

CSCI 4360/6360 Data Science II

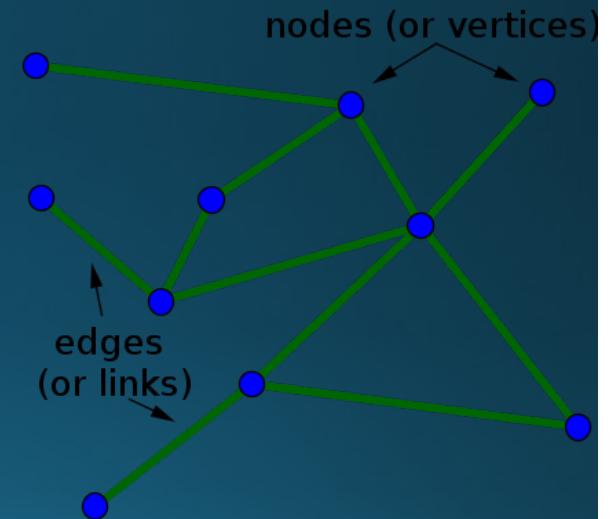
# Graphs

# Why graphs?

- Lots of data *is* graphs
  - Facebook, Twitter, citation data, and other social networks
  - The web, the blogosphere, the semantic web, Freebase, Wikipedia, Twitter, and other *information* networks
  - Text corpora (like RCV<sub>1</sub>), large datasets with discrete feature values, and other *bipartite* networks
    - nodes = documents or words
    - links connect document → word or word → document
  - Computer networks, biological networks (proteins, ecosystems, brains, ...), ...
  - Heterogeneous networks with multiple types of nodes
    - people, groups, documents

# Properties of Graphs

- Nodes & Edges
- Set  $V$  of vertices/nodes  $v_1, \dots$
- Set  $E$  of edges  $(u,v), \dots$ 
  - Can be weighted/directed/labeled
- *Degree of  $v$*  is # of edges on  $v$ 
  - *Indegree* and *outdegree* for weighted graphs
- *Path* is a sequence of edges  $(u_1,v_1), (u_2,v_2), \dots$
- *Geodesic path between  $u$  and  $v$*  is shortest path connecting them
  - Diameter is  $\max_{u,v \in V} \{\text{length of geodesic between } u,v\}$
  - Effective diameter is 90th percentile
  - Mean diameter is over connected pairs
- (*Connected*) *component* is subset of nodes that are all pairwise connected via paths
- *Clique* is subset of nodes that are all pairwise connected via *edges*
- *Triangle* is a clique of size three



# Properties of Graphs

- Descriptive statistics
- Number of connected components
- Diameter
- Degree distribution
- Centrality
- ...

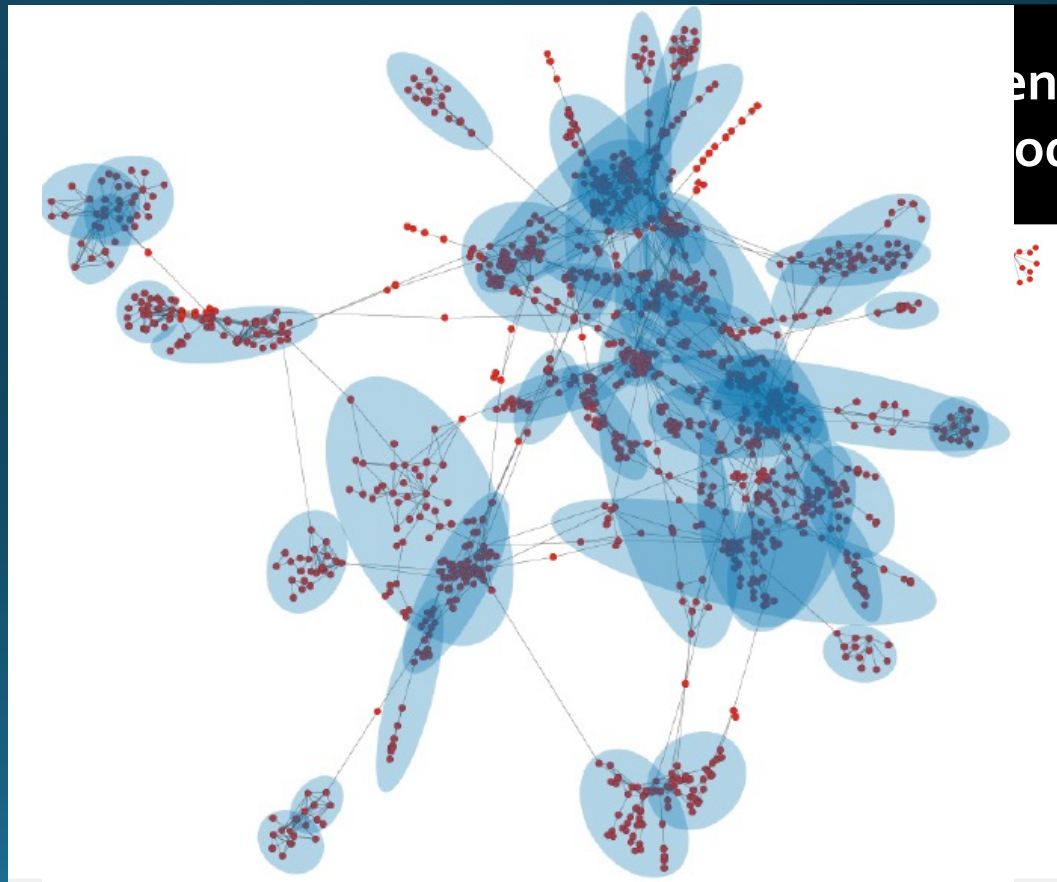


# Properties of Graphs

- Models of formation and growth
- Erdos-Rayni
- Watts-Strogatz
- Preferential attachment
- Stochastic block models
- ...

# Biology

- Protein-protein interaction networks
  - Nodes: proteins
  - Edges: interactions
- Functional modules

