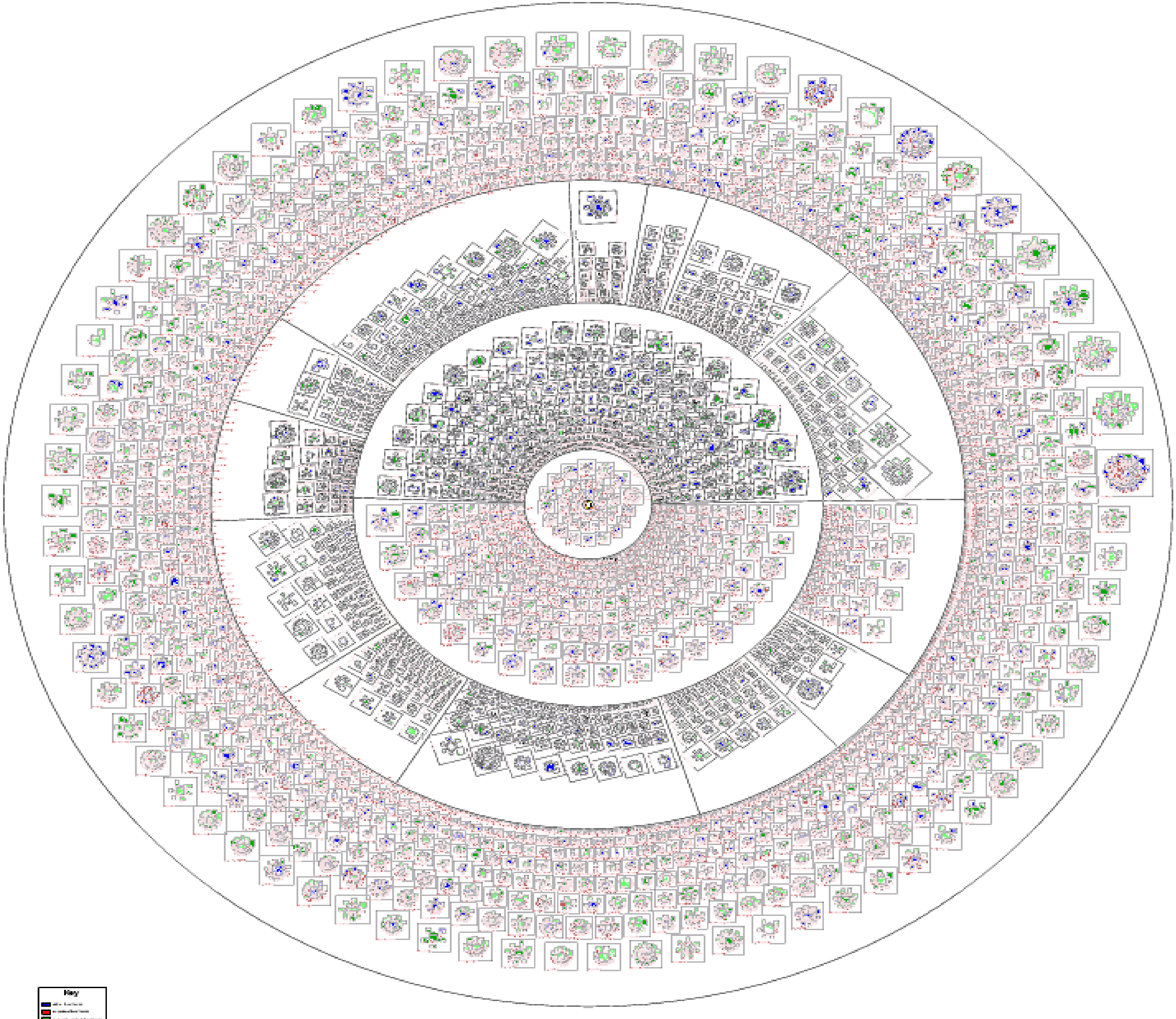
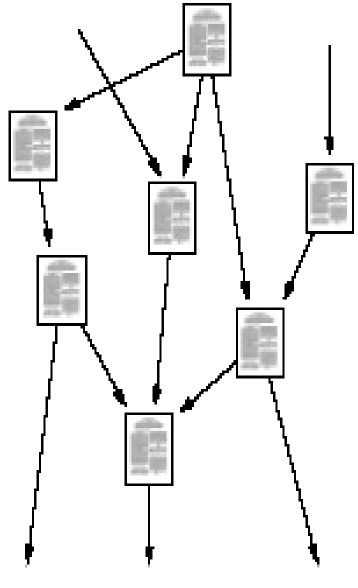
Network Analysis


Key

Degree Distribution: The case of Citation Networks

- **Papers (in almost all fields) refer to works done earlier on same/related topics – *Citations***
- **A network can be defined as**
  - **Each node is a paper**
  - **A directed edge from paper A to paper B indicates A cites B**

- **These networks are acyclic**
- **Edges point backward in time!**



citation network

Law of Scientific Productivity

- **Alfred Lotka (1926) did some analysis of such a citation network and made a statement**
  - ***the number of scientists who have k citations falls off as $k^{-\alpha}$ for some constant α.***
- **Considering each node in the citation network to be representative of scientists can you say what exactly did Lotka study???**

The distribution of the degree of the nodes !!!

Degree Distribution: Formal Definition

- **Let p_k be the fraction of vertices in the network that has a degree k**

- **Hence p_k is the probability that a vertex chosen uniformly at random has a degree k**

- **The k versus p_k plot is defined as the degree distribution of a network**

- **For most of the real world networks these distributions are right skewed with a long right tail showing up values far above the mean – p_k varies as $k^{-\alpha}$**

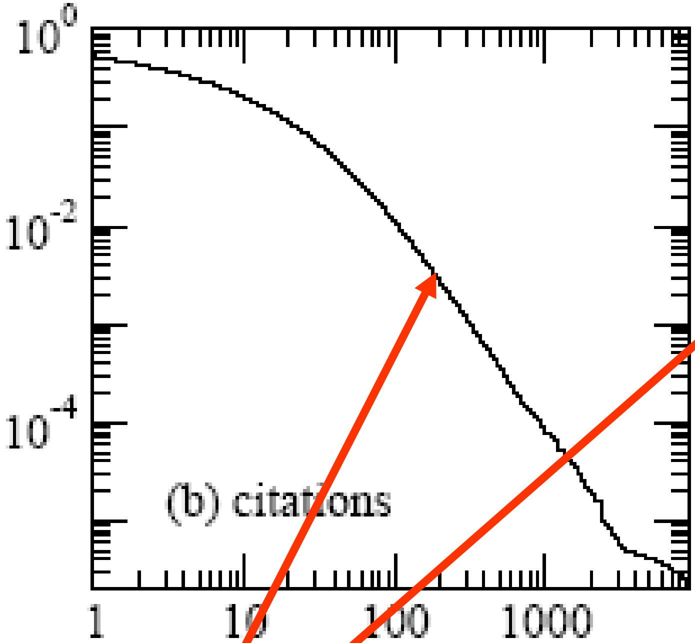
The Definition Slightly Modified

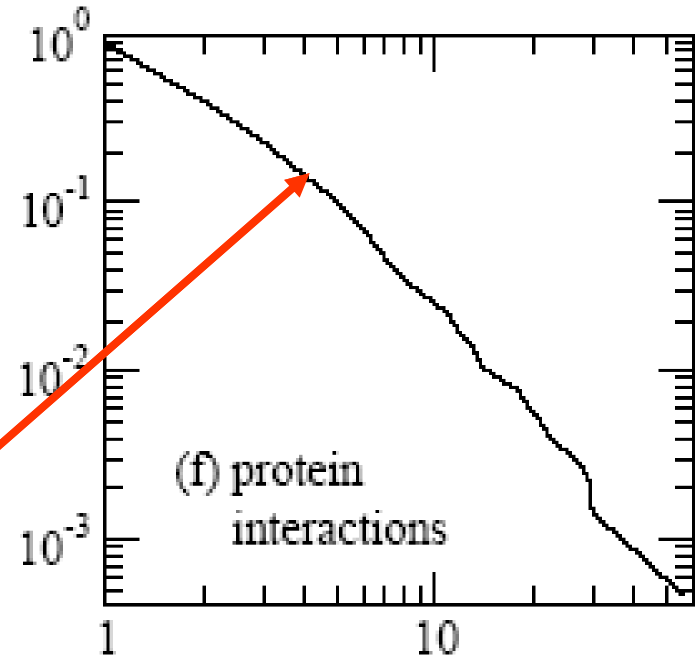
- **Due to noisy and insufficient data sometimes the definition is slightly modified**
  - **Cumulative degree distribution is plotted**

