Measuring Networks and the Random Graph Model

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

How the Class Fits Together

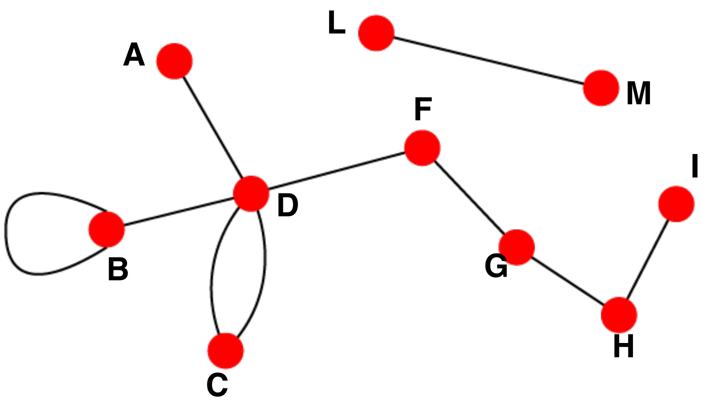
Measurements	Models	Algorithms
Small diameter, Edge clustering	Erdös-Renyi model, Small-world model	Decentralized search
Patterns of signed edge creation	Structural balance, Theory of status	Models for predicting edge signs
Viral Marketing, Blogosphere, Memetracking	Independent cascade model, Game theoretic model	Influence maximization, Outbreak detection, LIM
Scale-Free	Preferential attachment, Copying model	PageRank, Hubs and authorities
Densification power law, Shrinking diameters	Microscopic model of evolving networks	Link prediction, Supervised random walks
Strength of weak ties, Core-periphery	Kronecker Graphs	Community detection: Girvan-Newman, Modularity

Choice of the proper network representation of a given system determines our ability to use networks successfully

Directed vs. Undirected Graphs

Undirected

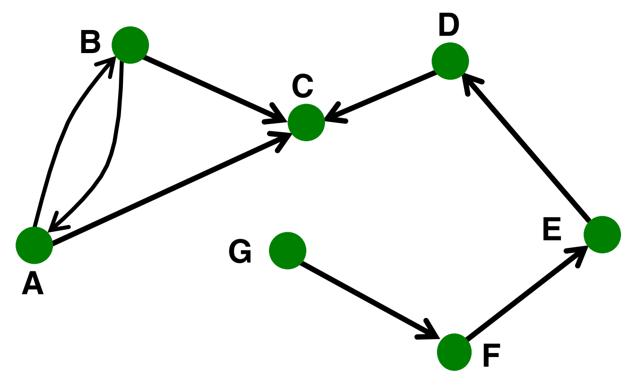
- **Links:** undirected (symmetrical, reciprocal)



- **Examples:**
  - Collaborations
  - Friendship on Facebook

Directed

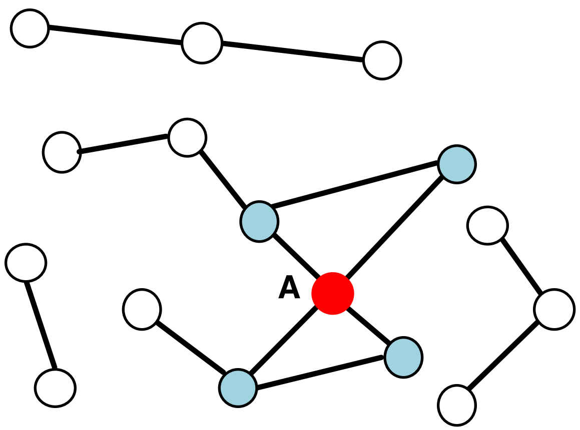
- **Links:** directed (arcs)



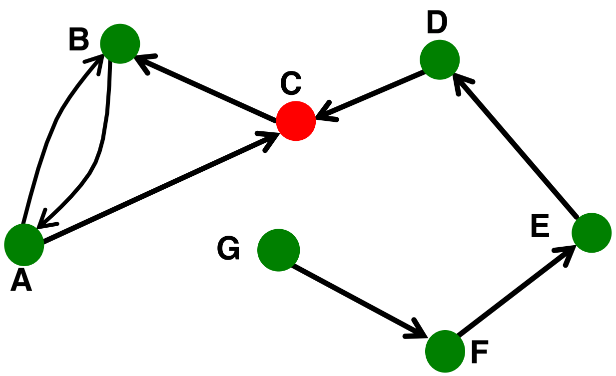
- **Examples:**
  - Phone calls
  - Following on Twitter

Node Degrees

Undirected



Directed



Source: Node with $k^{in} = 0$

Sink: Node with $k^{out} = 0$

Node degree, k_i: the number of edges adjacent to node i

$$k_A = 4$$

Avg. degree: $\bar{k} = \langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2E}{N}$

In directed networks we define an **in-degree** and **out-degree.**

The (total) degree of a node is the sum of in- and out-degrees.

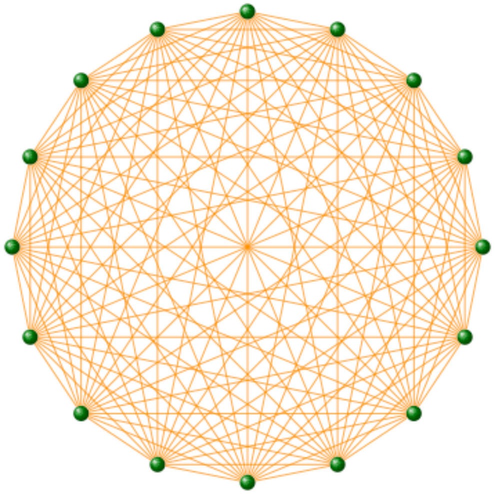
$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

$$\bar{k} = \frac{E}{N} \qquad \overline{k^{in}} = \overline{k^{out}}$$

Complete Graph

The **maximum number of edges** in an undirected graph on N nodes is

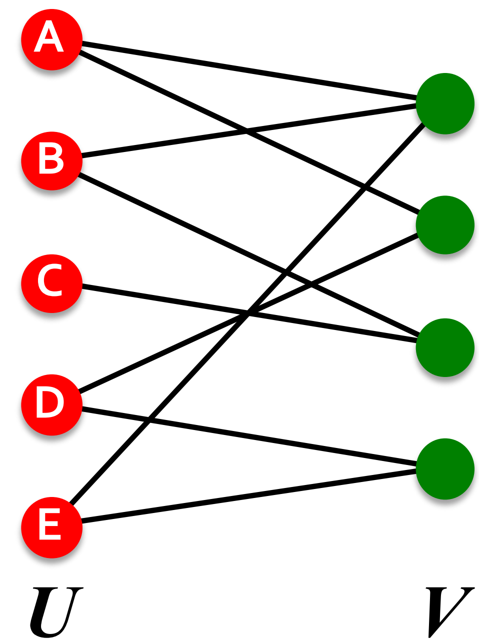
$$E_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

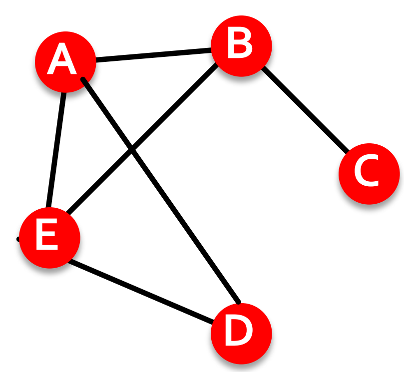


An undirected graph with the number of edges $E = E_{max}$ is called a **complete graph**, and its average degree is N-1

Bipartite Graph

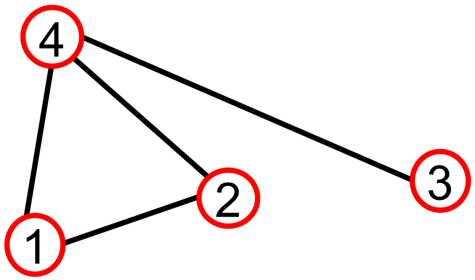
- **Bipartite graph** is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V; that is, U and V are **independent sets**
- **Examples:**
  - Authors-to-papers (they authored)
  - Actors-to-Movies (they appeared in)
  - Users-to-Movies (they rated)
- **"Folded" networks:**
  - Author collaboration networks
  - Movie co-rating networks

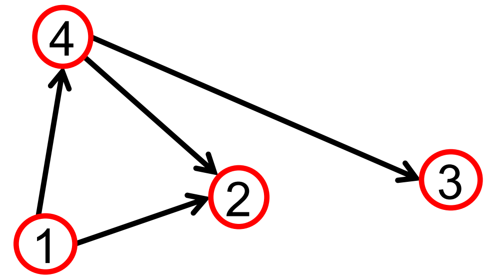




Folded version of the graph above

Representing Graphs: Adjacency Matrix





$A_{ij} = 1$ if there is a link from node ***i*** to node ***j***

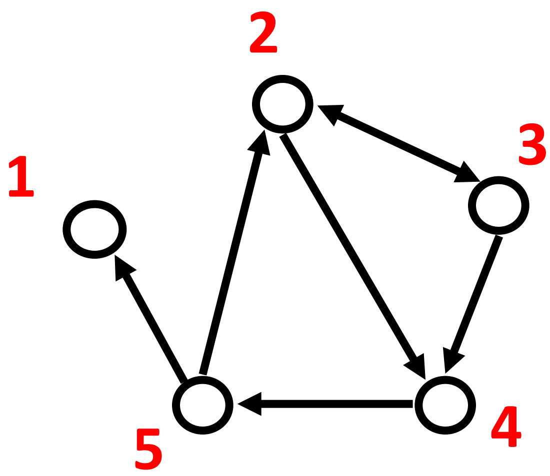
$A_{ij} = 0$ otherwise

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Note that for a directed graph (right) the matrix is not symmetric.

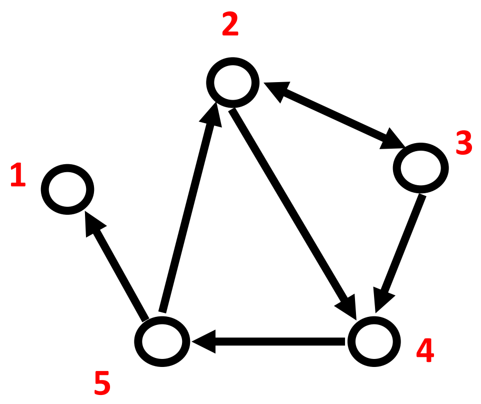
Representing Graphs: Edge list

- **Represent graph as a set of edges:**
  - (2, 3)
  - (2, 4)
  - (3, 2)
  - (3, 4)
  - (4, 5)
  - (5, 2)
  - (5, 1)



Representing Graphs: Adjacency list

- **Adjacency list:**
  - Easier to work with if network is
    - Large
    - Sparse
  - Allows us to quickly retrieve all neighbors of a given node
    - 1:
    - 2: 3, 4
    - 3: 2, 4
    - 4: 5
    - 5: 1, 2



Networks are Sparse Graphs

Most real-world networks are sparse

$$E \ll E_{max} \quad (\text{or } \bar{k} \ll N-1)$$

WWW (Stanford-Berkeley):	N=319,717	$\langle k \rangle$=9.65
Social networks (LinkedIn):	N=6,946,668	$\langle k \rangle$=8.87
Communication (MSN IM):	N=242,720,596	$\langle k \rangle$=11.1
Coauthorships (DBLP):	N=317,080	$\langle k \rangle$=6.62
Internet (AS-Skitter):	N=1,719,037	$\langle k \rangle$=14.91
Roads (California):	N=1,957,027	$\langle k \rangle$=2.82
Proteins (S. Cerevisiae):	N=1,870	$\langle k \rangle$=2.39

(Source: *Leskovec et al., Internet Mathematics, 2009*)

Consequence: Adjacency matrix is filled with zeros!

(**Density of the matrix (E/N^2):** WWW=1.51×10^{-5}, MSN IM = 2.27×10^{-8})

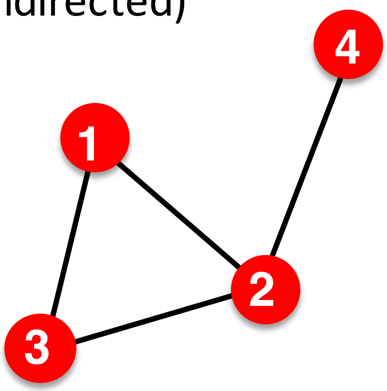
Edge Attributes

Possible options:

- Weight (e.g. frequency of communication)
- Ranking (best friend, second best friend…)
- Type (friend, relative, co-worker)
- Sign: Friend vs. Foe, Trust vs. Distrust
- Properties depending on the structure of the rest of the graph: number of common friends

More Types of Graphs

- **Unweighted**
  (undirected)



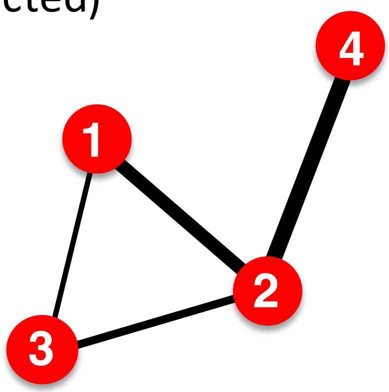
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \bar{k} = \frac{2E}{N}$$

Examples: Friendship, Hyperlink

- **Weighted**
  (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

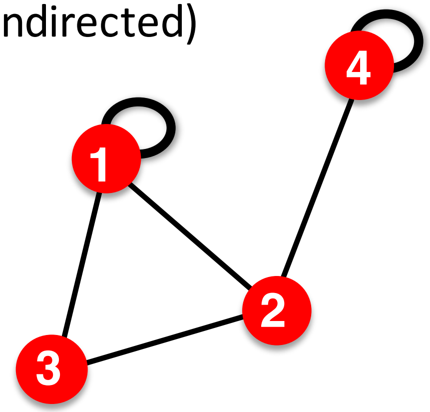
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \bar{k} = \frac{2E}{N}$$

Examples: Collaboration, Internet, Roads

More Types of Graphs

- **Self-edges (self-loops)**
  (undirected)



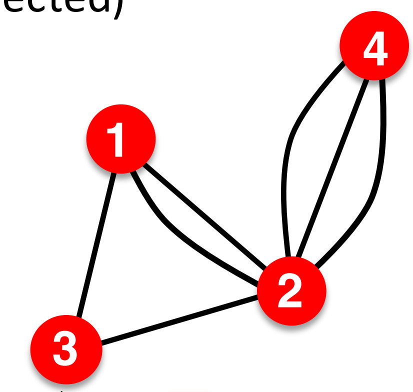
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2} \sum_{i,j=1, i \neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii}$$

Examples: Proteins, Hyperlinks

- **Multigraph**
  (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

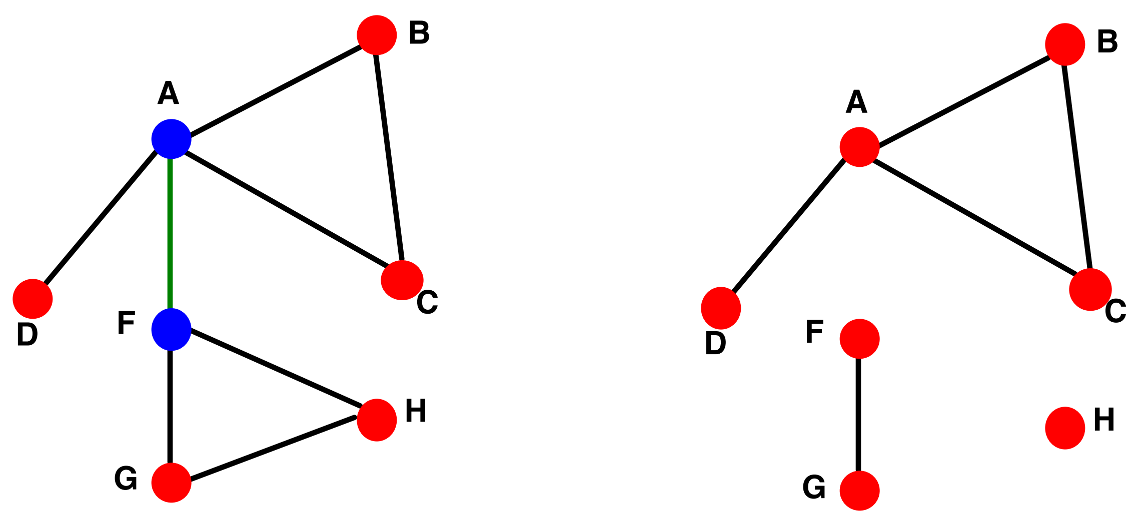
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2} \sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \bar{k} = \frac{2E}{N}$$

Examples: Communication, Collaboration

Connectivity of Undirected Graphs

- **Connected (undirected) graph:**
  - Any two vertices can be joined by a path
- A disconnected graph is made up by two or more connected components



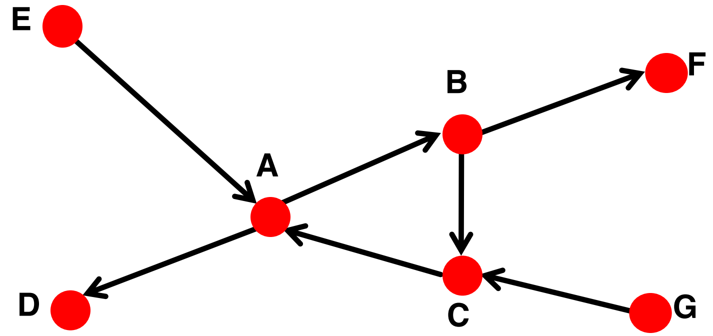
Largest Component:
Giant Component

Isolated node (node H)

Bridge edge: If we erase it, the graph becomes disconnected.
Articulation point: If we erase it, the graph becomes disconnected.