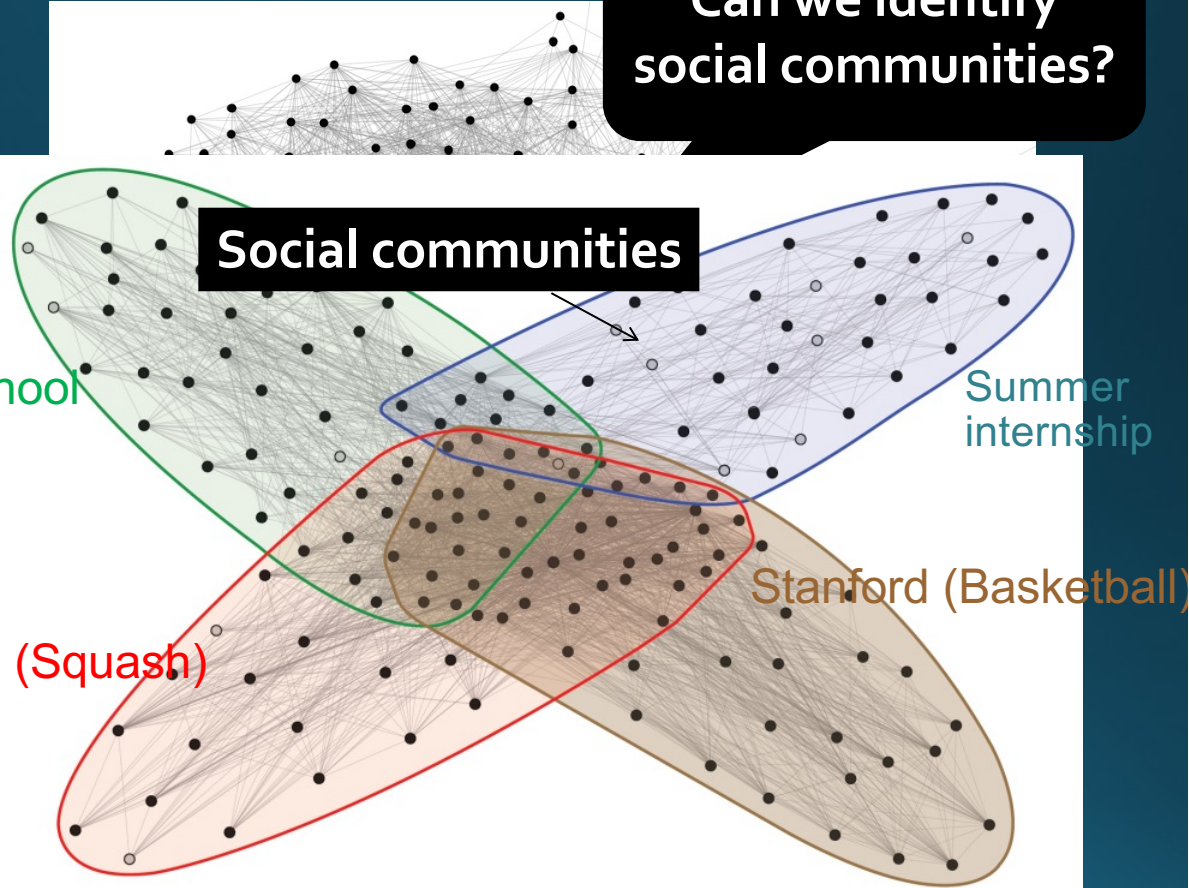


Identify  
modules?

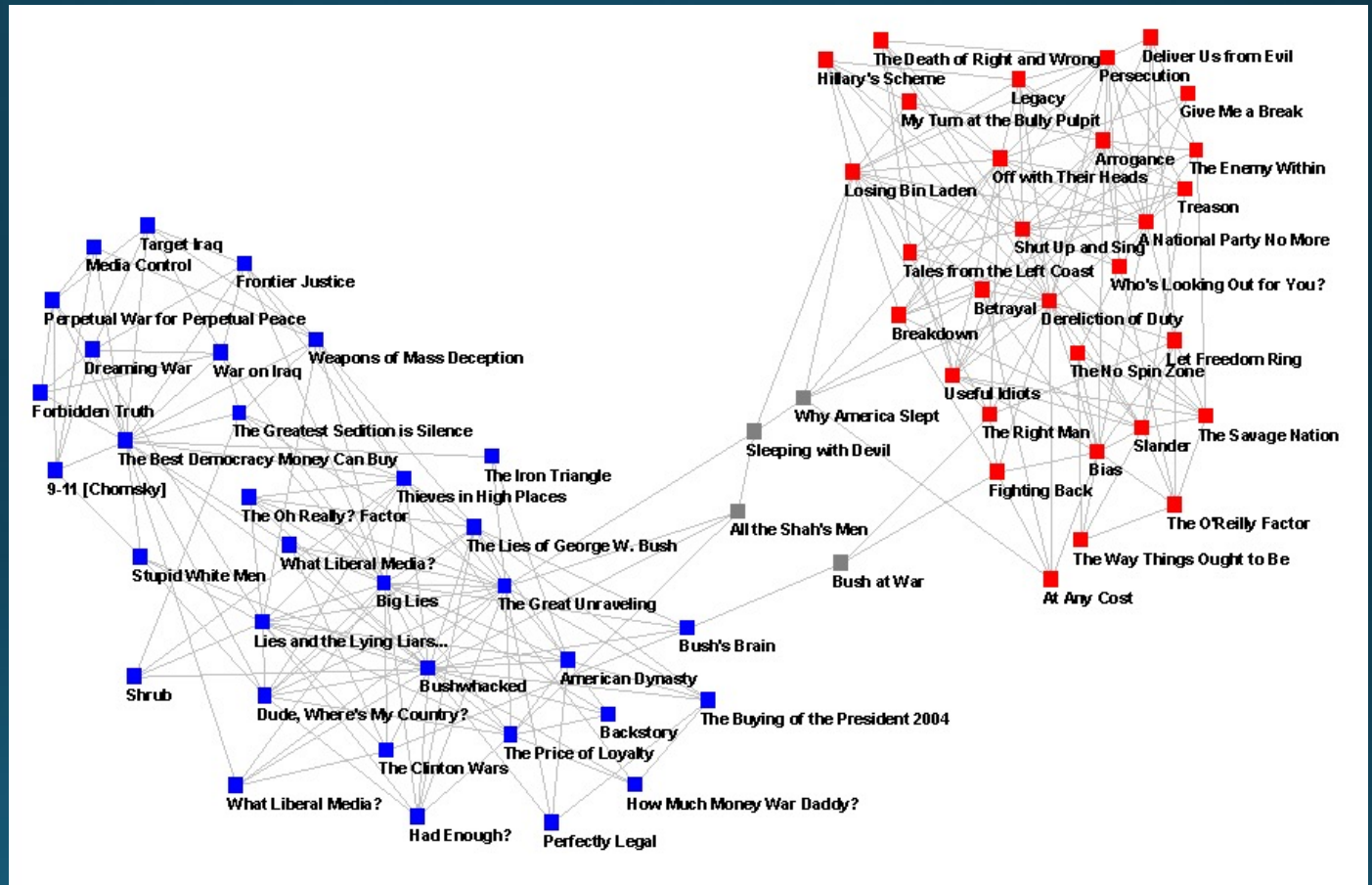
# Facebook

- Social communities
  - Nodes: Facebook users
  - Edges: Friendships

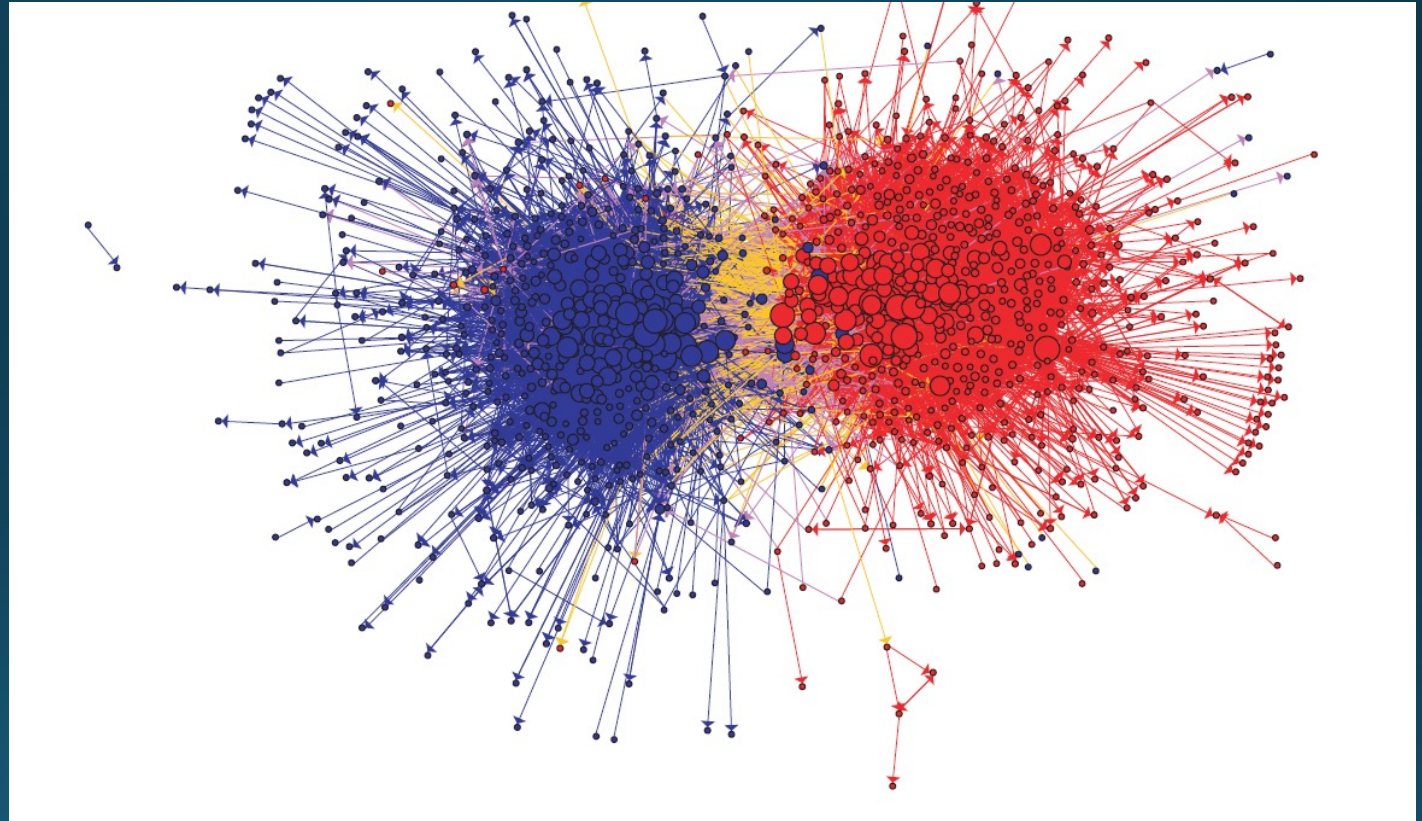
Can we identify social communities?



