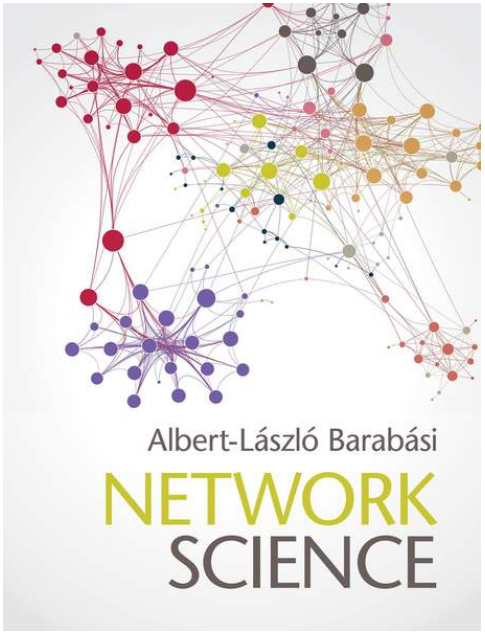
15-382 Collective Intelligence – S19

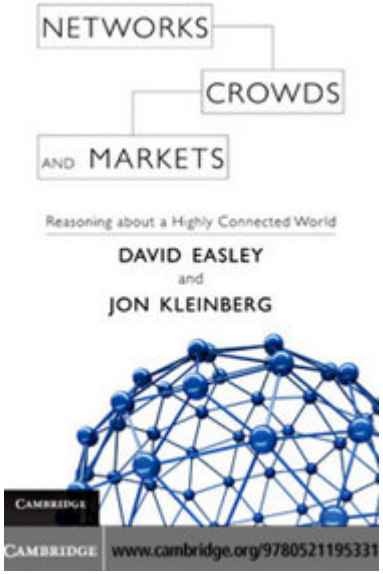
Lecture 34:
Networks 1

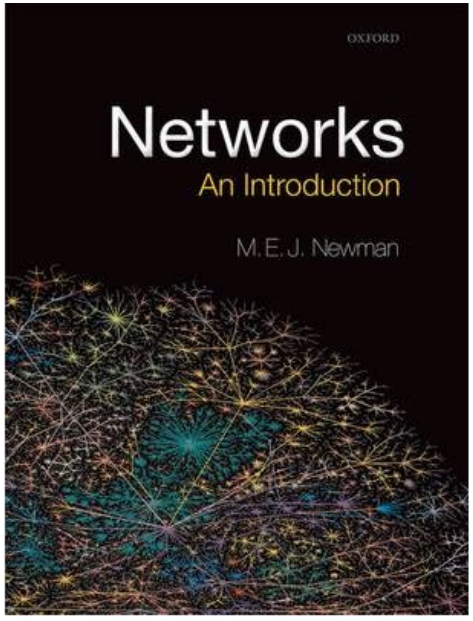
Teacher:
Gianni A. Di Caro

Network Science

- Barabasi, "Network Science"
- Easley & Kleinberg, "Networks, Crowds, and Markets: Reasoning about a Highly Connected World"
- Newman, "Networks"





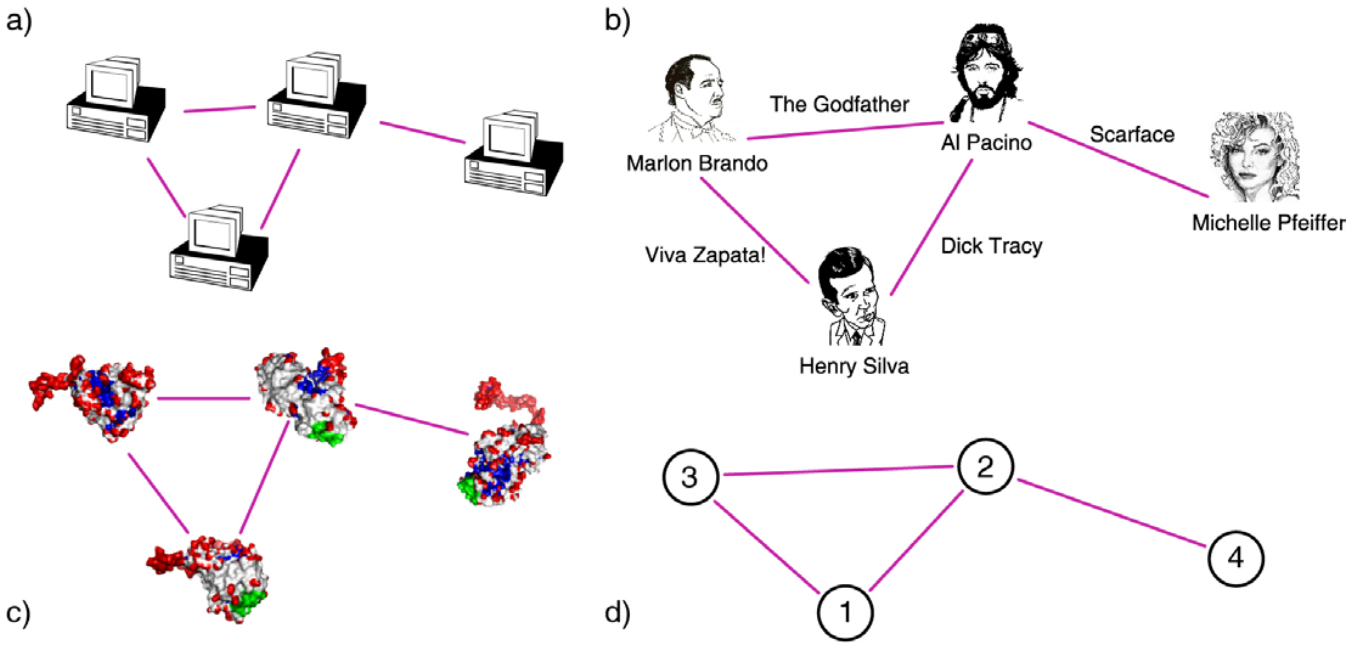


COMPLEX SYSTEMS AS NETWORKS

Many complex systems can be represented as networks

Any complex system has an associated network of communication / interaction among the components

- Nodes = components of the complex system
- Links = interactions between them

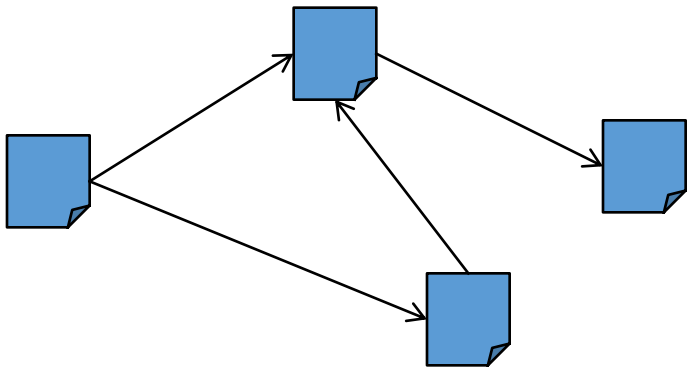


This and following slides are adapted from Kristina Lerman's slides

Directed vs. Undirected Networks

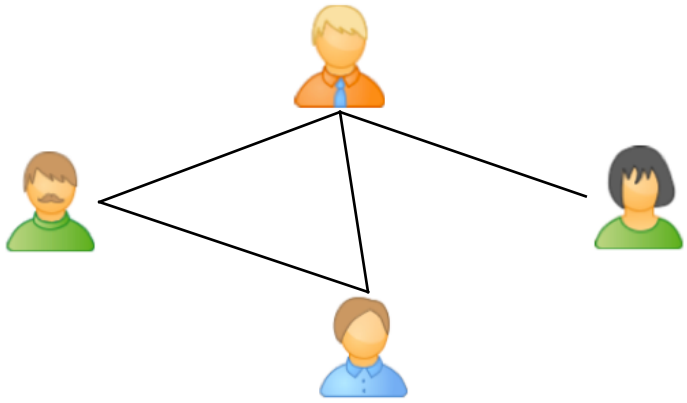
Directed

- Directed links
  - interaction flows one way
- Examples
  - WWW: web pages and hyperlinks
  - Twitter follower graph
  - Animal relations, prey-predator