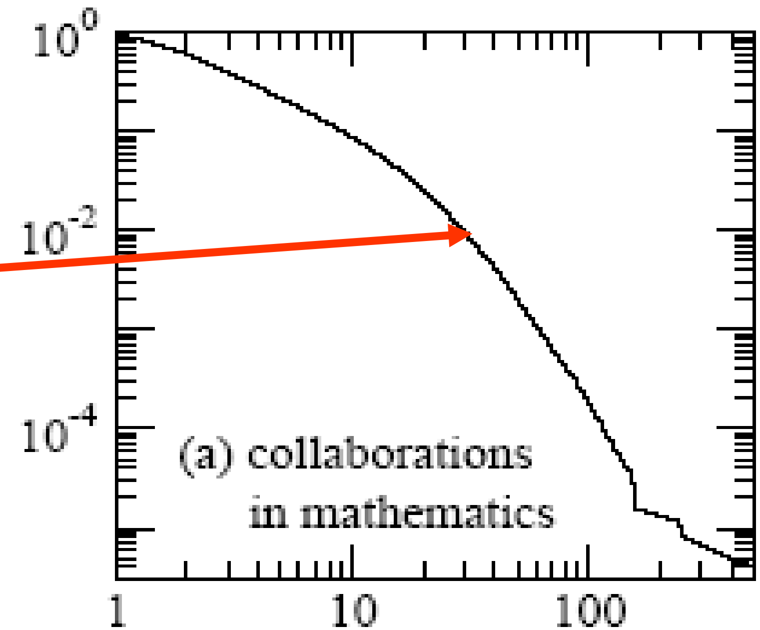
$$P_k = \sum_{k'=k}^{\infty} p_{k'},$$

- **Probability that the degree of a node is greater than or equal to k**

A Few Ex







Power law: $P_k \sim k^{-\alpha}$

Scale-free

For any function $f(x)$

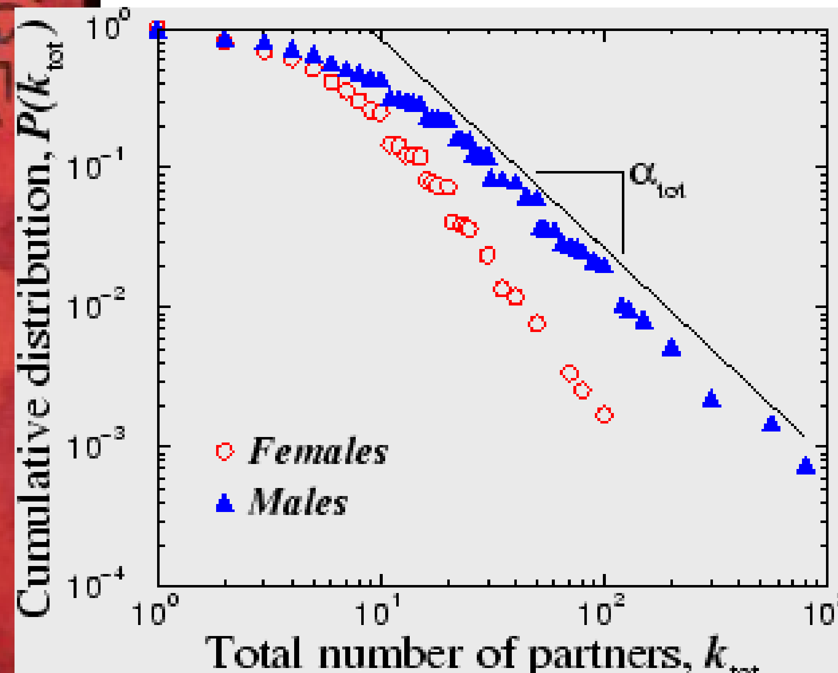
the independent variable when rescaled $f(ax)$

does not affect the functional form $bf(x)$

Power-laws – are they scale-free???

Swedish sex-web

Nodes: people (Females; Males)
Links: sexual relationships

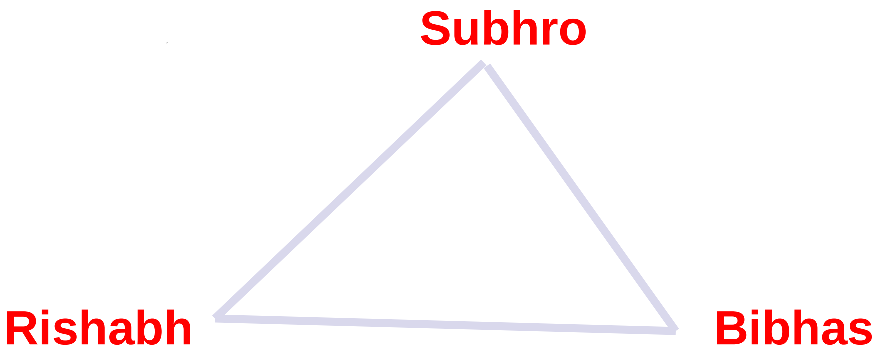


4781 Swedes; 18-74;
59% response rate.

Liljeros et al. *Nature* 2001

Friend of Friends are Friends

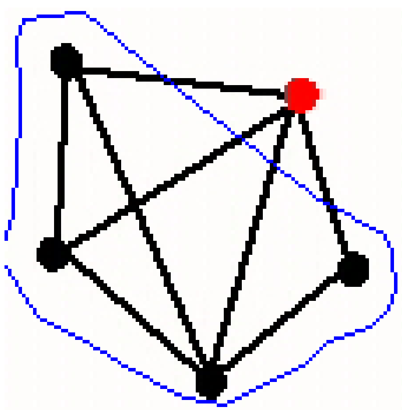
- **Consider the following scenario**
  - **Subhro and Rishabh are friends**
  - **Rishabh and Bibhas are friends**
  - **Are Subhro and Bibhas friends?**
  - **If so then …**

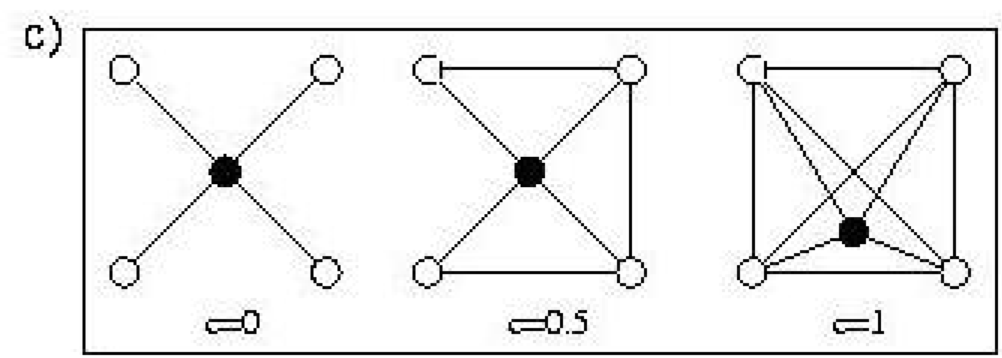


- **This property is known as transitivity**

Measuring Transitivity: Clustering Coefficient

- **The clustering coefficient for a vertex ‘*v*’ in a network is defined as the ratio between the total number of connections among the neighbors of ‘*v*’ to the total number of possible connections between the neighbors**