# Blogs



# Blogs





# Graph Models

- Fundamental graph types

# Erdos-Renyi graphs

- Take  $n$  nodes, and connect each pair with probability  $p$ 
  - Mean degree is  $z=p(n-1)$

$$\Pr[\text{degree}(v) = k] = p_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{z^k e^{-z}}{k!} \text{ for fixed } z, \text{ large } n$$

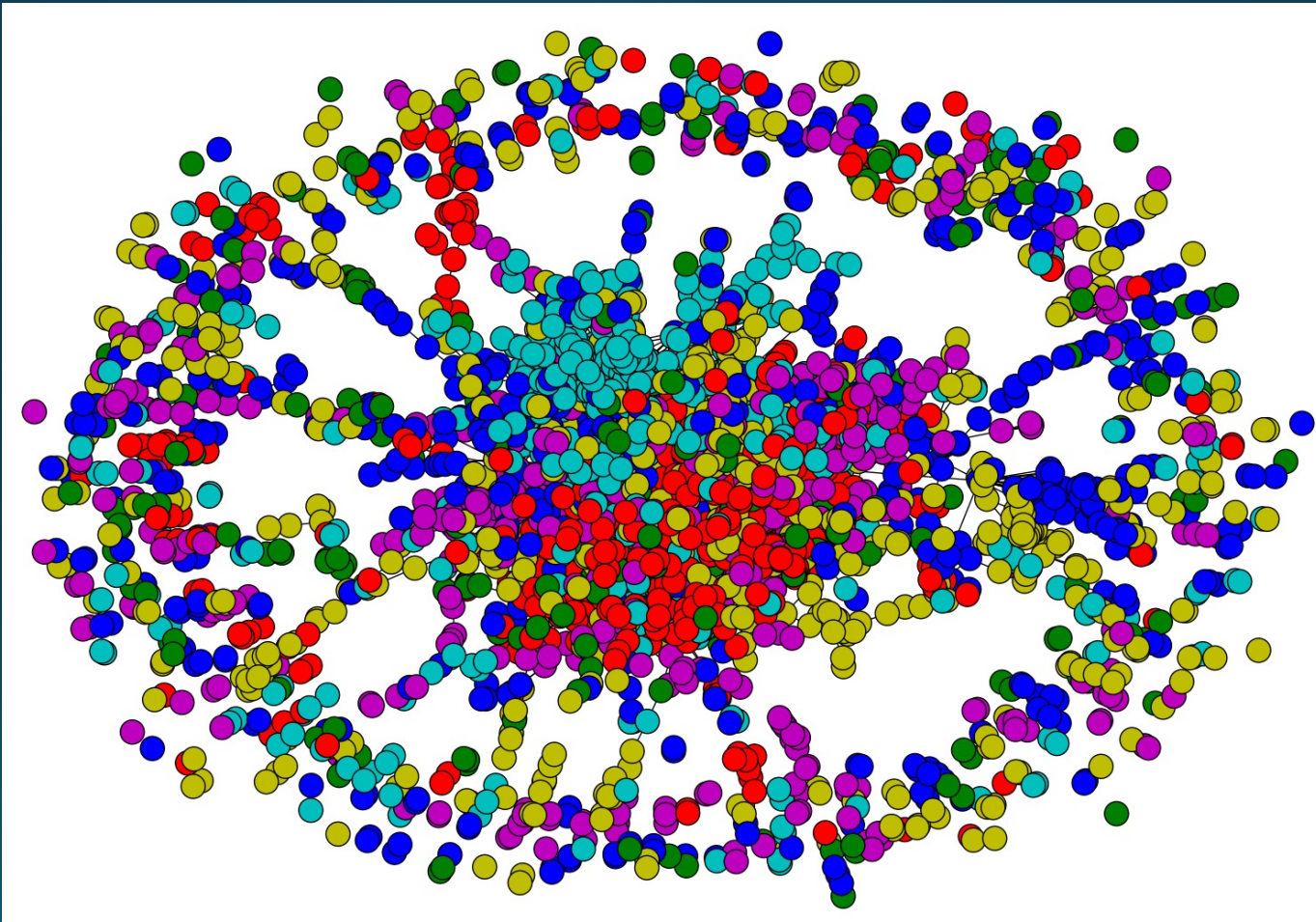
$$\binom{n}{k} p^k \approx \frac{n^k p^k}{k!} = \frac{(np)^k}{k!} \approx \frac{z^k}{k!}$$