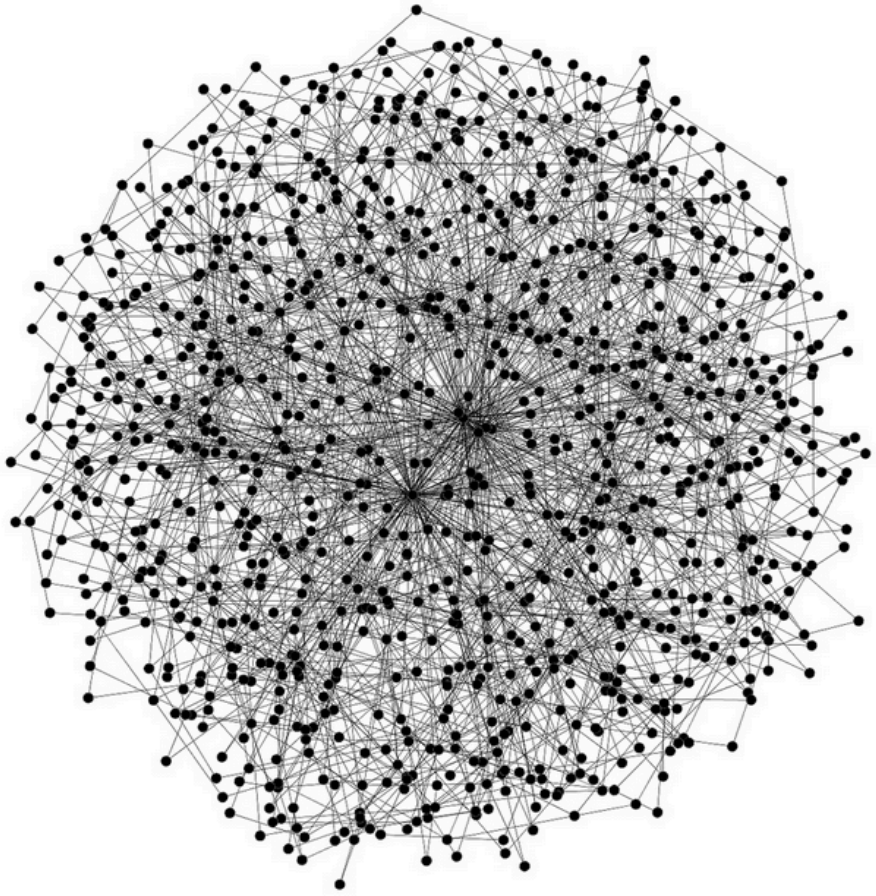
Undirected

- Undirected links
  - Interactions flow both ways
- Examples
  - Social networks: people and friendship
  - Atoms in a crystal
  - Countries in geographic maps

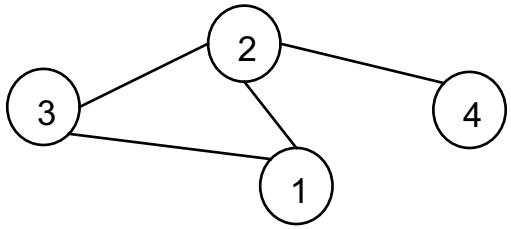


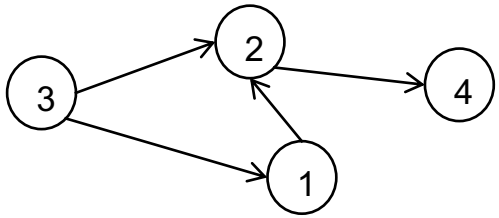
How do we characterize networks?

- Size
  - Number of nodes
  - Number of links
- Degree
  - Average degree
  - Degree distribution
- Diameter
- Clustering coefficient
- ...



Node degree





Undirected networks

- Node degree: number of links to other nodes

$$[k_1 = 2, k_2 = 3,\ k_3 = 2,\ k_4 = 1]$$

- Number of links

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i$$

- Average degree

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

Directed networks

- Indegree:

$$[k_1^{in} = 1\ \ k_2^{in} = 2, k_3^{in} = 0, k_4^{in} = 1]$$

- Outdegree:

$$[k_1^{out} = 1\ \ k_2^{out} = 1, k_3^{out} = 2, k_4^{out} = 0]$$

- Total degree = in + out
  - Number of links

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$$

  - Average degree: L/N

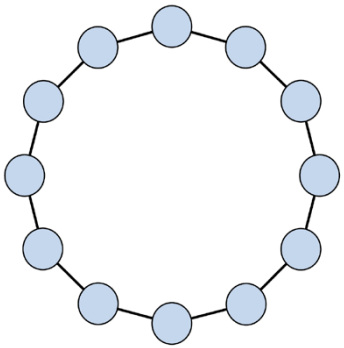
Degree distribution

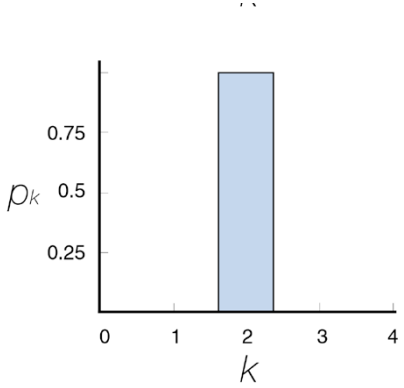
- Degree distribution p_k is the probability that a randomly selected node has degree k

$$p_k = N_k / N$$

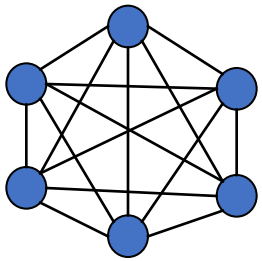
  - Where N_k is number of nodes of degree k

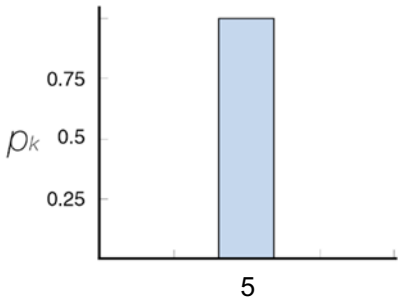
regular lattice



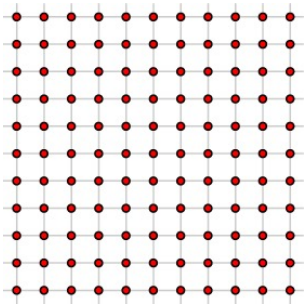


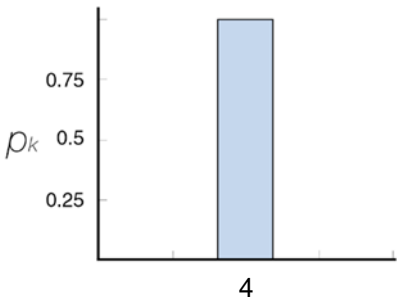
clique (fully connected graph)