- **The philosophy – High clustering coefficient means my friends know each other with high probability – a typical property of social networks**

Mathematically…

- The clustering index of a vertex *i* is

$$C_i = \frac{\text{\# of links between neighbors}}{n(n-1)/2}$$

- The clustering index of the whole network is the average

$$C = \frac{1}{N} \sum C_i$$

Network	C	C_{rand}	L	N
WWW	0.1078	0.00023	3.1	153127
Internet	0.18-0.3	0.001	3.7-3.76	3015-6209
Actor	0.79	0.00027	3.65	225226
Coauthorship	0.43	0.00018	5.9	52909
Metabolic	0.32	0.026	2.9	282
Foodweb	0.22	0.06	2.43	134
C. elegance	0.28	0.05	2.65	282

The World is Small!

- All late registrants in the Complex Networks course shall get 10 marks bonus!!!!!
- How long do you think the above information will take to spread among yourselves
- Experiments say it will spread very fast – within 6 hops from the initiator it would reach all
- This is the famous Milgram's six degrees of separation

Milgram’s Experiment

- Travers & Milgram 1969: classic study in early social science
  - **Source: Kharagpur stockbrokers**
  - **Destination: A Kolkata stockbroker (Kharagpur & Kolkata are “randoms”)**
  - **Job: Forward a letter to a friend “closer” to the target**
  - **Target information provided:**
    - **name, address, occupation, firm, college, wife’s name and hometown**

Findings

- Most of the letters in this experiment were lost…
- Nevertheless a quarter reached the target
- Strikingly those that reached the target passed through the hands of six people on an average
- In fact

- 64 of 296 chains reached the target

- average length of *completed* chains: 5.2
- Is our class a small-world???

Centrality

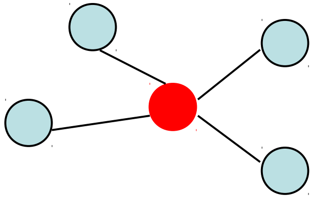
- Centrality measures are commonly described as indices of
  - prestige,
  - prominence,
  - importance,
  - and power -- the four Ps
- A measure indicating the importance of a vertex

Degree Centrality

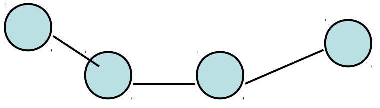
Degree Centrality – Immediate neighbors of a vertex (k) expressed as a fraction of the total number of neighbors possible

Variance of degree centrality – Centralization

Star network – an ideal centralized one



Line network – less centralized

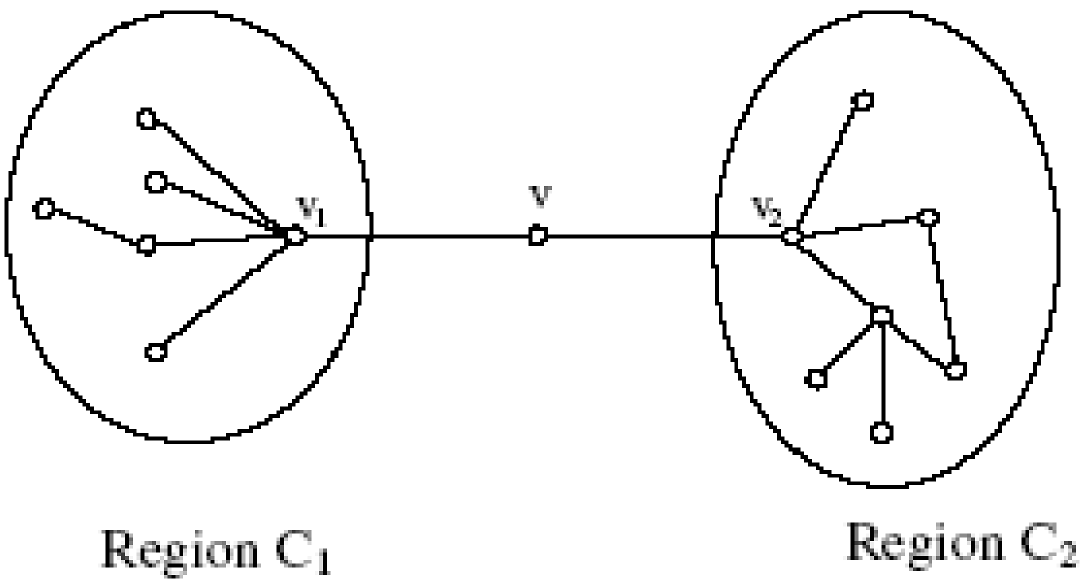


Betweenness Centrality

- Tries to determine how important is a node in a network
- Degree of a node doesn’t only determine its importance in the network – do you agree???

Betweenness Centrality

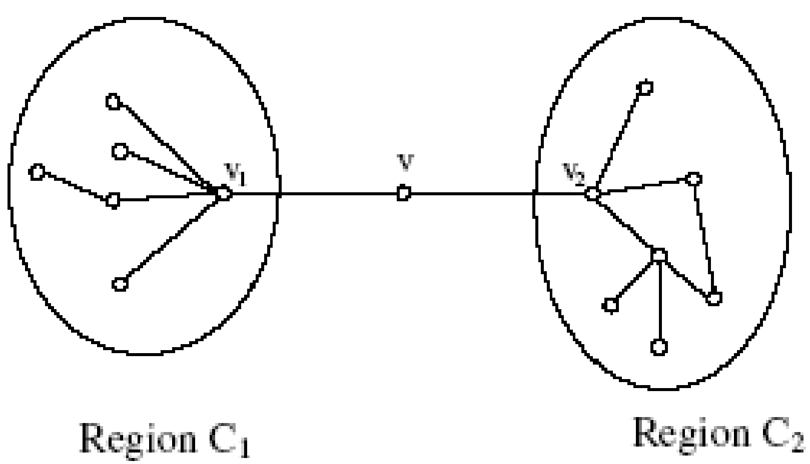
- Tries to determine how important is a node in a network
- Degree of a node doesn't only determine its importance in the network – do you agree???
- The node can be on a *bridge* centrally between two regions of the network!!



Betweenness Centrality

- Centrality of *v*: Geodesic path between *s* and *t* via *v* expressed as a fraction of total number of geodesic paths between *s* and *t*

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



$$g(v) = 2 \sum_{s \in C_1, t \in C_2} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$$= 2 \sum_{s \in C_1, t \in C_2} 1$$

$$= 2N_1N_2$$

Betweenness Centrality

- Removal – what can this lead to??

Betweenness Centrality

- Removal – what can this lead to??
- Increase in the geodesic path – extreme case is infinity (network gets disconnected)
- Can you visualize the impact of removal of the nodes with high betweenness in the following networks??
  - Epidemic network
  - Information network
  - Traffic network

Flow Betweenness

- What if the nodes with high betweenness behave as reluctant brokers and do not allow two other nodes (of different regions) to establish a relationship.
- They must find other ways to establish relationship (may not be cost effective)
  - Something like "wanting to propose someone via a third party (say his/her friends) who is also (kind of) your friend – but this common friend is reluctant to pursue the proposal!"
- This is the main idea of flow betweenness
- Takes into account all paths (not only the shortest ones) from s to t via v – computationally quite intractable for large networks.