Connectivity of Directed Graphs

- **Strongly connected directed graph**
  - has a path from each node to every other node and vice versa (e.g., A-B path and B-A path)
- **Weakly connected directed graph**
  - is connected if we disregard the edge directions



Graph on the left is connected but not strongly connected (e.g., there is no way to get from F to G by following the edge directions).

Network Representations

WWW >> directed multigraph with self-edges

Facebook friendships >> undirected, unweighted

Citation networks >> unweighted, directed, acyclic

Collaboration networks >> undirected multigraph or weighted graph

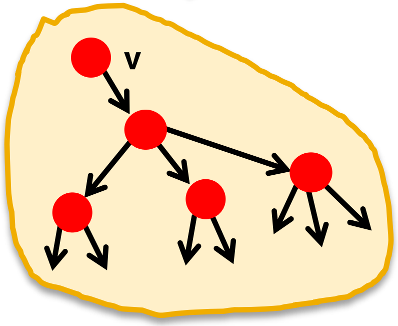
Mobile phone calls >> directed, (weighted?) multigraph

Protein Interactions >> undirected, unweighted with self-interactions

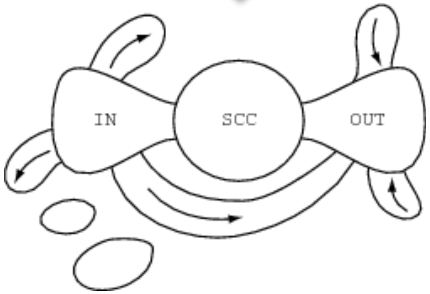
Web as a Graph

Structure of the Web

- **Today we will talk about observations and models for the Web graph:**
  - **1)** We will take a real system: **the Web**
  - **2)** We will represent it as a **directed graph**
  - **3)** We will use the language of graph theory
    - **Strongly Connected Components**
  - **4)** We will design a **computational experiment:**
    - Find In- and Out-components of a given node v
  - **5) We will learn something about the structure of the Web: BOWTIE!**



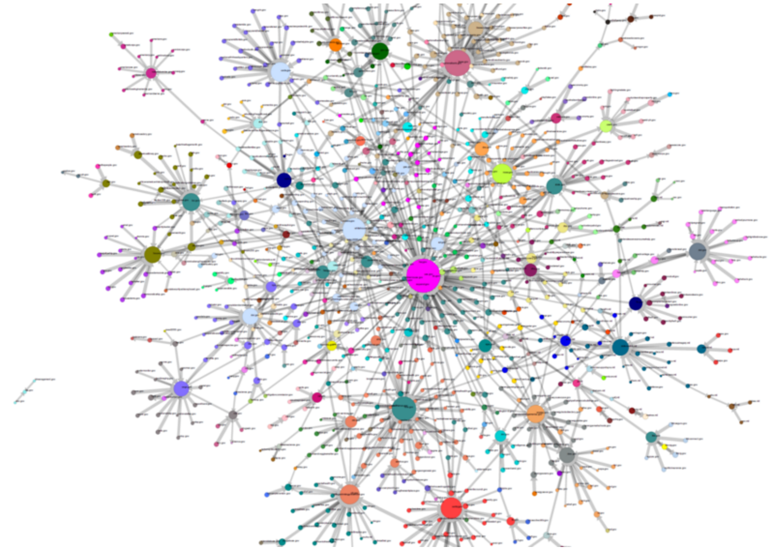
Out(v)



The Web as a Graph

Q: What does the Web "look like" at a global level?

- **Web as a graph:**
  - Nodes = web pages
  - Edges = hyperlinks
  - **Side issue:** What is a node?
    - Dynamic pages created on the fly
    - "dark matter" – inaccessible database generated pages



The Web as a Graph

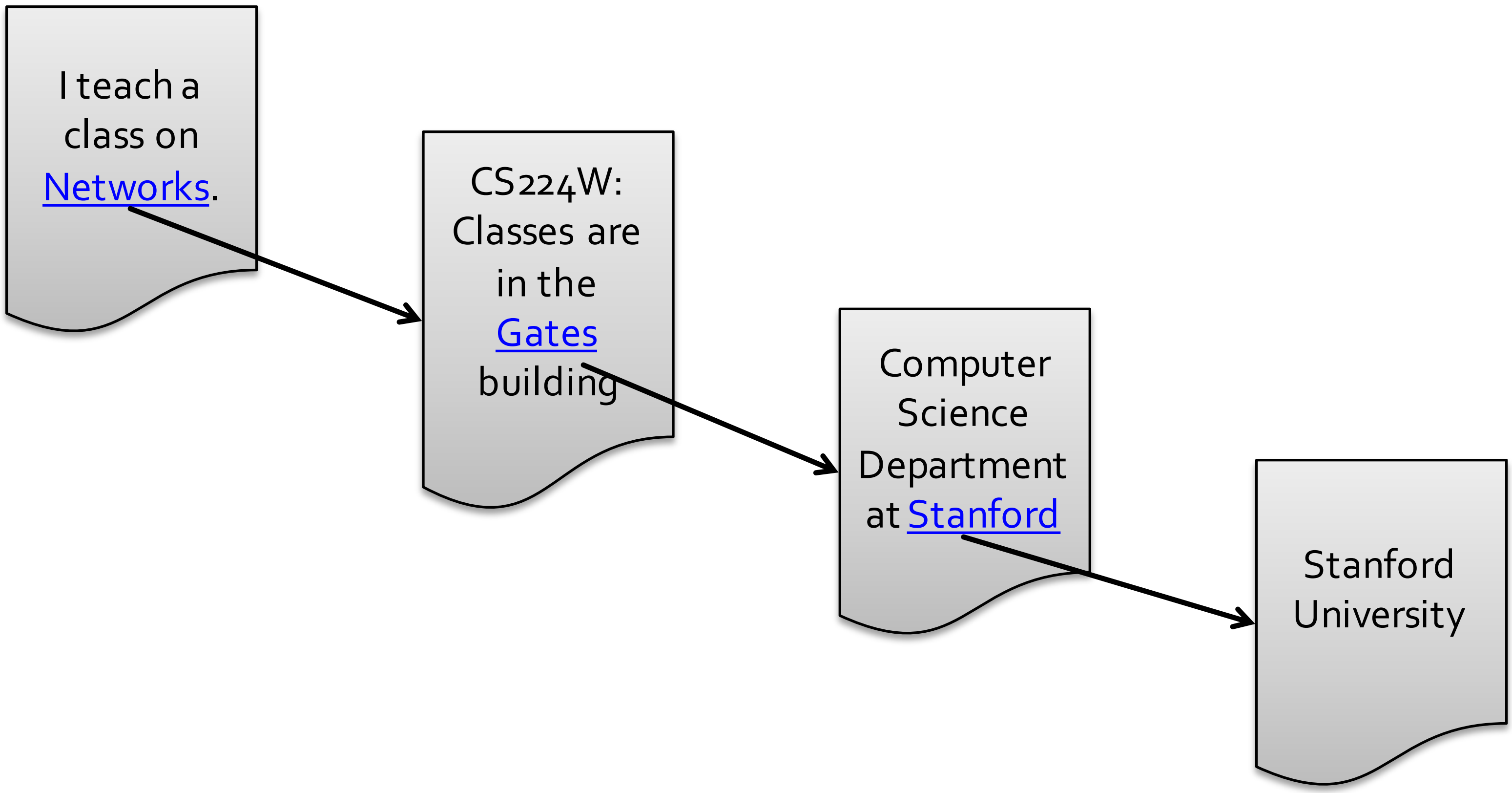
I teach a class on Networks.

CS224W: Classes are in the Gates building

Computer Science Department at Stanford

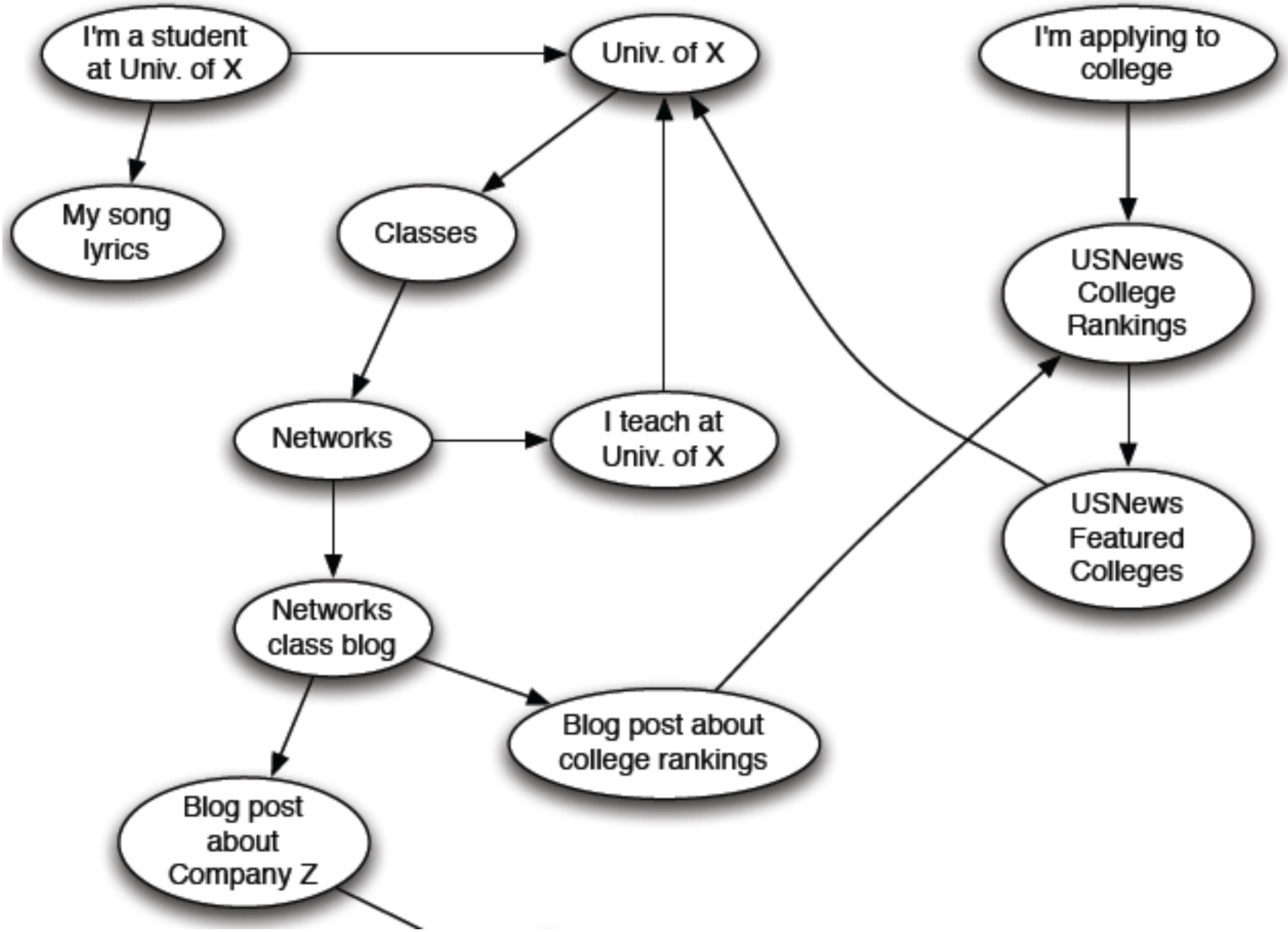
Stanford University

The Web as a Graph

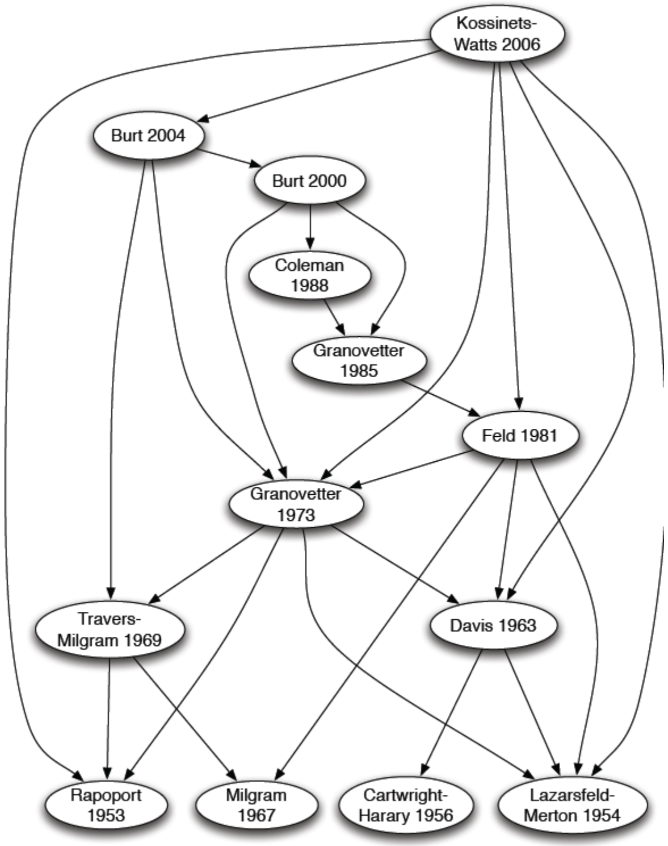


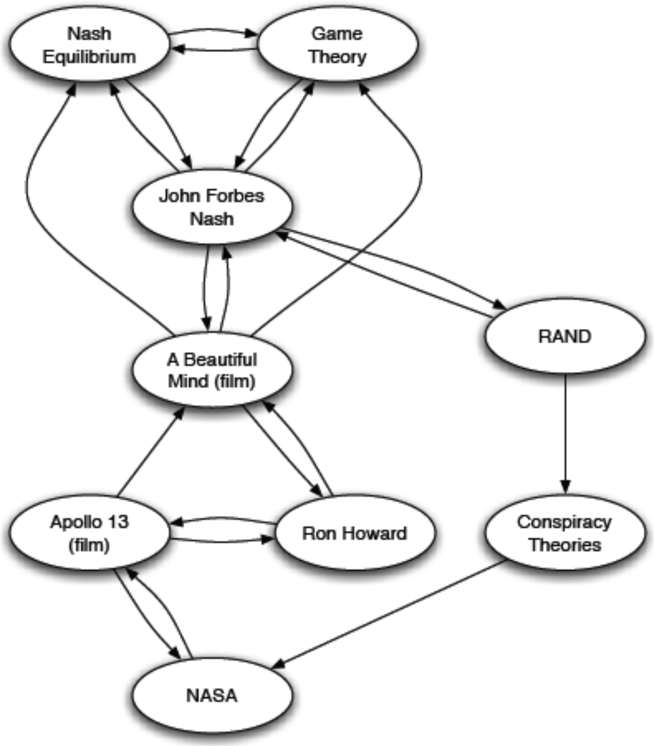
- In early days of the Web links were **navigational**
- Today many links are **transactional**