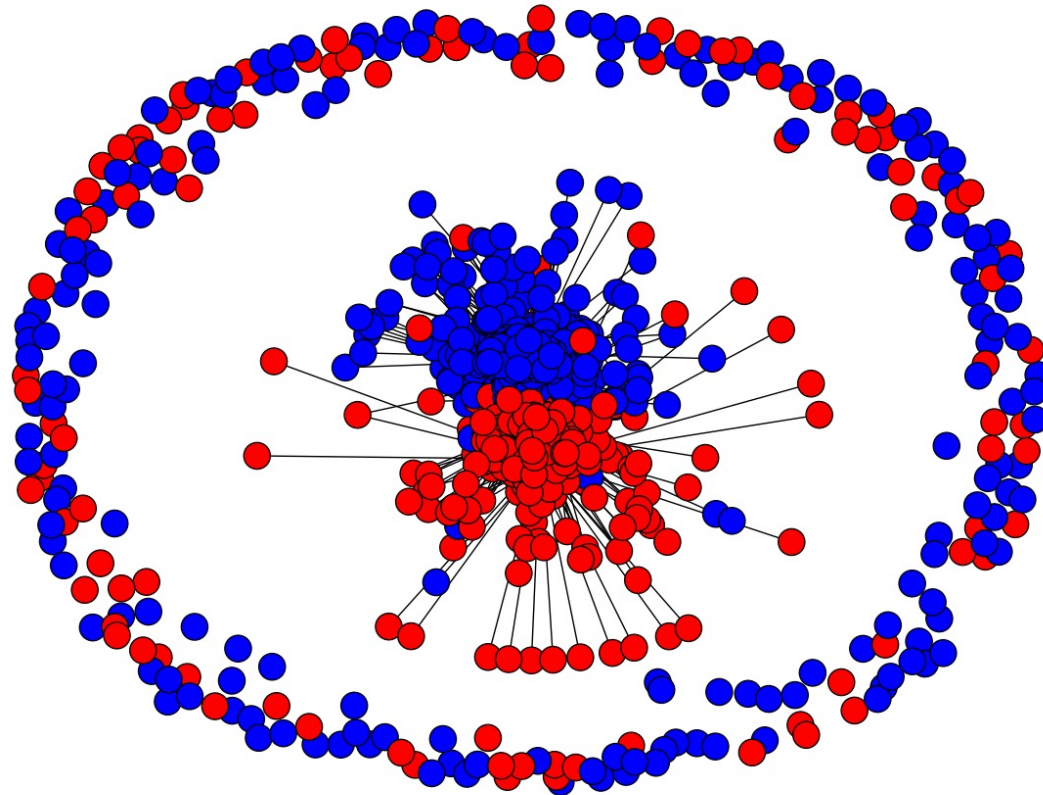
$$e^{-z}$$

# Erdos-Renyi graphs

- Take  $n$  nodes, and connect each pair with probability  $p$ 
  - Mean degree is  $z=p(n-1)$
  - Mean number of neighbors distance  $d$  from  $v$  is  $z^d$
  - How large does  $d$  need to be so that  $z^d \geq n$  ?
    - If  $z > 1$ ,  $d = \log(n)/\log(z)$
    - If  $z < 1$ , you can't do it
  - So:
    - *There tend to be either many small components ( $z < 1$ ) or one large one ( $z > 1$ ) giant connected component*
  - Another intuition:
    - If there are a two large connected components, then with high probability a few random edges will link them up.







# Erdos-Renyi graphs

- Take  $n$  nodes, and connect each pair with probability  $p$ 
  - Mean degree is  $z=p(n-1)$
  - Mean number of neighbors distance  $d$  from  $v$  is  $z^d$
  - How large does  $d$  need to be so that  $z^d \geq n$  ?
    - If  $z > 1$ ,  $d = \log(n)/\log(z)$
    - If  $z < 1$ , you can't do it
  - So:
    - If  $z > 1$ , diameters tend to be small (relative to  $n$ )

*Sociometry*, Vol. 32, No. 4. (Dec., 1969), pp. 425-443.

64 of 296 chains  
succeed, avg chain  
length is 6.2

## An Experimental Study of the Small World Problem \*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

*Arbitrarily selected individuals ( $N=296$ ) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target*

# Illustrations of the Small World

- Milgram's experiment
- Erdős numbers
  - <http://www.ams.org/mathscinet/searchauthors.html>
- Bacon numbers
  - <http://oracleofbacon.org/>
- LinkedIn
  - <http://www.linkedin.com/>
  - Privacy issues: the whole network is *not* visible to all



THE APOCALYPSE! THE SKIES BURN, THE SEAS TURN TO BLOOD, AND THE DEAD WALK THE EARTH!



|   |  |  |  |
|---|--|--|--|
| <p>THE DEAD WHAT?<br/>WALK THE EARTH!</p>   | <p>I HAVE TO GO.</p>                         | <p><math>\sum_{i=1}^n i^k \log(n)</math><br/><math>\sqrt{163}</math><br/>SCRIBBLE<br/>SCRIBBLE</p> |  |
| <p>THE DEAD RETURN!<br/>EVERYONE, QUICK,<br/>GET YOUR NAMES<br/>ON HERE!</p> <p>MATH DEPT</p> | <p>AT LAST!<br/>I HOPE THERE'S<br/>TIME!</p> |  |  |
| <p>HURRHHH</p> <p>CEMETERY</p>  |  | <p>PAUL ERDŐS?<br/>WE NEED YOU<br/>TO SIGN THIS.</p> <p>YES?</p>                                   |  |



|               | network                 | type       | $n$         | $m$           | $z$    | $\ell$ | $\alpha$ | $C^{(1)}$ | $C^{(2)}$ |
|---------------|-------------------------|------------|-------------|---------------|--------|--------|----------|-----------|-----------|
| social        | film actors             | undirected | 449 913     | 25 516 482    | 113.43 | 3.8    | 2.3      | 0.20      | 0.78      |
|               | company directors       | undirected | 7 673       | 55 392        | 14.44  | 4.0    | –        | 0.59      | 0.88      |
|               | math coauthorship       | undirected | 253 339     | 496 489       | 3.92   | 7.7    | –        | 0.15      | 0.34      |
|               | physics coauthorship    | undirected | 52 909      | 245 300       | 9.27   | 6.9    | –        | 0.45      | 0.56      |
|               | biology coauthorship    | undirected | 1 520 251   | 11 803 061    | 15.53  | 4.2    | –        | 0.088     | 0.60      |
|               | telephone call graph    | undirected | 47 000 000  | 80 000 000    | 3.16   |        | 2.1      |           |           |
|               | email messages          | directed   | 59 912      | 86 300        | 1.44   | 4.5    | 1.5/2.0  |           | 0.16      |
|               | email address books     | directed   | 16 881      | 57 029        | 3.38   | 5.2    | –        | 0.17      | 0.13      |
|               | student relationships   | undirected | 573         | 477           | 1.66   | 16.1   | –        | 0.005     | 0.001     |
|               | sexual contacts         | undirected | 2 810       |               |        |        | 3.2      |           |           |
| information   | WWW <code>nd.edu</code> | directed   | 269 504     | 1 497 135     | 5.55   | 11.7   | 2.1/2.4  | 0.11      | 0.29      |
|               | WWW Altavista           | directed   | 203 549 046 | 2 130 000 000 | 10.46  | 16.8   | 2.1/2.7  |           |           |
|               | citation network        | directed   | 783 339     | 6 716 198     | 8.57   |        | 3.0/–    |           |           |
|               | Roget's Thesaurus       | directed   | 1 022       | 5 103         | 4.99   | 4.7    | –        | 0.13      | 0.15      |
|               | word co-occurrence      | undirected | 460 902     | 17 000 000    | 70.13  |        | 2.7      |           | 0.44      |
| technological | Internet                | undirected | 10 697      | 31 992        | 5.98   | 3.1    | 2.5      | 0.035     | 0.39      |
|               | power grid              | undirected | 4 941       | 6 591         | 2.67   | 18.9   | –        | 0.10      | 0.080     |
|               | train routes            | undirected | 587         | 19 603        | 66.79  | 2.6    | –        |           | 0.69      |
|               | software packages       | directed   | 1 439       | 1 723         | 1.20   | 2.2    | 1.6/1.4  | 0.070     | 0.082     |
|               | software classes        | directed   | 1 377       | 2 213         | 1.61   | 1.1    | –        | 0.033     | 0.012     |
|               | electronic circuits     | undirected | 24 097      | 53 248        | 4.34   | 11.5   | 3.0      | 0.010     | 0.030     |
|               | peer-to-peer network    | undirected | 880         | 1 295         | 1.47   | 4.8    | 2.1      | 0.012     | 0.011     |
| biological    | metabolic network       | undirected | 765         | 3 685         | 9.64   | 2.6    | 2.2      | 0.090     | 0.67      |
|               | protein interactions    | undirected | 2 115       | 2 240         | 2.12   | 6.0    | 2.4      | 0.072     | 0.071     |
|               | marine food web         | directed   | 135         | 598           | 4.43   | 2.5    | –        | 0.16      | 0.23      |
|               | freshwater food web     | directed   | 92          | 997           | 10.84  | 1.0    | –        | 0.20      | 0.087     |
|               | neural network          | directed   | 307         | 2 350         | 7.68   | 3.7    | –        | 0.18      | 0.28      |

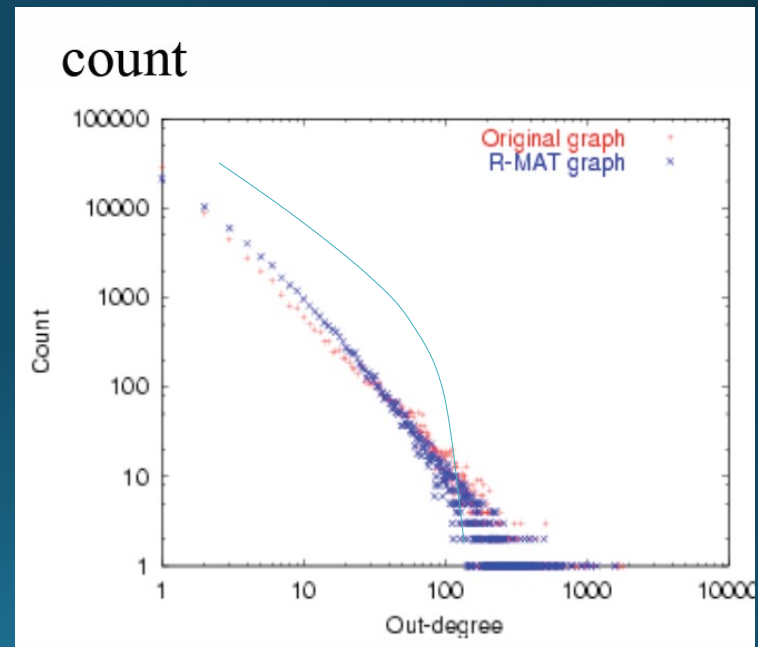
# Erdos-Renyi graphs

- A good model of degree distribution in “natural” networks?