Eigenvector Centrality (Bonacich 1972)

- In context of HIV transmission – A person x with one sex partner is less prone to the disease than a person y with multiple partners

Eigenvector Centrality (Bonacich 1972)

- In context of HIV transmission – A person x with one sex partner is less prone to the disease than a person y with multiple partners
- But imagine what happens if the partner of x has multiple partners
- It is not just how many people knows me counts to my popularity (or power) but how many people knows people who knows me – this is recursive!
- The basic idea of eigenvector centrality

Eigenvector Centrality

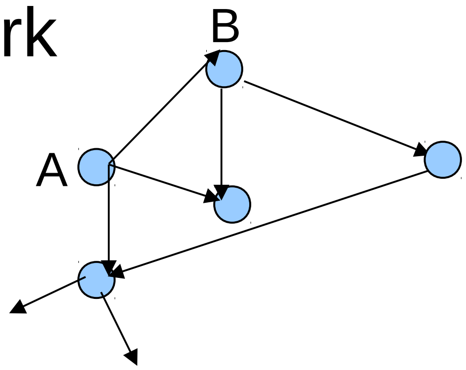
- Idea is to define centrality of vertex as sum of centralities of neighbors.
- Suppose we guess initially vertex i has centrality $x_i(0)$
- Improvement is $x_i(1) = \sum_j A_{ij}\, x_j(0)$
- Continue until there is no more improvement observed
- So, $x(t) = \mathbf{A}x(t\text{-}1) \Rightarrow x(t) = \mathbf{A}^t x(0)$ [Power iteration method proposed by Hotelling]

Eigenvector Centrality

- Express $x(0)$ as linear combination of eigenvectors v_i of adjacency matrix A
- $x(0) = \sum_i c_i v_i \Rightarrow x(t) = \mathbf{A}^t \sum_i c_i v_i \Rightarrow x(t) = \sum_i \lambda_i^t c_i v_i$
- *Or,* $x(t) = \lambda_1^t \sum_i (\lambda_i/\lambda_1)^t c_i v_i$
- In the limit of large number of iterations,
- $Lt_{t \to \infty} (1/\lambda_1^t)\, x(t) = c_1 v_1$
- Limiting centrality should be proportional to leading eigenvector v_1

Eigenvector centrality for directed networks

- Can be recast for directed networks (e.g., the link structure of the Web)
- Problem of *zero* centrality in directed network
  - A has centrality 0 as there are no incoming edges (seems reasonable for web page)
  - But B has one incoming edge from A; centrality of B is 0 because A has centrality 0
  - centralities all 0 in acyclic network



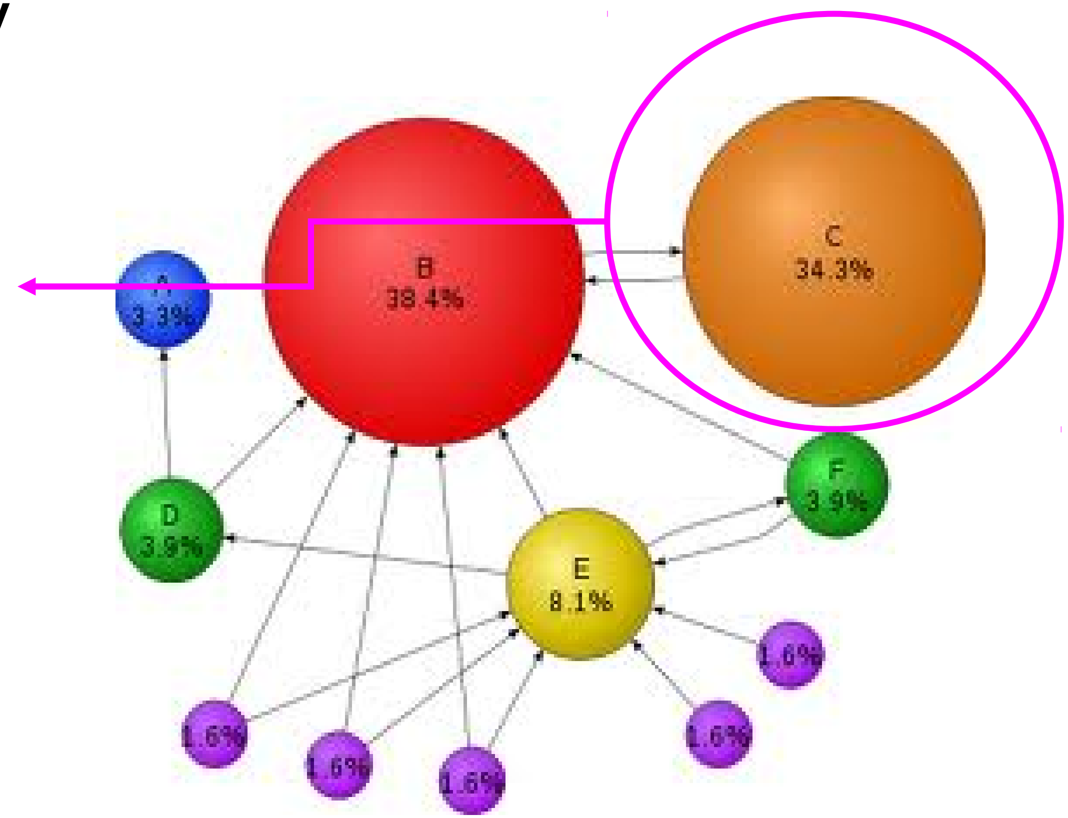
Katz Centrality

- Give every node small amount of centrality for free $\alpha, \beta > 0$
- $x_i = \alpha \sum_j A_{ij} x_j + \beta$
- Avoids problem of zero centrality
- In matrix terms, $\mathbf{x} = \alpha \mathbf{A}\mathbf{x} + \beta \mathbf{1}$ where $\mathbf{1} = (1,1,\ldots,1)^T$
- So, $\mathbf{x} = \beta (\mathbf{I} - \alpha \mathbf{A})^{-1} \mathbf{1}$
- Katz centrality: set $\beta = 1$ => $\mathbf{x} = (\mathbf{I} - \alpha \mathbf{A})^{-1} \mathbf{1}$
- Compute Katz centrality by iterating $x(t) = \alpha \mathbf{A} x(t\text{-}1) + \beta$
- => avoid inverting the matrix directly

PageRank

- Link analysis algorithm → Assigns *link popularity*
- Named after Larry Page
- Google trademark
- Variant of Katz similarity

C has less links than E but more popularity (derived from the popularity of B due to the in-link)

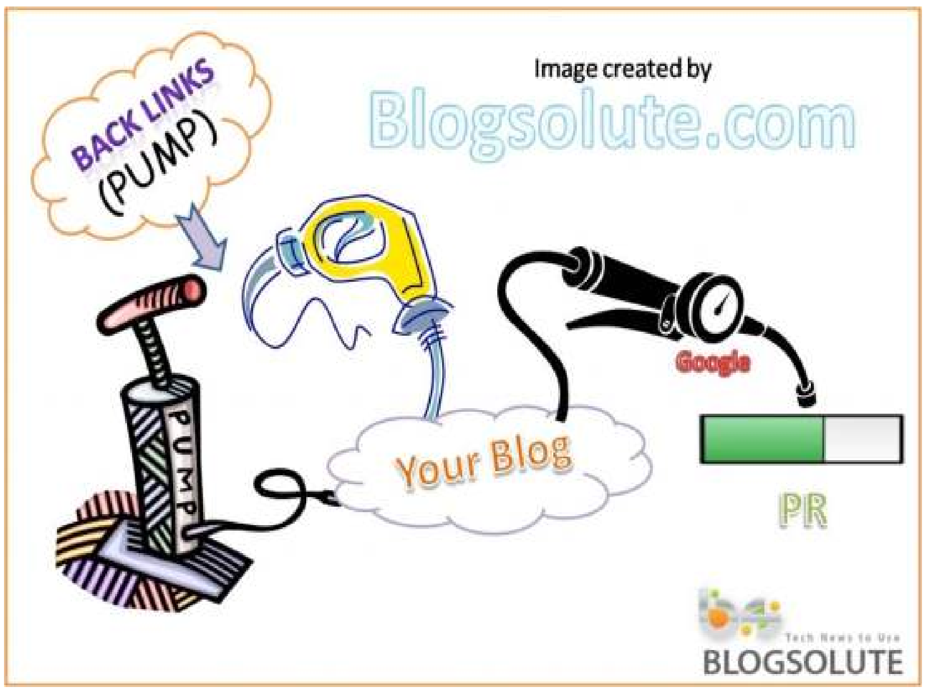


PageRank

- How can you make yourself popular??