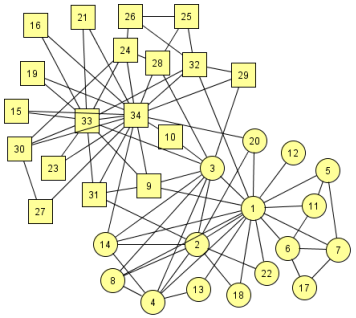


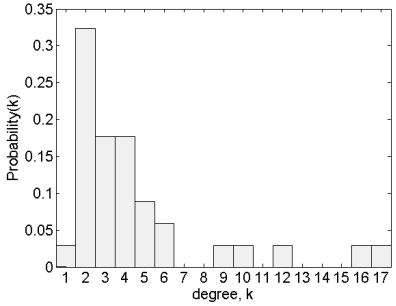
regular lattice





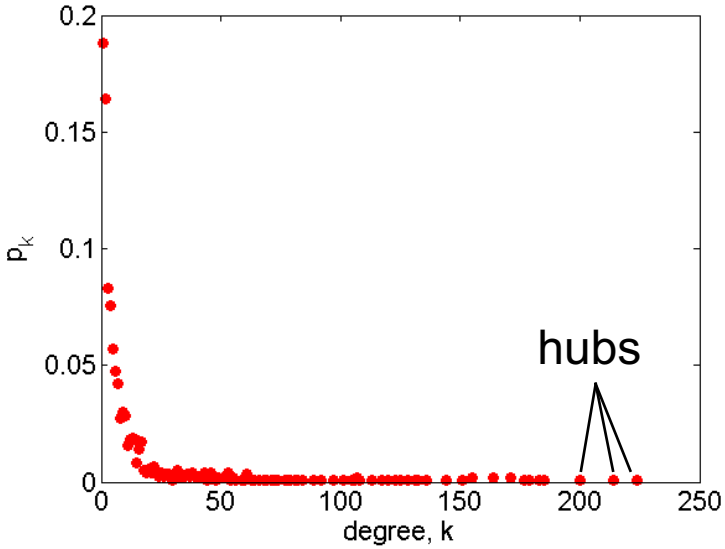
karate club friendship network

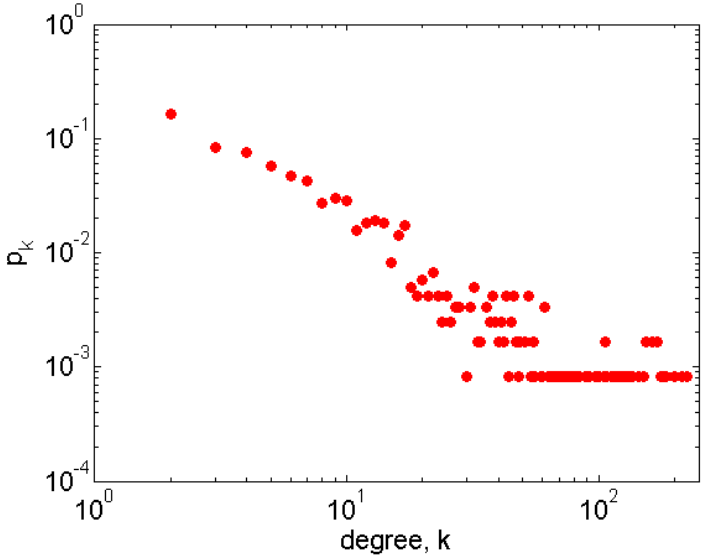




Degree distribution in real networks

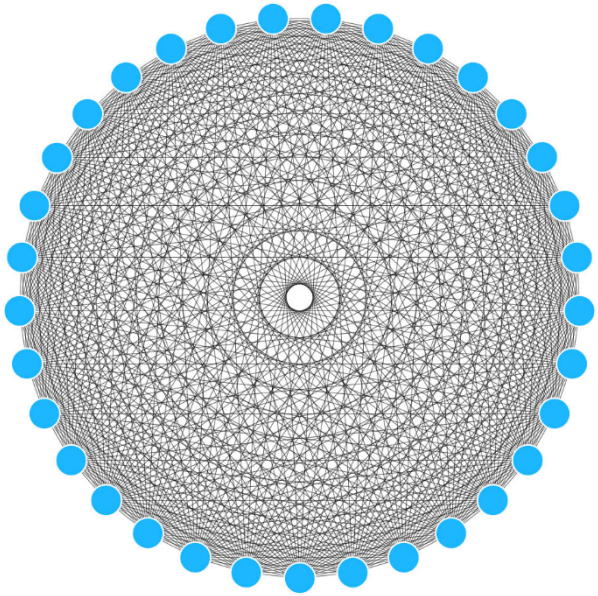
Degree distribution of real-world networks is highly heterogeneous, i.e., it can vary significantly





Real networks are sparse

- Complete graph



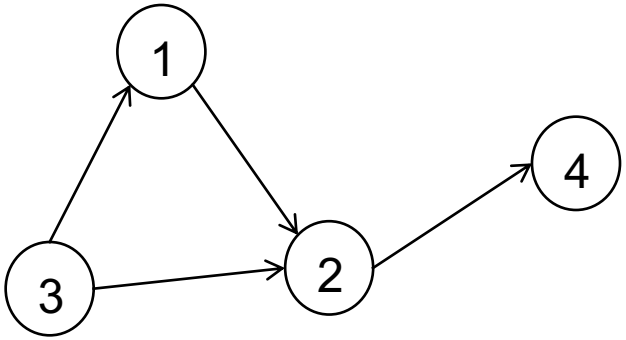
- Real network

$$L \ll N(N-1)/2$$

Mathematical representation of directed graphs

- Adjacency list
  - List of links

  [(1,2), (2,4), (3,1), (3,2)]



- Adjacency matrix

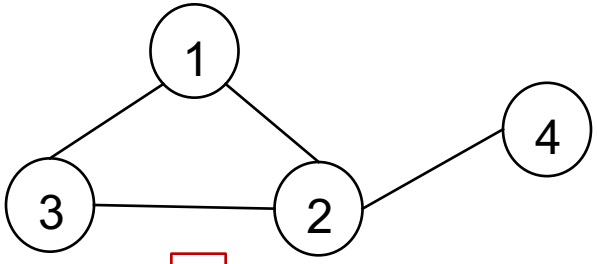
  $N \times N$ matrix $\boldsymbol{A}$ such that

  - $A_{ij} = 1$ if link (i, j) exists
  - $A_{ij} = 0$ if there is no link

$A_{ij} =$

	j			
i	0	1	0	0
	0	0	0	1
	1	1	0	0
	0	0	0	0

UNDIRECTED VS. DIRECTED GRAPHS



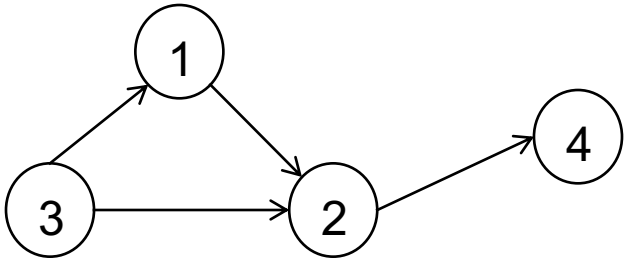
$A_{ij} =$

0	1	1	0
1	0	1	1
1	1	0	0
0	1	0	0

$$k_2 = \sum_{j=1}^{4} A_{2j} = \sum_{i=1}^{4} A_{i2} = 3$$

Symmetric $A_{ij} = A_{ji}$

$$L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij}$$



$A_{ij} =$

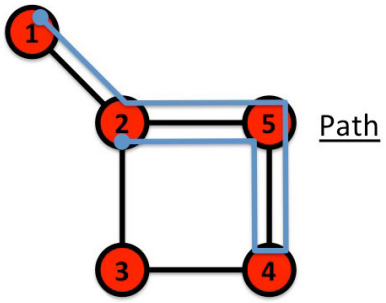
0	1	0	0
0	0	0	1
1	1	0	0
0	0	0	0

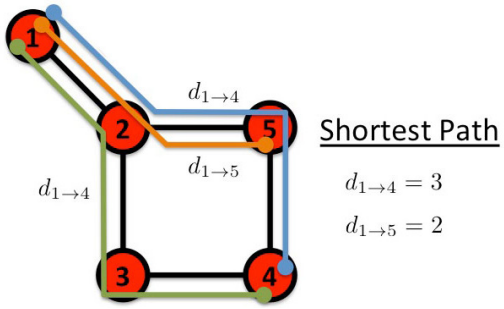
$$k_2^{in} = \sum_{j=1}^{4} A_{2j} = 2 \qquad k_2^{out} = \sum_{i=1}^{4} A_{i2} = 1$$

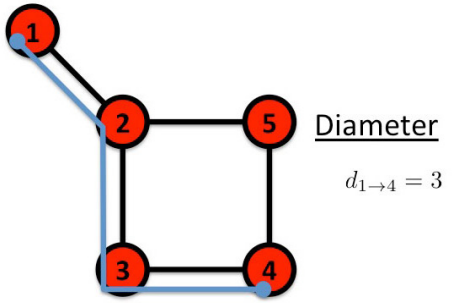
$$L = \sum_{i,j=1}^{N} A_{ij}$$

Paths and distances in networks

- Path: sequence of links (or nodes) from one node to another
- Walk: a Path of length n from one node to another, that can include repeated nodes / links (e.g., [1-2-1])
- Shortest Path: path with the shortest distance between two nodes
- Diameter: Shortest paths between most distant nodes







COMPUTING PATHS/DISTANCES

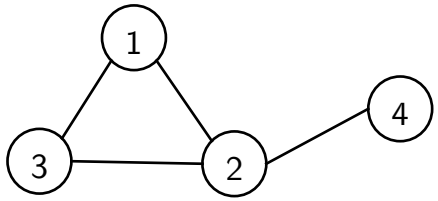
Number of *walks* N_{ij} between nodes i and j can be calculated using the adjacency matrix

- A_{ij} gives paths of length $d = 1$
- $(A^2)_{ij}$ gives #walks of length $d = 2$
- $\left(A^l\right)_{ij}$ gives #walks of length $d = l$

- The minimum l such that $\left(A^l\right)_{ij} > 0$ gives the distance (in hops) between i and j

$A_{ij} =$

0	1	1	0
1	0	1	1
1	1	0	0
0	1	0	0



$(A^2)_{ij} =$

2	1	1	1
1	3	1	0
1	1	2	1
1	0	1	1

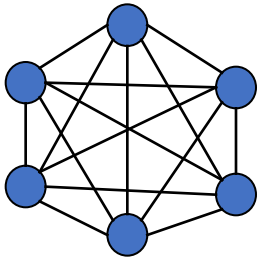
$(A^3)_{ij} =$

2	4	3	1
4	2	4	3
3	4	2	1
1	3	1	0

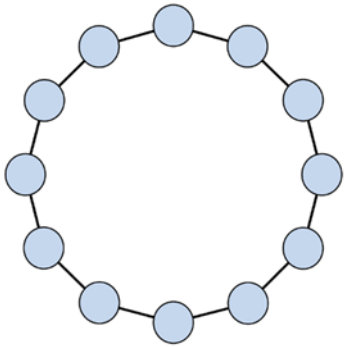
- $A_{ij} = (a_{ij})$
- $(A^2)_{34} = a_{31}a_{14} + a_{32}a_{24} + a_{33}a_{34} + a_{34}a_{44} + a_{35}a_{54} + a_{36}a_{64}$
- $a_{34}a_{44}$ is the # of walks from 3 to 1 multiplied by the # of walks from 1 to 4 → # of walks from 3 to 4 through 1
- $a_{3k}a_{k4}$ is the # of walks from 3 to k multiplied by the # of walks from k to 4 → # of walks from 3 to 4 through k
- Sum of all two-steps walks between 3 and 4