The Web as a Directed Graph



Other Information Networks





Citations

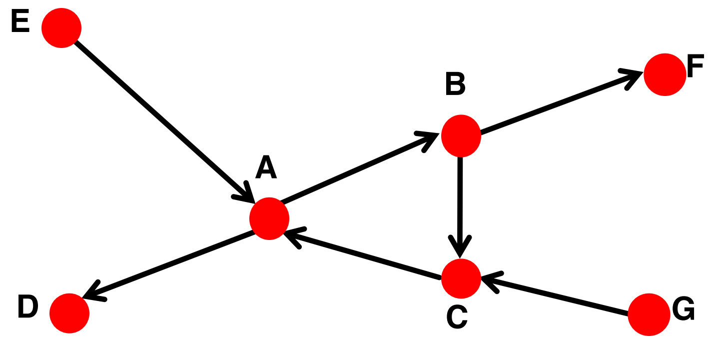
References in an Encyclopedia

What Does the Web Look Like?

- **How is the Web linked?**
- **What is the "map" of the Web?**

Web as a <u>directed graph</u> [Broder et al. 2000]:

- Given node v, what can v reach?
- What other nodes can reach v?

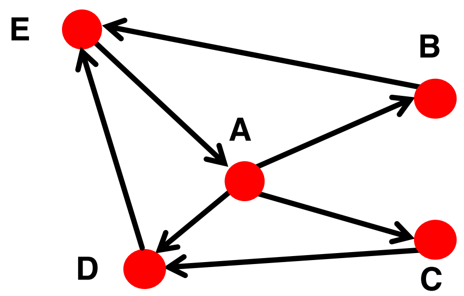


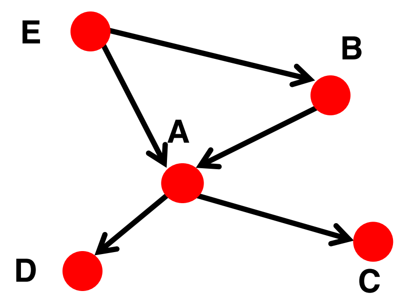
$$In(v) = \{w \mid w \text{ can reach } v\}$$
$$Out(v) = \{w \mid v \text{ can reach } w\}$$

For example:
In(A) = {A,B,C,E,G}
Out(A)={A,B,C,D,F}

Directed Graphs

- **Two types of directed graphs:**
  - **Strongly connected:**
    - Any node can reach any node via a directed path
      $In(A)=Out(A)=\{A,B,C,D,E\}$
  - **Directed Acyclic Graph (DAG):**
    - Has no cycles: if $\boldsymbol{u}$ can reach $\boldsymbol{v}$, then $\boldsymbol{v}$ cannot reach $\boldsymbol{u}$

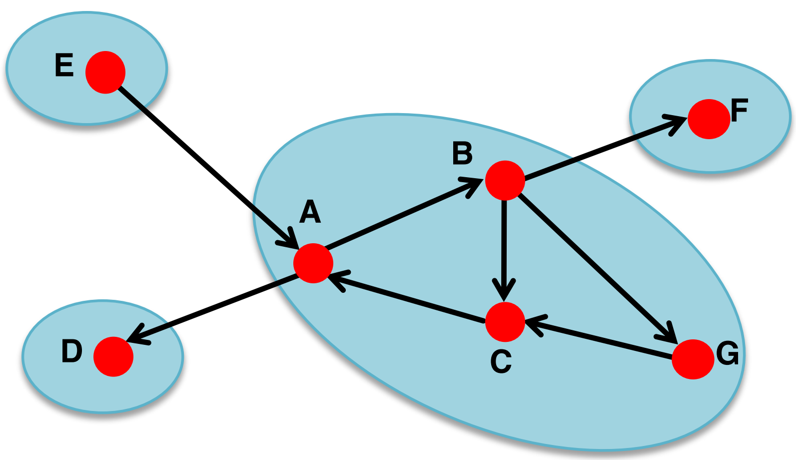




- **Any directed graph can be expressed in terms of these two types!**

Strongly Connected Component

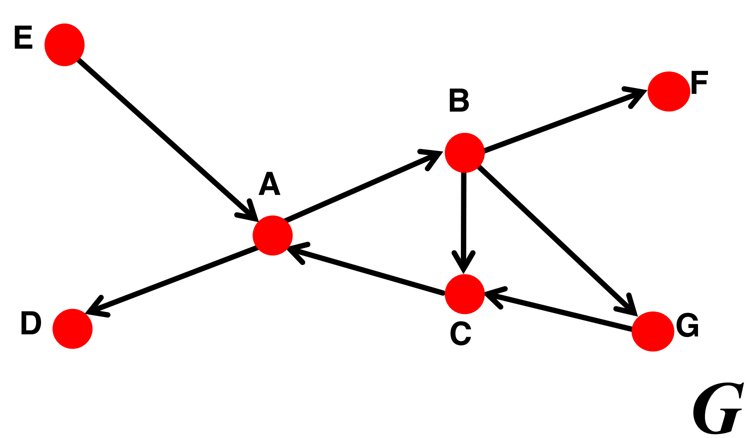
- **A Strongly Connected Component (SCC)** is a set of nodes S so that:
  - Every pair of nodes in S can reach each other
  - There is no larger set containing S with this property



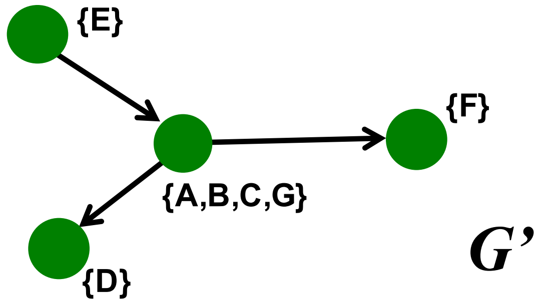
Strongly connected components of the graph: {A,B,C,G}, {D}, {E}, {F}

Strongly Connected Component

- **Fact:** **Every directed graph is a DAG on its SCCs**
  - **(1)** SCCs partitions the nodes of G
    - That is, each node is in exactly one SCC
  - **(2)** If we build a graph G' whose nodes are SCCs, and with an edge between nodes of G' if there is an edge between corresponding SCCs in G, then G' is a DAG

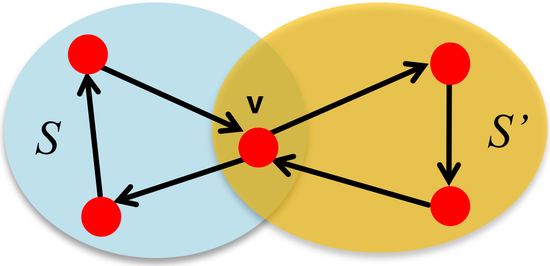


(1) Strongly connected components of graph G: {A,B,C,G}, {D}, {E}, {F}
(2) G' is a DAG:



Proof of (1)

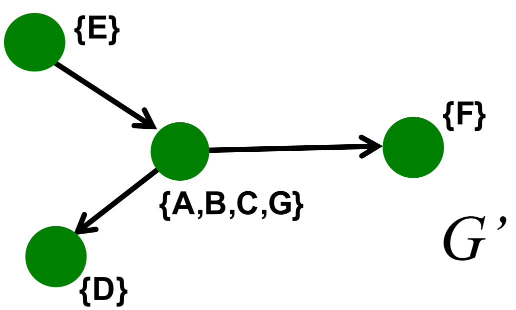
- **Claim: SCCs partition nodes of G.**
  - This means: Each node is member of exactly 1 SCC
- Proof by contradiction:
  - Suppose there exists a node v which is a member of two SCCs S and S'

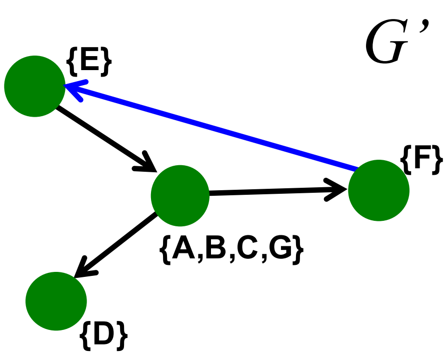


  - But then $S \cup S'$ is one large SCC!
    - **Contradiction:** By definition SCC is a maximal set with the SCC property, so S and S' are not two SCCs.

Proof of (2)

- **Claim: *G'* (graph of SCCs) is a DAG.**
  - This means: ***G'*** has no cycles
- Proof by contradiction:
  - Assume ***G'*** is <u>not</u> a DAG
  - Then ***G'*** has a directed cycle
  - Now all nodes on the cycle are mutually reachable, and all are part of the same SCC
  - But then ***G'*** is not a graph of connections between SCCs (SCCs are defined as maximal sets)
    - **Contradiction!**



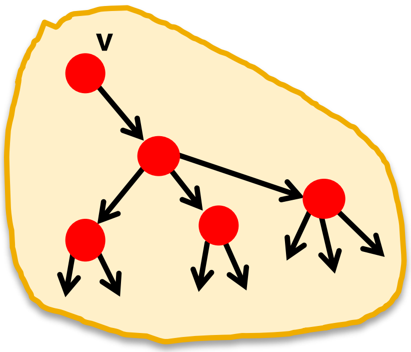


Now {A,B,C,G,E,F} is a SCC!

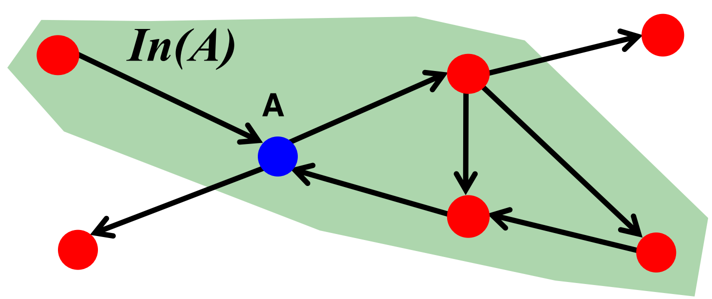
Graph Structure of the Web

- **Goal:** **Take a large snapshot of the Web and try to understand how its SCCs "fit together" as a DAG**
- **Computational issue:**
  - Want to find a SCC containing node v?
  - **Observation:**
    - $Out(v)$ ... nodes that can be reached from v
    - **SCC containing v is:** $Out(v) \cap In(v)$

$$= Out(v,G) \cap Out(v,\overline{G}),$$ where $\overline{G}$ is G with all edge directions flipped

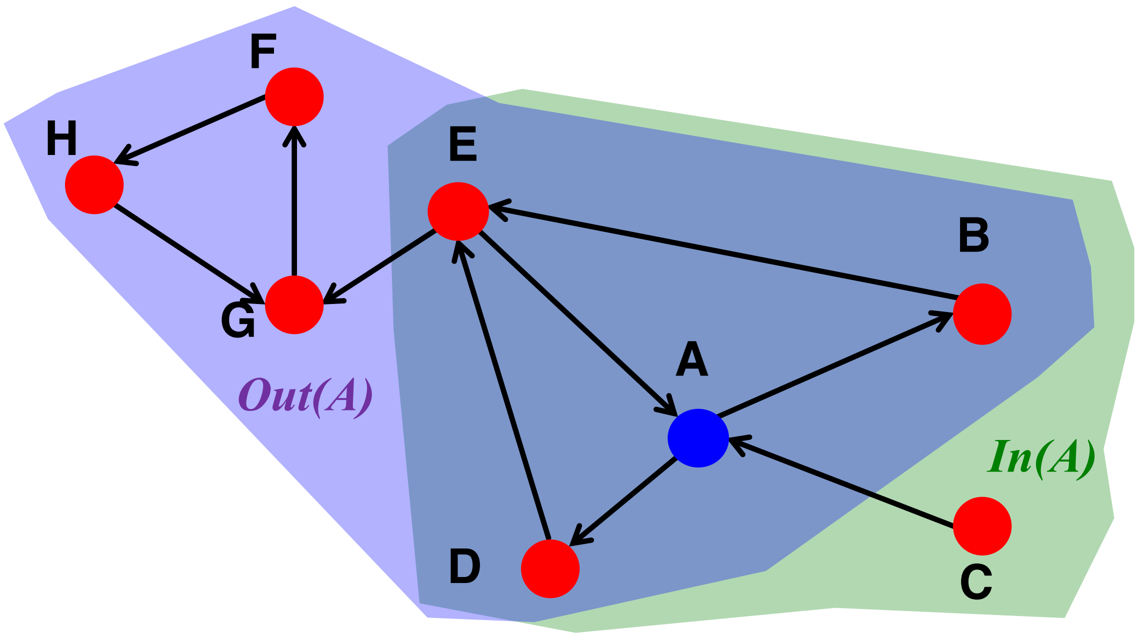


$Out(v)$



Out(A) ∩ In(A) = SCC

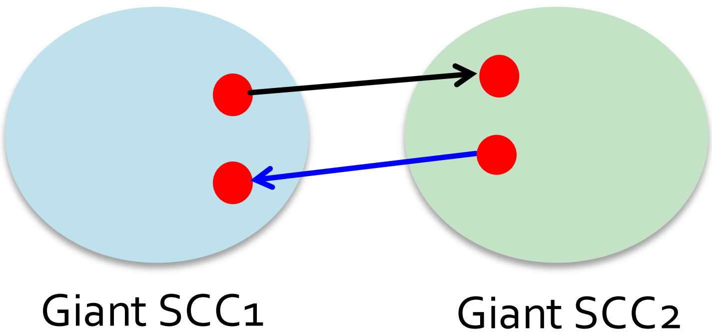
- **Example:**



  - Out(A) = {A, B, D, E, F, G, H}
  - In(A) = {A, B, C, D, E}
  - So, SCC(A) = Out(A) ∩ In(A) = {A, B, D, E}

Graph Structure of the Web

- **There is a single giant SCC**
  - That is, there won't be two SCCs
- **Heuristic argument:**
  - It just takes 1 page from one SCC to link to the other SCC
  - If the 2 SCCs have millions of pages the likelihood of this not happening is very very small