PageRank

- How can you make yourself popular??



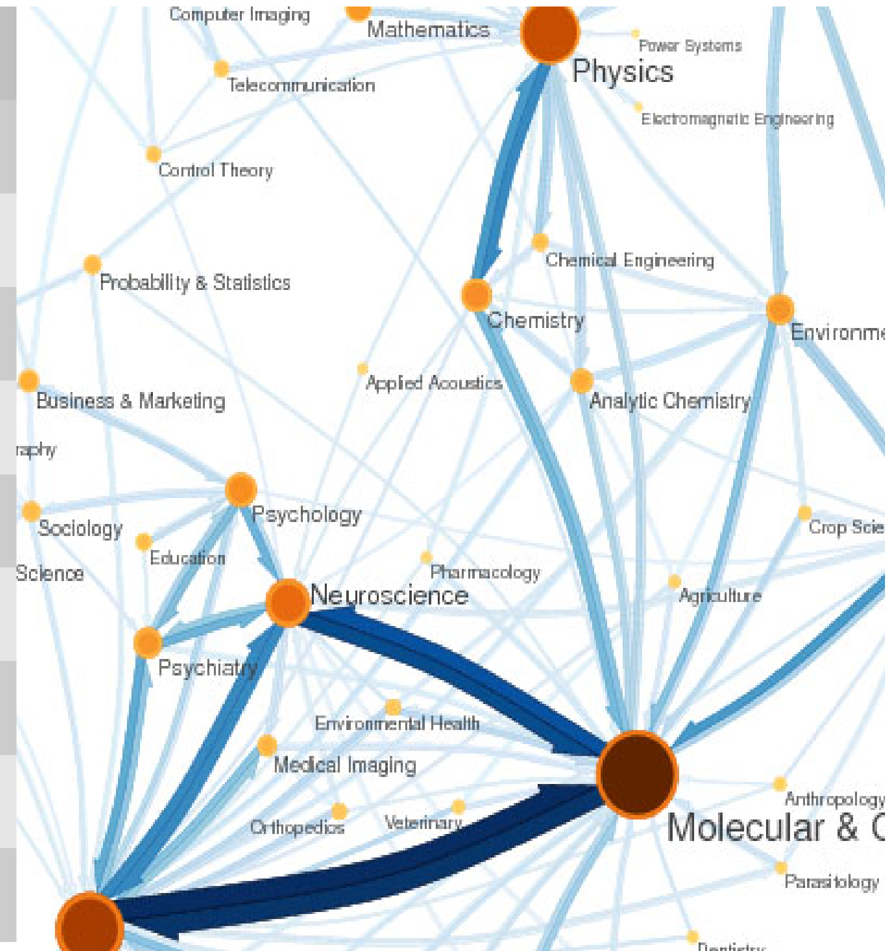
PageRank: Calculation

- Variant of Katz similarity
- $x_i = \alpha \sum_j A_{ij} x_j / k_j^{out} + \beta$
- But if k_j^{out} is 0??
- Easy fix: since vertex with zero out-degree contributes zero to centralities of other vertices, set $k_j^{out} = 1$ in above calculation
- Matrix terms, $\mathbf{x} = \alpha \mathbf{A}\mathbf{D}^{-1}\mathbf{x} + \beta \mathbf{1} \Rightarrow \mathbf{x} = \beta (\mathbf{I} - \alpha \mathbf{A}\mathbf{D}^{-1})^{-1} \mathbf{1}$
- $\mathbf{D}$ is the diagonal matrix such that $D_{ii} = \max(k_j^{out}, 1)$

PageRank

- Google uses $\beta = 1$, $\alpha = 0.85$ (no theory behind this choice)
- Used in measuring impact factor → a measure reflecting the average number of *citations* to articles published in science and social science journals
- Eigenfactor → journals are rated according to the number of incoming citations, highly ranked journals make larger contribution to the eigenfactor than the poorly ranked journals

16.78	Nature
16.39	Journal of Biological Chemistry
16.38	Science
14.49	PNAS ←
8.41	PHYS REV LETT
5.76	Cell
5.70	New England Journal of Medicine
4.67	Journal of the American Chemical Society
4.46	J IMMUNOL
4.28	APPL PHYS LETT



Interpreting web surfing

- Iinitially, every web page chosen uniformly at random
- With probability α, perform random walk on web by randomly choosing hyperlink in page
- With probability 1 - α, stop random walk and restart web surfing
- PageRank $\rightarrow$ steady state probability that a web page is visited through web surfing

Interpreting web surfing

- Iinitially, every web page chosen uniformly at random
- With probability α, perform random walk on web by randomly choosing hyperlink in page ??
- With probability 1 - α, stop random walk and restart web surfing
- PageRank → steady state probability that a web page is visited through web surfing ??

Transition matrix

$$\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{matrix}$$

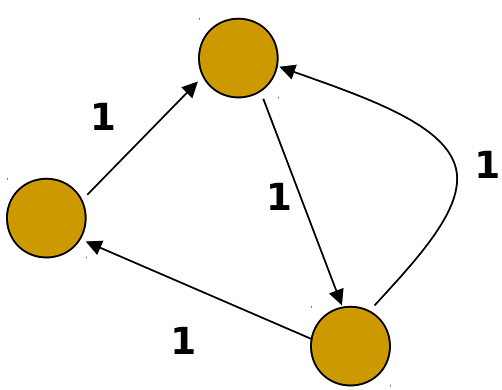
Adjacency matrix A

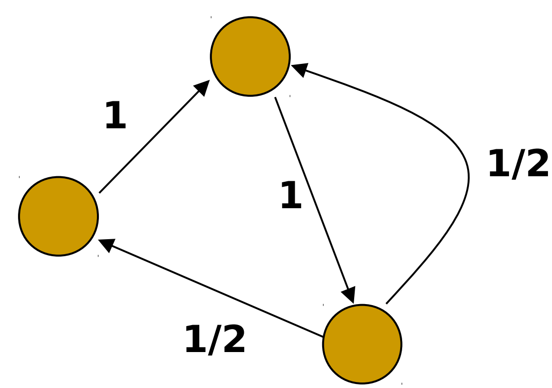
$$\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{matrix}$$

