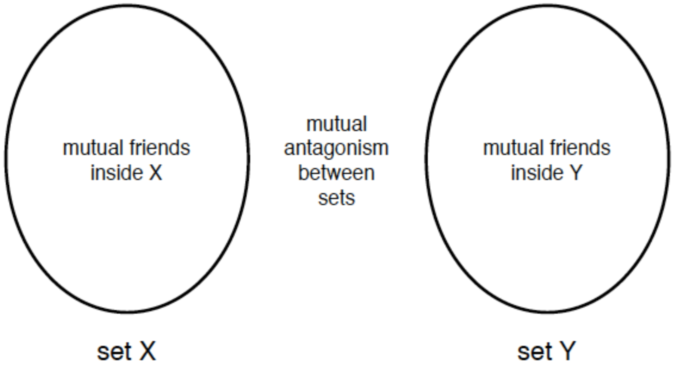


Structural Balance

- Theorem: A complete graph is balanced if and only if it can be decomposed into sets X and Y as below.
  - Assume we have a balanced complete graph.
    - If no negative edge, put all nodes in set X!
    - Otherwise, pick a vertex A and let X be the set A with all of its friends, and let Y be the set of all of its enemies.
      - Two nodes in X must be friends
      - Two nodes in Y must be friends
      - A node in X and a node in Y must be enemies.
  - True without the completeness assumption?



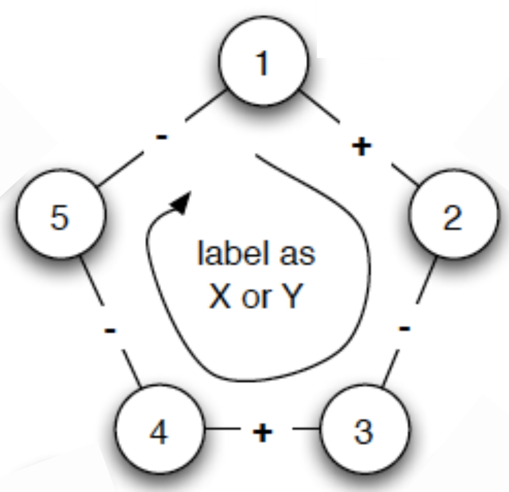
Structural Balance



- Theorem: Any graph is balanced if and only if it can be decomposed into sets X and Y as above.
- Lemma: If a network has a negative cycle of odd length, then it cannot be decomposed into X and Y.
  - Corollary: If it cannot be decomposed into X an Y then it is not balanced (adding edges won't help!).
- Suffices to show: A graph is balanced if and only if it contains no cycle with an odd number of negative edges.
  - We will conduct a process that always either
    - Results in an X/Y decomposition (which implied balancedness from the complete graph theorem), or
    - Results in an odd-length negative cycle (which implies non-balancedness from the above Corollary.

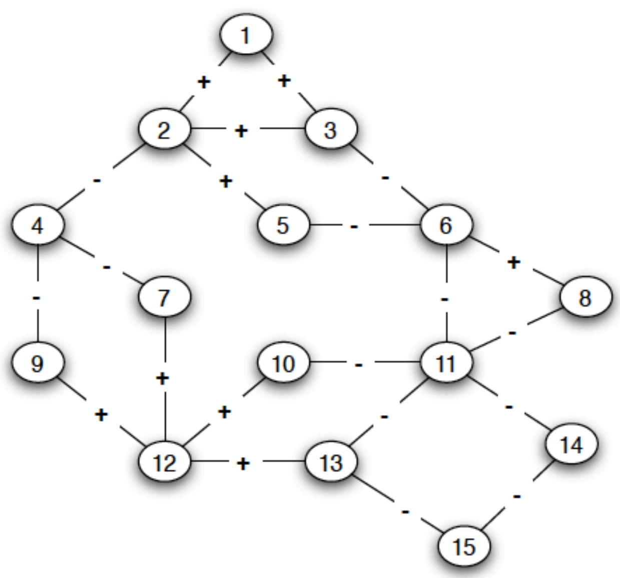
Structural Balance

- Lemma: If a network has a negative cycle of odd length, then it cannot be decomposed into X and Y.
  - Corollary: If it cannot be decomposed into an X and Y then it is not balanced (adding edges won't help!).