AVERAGE DISTANCE IN NETWORKS

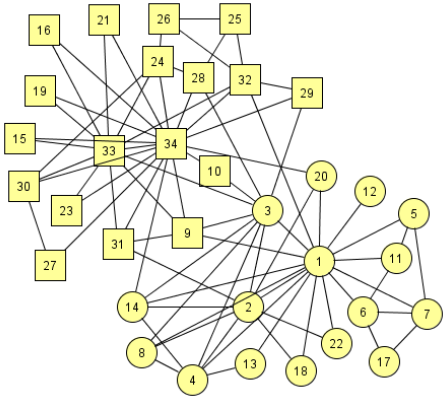
clique: $d = 1$

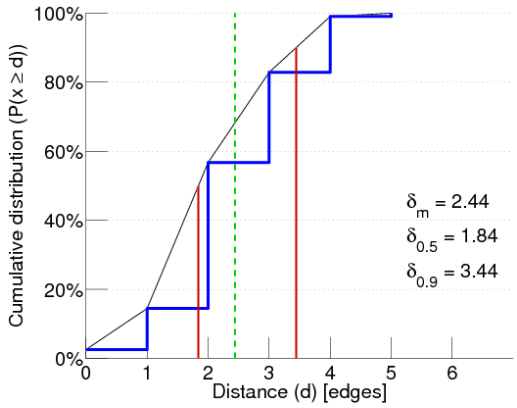


regular lattice (ring): $d = O(N)$

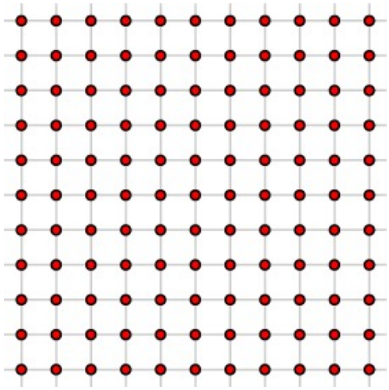


karate club friendship network: $d = 2.44$





regular lattice (square): $d = O(\sqrt{N})$



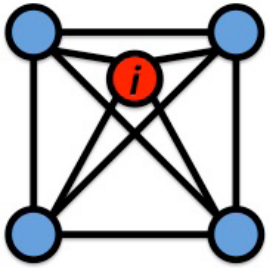
CLUSTERING

- **Clustering coefficient** captures the probability of neighbors of a given node i to be linked
- **Local clustering coefficient** of a vertex i in a graph quantifies how close its neighbors are to being a *clique*

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

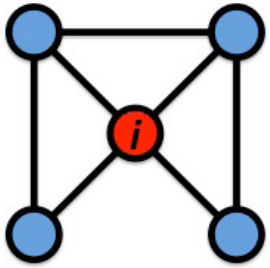
L_i is number of links between neighbors of I

In a clique: $\frac{k_i(k_i-1)}{2}$ links for k_i neighbors



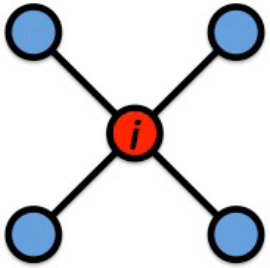
$C_i = 1$

$C = 1$



$C_i = 1/2$

$C = 9/14$



$C_i = 0$

$C = 0$

Properties of real world networks

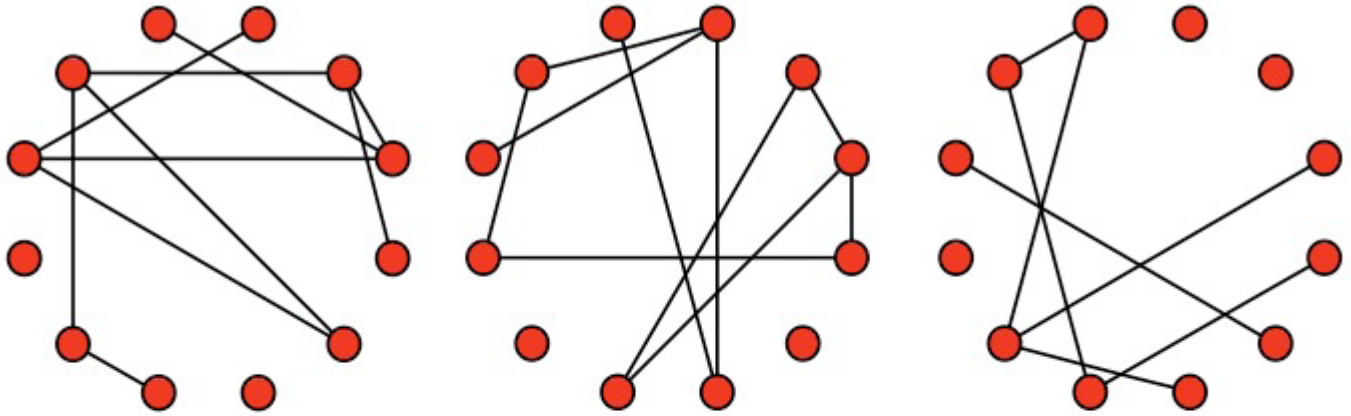
- Real networks are fundamentally different from what we'd expect
  - *Degree distribution*
    - Real networks are scale-free
  - *Average distance between nodes*
    - Real networks are small world
  - *Clustering*
    - Real networks are locally dense
- What do we expect?
  - Create a model of a network. Useful for calculating network properties and thinking about networks.

Random network model

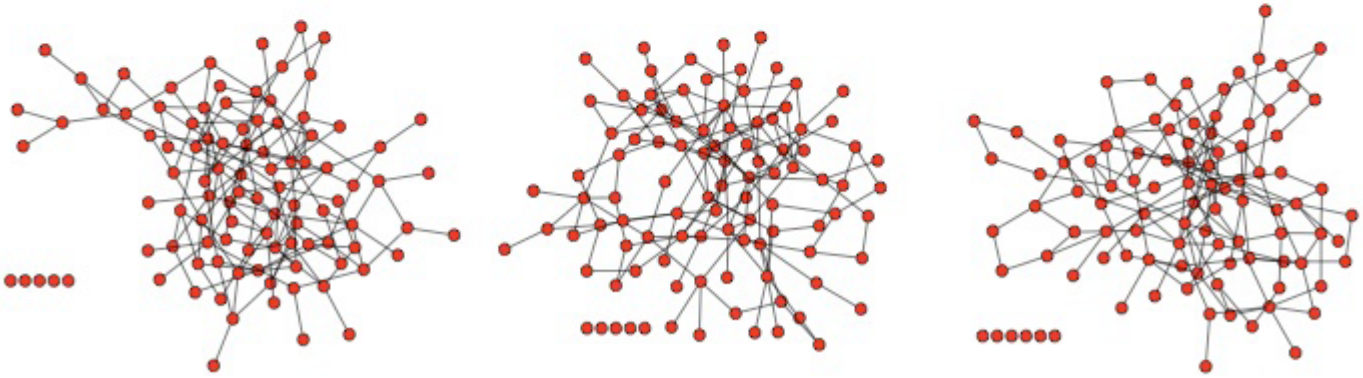
- Networks do not have a regular structure
- Given N nodes, how can we link them in a way that reproduces the observed complexity of real networks?
- Let connect nodes at random!
- *Erdos-Renyi model of a random network*
  - Given N isolated nodes
  - Select a pair of nodes. Pick a random number between 0 and 1. If the number $> p$, create a link
  - Repeat previous step for each remaining node pair
  - Average degree: $\langle k \rangle = p(N-1)$
- Easy to compute properties of random networks