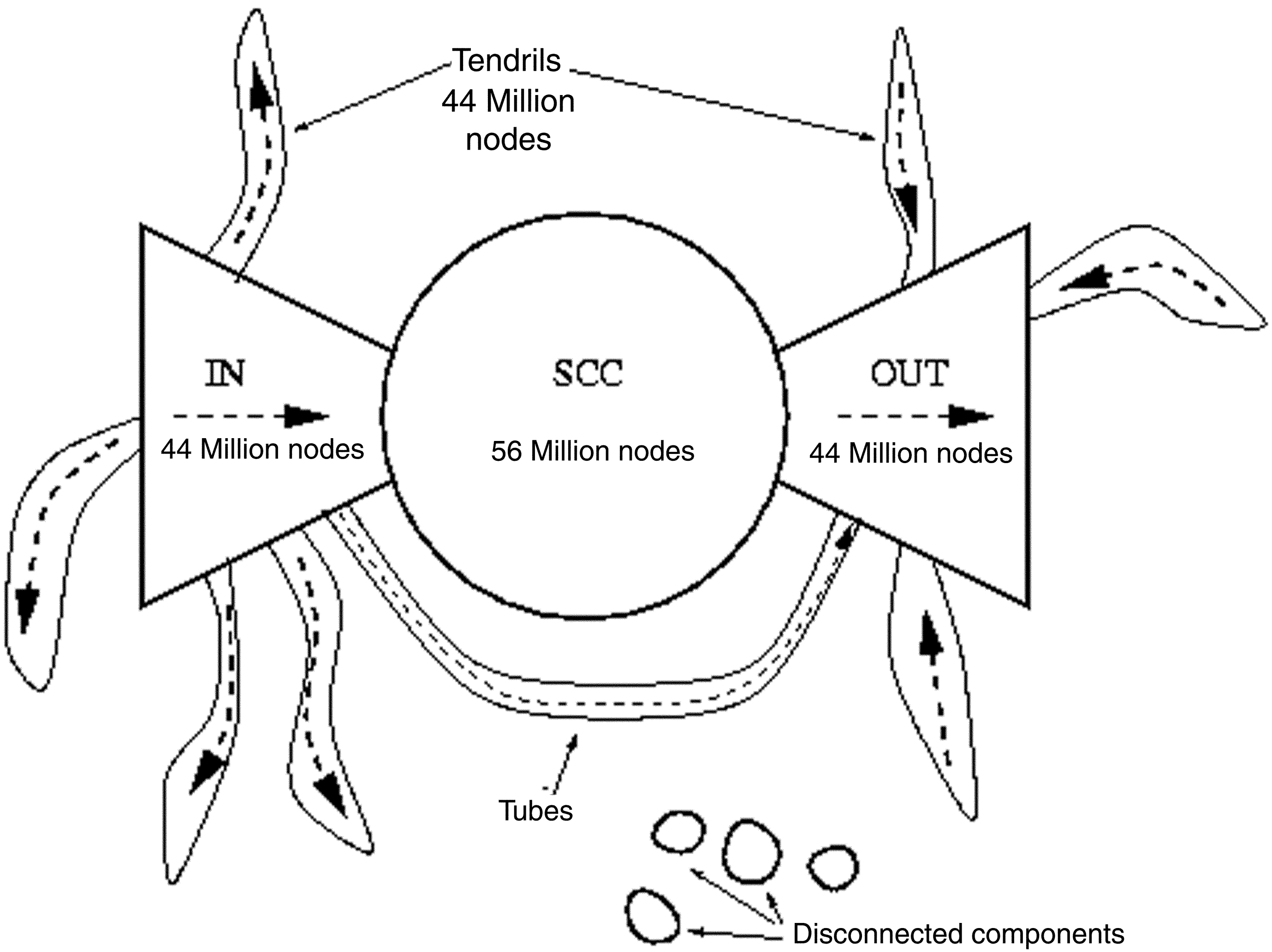
Structure of the Web

- **Broder et al., 2000:**
  - Altavista crawl from October 1999
    - 203 million URLS
    - 1.5 billion links
  - Computer: Server with 12GB of memory
- **Undirected version of the Web graph:**
  - 91% nodes in the largest weakly conn. component
  - Are hubs making the web graph connected?
    - Even if they deleted links to pages with in-degree >10 WCC was still ≈50% of the graph

Structure of the Web

- **Directed version of the Web graph:**
  - **Largest SCC:** 28% of the nodes (56 million)
  - Taking a random node v
    - **Out(v) ≈ 50%** (100 million)
    - **In(v) ≈ 50%** (100 million)
- **What does this tell us about the conceptual picture of the Web graph?**

Bowtie Structure of the Web



203 million pages, 1.5 billion links [Broder et al. 2000]

What did We Learn/Not Learn ?

- **What did we learn:**
  - Conceptual organization of the Web (i.e., the bowtie)
- **What did we not learn:**
  - **Treats all pages as equal**
    - Google's homepage == my homepage
  - **What are the most important pages**
    - How many pages have k in-links as a function of k?
      The degree distribution: $\sim k^{-2}$
  - **Internal structure inside giant SCC**
    - Clusters, implicit communities?
  - **How far apart are nodes in the giant SCC:**
    - Distance = # of edges in shortest path
    - Avg. = 16 [Broder et al.]

Network Properties:
How to Measure a Network?

Plan: Key Network Properties

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

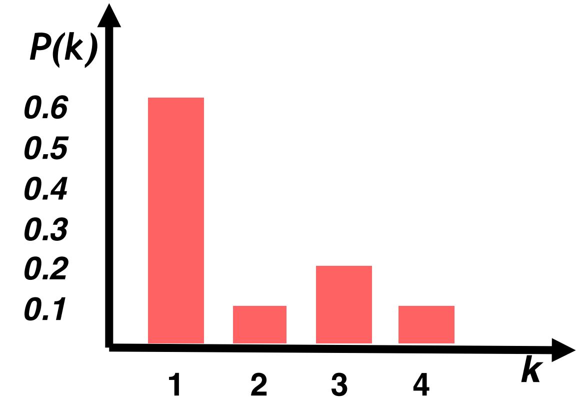
(1) Degree Distribution

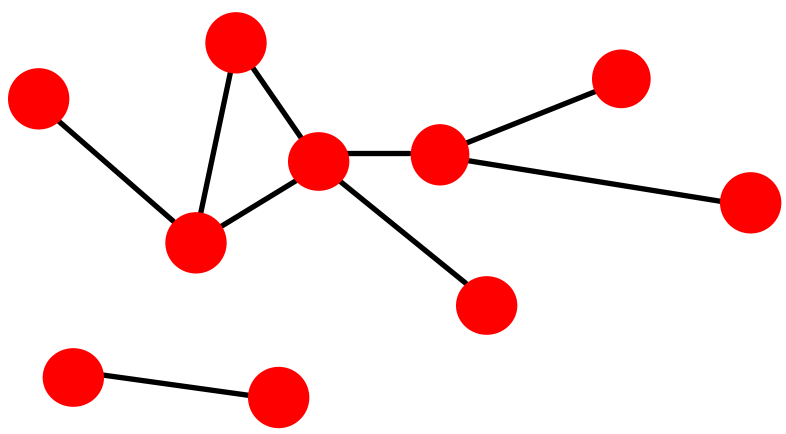
- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

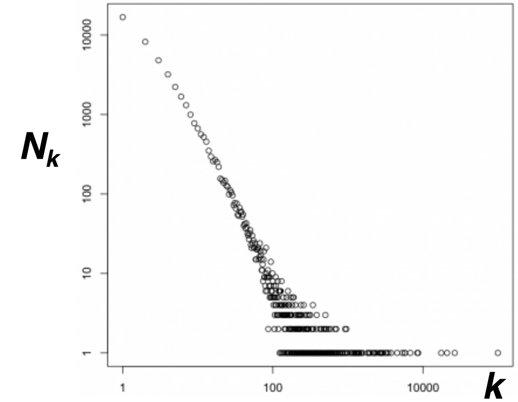
  N_k = # nodes with degree k

- Normalized histogram:

  $P(k) = N_k / N$ ➔ plot





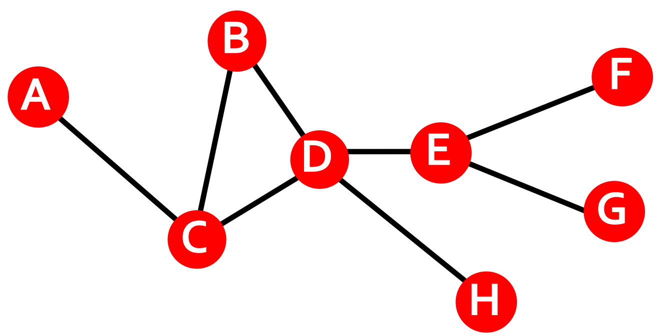


(2) Paths in a Graph

- A ***path*** is a sequence of nodes in which each node is linked to the next one

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

- Path can intersect itself and pass through the same edge multiple times
  - E.g.: ACBDCDEG
  - In a directed graph a path can only follow the direction of the "arrow"



Number of Paths

Extra

- **Number of paths between nodes u and v :**
  - **Length $h=1$:** If there is a link between u and v, $A_{uv}=1$ else $A_{uv}=0$
  - **Length $h=2$:** If there is a path of length two between u and v then $A_{uk}A_{kv}=1$ else $A_{uk}A_{kv}=0$

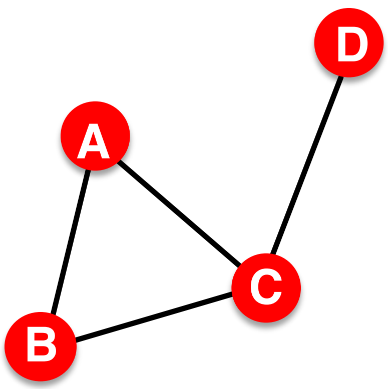
$$H_{uv}^{(2)} = \sum_{k=1}^{N} A_{uk} A_{kv} = [A^2]_{uv}$$

  - **Length h:** If there is a path of length h between u and v then $A_{uk} \ldots. A_{kv}=1$ else $A_{uk} \ldots. A_{kv}=0$
    So, the no. of paths of length h between u and v is

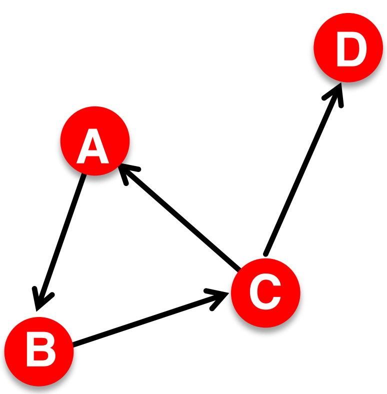
$$H_{uv}^{(h)} = [A^h]_{uv}$$

(holds for both directed and undirected graphs)

Distance in a Graph



$h_{B,D} = 2$



$h_{B,C} = 1$, $h_{C,B} = 2$

- **Distance (shortest path, geodesic)** between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
  - *If the two nodes are disconnected, the distance is usually defined as infinite
- In **directed graphs** paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric:** $h_{A,C} \neq h_{C,A}$

Network Diameter

- **Diameter:** the maximum (shortest path) distance between any pair of nodes in a graph

- **Average path length** for a connected graph (component) or a strongly connected (component of a) directed graph

$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{ij}$$

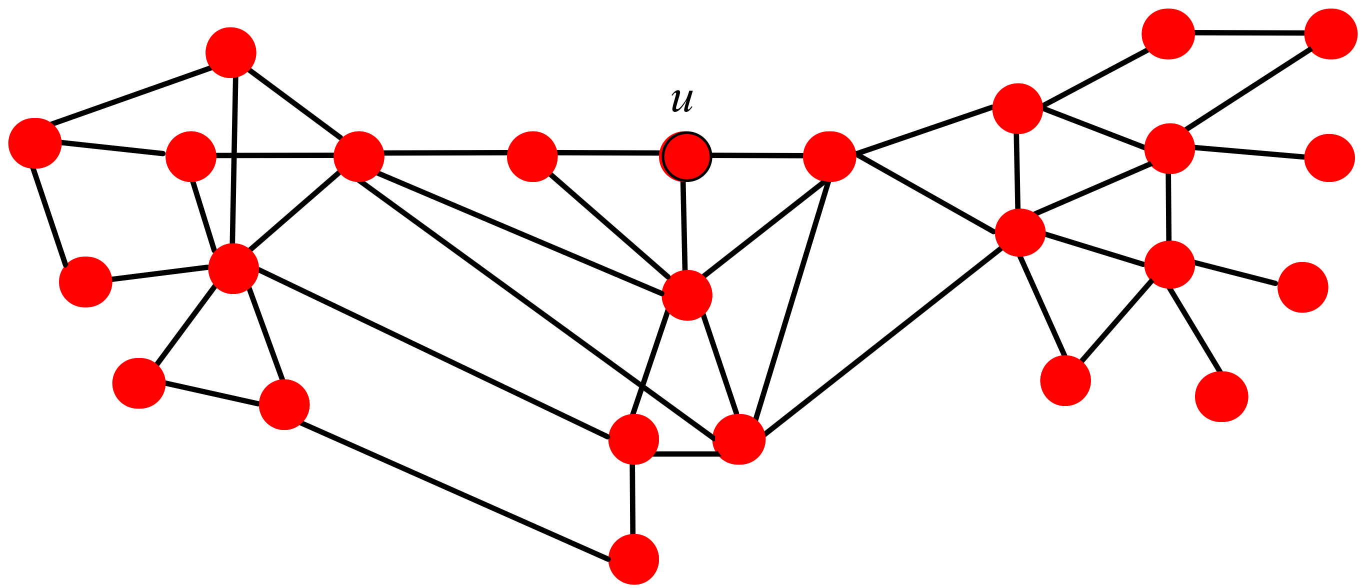
where h_{ij} is the distance from node i to node j

  - Many times we compute the average only over the connected pairs of nodes (that is, we ignore "infinite" length paths)

Finding Shortest Paths

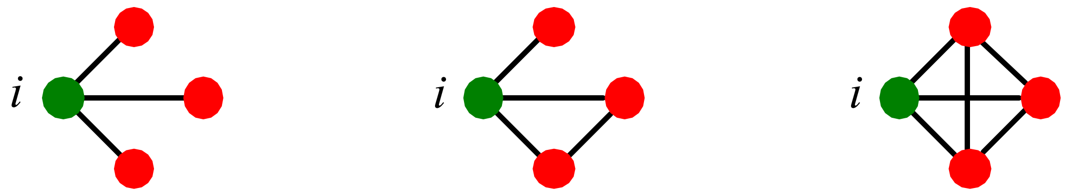
Extra

- **Breadth First Search:**
  - Start with node u, mark it to be at distance $h_u(u)=0$, add u to the queue
  - While the queue not empty:
    - Take node v off the queue, put its unmarked neighbors w into the queue and mark $h_u(w)=h_u(v)+1$



(3) Clustering Coefficient

- **Clustering coefficient:**
  - What portion of i's neighbors are connected?
  - Node i with degree k_i
  - $C_i \in [0,1]$
  - $C_i = \frac{2e_i}{k_i(k_i - 1)}$ where e_i is the number of edges between the neighbors of node i