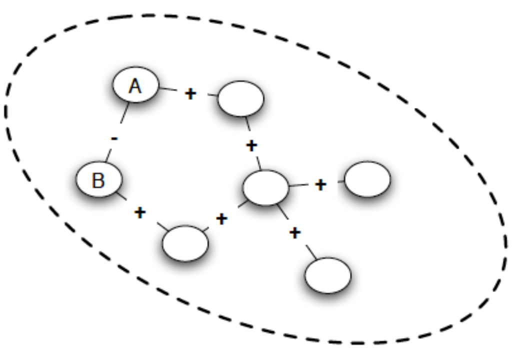
Structural Balance

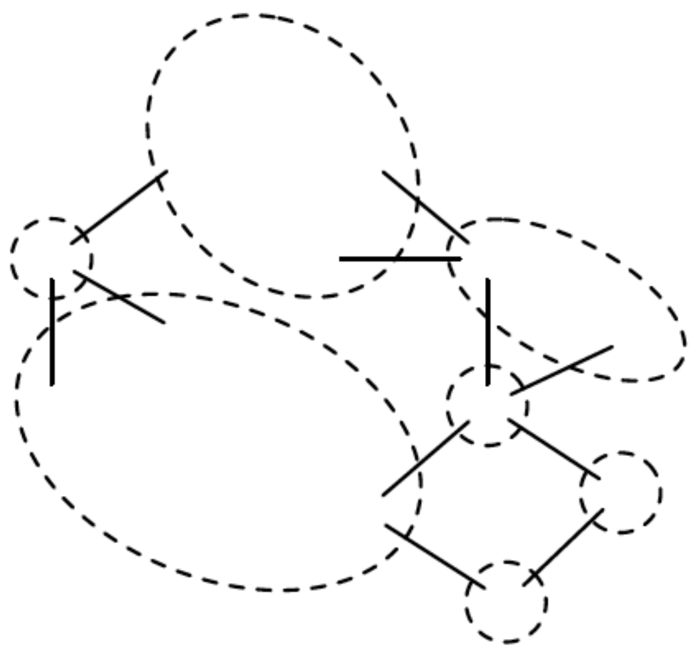
- Lemma: A graph is balanced if and only if it contains no cycle with an odd number of negative edges.
  - We will conduct a process that either
    - Results in an X/Y decomposition (which implied balancedness from the complete graph theorem), or
    - Results in an odd-length negative cycle (which implies non-balancedness from the above Corollary.



Structural Balance

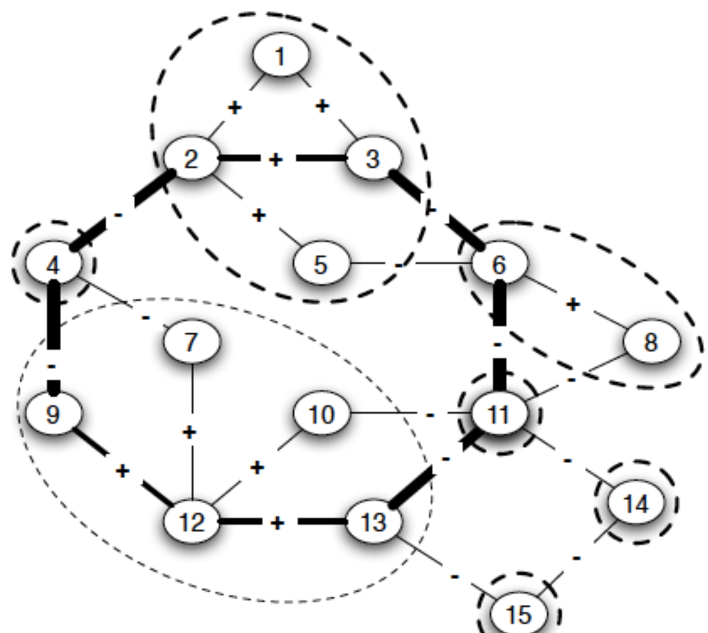
- Lemma: A graph is balanced if and only if it contains no cycle with an odd number of negative edges.
  - We will conduct a process that either
    - Results in an X/Y decomposition (which implied balancedness from the complete graph theorem), or
    - Results in an odd-length negative cycle (which implies non-balancedness from the above Corollary).





Structural Balance

- Lemma: A graph is balanced if and only if it contains no cycle with an odd number of negative edges.
  - We will conduct a process that either
    - Results in an X/Y decomposition (which implied balancedness from the complete graph theorem), or
    - Results in an odd-length negative cycle (which implies non-balancedness from the above Corollary).



Structural Balance

- Theorem: Any graph is balanced if and only if it can be decomposed into sets X and Y as below.