RANDOM NETWORKS ARE TRULY RANDOM

N=12, p=1/6



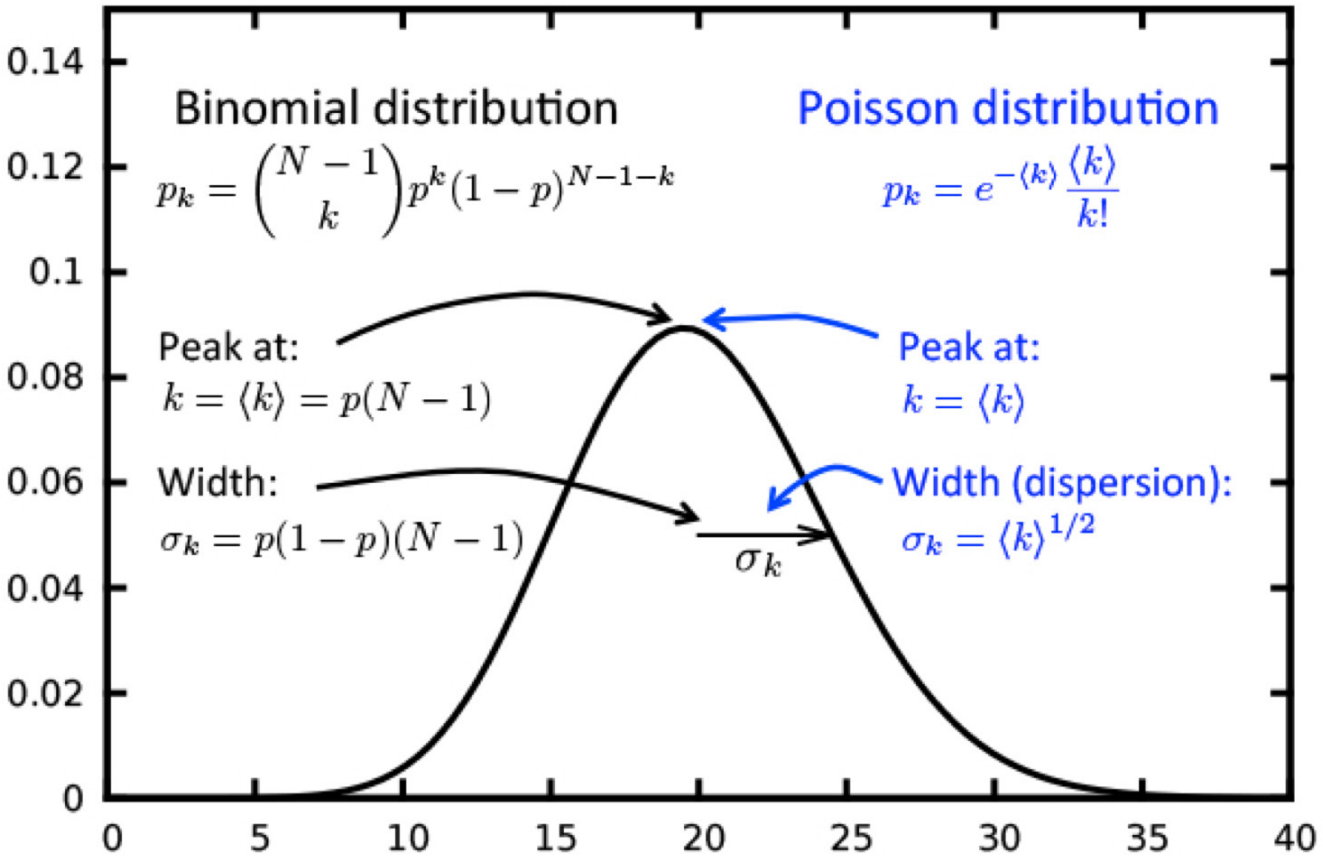
N=100, p=1/6



Average degree: $\langle k \rangle = p(N-1)$

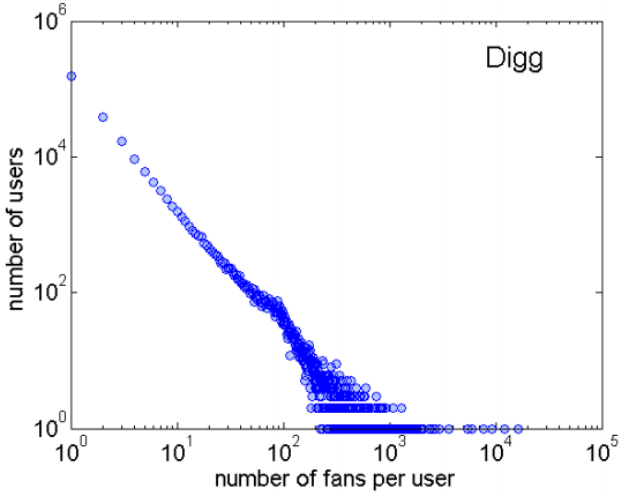
Degree distribution in random network

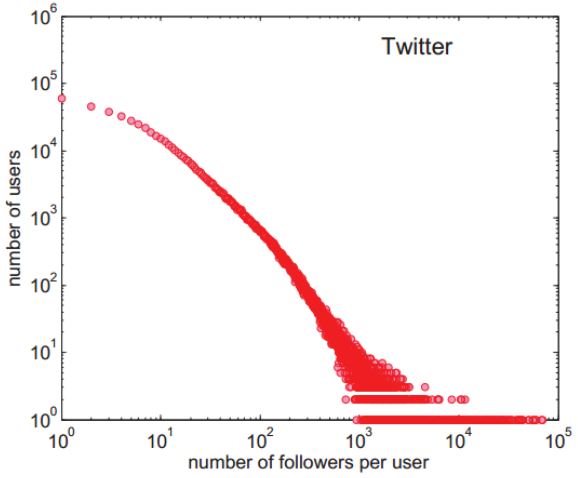
- Follows a binomial distribution
- For sparse networks, <k> << N, Poisson distribution.
  - Depends only on <k>, not network size N



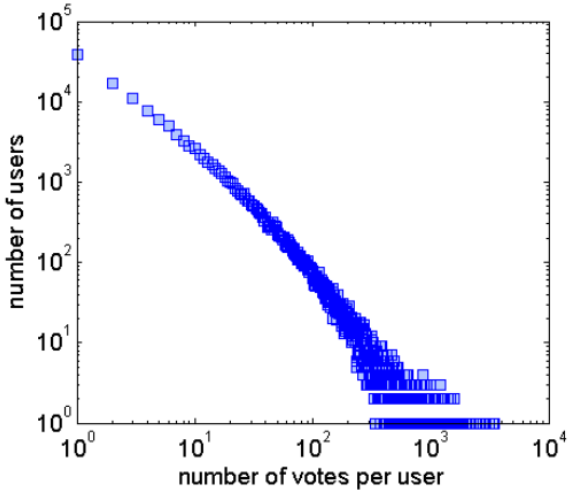
Real networks do not have Poisson degree distribution