Only probabilities
--> stochastic
matrix

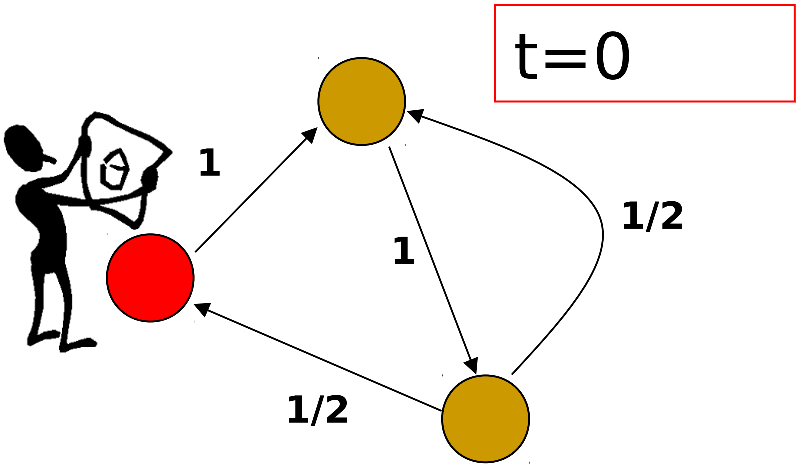
Transition matrix P

$$A_{ij}/\sum_{i} A_{ij}$$

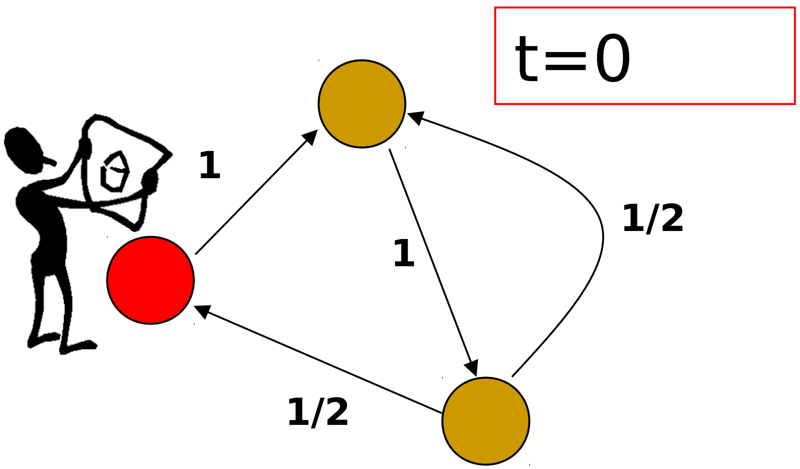


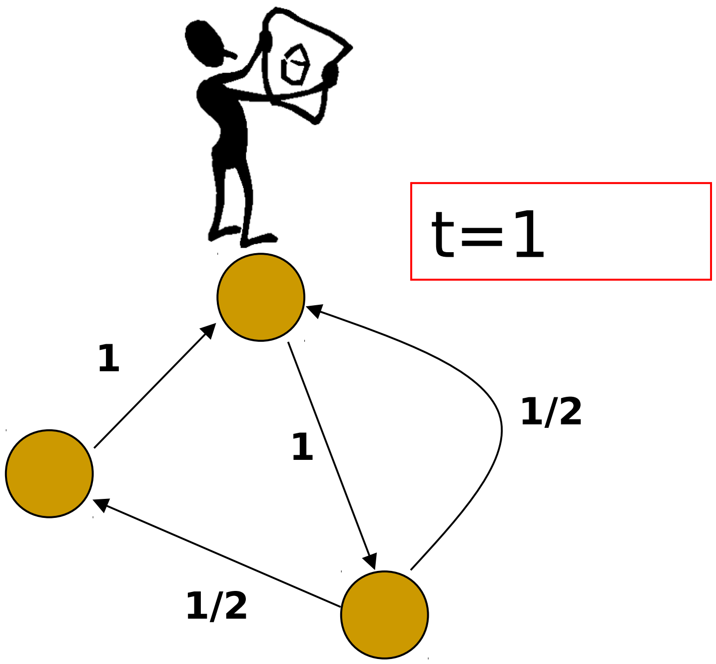


What is a random walk

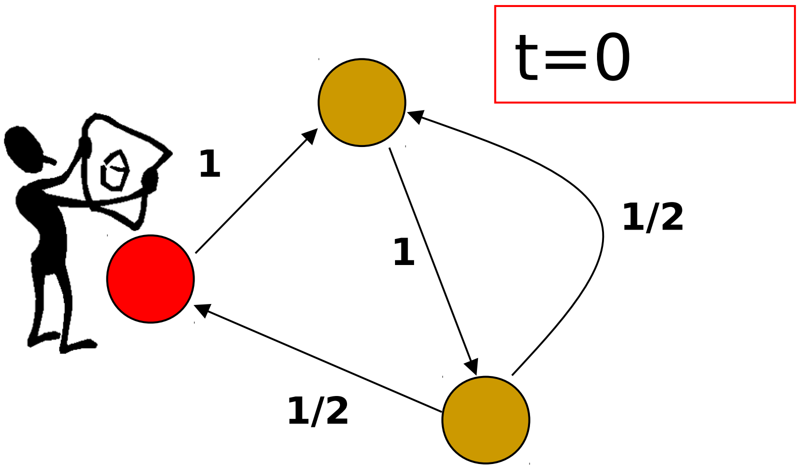


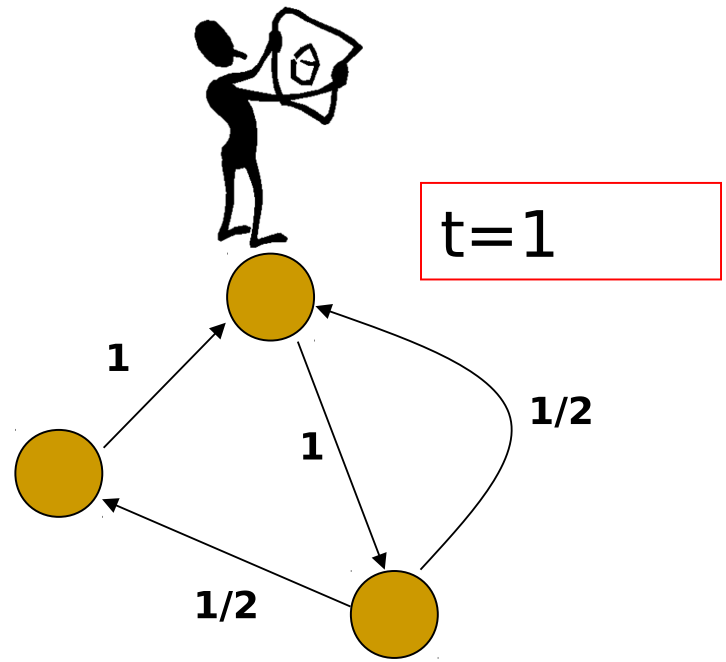
What is a random walk

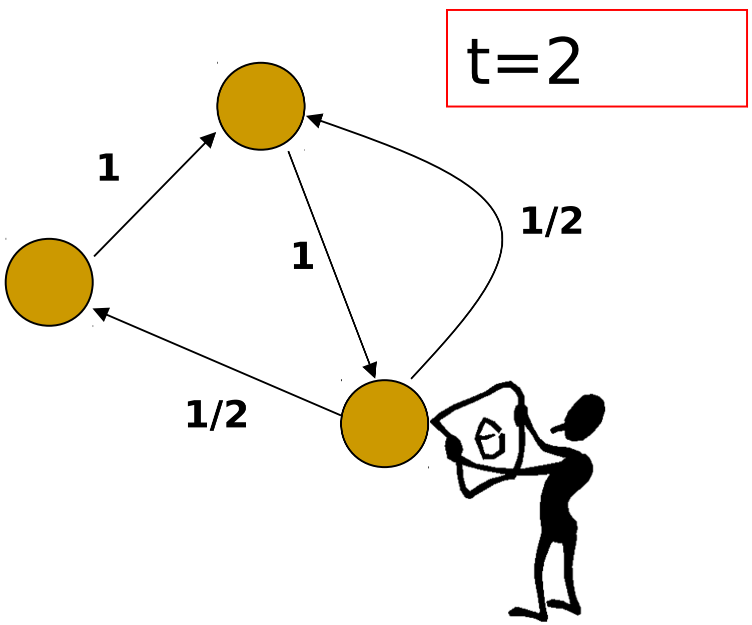




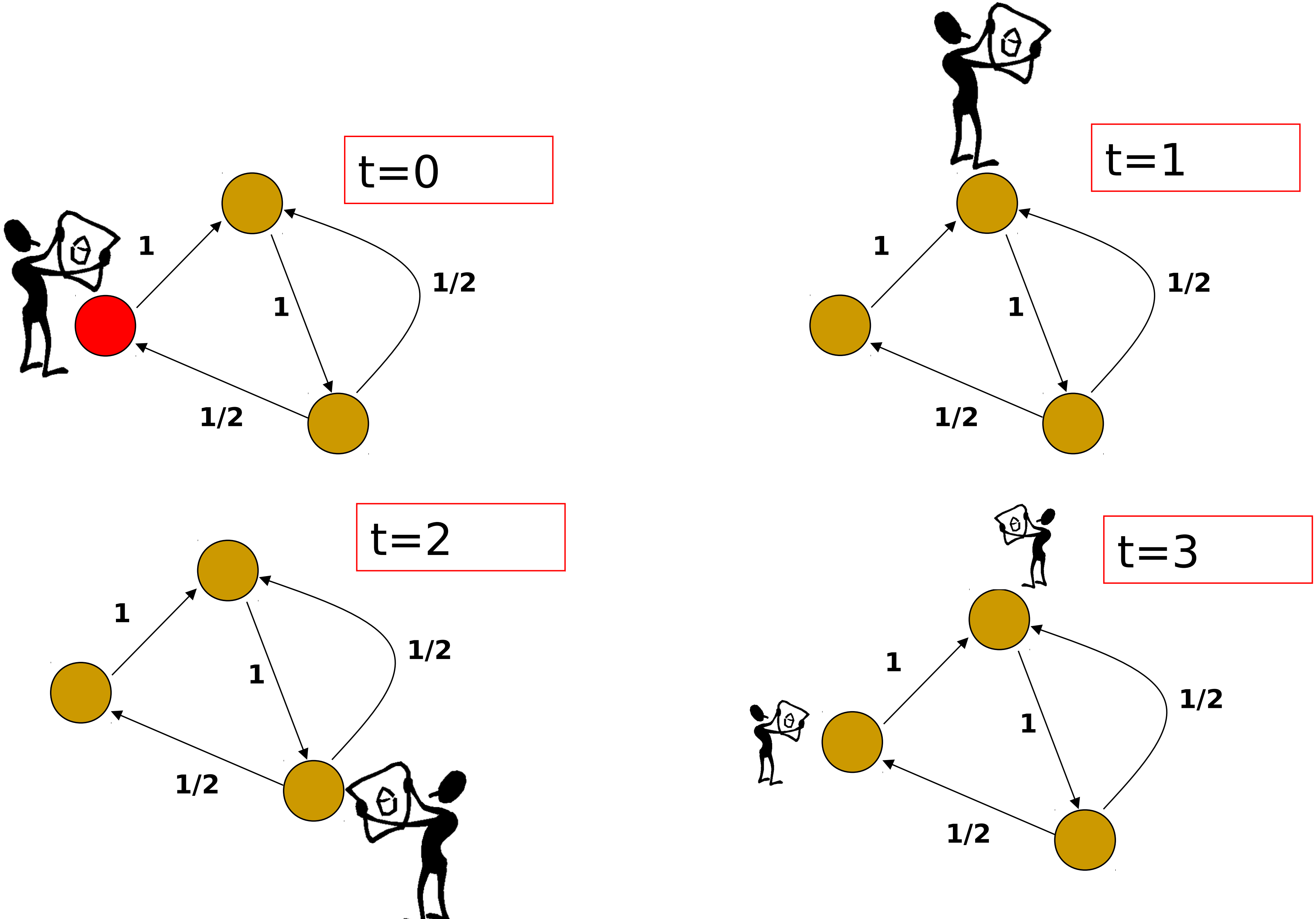
What is a random walk







What is a random walk

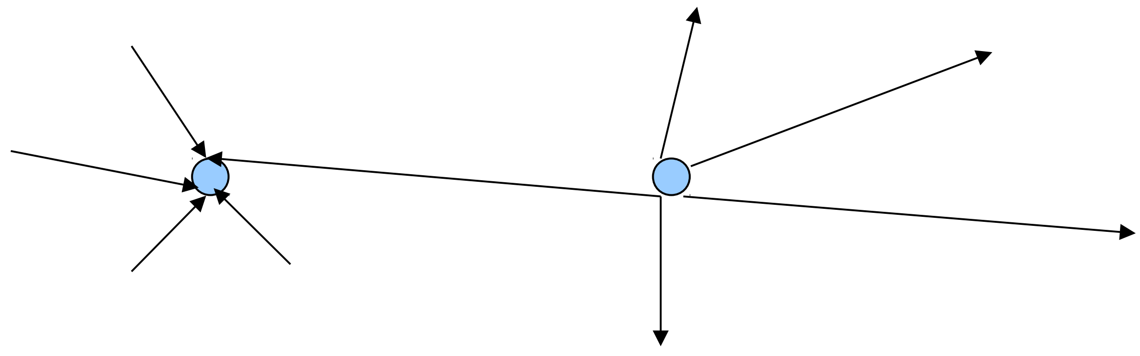


Steady State Calculations

- Set $\beta = 1 - \alpha$ in the PageRank expression
- $\mathbf{x}(t) = \alpha\mathbf{A}\mathbf{D}^{-1}\mathbf{x}(t\text{-}1) + (1\text{-}\alpha)\mathbf{1}$
- Further, $\sum_{i=1\ldots n} x_i(t) = 1$
- So, $\mathbf{x}(t) = \alpha\mathbf{A}\mathbf{D}^{-1}\mathbf{x}(t\text{-}1) + (1\text{-}\alpha)\mathbf{1}\mathbf{1}^{T}\mathbf{x}(t\text{-}1) = \mathbf{P}\mathbf{x}(t\text{-}1)$
- Where, $\mathbf{P} = \alpha\mathbf{A}\mathbf{D}^{-1} + (1\text{-}\alpha)\mathbf{1}\mathbf{1}^{T}$
- $\mathbf{P}^{T}$ is called the probability transition matrix (remember Mark Chain??)
- Steady state probabilities: $Lt_{m \to \infty}(\mathrm{P}^{T})^{m}$

Hubs and Authorities

- Each node has two types of centralities: hub centrality, authority centrality