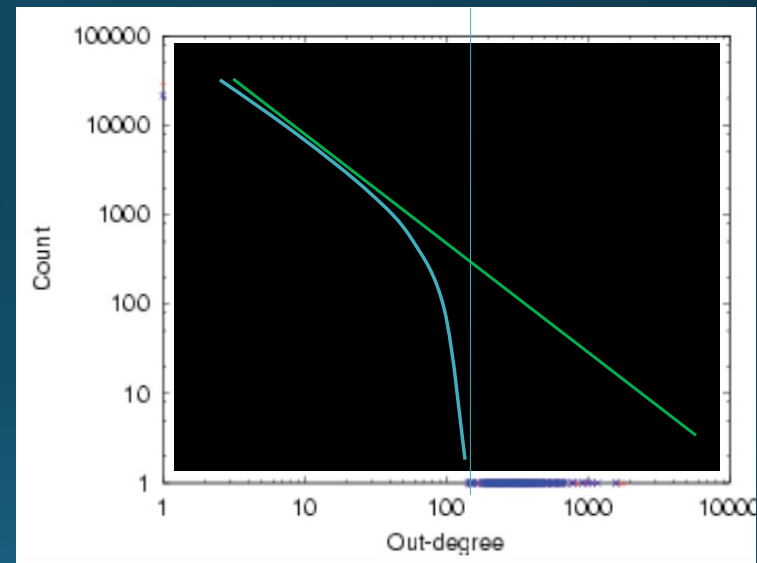
# Degree distribution

- Plot cumulative degree
  - X axis is degree
  - Y axis is #nodes that have degree at least  $k$
- Typically use a log-log scale
  - Straight lines are a power law; normal curve dives to zero at some point
  - Right: trust network in opinions web site from Richardson & Domingos



# Degree distribution

- Plot cumulative degree
  - X axis is degree
  - Y axis is #nodes that have degree at least  $k$
- Typically use a log-log scale
  - Straight lines are a power law; normal curve dives to zero at some point
    - This defines a “scale” for the network
  - Right: trust network in opinions web site from Richardson & Domingos



$$p_k \propto k^{-\alpha}$$

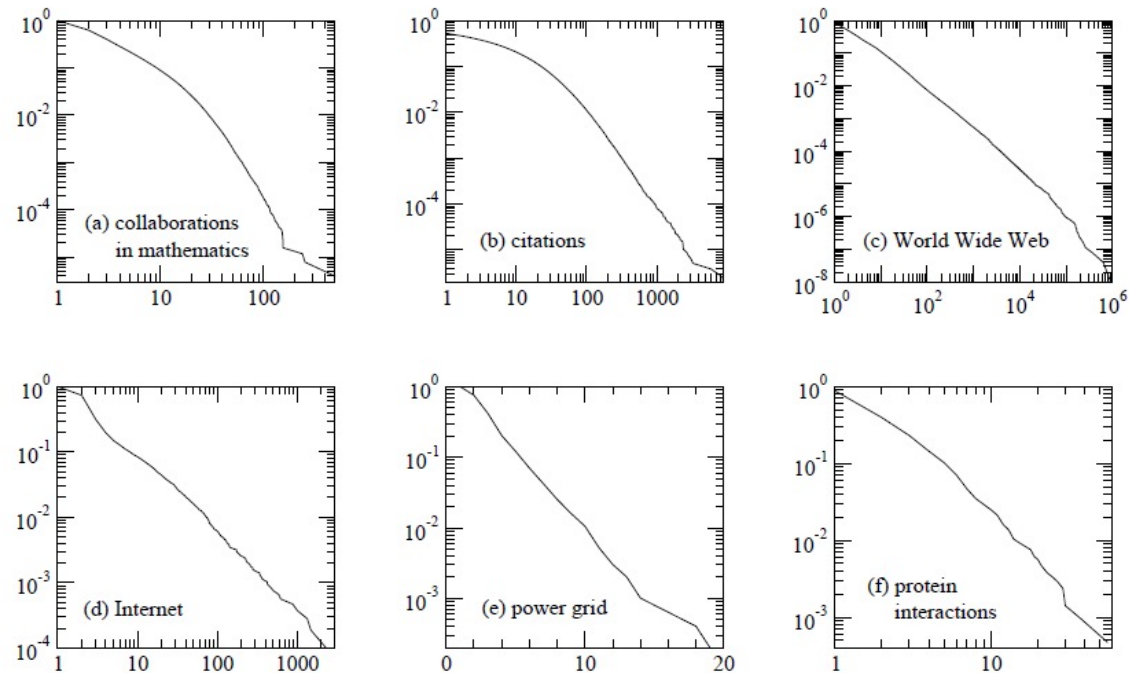



FIG. 6 Cumulative degree distributions for six different networks. The horizontal axis for each panel is vertex degree  $k$  (or in-degree for the citation and Web networks, which are directed) and the vertical axis is the cumulative probability distribution of degrees, i.e., the fraction of vertices that have degree greater than or equal to  $k$ . The networks shown are: (a) the collaboration network of mathematicians [182]; (b) citations between 1981 and 1997 to all papers cataloged by the Institute for Scientific Information [351]; (c) a 300 million vertex subset of the World Wide Web, circa 1999 [74]; (d) the Internet at the level of autonomous systems, April 1999 [86]; (e) the power grid of the western United States [416]; (f) the interaction network of proteins in the metabolism of the yeast *S. Cerevisiae* [212]. Of these networks, three of them, (c), (d) and (f), appear to have power-law degree distributions, as indicated by their approximately straight-line forms on the doubly logarithmic scales, and one (b) has a power-law tail but deviates markedly from power-law behavior for small degree. Network (e) has an exponential degree distribution (note the log-linear scales used in this panel) and network (a) appears to have a truncated power-law degree distribution of some type, or possibly two separate power-law regimes with different exponents.