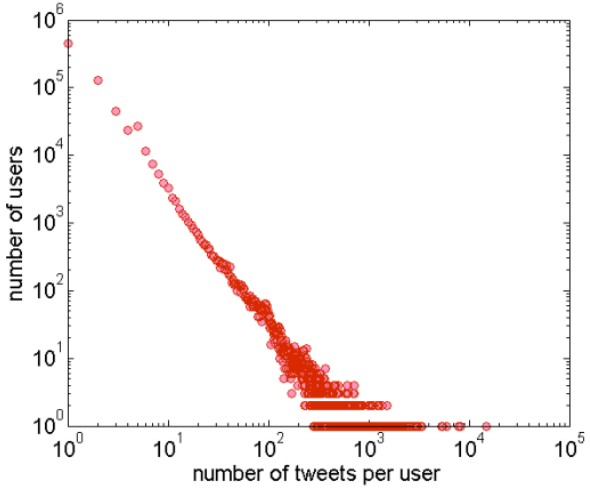
degree (followers) distribution





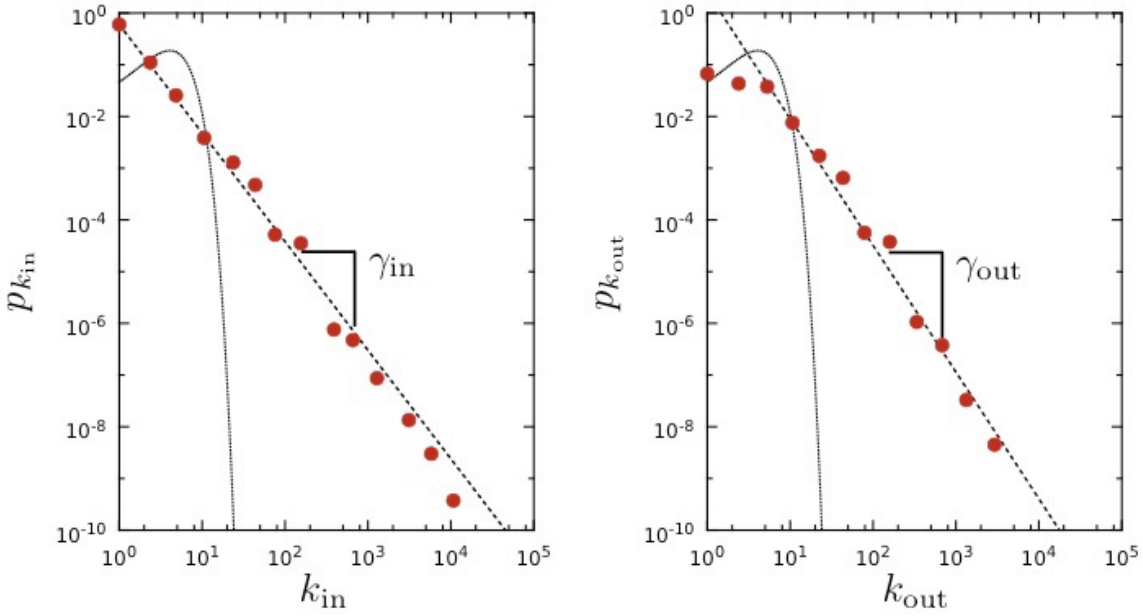
activity (num posts) distribution





Scale free property

WWW hyperlinks distribution

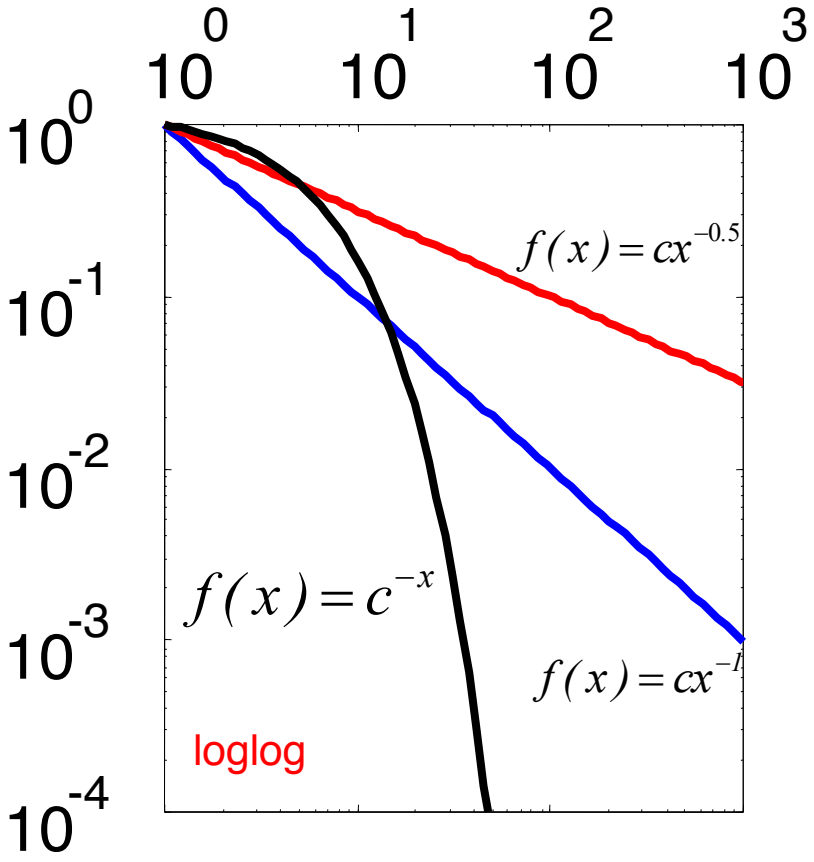


Power-law distribution: $p_k \sim k^{-\gamma}$

- Networks whose degree distribution follows a power-law distribution are called **scale free** networks
- Real network have hubs

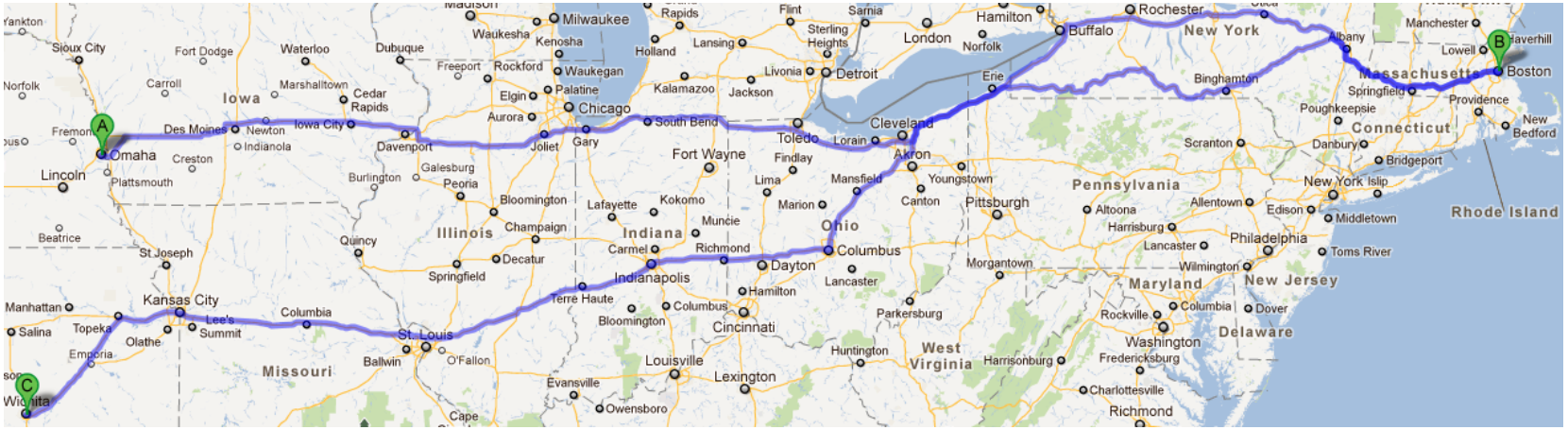
Random vs scale-free networks

Random networks and scale-free networks are very different. Differences are apparent when degree distribution is plotted on log scale.



The Milgram experiment

- In 1960's, Stanley Milgram asked 160 randomly selected people in Kansas and Nebraska to deliver a letter to a stock broker in Boston.
  - Rule: can only forward the letter to a friend who is more likely to know the target person
- How many steps would it take?



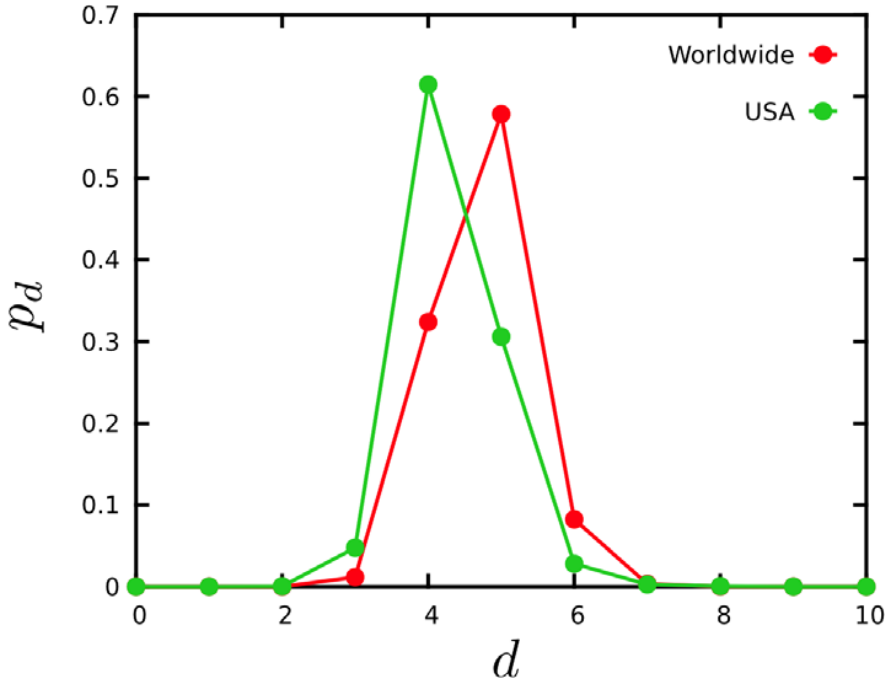
The Milgram experiment

- Within a few days the first letter arrived, passing through only two links.
- Eventually 42 of the 160 letters made it to the target, some requiring close to a dozen intermediates.
- The median number of steps in completed chains was 5.5

→"**six degrees of separation**"

FACEBOOK IS A VERY SMALL WORLD

- Ugander et al. directly measured distances between nodes in the Facebook social graph (May 2011)
  - 721 million active users
  - 68 billion symmetric friendship links
  - the average distance between the users was 4.74