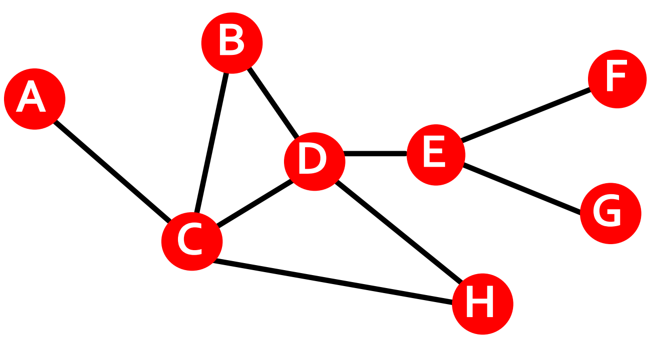


$C_i=0$ $C_i=1/3$ $C_i=1$

- **Average clustering coefficient:** $C = \frac{1}{N}\sum_{i}^{N} C_i$

Clustering Coefficient

- **Clustering coefficient:**
  - What portion of $\boldsymbol{i}$'s neighbors are connected?
  - Node i with degree $\boldsymbol{k_i}$
  - $C_i = \dfrac{2e_i}{k_i(k_i-1)}$ where e_i is the number of edges between the neighbors of node i



$k_B=2,\ e_B=1,\ C_B=2/2=1$

$k_D=4,\ e_D=2,\ C_D=4/12=1/3$

Summary: Key Network Properties

Degree distribution: $P(k)$

Path length: h

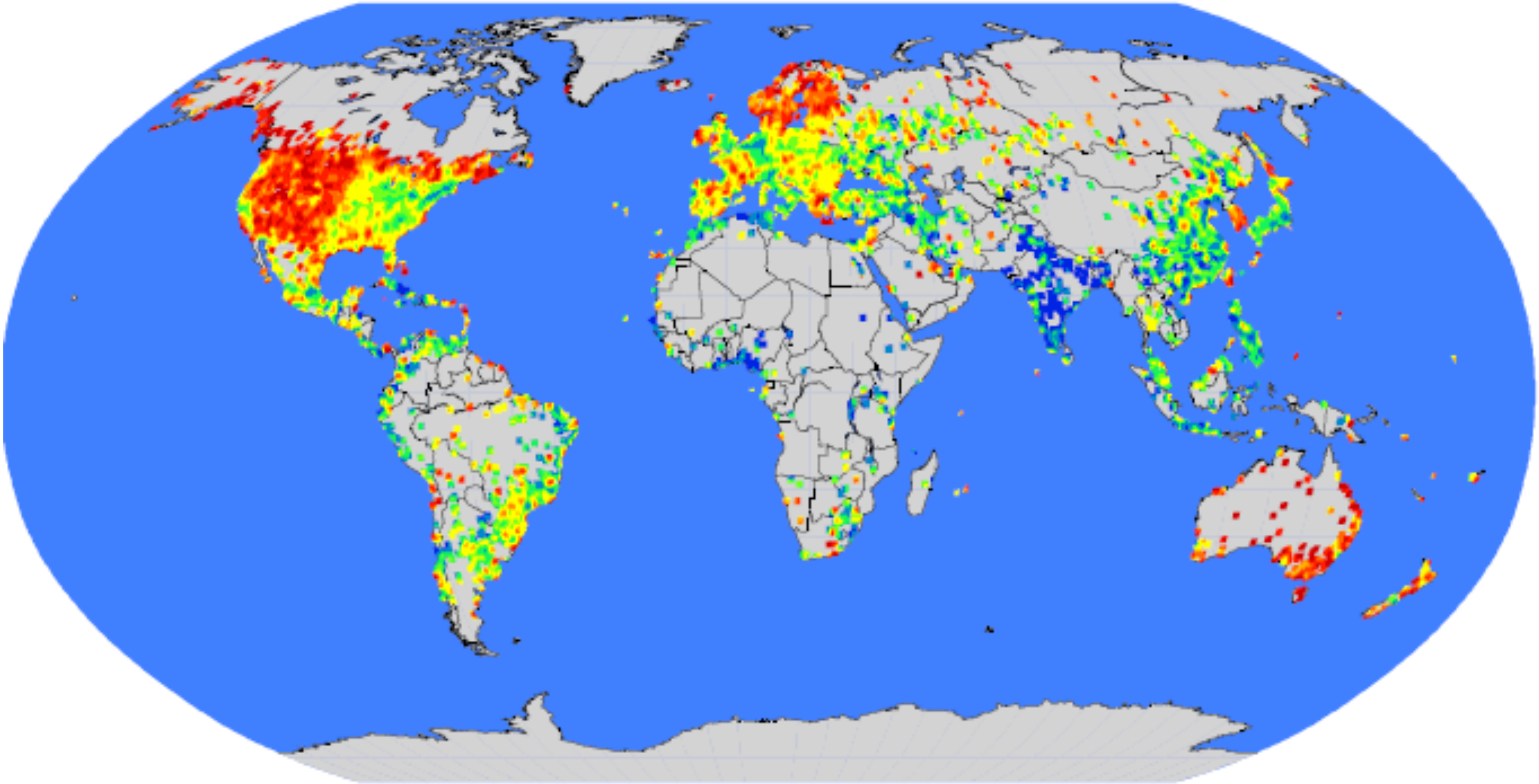
Clustering coefficient: C

Let's measure $P(k)$, h and C on a real-world network!

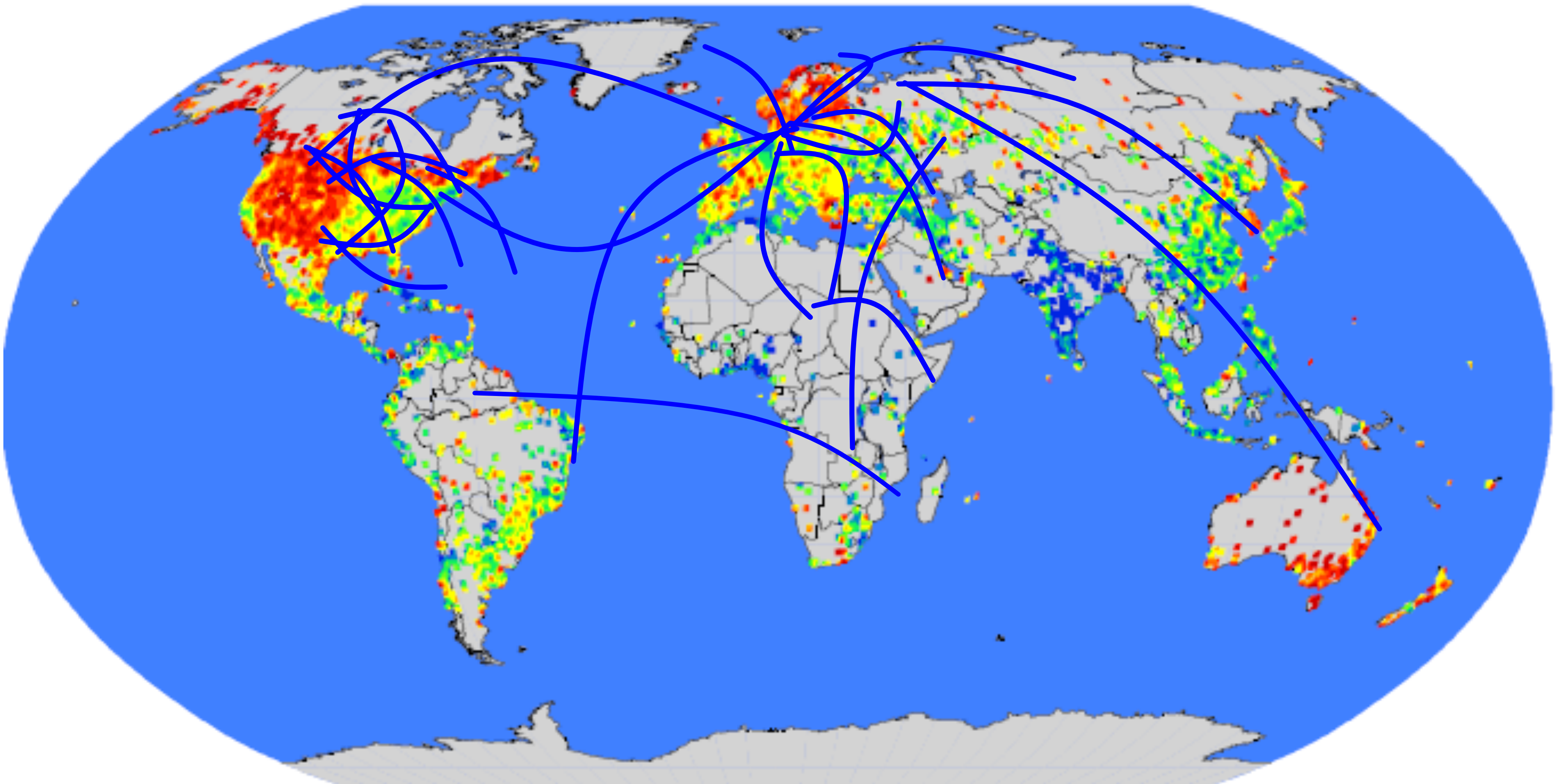
The MSN Messenger

- **MSN Messenger activity in June 2006:**
  - 245 million users logged in
  - 180 million users engaged in conversations
  - More than 30 billion conversations
  - More than 255 billion exchanged messages