|               | network               | type       | $n$         | $m$           | $z$    | $\ell$ | $\alpha$ | $C^{(1)}$ | $C^{(2)}$ |
|---------------|-----------------------|------------|-------------|---------------|--------|--------|----------|-----------|-----------|
| social        | film actors           | undirected | 449 913     | 25 516 482    | 113.43 | 3.48   | 2.3      | 0.20      | 0.78      |
|               | company directors     | undirected | 7 673       | 55 392        | 14.44  | 4.60   | –        | 0.59      | 0.88      |
|               | math coauthorship     | undirected | 253 339     | 496 489       | 3.92   | 7.57   | –        | 0.15      | 0.34      |
|               | physics coauthorship  | undirected | 52 909      | 245 300       | 9.27   | 6.19   | –        | 0.45      | 0.56      |
|               | biology coauthorship  | undirected | 1 520 251   | 11 803 064    | 15.53  | 4.92   | –        | 0.088     | 0.60      |
|               | telephone call graph  | undirected | 47 000 000  | 80 000 000    | 3.16   |        | 2.1      |           |           |
|               | email messages        | directed   | 59 912      | 86 300        | 1.44   | 4.95   | 1.5/2.0  |           | 0.16      |
|               | email address books   | directed   | 16 881      | 57 029        | 3.38   | 5.22   | –        | 0.17      | 0.13      |
|               | student relationships | undirected | 573         | 477           | 1.66   | 16.01  | –        | 0.005     | 0.001     |
|               | sexual contacts       | undirected | 2 810       |               |        |        | 3.2      |           |           |
| information   | WWW nd.edu            | directed   | 269 504     | 1 497 135     | 5.55   | 11.27  | 2.1/2.4  | 0.11      | 0.29      |
|               | WWW Altavista         | directed   | 203 549 046 | 2 130 000 000 | 10.46  | 16.18  | 2.1/2.7  |           |           |
|               | citation network      | directed   | 783 339     | 6 716 198     | 8.57   |        | 3.0/–    |           |           |
|               | Roget's Thesaurus     | directed   | 1 022       | 5 103         | 4.99   | 4.87   | –        | 0.13      | 0.15      |
|               | word co-occurrence    | undirected | 460 902     | 17 000 000    | 70.13  |        | 2.7      |           | 0.44      |
| technological | Internet              | undirected | 10 697      | 31 992        | 5.98   | 3.31   | 2.5      | 0.035     | 0.39      |
|               | power grid            | undirected | 4 941       | 6 594         | 2.67   | 18.99  | –        | 0.10      | 0.080     |
|               | train routes          | undirected | 587         | 19 603        | 66.79  | 2.16   | –        |           | 0.69      |
|               | software packages     | directed   | 1 439       | 1 723         | 1.20   | 2.42   | 1.6/1.4  | 0.070     | 0.082     |
|               | software classes      | directed   | 1 377       | 2 213         | 1.61   | 1.51   | –        | 0.033     | 0.012     |
|               | electronic circuits   | undirected | 24 097      | 53 248        | 4.34   | 11.05  | 3.0      | 0.010     | 0.030     |
|               | peer-to-peer network  | undirected | 880         | 1 296         | 1.47   | 4.28   | 2.1      | 0.012     | 0.011     |
| biological    | metabolic network     | undirected | 765         | 3 686         | 9.64   | 2.56   | 2.2      | 0.090     | 0.67      |
|               | protein interactions  | undirected | 2 115       | 2 240         | 2.12   | 6.80   | 2.4      | 0.072     | 0.071     |
|               | marine food web       | directed   | 135         | 598           | 4.43   | 2.05   | –        | 0.16      | 0.23      |
|               | freshwater food web   | directed   | 92          | 997           | 10.84  | 1.90   | –        | 0.20      | 0.087     |
|               | neural network        | directed   | 307         | 2 359         | 7.68   | 3.97   | –        | 0.18      | 0.28      |

# Graphs

- Some common properties of graphs:
  - **Distribution of node degrees**
  - Distribution of cliques (e.g., triangles)
  - **Distribution of paths**
    - **Diameter** (max shortest-path)
    - Effective diameter (90<sup>th</sup> percentile)
    - **Connected components**
  - ...
- Some types of graphs to consider:
  - Real graphs (social & otherwise)
  - Generated graphs:
    - **Erdos-Renyi** "Bernoulli" or "Poisson"
    - Watts-Strogatz "small world" graphs
    - Barabasi-Albert "preferential attachment"
    - ...

# Graphs

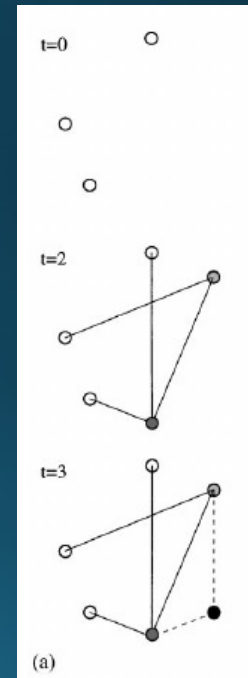
- Some common properties of graphs:
    - **Distribution of node degrees:**  
**often scale-free**
    - Distribution of cliques (e.g., triangles)
    - **Distribution of paths**
      - **Diameter** (max shortest-path)
      - Effective **diameter** (90<sup>th</sup> percentile)  
**often small**
      - **Connected components usually one giant CC**
    - ...
  - Some types of graphs to consider:
    - Real graphs (social & otherwise)
    - Generated graphs:
      - **Erdos-Renyi**  
**"Bernoulli" or "Poisson"**
      - Watts-Strogatz "small world" graphs
      - Barbosi-Albert "preferential attachment" **generates scale-free graphs**
      - ...
- 

# Barabasi-Albert Networks

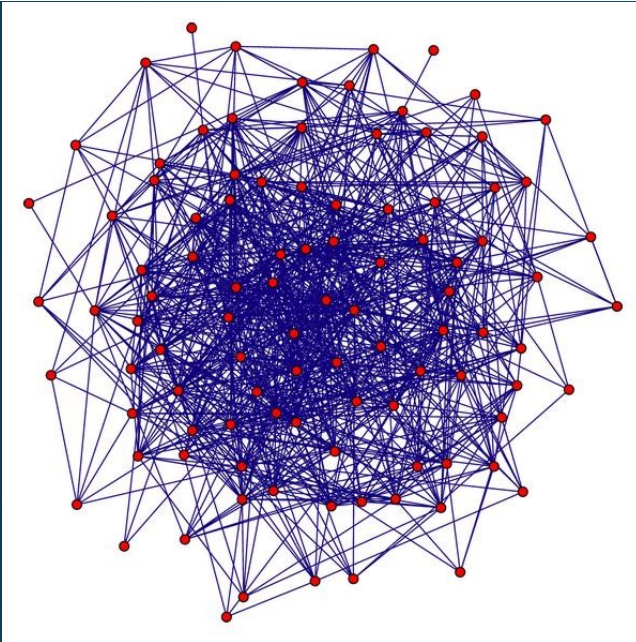
- *Science* 286 (1999)
- Start from a small number of nodes, add a new node with  $m$  links
- **Preferential Attachment**
  - Probability of these links to connect to existing nodes is proportional to the node's degree

$$P(k_i) = \frac{k_i}{\sum_j k_j}$$

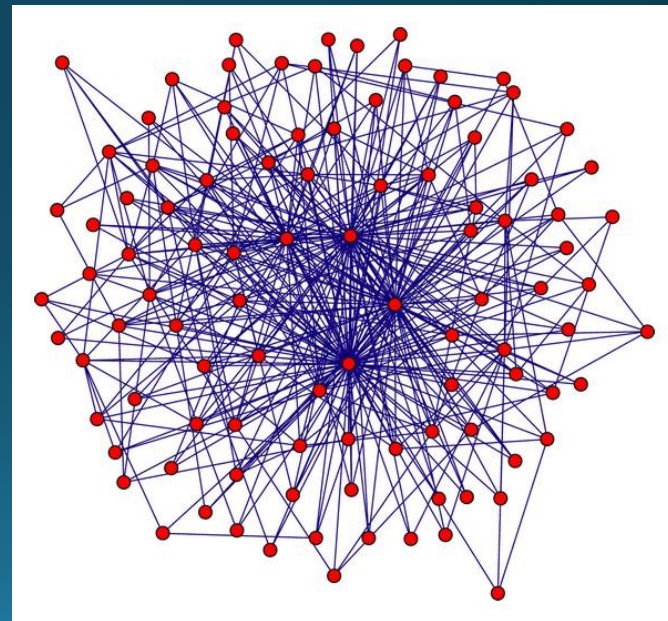
- 'Rich gets richer'
- This creates 'hubs': few nodes with very large degrees








Random graph  
(Erdos Renyi)