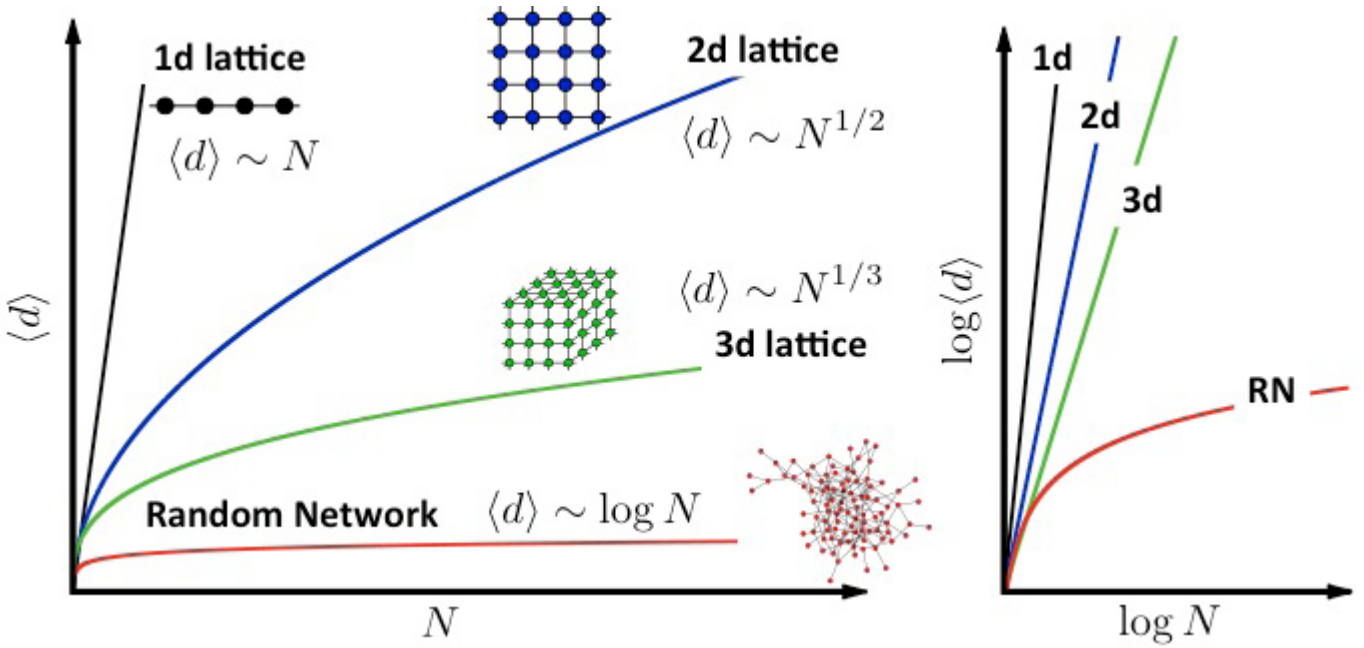


Small world property

- *Distance between any two nodes in a network is surprisingly short*
  - "six degrees of separation": you can reach any other individual in the world through a short sequence of intermediaries
- What is small?
  - Consider a random network with average degree $\langle k \rangle$
  - Expected number of nodes a distance d is $N(d) \sim \langle k \rangle^d$
  - Diameter $d_{max} \sim \log N / \log \langle k \rangle$
  - *Random networks are small*

What is it surprising?

- **Regular lattices (e.g., physical geography) do *not* have the small world property**
  - Distances grow *polynomially* with system size
  - In networks, distances *grow logarithmically* with network size



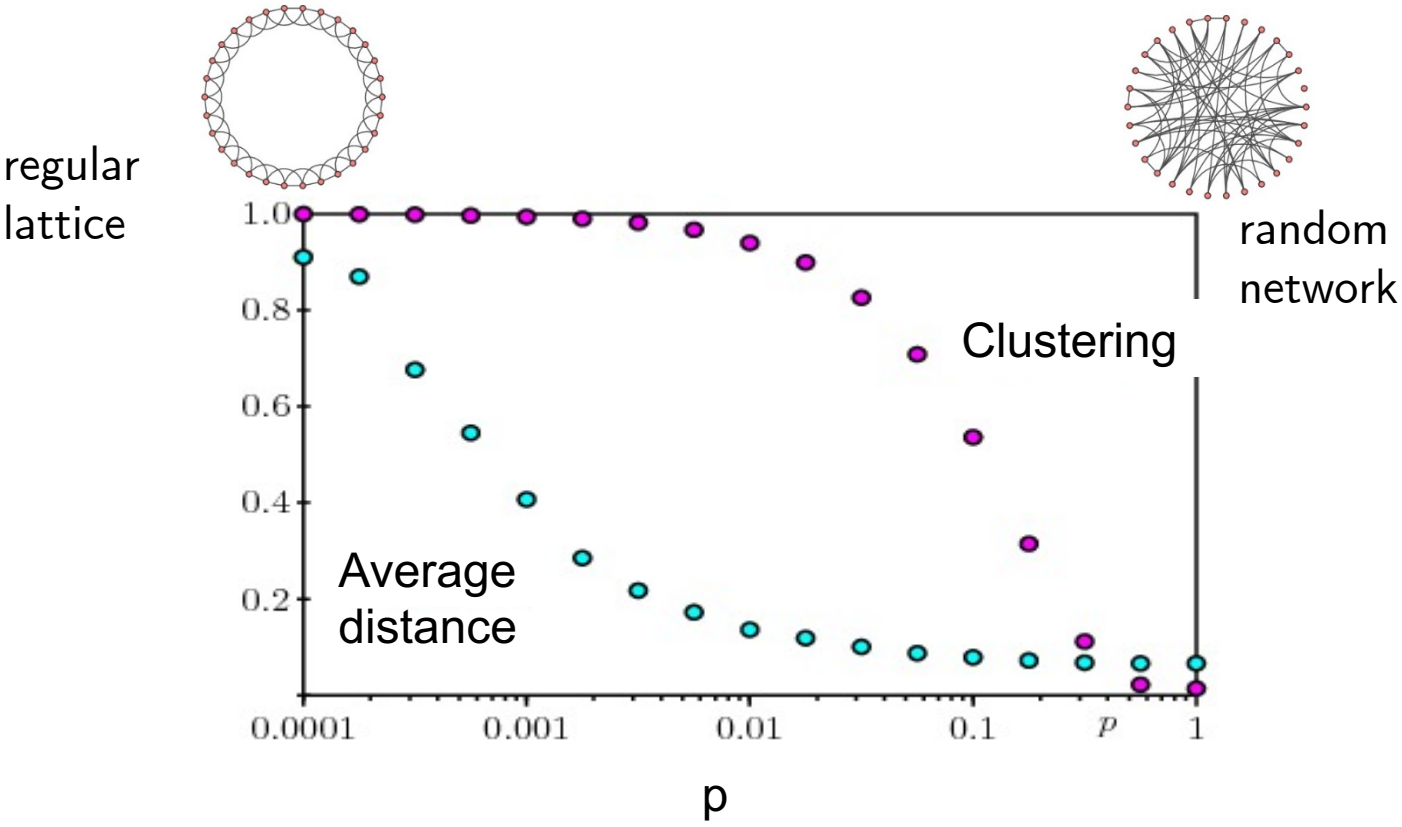
SMALL WORLD EFFECT IN RANDOM NETWORKS

Watts-Strogatz model

- Start with a regular lattice, e.g., a ring where each node is connected to immediate and next neighbors.
  - Local clustering is $C = 3/4$
- With probability p, rewire link to a randomly chosen node
  - For small p, clustering remains high, but diameter shrinks
  - For large p, becomes random network

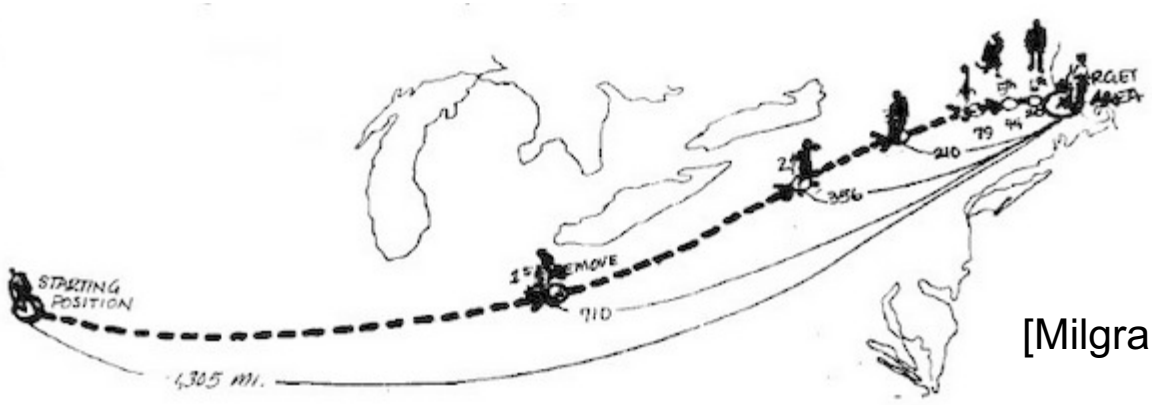
SMALL WORLD NETWORKS

- **Small world networks** constructed using Watts-Strogatz model have small average distance and high clustering, just like real networks
- *Long-distance links*, joining distant local clusters



Social networks are searchable

- Milgram experiments showed that
  - Short chains exist!
  - People can find them!
    - Using only local knowledge (who their friends are, their location and profession)
    - How are short chains discovered with this limited information?
    - Hint: geographic information?