Communication: Geography

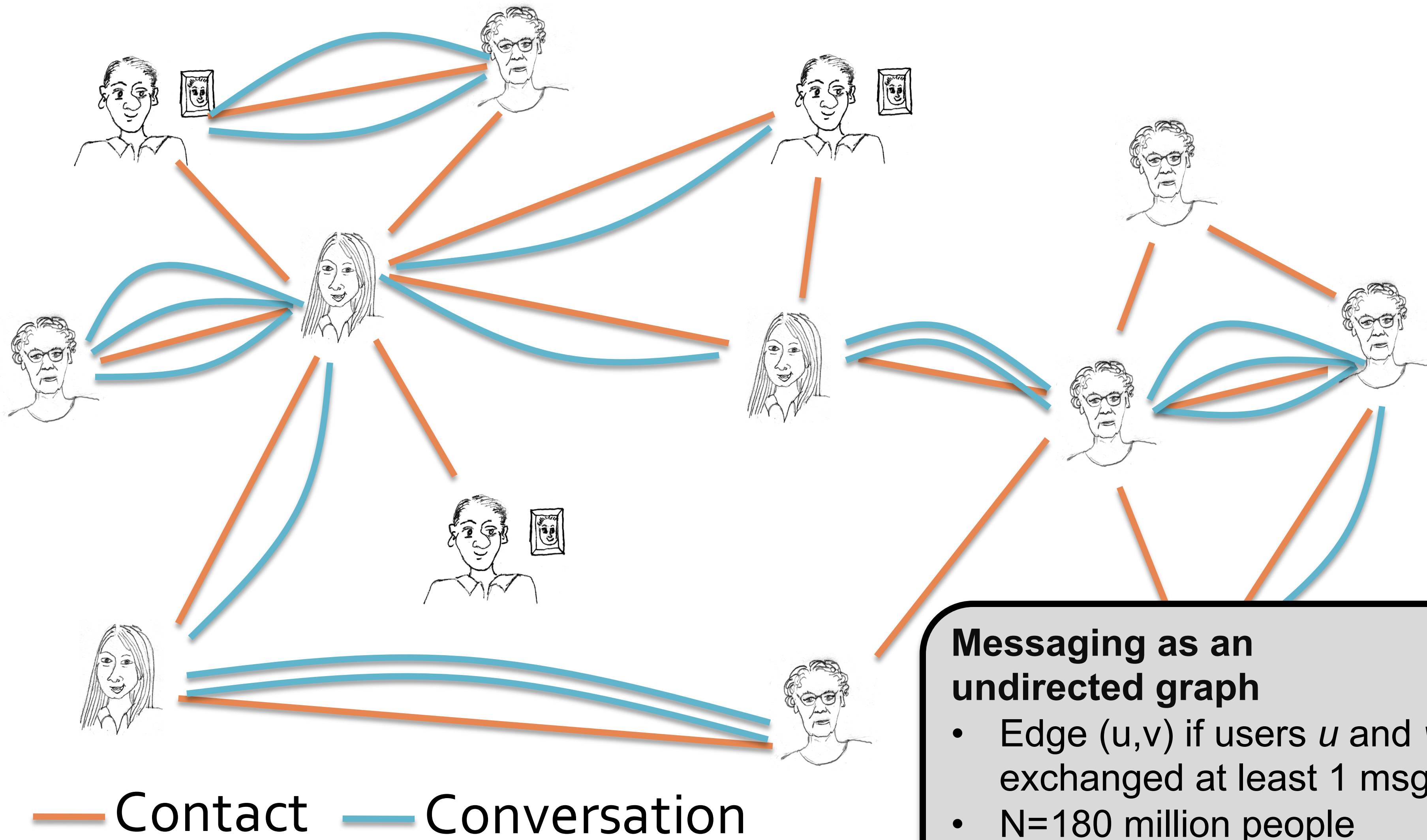


Communication Network



Network: 180M people, 1.3B edges

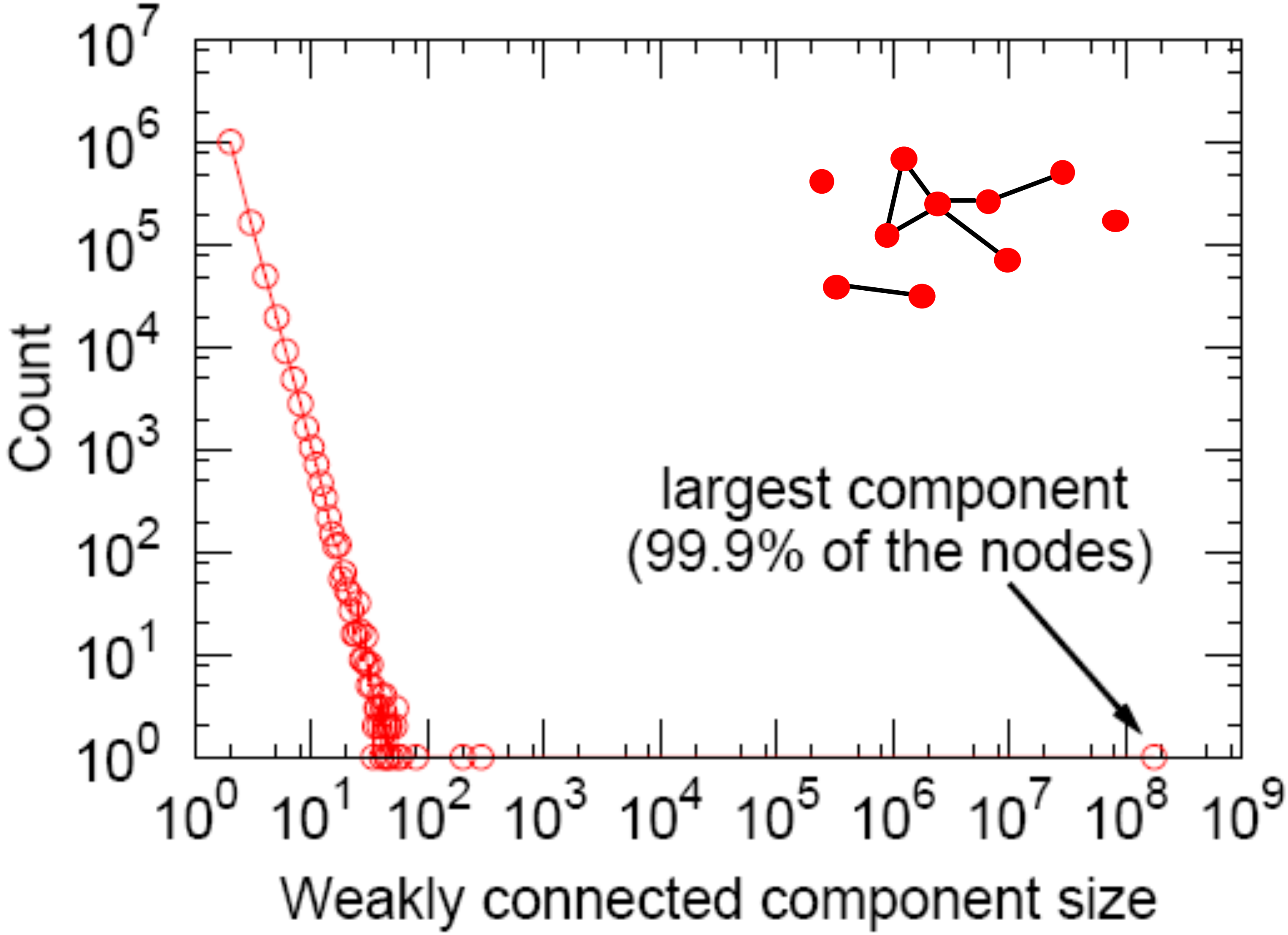
Messaging as a Multigraph



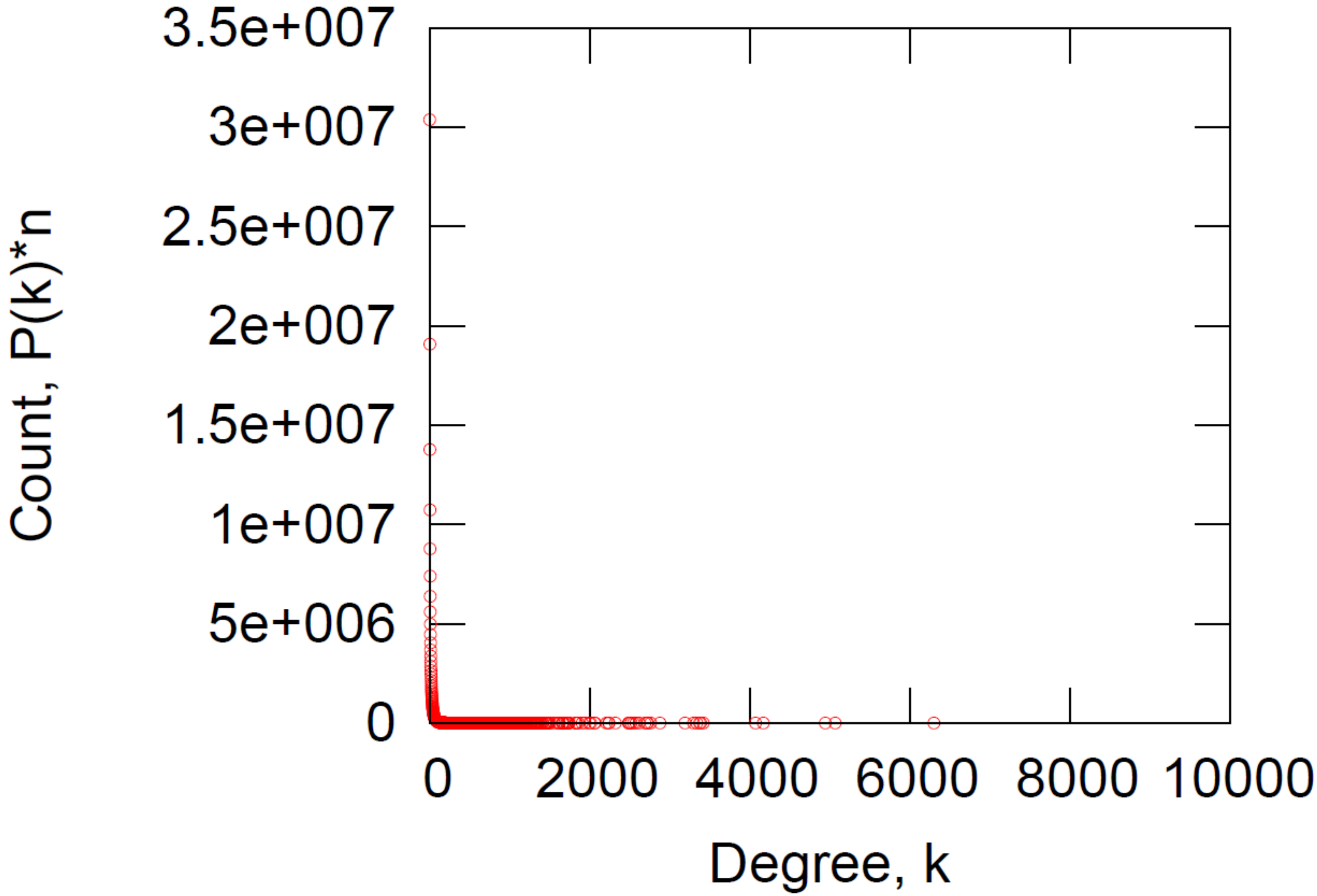
Messaging as an undirected graph

- Edge (u,v) if users *u* and *v* exchanged at least 1 msg
- N=180 million people
- E=1.3 billion edges

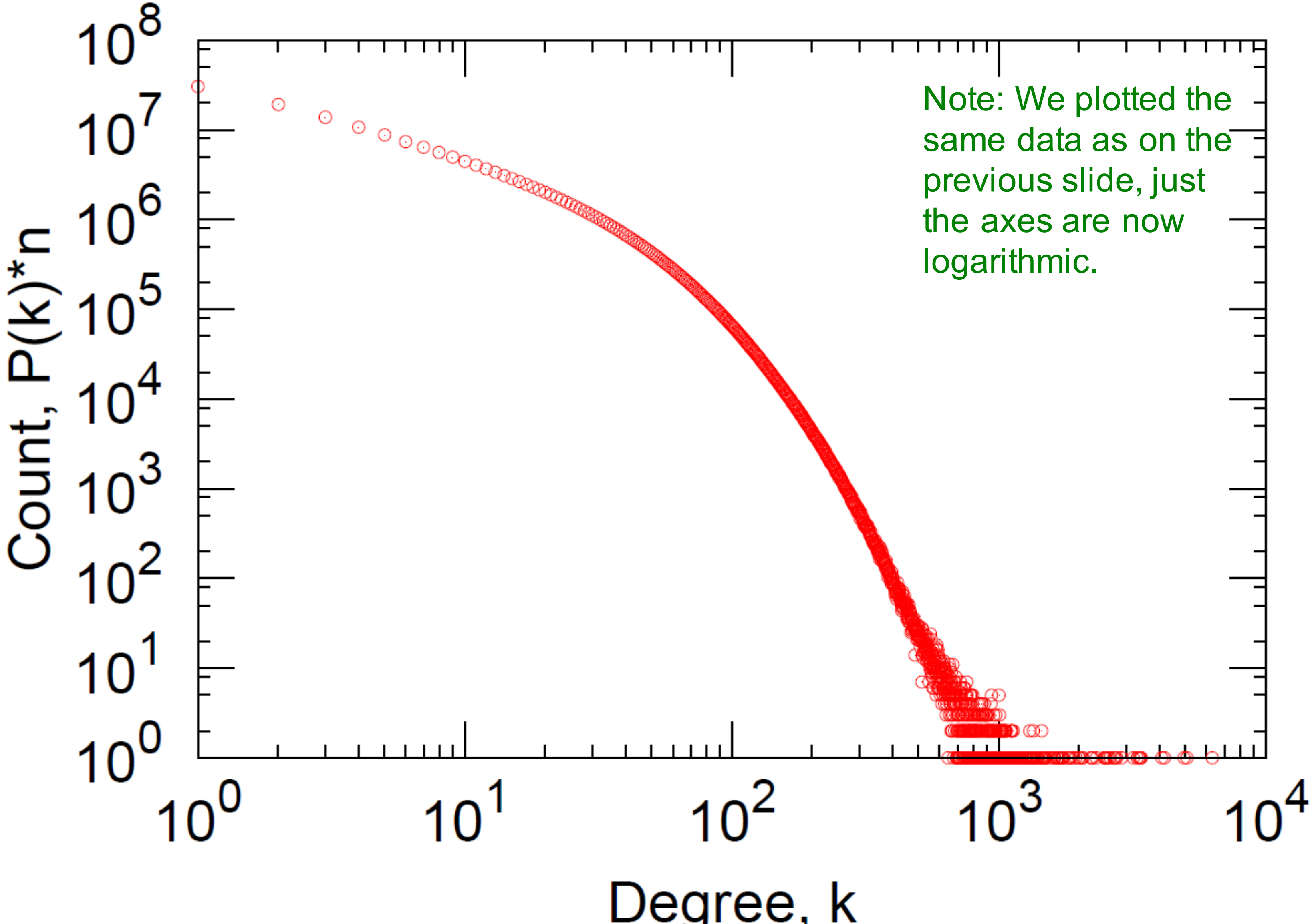
MSN: (1) Connectivity



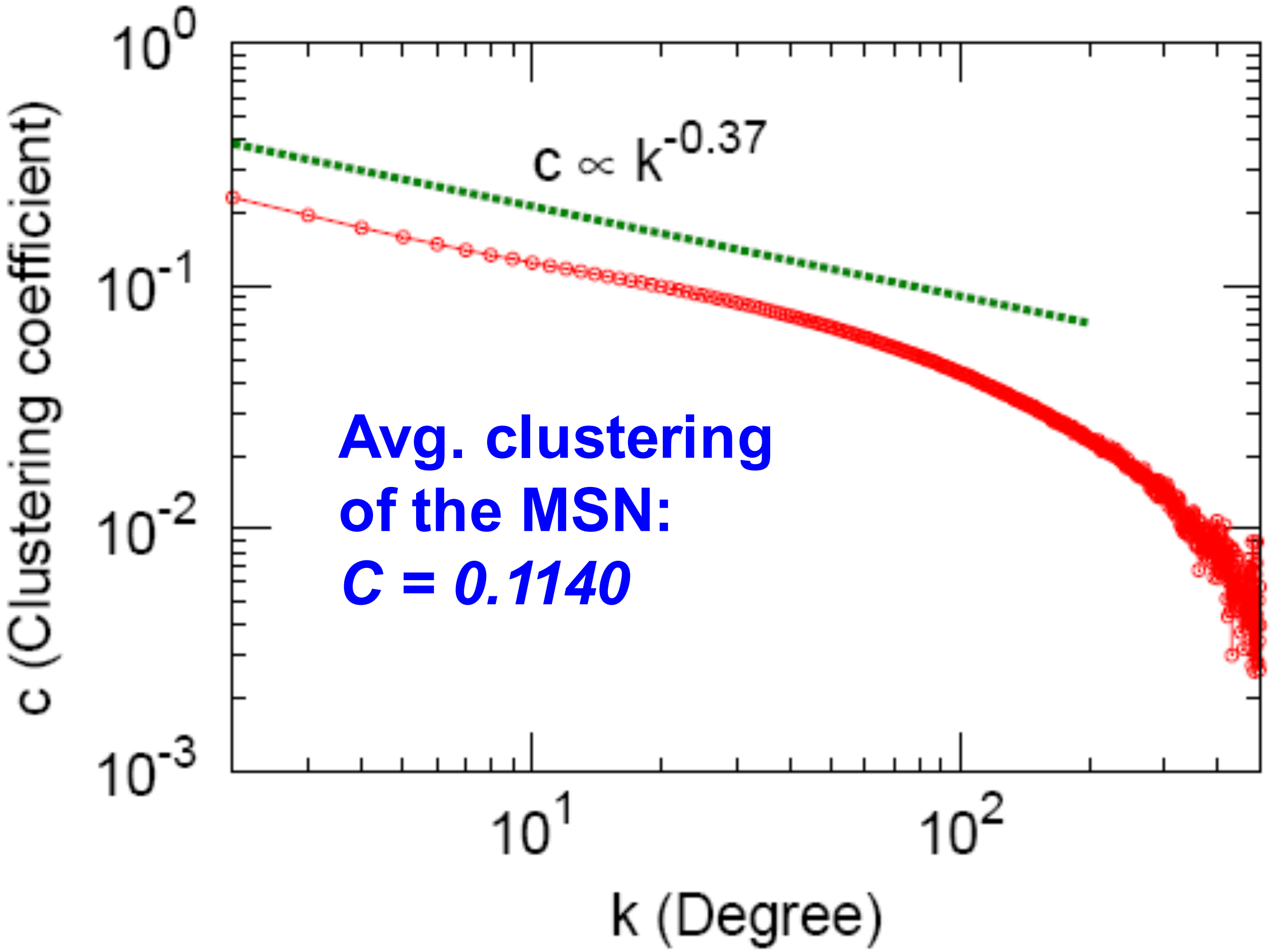
MSN: (2) Degree Distribution



MSN: Log-Log Degree Distribution

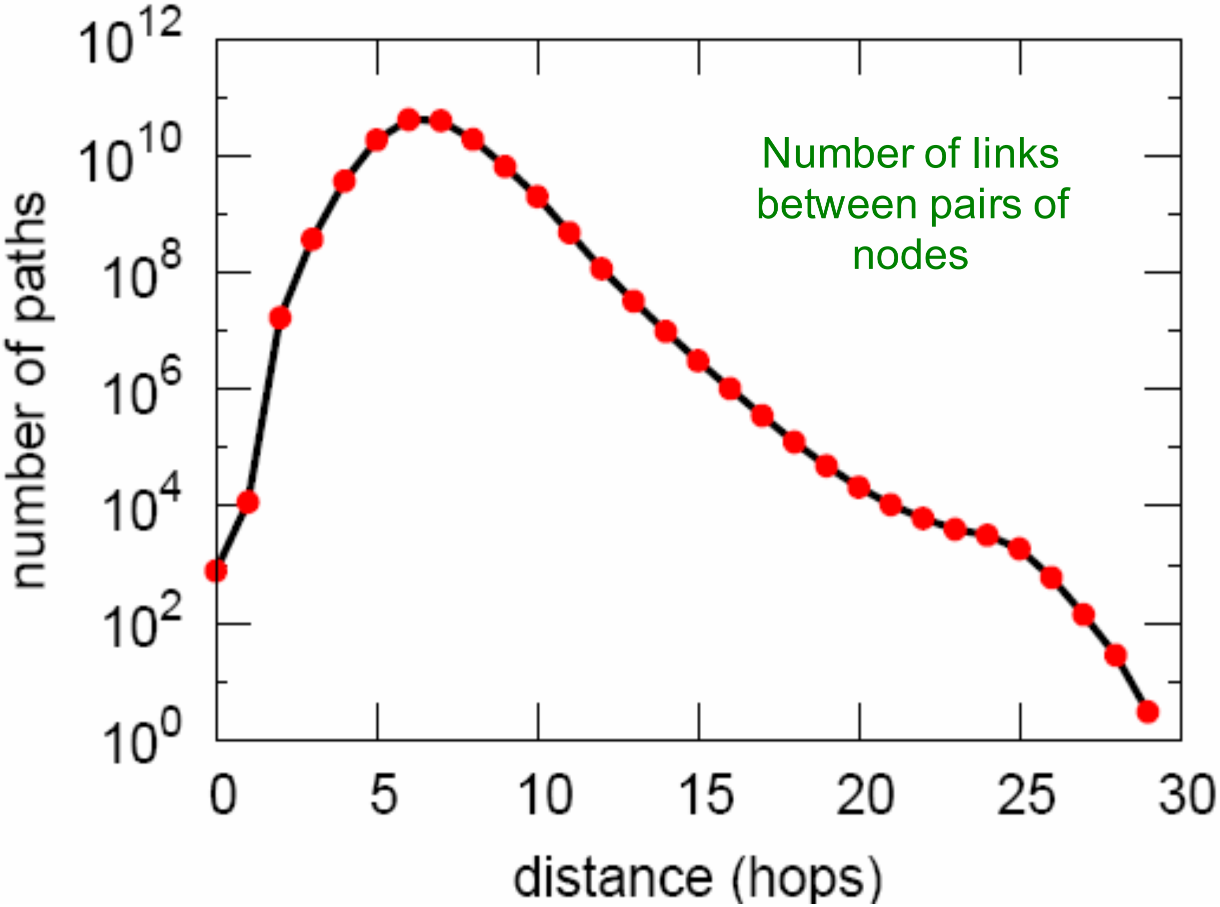


MSN: (3) Clustering



C_k: average C_i of nodes i of degree k: $C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i$

MSN: (4) Diameter



Avg. path length **6.6**
90% of the nodes can be reached in < 8 hops

Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

nodes as we do BFS out of a random node

MSN: Key Network Properties

Degree distribution: *Heavily skewed*
avg. degree= 14.4

Path length: *6.6*

Clustering coefficient: *0.11*

Are these values "expected"?
Are they "surprising"?