Preferential attachment  
(Barabasi-Albert)



# Graphs

- Some common properties of graphs:
    - **Distribution of node degrees: often scale-free**
    - Distribution of cliques (e.g., triangles)
    - **Distribution of paths**
      - **Diameter** (max shortest-path)
      - Effective **diameter** (90<sup>th</sup> percentile) **often small**
      - **Connected components usually one giant CC**
    - ...
  - Some types of graphs to consider:
    - Real graphs (social & otherwise)
    - Generated graphs:
      - **Erdos-Renyi "Bernoulli" or "Poisson"**
      - Watts-Strogatz "small world" graphs
      - Barbosi-Albert "preferential attachment" **generates scale-free graphs**
      - ...
- 

# Homophily

- One definition: excess edges between similar nodes
  - E.g., assume nodes are male and female and  $\Pr(\text{male})=p$ ,  $\Pr(\text{female})=q$ .
  - Is  $\Pr(\text{gender}(u) \neq \text{gender}(v) \mid \text{edge}(u,v)) \geq 2pq$ ?
- Another definition: excess edges between common neighbors of  $v$

$$CC(v) = \frac{\# \text{triangles connected to } v}{\# \text{pairs connected to } v}$$

$$CC(V, E) = \frac{1}{|V|} \sum_v CC(v)$$

# Homophily

- Another definition: excess edges between common neighbors of  $v$

$$CC(v) = \frac{\# \text{triangles connected to } v}{\# \text{pairs connected to } v}$$

$$CC(V, E) = \frac{1}{|V|} \sum_v CC(v)$$

$$CC'(V, E) = \frac{\# \text{triangles in graph}}{\# \text{length 3 paths in graph}}$$

# Homophily

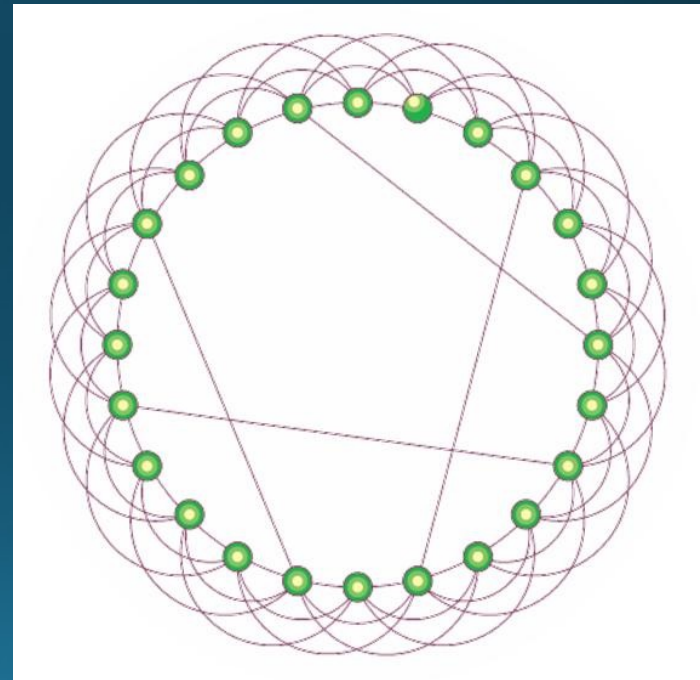
- In a random Erdos-Renyi graph:

$$CC'(V, E) = \frac{\# \text{triangles in graph}}{\# \text{length 3 paths in graph}} \approx \frac{1}{n} \text{ for large } n$$

- Probably not realistic!
- In a natural graph, two of your mutual friends might also be friends
  - Both in the same class or organization
  - You introduced them
  - They introduced you

# Watts-Strogatz model

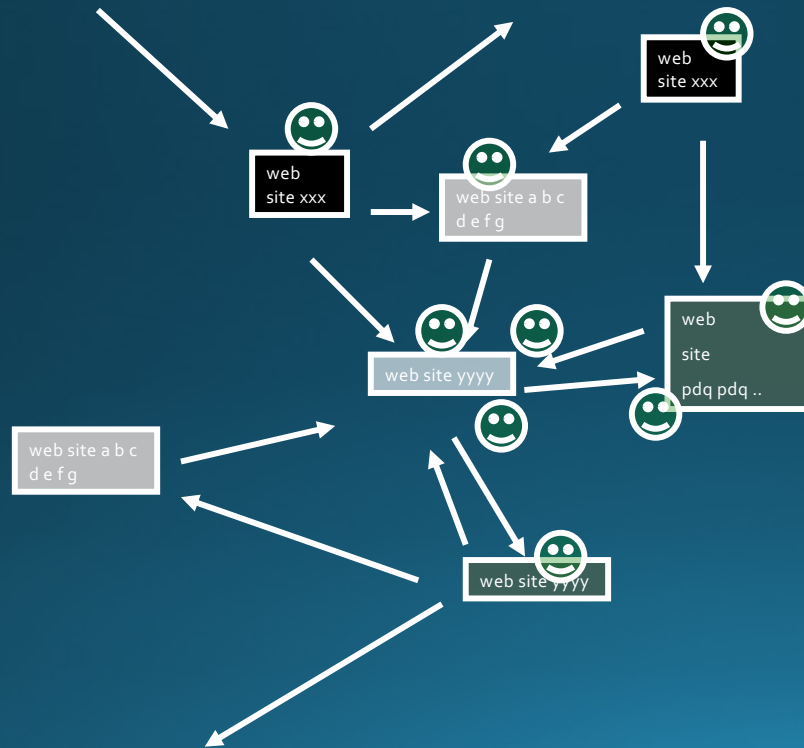
- Start with a ring
- Connect each node to  $k$  nearest neighbors
  - → homophily
- Add some random shortcuts from one point to another
  - → small diameter
- Degree distribution *not* scale-free
- Generalizes to  $d$  dimensions



|               | network               | type       | $n$         | $m$           | $z$    | $\ell$ | $\alpha$ | $C^{(1)}$ | $C^{(2)}$ |
|---------------|-----------------------|------------|-------------|---------------|--------|--------|----------|-----------|-----------|
| social        | film actors           | undirected | 449 913     | 25 516 482    | 113.43 | 3.48   | 2.3      | 0.20      | 0.78      |
|               | company directors     | undirected | 7 673       | 55 392        | 14.44  | 4.60   | –        | 0.59      | 0.88      |
|               | math coauthorship     | undirected | 253 339     | 496 489       | 3.92   | 7.57   | –        | 0.15      | 0.34      |
|               | physics coauthorship  | undirected | 52 909      | 245 300       | 9.27   | 6.19   | –        | 0.45      | 0.56      |
|               | biology coauthorship  | undirected | 1 520 251   | 11 803 064    | 15.53  | 4.92   | –        | 0.088     | 0.60      |
|               | telephone call graph  | undirected | 47 000 000  | 80 000 000    | 3.16   |        | 2.1      |           |           |
|               | email messages        | directed   | 59 912      | 86 300        | 1.44   | 4.95   | 1.5/2.0  |           | 0.16      |
|               | email address books   | directed   | 16 881      | 57 029        | 3.38   | 5.22   | –        | 0.17      | 0.13      |
|               | student relationships | undirected | 573         | 477           | 1.66   | 16.01  | –        | 0.005     | 0.001     |
|               | sexual contacts       | undirected | 2 810       |               |        |        | 3.2      |           |           |
| information   | WWW nd.edu            | directed   | 269 504     | 1 497 135     | 5.55   | 11.27  | 2.1/2.4  | 0.11      | 0.29      |
|               | WWW Altavista         | directed   | 203 549 046 | 2 130 000 000 | 10.46  | 16.18  | 2.1/2.7  |           |           |
|               | citation network      | directed   | 783 339     | 6 716 198     | 8.57   |        | 3.0/–    |           |           |
|               | Roget's Thesaurus     | directed   | 1 022       | 5 103         | 4.99   | 4.87   | –        | 0.13      | 0.15      |
|               | word co-occurrence    | undirected | 460 902     | 17 000 000    | 70.13  |        | 2.7      |           | 0.44      |
| technological | Internet              | undirected | 10 697      | 31 992        | 5.98   | 3.31   | 2.5      | 0.035     | 0.39      |
|               | power grid            | undirected | 4 941       | 6 594         | 2.67   | 18.99  | –        | 0.10      | 0.080     |
|               | train routes          | undirected | 587         | 19 603        | 66.79  | 2.16   | –        |           | 0.69      |
|               | software packages     | directed   | 1 439       | 1 723         | 1.20   | 2.42   | 1.6/1.4  | 0.070     | 0.082     |
|               | software classes      | directed   | 1 377       | 2 213         | 1.61   | 1.51   | –        | 0.033     | 0.012     |
|               | electronic circuits   | undirected | 24 097      | 53 248        | 4.34   | 11.05  | 3.0      | 0.010     | 0.030     |
|               | peer-to-peer network  | undirected | 880         | 1 296         | 1.47   | 4.28   | 2.1      | 0.012     | 0.011     |
| biological    | metabolic network     | undirected | 765         | 3 686         | 9.64   | 2.56   | 2.2      | 0.090     | 0.67      |
|               | protein interactions  | undirected | 2 115       | 2 240         | 2.12   | 6.80   | 2.4      | 0.072     | 0.071     |
|               | marine food web       | directed   | 135         | 598           | 4.43   | 2.05   | –        | 0.16      | 0.23      |
|               | freshwater food web   | directed   | 92          | 997           | 10.84  | 1.90   | –        | 0.20      | 0.087     |
|               | neural network        | directed   | 307         | 2 359         | 7.68   | 3.97   | –        | 0.18      | 0.28      |