- authorities: nodes with useful (important) information (e.g., important scientific paper)
- hubs: nodes that tell where best authorities are (e.g., good review paper)
- Hyperlink-induced topic search (HITS) proposed by Kleinberg 1999 in J. ACM



Hubs and Authorities

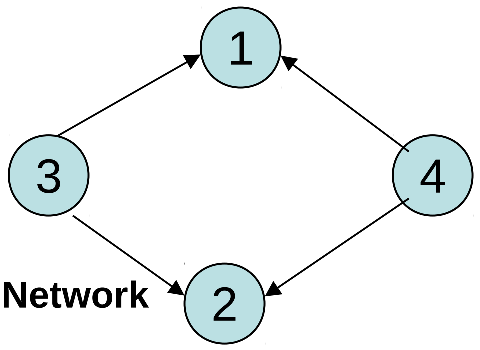
- Authority centrality of node (denoted by x_i) proportional to sum of hub centralities of nodes (denoted by y_j) pointing to it
  - $x_i = \alpha \sum_j A_{ij} y_j$
- Hub centrality of node proportional to sum of authority centralities of nodes pointing to it
  - $y_i = \beta \sum_j A_{ij} x_j$

Hubs and Authorities

- In matrix terms, $\mathbf{x} = \alpha\mathbf{A}^{T}\mathbf{y}$, $\mathbf{y} = \beta\mathbf{A}\mathbf{x}$
- => $\mathbf{x} = \alpha\beta\mathbf{A}^{T}\mathbf{A}\mathbf{x}$ (converges to the principal eigenvector of $\mathbf{A}^{T}\mathbf{A}$)
- => $\mathbf{y} = \alpha\beta\mathbf{A}\mathbf{A}^{T}\mathbf{y}$ (converges to the principal eigenvector of $\mathbf{A}\mathbf{A}^{T}$)
- Assemble the target subset of web pages, form the graph induced by their hyperlinks and compute $\mathbf{A}\mathbf{A}^{T}$ and $\mathbf{A}\mathbf{A}^{T}$.
- Compute the principal eigenvectors of $\mathbf{A}\mathbf{A}^{T}$ and $\mathbf{A}\mathbf{A}^{T}$ to form the vector of hub and authority scores .
- Output the top-scoring hubs authorities.