To answer this we need a null-model!

Erdös-Renyi Random Graph Model

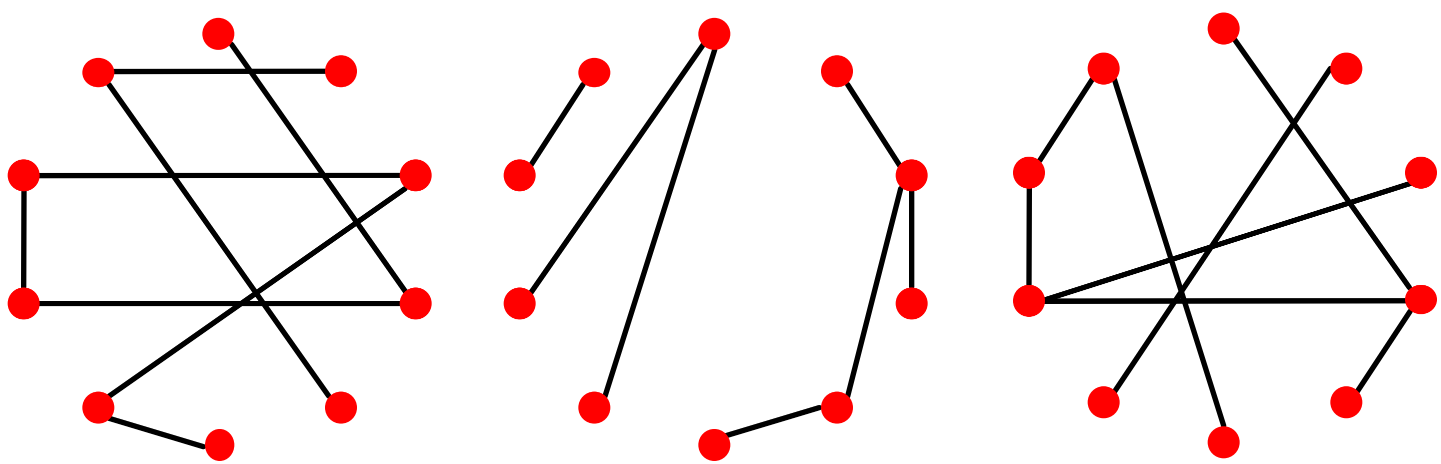
Simplest Model of Graphs

- **Erdös-Renyi Random Graphs** [Erdös-Renyi, '60]
- **Two variants:**
  - $G_{n,p}$: undirected graph on n nodes and each edge (u,v) appears i.i.d. with probability p
  - $G_{n,m}$: undirected graph with n nodes, and m uniformly at random picked edges

What kinds of networks does such model produce?

Random Graph Model

- ***n* and *p* do not uniquely determine the graph!**
  - The graph is a result of a random process
- We can have many different realizations given the same *n* and *p*



n = 10
p= 1/6

Random Graph Model: Edges

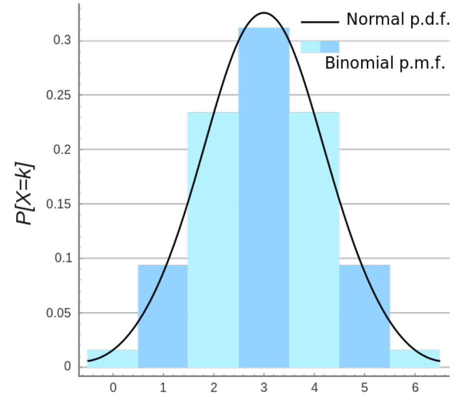
- **How likely is a graph on E edges?**
- $P(E)$: the probability that a given G_{np} generates a graph on exactly E edges:

$$P(E) = \binom{E_{\max}}{E} p^{E} (1-p)^{E_{\max}-E}$$

where $E_{max}=n(n-1)/2$ is the maximum possible number of edges in an undirected graph of n nodes

P(E) is exactly the Binomial distribution >>>

Number of successes in a sequence of E_{max} independent yes/no experiments



Node Degrees in a Random Graph

- **What is expected degree of a node?**
  - Let X_v be a rnd. var. measuring the degree of node v
  - **We want to know:** $E[X_v] = \sum_{j=0}^{n-1} j\, P(X_v = j)$
    - **For the calculation we will need: Linearity of expectation**
      - For any random variables $Y_1, Y_2, \ldots, Y_k$
      - If $Y = Y_1 + Y_2 + \ldots Y_k$, then $E[Y] = \sum_i E[Y_i]$
- **An easier way:**
  - Decompose X_v to $X_v = X_{v,1} + X_{v,2} + \ldots + X_{v,n-1}$
    - where $X_{v,u}$ is a $\{0,1\}$-random variable which tells if edge (v,u) exists or not

$$E[X_v] = \sum_{u=1}^{n-1} E[X_{vu}] = (n-1)p$$

How to think about this?

- Prob. of node u linking to node v is p
- u can link (flips a coin) to all other $(n-1)$ nodes
- Thus, the expected degree of node u is: $p(n-1)$

Properties of G_{np}

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

What are values of these properties for G_{np}?

Degree Distribution

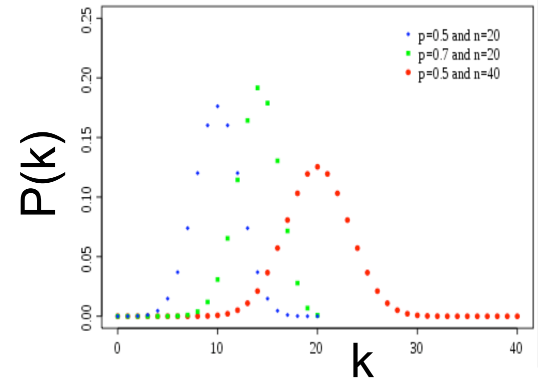
- **Fact:** **Degree distribution of G_{np} is Binomial.**
- Let $P(k)$ denote a fraction of nodes with degree k:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Select k nodes out of n-1

Probability of having k edges

Probability of missing the rest of the n-1-k edges



Mean, variance of a binomial distribution

$$\bar{k} = p(n-1)$$

$$\sigma^2 = p(1-p)(n-1)$$

$$\frac{\sigma}{\bar{k}} = \left[\frac{1-p}{p}\frac{1}{(n-1)}\right]^{1/2} \approx \frac{1}{(n-1)^{1/2}}$$

By the law of large numbers, as the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of k.

Clustering Coefficient of G_{np}

- **Remember:** $C_i = \dfrac{2e_i}{k_i(k_i-1)}$

Where e_i is the number of edges between i's neighbors

- Edges in G_{np} appear i.i.d. with prob. p

- **So:** $e_i = p\dfrac{k_i(k_i-1)}{2}$

Each pair is connected with prob. p

Number of distinct pairs of neighbors of node i of degree k_i

- **Then:** $C = \dfrac{p \cdot k_i(k_i-1)}{k_i(k_i-1)} = p = \dfrac{\bar{k}}{n-1} \approx \dfrac{\bar{k}}{n}$

Clustering coefficient of a random graph is small.
For a fixed avg. degree (that is p=1/n), C decreases with the graph size n.

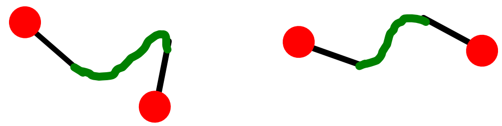
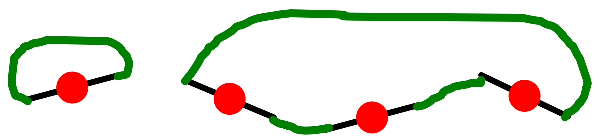
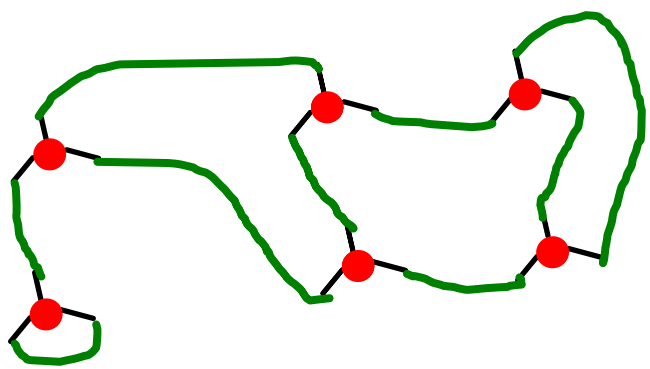
Network Properties of G_{np}

Degree distribution: $$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Clustering coefficient: $$C = p = \bar{k}/n$$

Path length: ***next!***

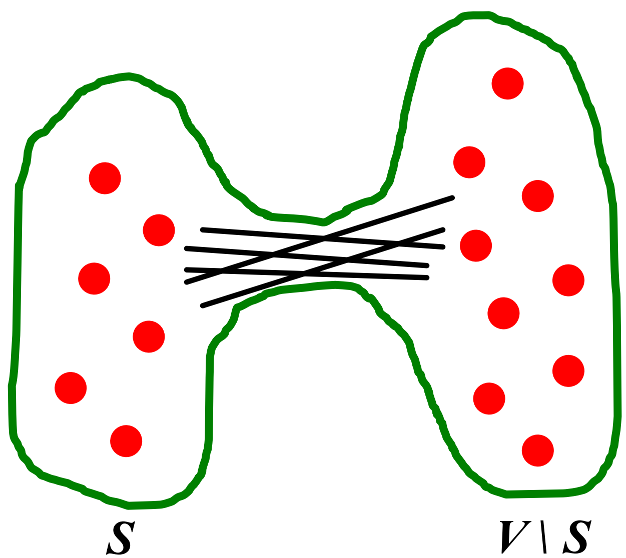
Def: Random k-Regular Graphs

- To prove the diameter of a G_{np} we define few concepts
- **Define: Random k-Regular graph**
  - **Assume each node has k spokes (half-edges)**
    - **k=1:**  **Graph is a set of pairs**
    - **k=2:**  **Graph is a set of cycles**
    - **k=3:**  **Arbitrarily complicated graphs**
  - **Randomly pair them up!**

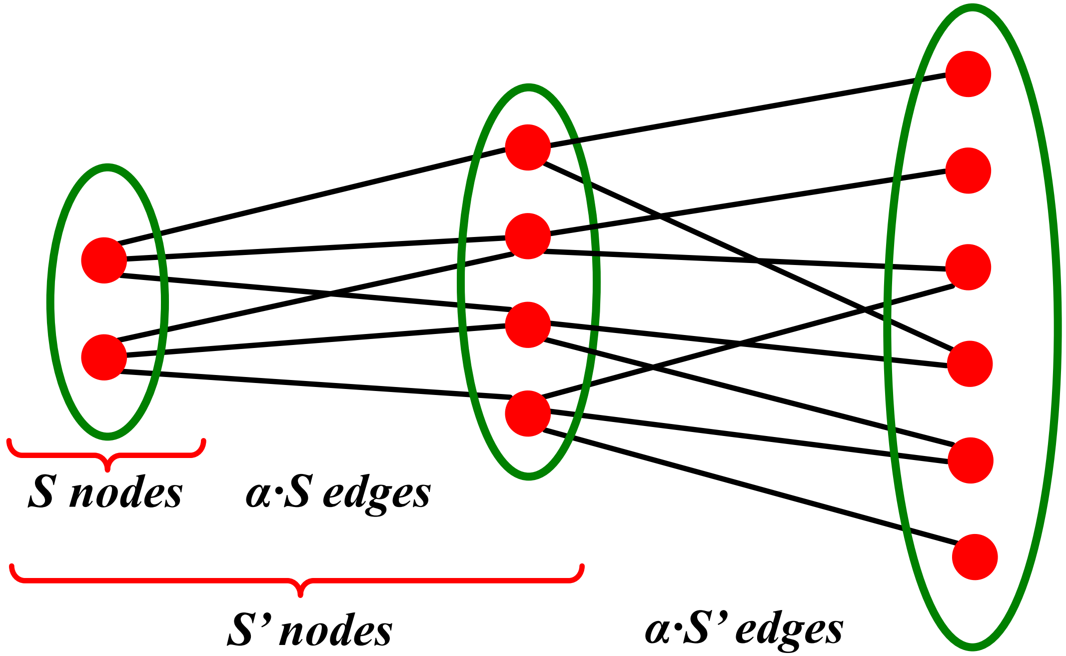
Def: Expansion

- Graph $G(V, E)$ has **expansion α**: if $\forall\, S \subseteq V$:
  $\#$ of edges leaving $S \geq \alpha \cdot \min(|S|, |V \setminus S|)$
- **Or equivalently:**

$$\alpha = \min_{S \subseteq V} \frac{\#\, edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$



Expansion: Intuition



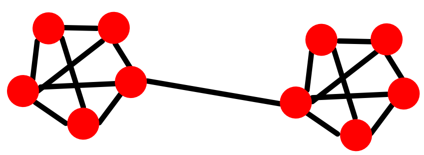
$$\alpha = \min_{S \subseteq V} \frac{\# \, edges \; leaving \; S}{\min(|S|, |V \setminus S|)}$$

(A big) graph with "good" expansion

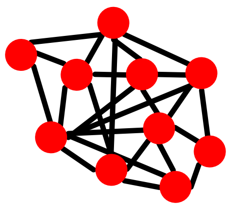
Expansion: Measures Robustness

$$\alpha = \min_{S \subseteq V} \frac{\#\,edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$

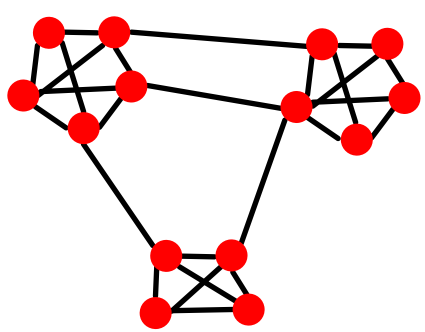
- Expansion is **measure of robustness:**
  - To disconnect l nodes, we need to cut $\geq \alpha \cdot l$ edges
- Low expansion:



- High expansion:



- Social networks:
  - "Communities"



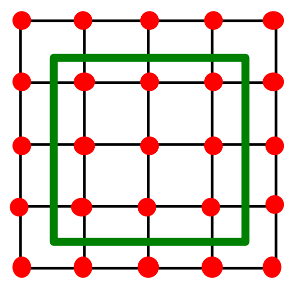
Expansion: k-Regular Graphs

$$\alpha = \min_{S \subseteq V} \frac{\#\, edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$