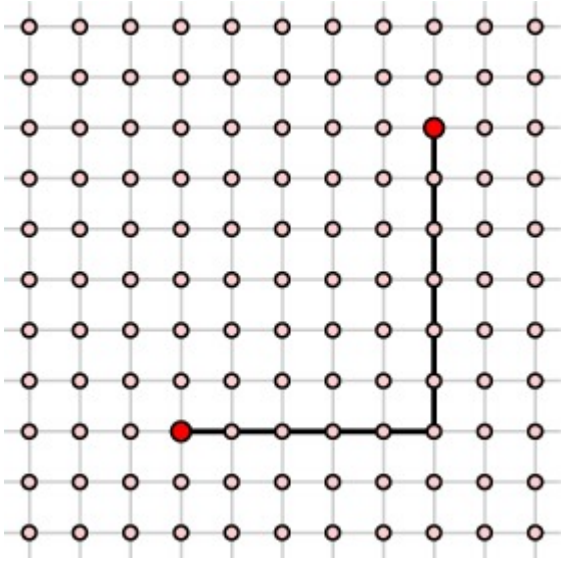


[Milgram]

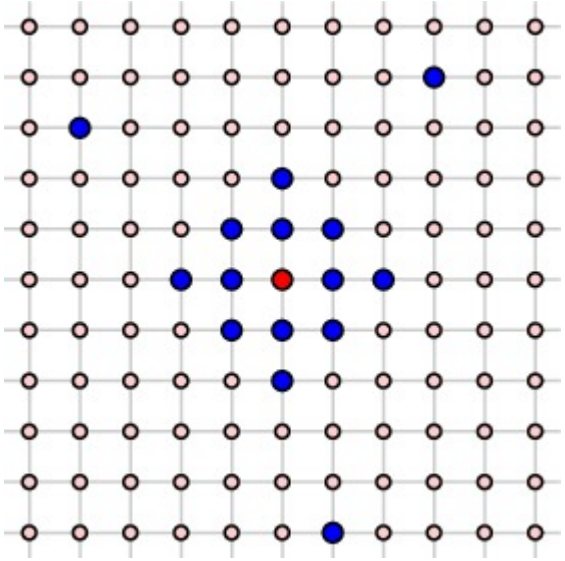
Kleinberg model of geographic links

- Incorporate geographic distance in the distribution of links

Distance between nodes is d

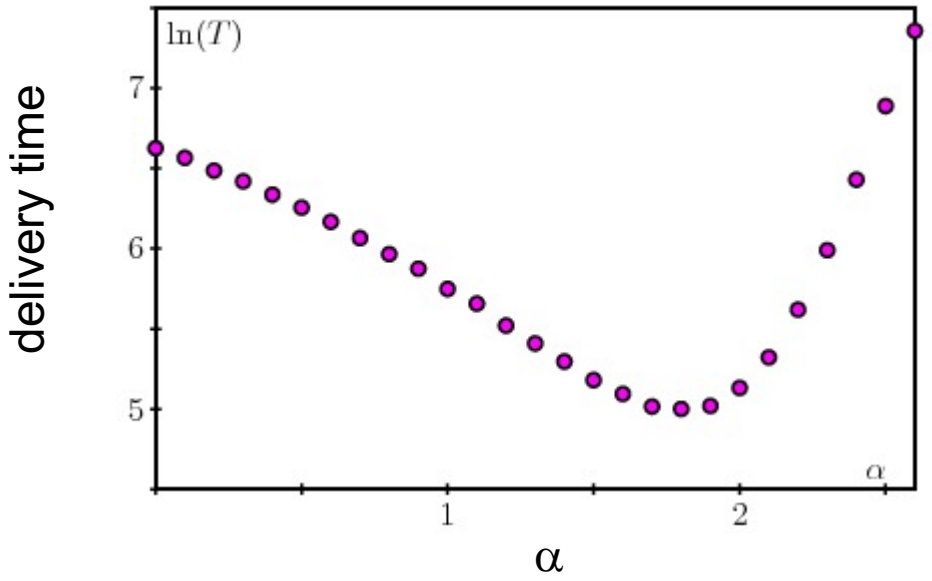


Link to all nodes within distance *r*, then add *q* long range links with probability d^{-a}



How does this affect short chains?

- Simulate Milgram experiment
  - at each time step, a node selects a friend who is closer to the target (in lattice space) and forwards the letter to it
    - Each node uses only local information about its own social network and not the entire structure of the network
  - delivery time T is the time for the letter to reach the target

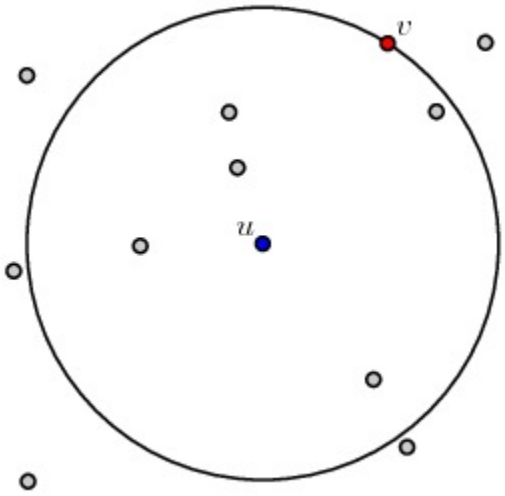


Kleinberg's analysis

- Network is only searchable when a=2
  - i.e., probability to form a link drops as square of distance
  - Average delivery time is at most proportional to (log N)2
- For other values of a, the average chain length produced by search algorithm is at least Nb.

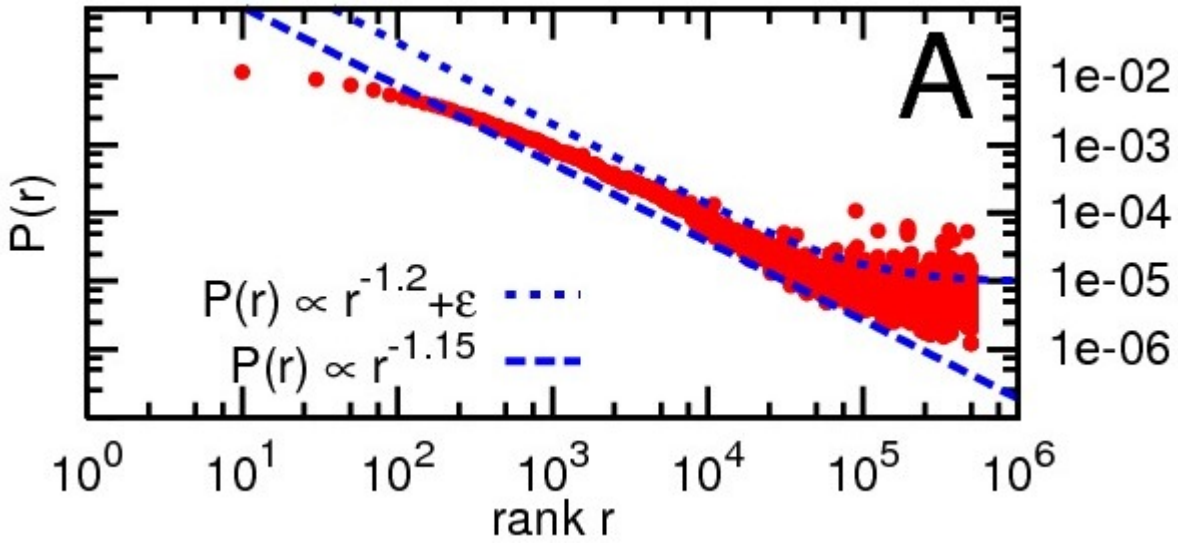
Does this hold for real networks?

- Liben-Nowell et al. tested Kleinberg's prediction for the LiveJournal network of 1M+ bloggers
  - Blogger's geographic information in profile
  - How does friendship probability in LiveJournal network depend on distance between people?
- People are not uniformly distributed spatially
  - Coasts, cities are denser



Use rank, instead of distance d(u,v)
$rank_u(v) = 6$

Since $rank_u(v) \sim d(u,v)^2$, and link probability $Pr(u \rightarrow v) \sim d(u,v)^{-2}$, we expect that $Pr(u \rightarrow v) \sim 1/rank_u(v)$

LiveJournal is a searchable network

- Probability that a link exists between two people as a function of the rank between them
  - LiveJournal is a rank-based network → it is searchable