# Google's PageRank

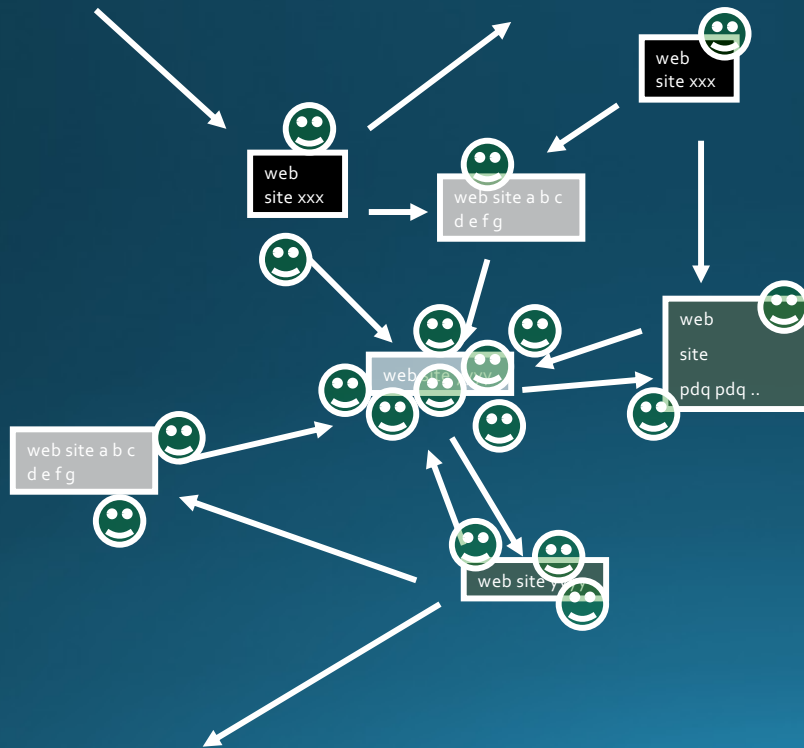


- Imagine a "pagehopper" that always either ☺
- follows a random link, or
  - jumps to random page



# Google's PageRank

(Brin & Page, <http://www-db.stanford.edu/~backrub/google.html>)



Imagine a "pagehopper" that always either

- follows a random link, or
- jumps to random page

PageRank ranks pages by the amount of time the pagehopper spends on a page:

- or, if there were many pagehoppers, PageRank is the expected "crowd size"

# Random Walks

$G$  : a graph

$P$  : transition probability matrix

$$P(u, v) = \begin{cases} \frac{1}{d_u} & \text{if } u : v, \quad d_u := \text{the degree of } u. \\ 0 & \text{otherwise.} \end{cases}$$

A lazy walk:  $W = \frac{I + P}{2}$  avoids messy "dead ends"....

# Random Walks: PageRank

## A (bored) surfer

- either surf a random webpage  
with probability  $\alpha$
- or surf a linked webpage  
with probability  $1 - \alpha$



$\alpha$  : the jumping constant

$$p = \alpha \left( \frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n} \right) + (1 - \alpha) pW$$

# Random Walks: PageRank

Two equivalent ways to define PageRank  $p = pr(\alpha, s)$

$$(1) \quad p = \alpha s + (1 - \alpha) pW$$

$$(2) \quad p = \alpha \sum_{t=0}^{\infty} (1 - \alpha)^t (sW^t)$$

$s = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$   $\longrightarrow$  the (original) PageRank

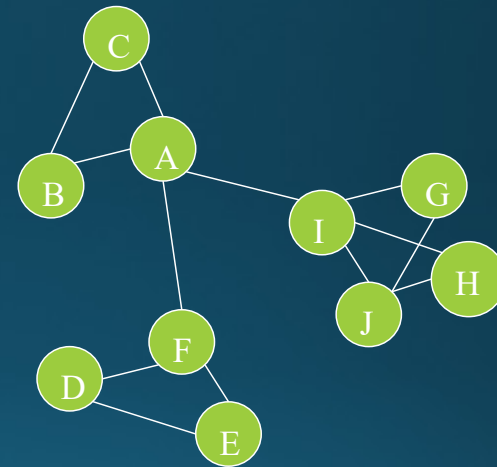
$s =$  some "seed", e.g.,  $(1, 0, \dots, 0)$

$\longrightarrow$  personalized PageRank

# Graph = Matrix

Vector = Node  $\rightarrow$  Weight

| M |   |   |   |   |   |   |   |   |   |   | V |
|---|---|---|---|---|---|---|---|---|---|---|---|
|   | A | B | C | D | E | F | G | H | I | J | A |
| A | - | 1 | 1 |   |   | 1 |   |   |   |   | 3 |
| B | 1 | - | 1 |   |   |   |   |   |   |   | 2 |
| C | 1 | 1 | - |   |   |   |   |   |   |   | 3 |
| D |   |   |   | - | 1 | 1 |   |   |   |   |   |
| E |   |   |   | 1 | - | 1 |   |   |   |   |   |
| F | 1 |   |   | 1 | 1 | - |   |   |   |   |   |
| G |   |   |   |   |   |   | - |   | 1 | 1 |   |
| H |   |   |   |   |   |   |   | - | 1 | 1 |   |
| I |   |   |   |   |   |   | 1 | 1 | - | 1 |   |
| J |   |   |   |   |   |   | 1 | 1 | 1 | - |   |



# PageRank

- Let  $\mathbf{u} = (1/N, \dots, 1/N)$ 
  - dimension = #nodes  $N$
- Let  $A$  = adjacency matrix: [ $a_{ij}=1 \Leftrightarrow i$  links to  $j$ ]
- Let  $W = [w_{ij} = a_{ij}/\text{outdegree}(i)]$ 
  - $w_{ij}$  is probability of jump from  $i$  to  $j$
- Let  $\mathbf{v}^0 = (1, 1, \dots, 1)$ 
  - or anything else you want
- Repeat until converged:
  - Let  $\mathbf{v}^{t+1} = c\mathbf{u} + (1-c)\mathbf{W}\mathbf{v}^t$ 
    - $c$  is probability of jumping “anywhere randomly”