Co-citation Index

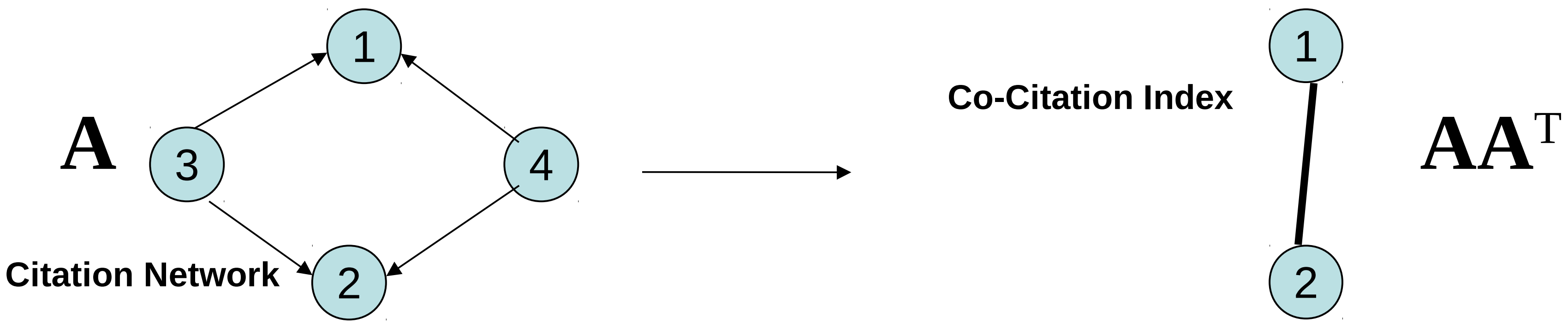
- Consider the following (co-citation)
  - Author 1 is cited by author 3
  - Author 2 is cited by author 3
- Either of 1 or 2 has never cited each other
- Can there be any relationship between author 1 and author 2??



Citation Network

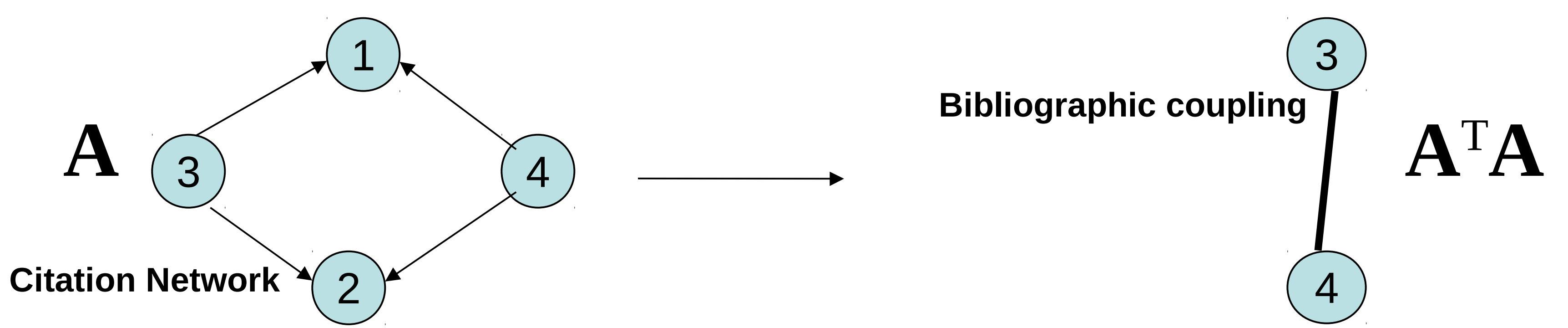
Co-citation Index

- Consider the following (co-citation)
  - Author 1 is cited by author 3
  - Author 2 is cited by author 3
- Either of 1 or 2 has never cited each other
- Can there be any relationship between author 1 and author 2?? Seems to be!! If you are not convinced consider that there are 1000 others like author 3
- There is a high chance that 1 and 2 work in related fields



Bibliographic coupling

- Mirror Image: Consider the following
- Author 3 cites author 1
- Author 4 cites author 1
- Either of 3 or 4 has never cited each other
- Can there be any relationship between author A and author B?? Agian it seems to be so!!
- 3 and 4 possibly works in the same field

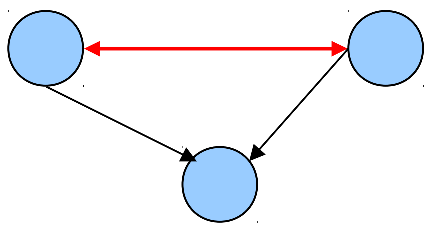


Closeness Centrality

- Measure of *mean distance* from node i to other nodes
- d_{ij} - length of geodesic path from i to j
- Mean geodesic distance from vertex i to other nodes $l_i = (N)^{-1}\sum_j d_{ij}$
- When $j = i$, $d_{ii} = 0$, better to use $l_i = (N\text{-}1)^{-1}\sum_{j\neq i} d_{ij}$
- mean geodesic distance gives low values for more central vertices
- $=> C_i = l_i = N(\sum_j d_{ij})^{-1}$ $\rightarrow$ values sparsely placed, problem with disconnected network $\rightarrow$ take harmonic mean $\rightarrow$ $C_i' = (N\text{-}1)^{-1}\sum_j (d_{ij})^{-1}$

Reciprocity

- If there is directed edge from node i to node j in directed network and there is also edge from node j to i, then edge from i to j is reciprocated.
- pairs of reciprocated edges called co-links.



- reciprocity r defined as fraction of edges that are reciprocated => $r = m^{-1}\sum_{ij} A_{ij} A_{ji}$

Rich-club Coefficient

- In science, influential researchers sometimes co-author a paper together (something strongly impactful)
- Hubs (usually high degree nodes) in a network are densely connected → A "rich club"
- The rich-club of degree k of a network $G = (V, E)$ is the set of vertices with degree greater than k, $R(k) = \{v \in V \mid k_v > k\}$. The rich-club coefficient of degree k is given by:

  $(\#\text{edge}(i,j) \mid (i,j) \in R(k))\ (|R(k)||R(k) - 1|)^{-1}$

Entropy of degree distribution

- The entropy of the degree distribution provides an average measurement of the heterogeneity of the network
- $\mathrm{H} = \sum_k P(k) log P(k)$
- What is the H of a regular graph?
- What if $P(k)$ is uniform?

Matching Index

- A *matching index* can be assigned to each edge in a network in order to quantify the similarity between the connectivity pattern of the two vertices adjacent to that edge
- Low value → Dis-similar regions of the network → a shortcut to distant regions
- Matching Index of edge(i,j):

$$\mu_{ij} = (\sum_{k\neq i,j} a_{ik} a_{kj})(\sum_{k\neq j} a_{ik} + \sum_{k\neq i} a_{jk})^{-1}$$