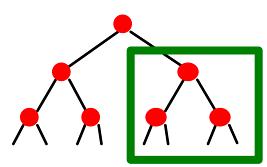
- **k-regular graph** (every node has degree k):
  - Expansion is at most k (when S is a single node)
- **Is there a graph on n nodes ($n \to \infty$), of fixed max deg. k, so that expansion α remains const?**

Examples:

- **n×n grid:** $k=4$: $\alpha = 2n/(n^2/4) \to 0$
  ($S = n/2 \times n/2$ square in the center)



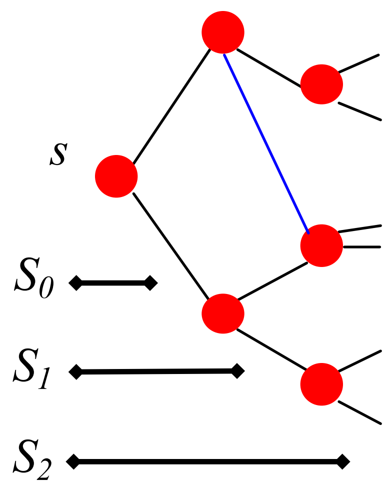
- **Complete binary tree:**
  $\alpha \to 0$ for $|S| = (n/2)-1$



- **Fact:** For a random **3-regular graph** on n nodes, there is some const α ($\alpha > 0$, independent. of n) such that w.h.p. the expansion of the graph is $\geq \alpha$

Diameter of 3-Regular Rnd. Graph

- **Fact:** In a graph on n nodes with expansion α for all pairs of nodes s and t there is a path of $O((\log n) / \alpha)$ edges connecting them.
- Proof:
  - Proof strategy:
    - We want to show that from any node s there is a path of length $O((\log n)/\alpha)$ to any other node t
  - Let S_j be a set of all nodes found within j steps of BFS from s.
  - **How does S_j increase as a function of j?**

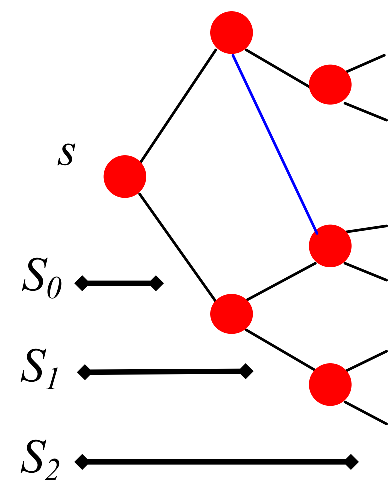


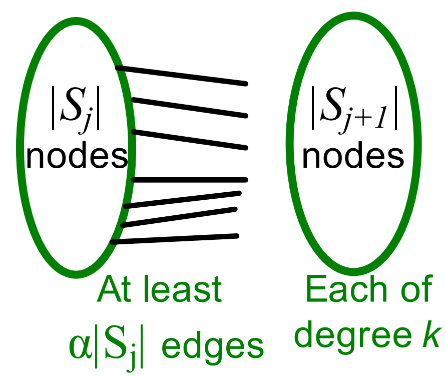
Diameter of 3-Regular Rnd. Graph

- <u>Proof (continued):</u>
  - Let S_j be a set of all nodes found within j steps of BFS from s.
  - **We want to relate S_j and S_{j+1}**

$$|S_{j+1}| \ge |S_j| + \frac{\overbrace{\alpha|S_j|}^{\text{Expansion}}}{\underbrace{k}_{\substack{\text{At most } k \text{ edges} \\ \text{“collide” at a node}}}} =$$

$$|S_{j+1}| \ge |S_j|\left(1+\frac{\alpha}{k}\right) = \left(1+\frac{\alpha}{k}\right)^{j+1}$$





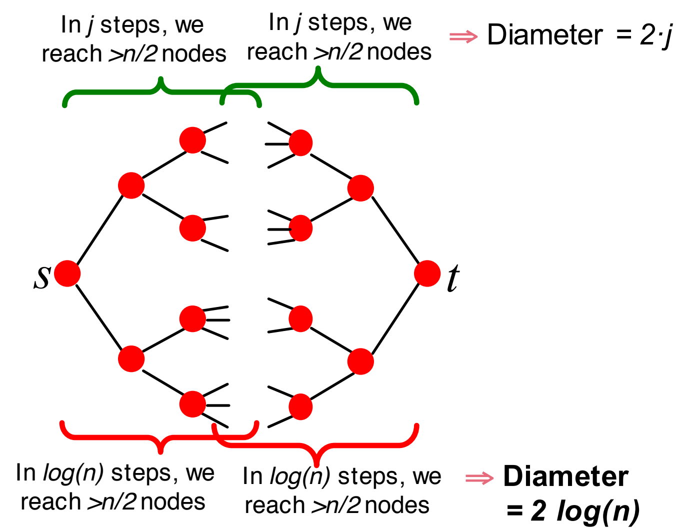
Diameter of 3-Regular Rnd. Graph

$$e = \lim_{x \to \infty}\left(1 + \frac{1}{x}\right)^x$$

- <u>Proof (continued):</u>
  - **In how many steps of BFS do we reach $>n/2$ nodes?**
  - Need j so that: $S_j = \left(1 + \frac{\alpha}{k}\right)^j \geq \frac{n}{2}$
  - Let's set: $j = \frac{k \log_2 n}{\alpha}$
  - Then:

$$\left(1 + \frac{\alpha}{k}\right)^{\frac{k \log_2 n}{\alpha}} \geq 2^{\log_2 n} = n > \frac{n}{2}$$

  - In $2k/\alpha \cdot log\ n$ steps $|S_j|$ grows to $\Theta(n)$. So, **the diameter of G is $O(\log(n)/\ \alpha)$**



Claim:

$$\left(1 + \frac{\alpha}{k}\right)^{\frac{k \log_2 n}{\alpha}} \geq 2^{\log_2 n}$$

Remember $n>0$, $\alpha \leq k$ then:

$$\text{if } \alpha = \text{k} : (1+1)^{\frac{1}{1}\log_2 n} = 2^{\log_2 n}$$

$$\text{if } \alpha \to 0 \text{ then } \frac{k}{\alpha} = x \to \infty :$$

$$\text{and } \left(1 + \frac{1}{x}\right)^{x \log_2 n} = e^{\log_2 n} > 2^{\log_2 n}$$

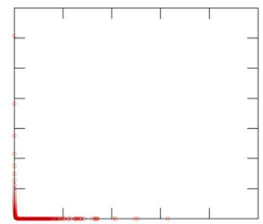
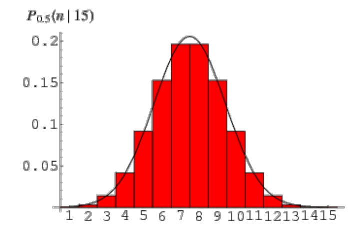
Network Properties of G_{np}

Degree distribution: $$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Path length: $$O(\log n)$$

Clustering coefficient: $$C = p = \bar{k} / n$$

MSN vs. G_{np}

	MSN	**G_{np}**
Degree distribution:		
Path length:	**6.6**	$O(\log n)$ ≈ 8.2
Clustering coefficient:	0.11	$\bar{k} / n$ ≈ $8 \cdot 10^{-8}$

Real Networks vs. G_{np}

- **Are real networks like random graphs?**
  - Giant connected component: ☺
  - Average path length: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- **Problems with the random networks model:**
  - Degreed distribution differs from that of real networks
  - Giant component in most real network does NOT emerge through a phase transition
  - No local structure – clustering coefficient is too low
- **Most important: Are real networks random?**
  - The answer is simply: **NO!**

Real Networks vs. G_{np}

- **If G_{np} is wrong, why did we spend time on it?**
  - It is the reference model for the rest of the class.
  - It will help us calculate many quantities, that can then be compared to the real data
  - It will help us understand to what degree is a particular property the result of some random process

So, while G_{np} is WRONG, it will turn out to be extremly USEFUL!

EXTRA: "Evolution" of the G_{np}

What happens to G_{np} when we vary p?

Back to Node Degrees of G_{np}

- **Remember, expected degree** $E[X_v] = (n-1)p$
- **We want *E[X$_v$]* be independent of *n***
  So let: $p=c/(n-1)$
- **Observation:** If we build random graph G_{np} with $p=c/(n-1)$ we have many isolated nodes
- **Why?**

$$P[v \text{ has degree } 0] = (1-p)^{n-1} = \left(1-\frac{c}{n-1}\right)^{n-1} \xrightarrow[n\to\infty]{} e^{-c}$$

$$\lim_{n\to\infty}\left(1-\frac{c}{n-1}\right)^{n-1} = \left(1-\frac{1}{x}\right)^{-x\cdot c} = \left[\underbrace{\lim_{x\to\infty}\left(1-\frac{1}{x}\right)^{-x}}_{e}\right]^{-c} = e^{-c}$$

Use substitution $\frac{1}{x} = \frac{c}{n-1}$

By definition:

$$e = \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x}$$

No Isolated Nodes

- **How big do we have to make p before we are likely to have no isolated nodes?**
- We know: $P[v \text{ has degree } 0] = e^{-c}$
- Event we are asking about is:
  - I = some node is isolated
  - $I = \bigcup_{v \in N} I_v$ where I_v is the event that v is isolated
- **We have:**

$$P(I) = P\left(\bigcup_{v \in N} I_v\right) \leq \sum_{v \in N} P(I_v) = ne^{-c}$$

Union bound

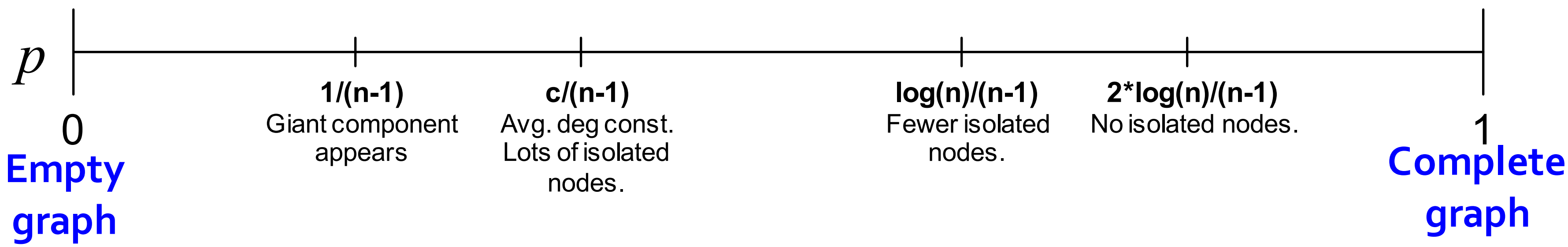
$$\left|\bigcup_i A_i\right| \leq \sum_i |A_i|$$

No Isolated Nodes

- **We just learned:** $\boldsymbol{P(I) = n\, e^{-c}}$
- Let's try:
  - $c = \ln n$ then: $n\, e^{-c} = n\, e^{-\ln n} = n \cdot 1/n = 1$
  - $c = 2 \ln n$ then: $n\, e^{-2 \ln n} = n \cdot 1/n^2 = 1/n$
- **So if:**
  - $p = \ln n$ then: $P(I) = 1$
  - $p = 2 \ln n$ then: $P(I) = 1/n \rightarrow 0$ as $n \rightarrow \infty$

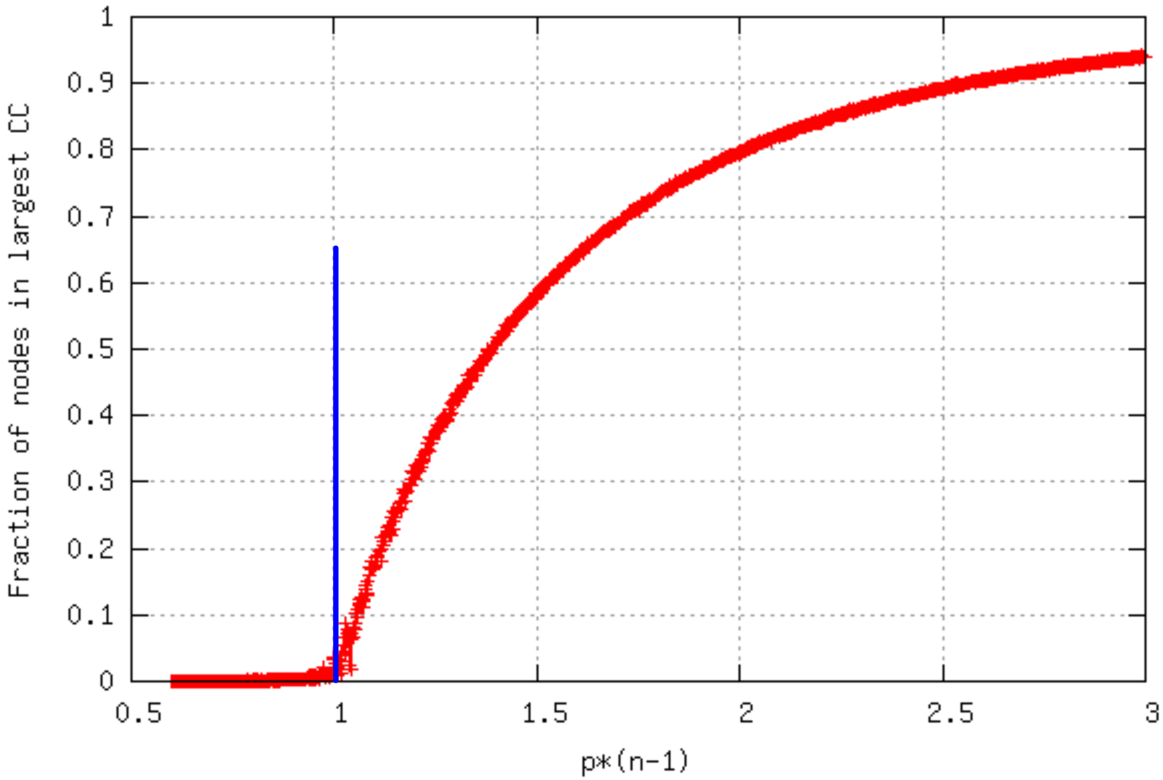
"Evolution" of a Random Graph

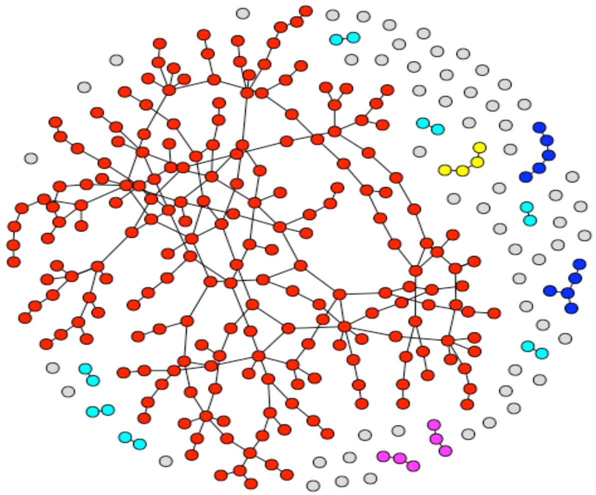
- **Graph structure of G_{np} as p changes:**



- **Emergence of a Giant Component:**
  avg. degree $k=2E/n$ or $p=k/(n-1)$
  - $k=1-\varepsilon$: all components are of size $\Omega(\log n)$
  - $k=1+\varepsilon$: 1 component of size $\Omega(n)$, others have size $\Omega(\log n)$

G_{np} Simulation Experiment





Fraction of nodes in the largest component

- G_{np}, n=100k, $p(n-1)$ = 0.5 … 3