Web as a Graph, Measuring Networks, and the Random Graph Model

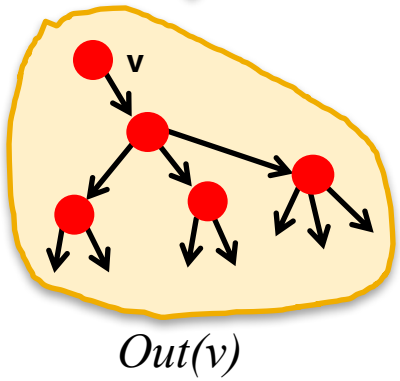
How the Class Fits Together

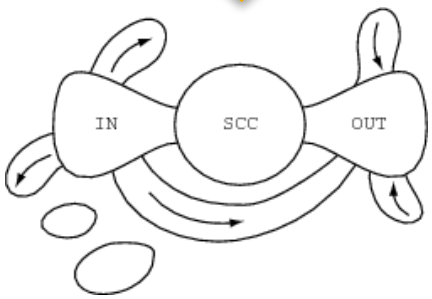
Measurements	Models	Algorithms
Small diameter, Edge clustering	Erdös-Renyi model, Small-world model	Decentralized search
Patterns of signed edge creation	Structural balance, Theory of status	Models for predicting edge signs
Viral Marketing, Blogosphere, Memetracking	Independent cascade model, Game theoretic 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

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!**

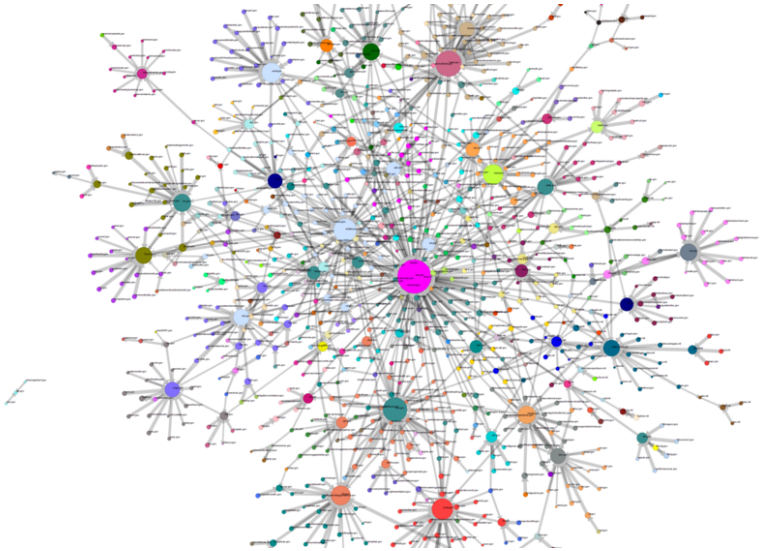




The Web as a Graph

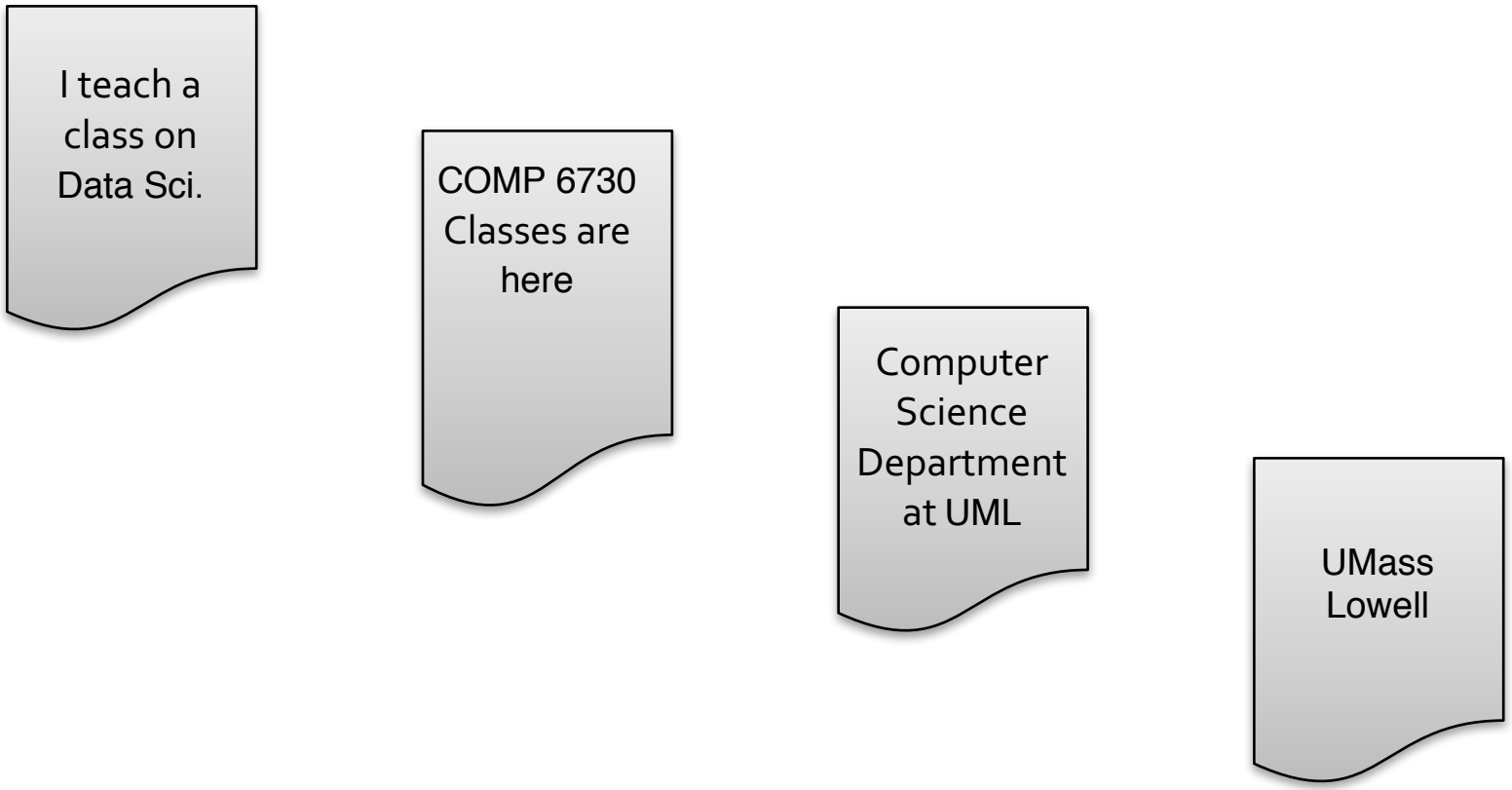
Q: What does the Web "look like" at a global level?

- **Web as a graph:**
  - Nodes = web pages
  - Edges = hyperlinks

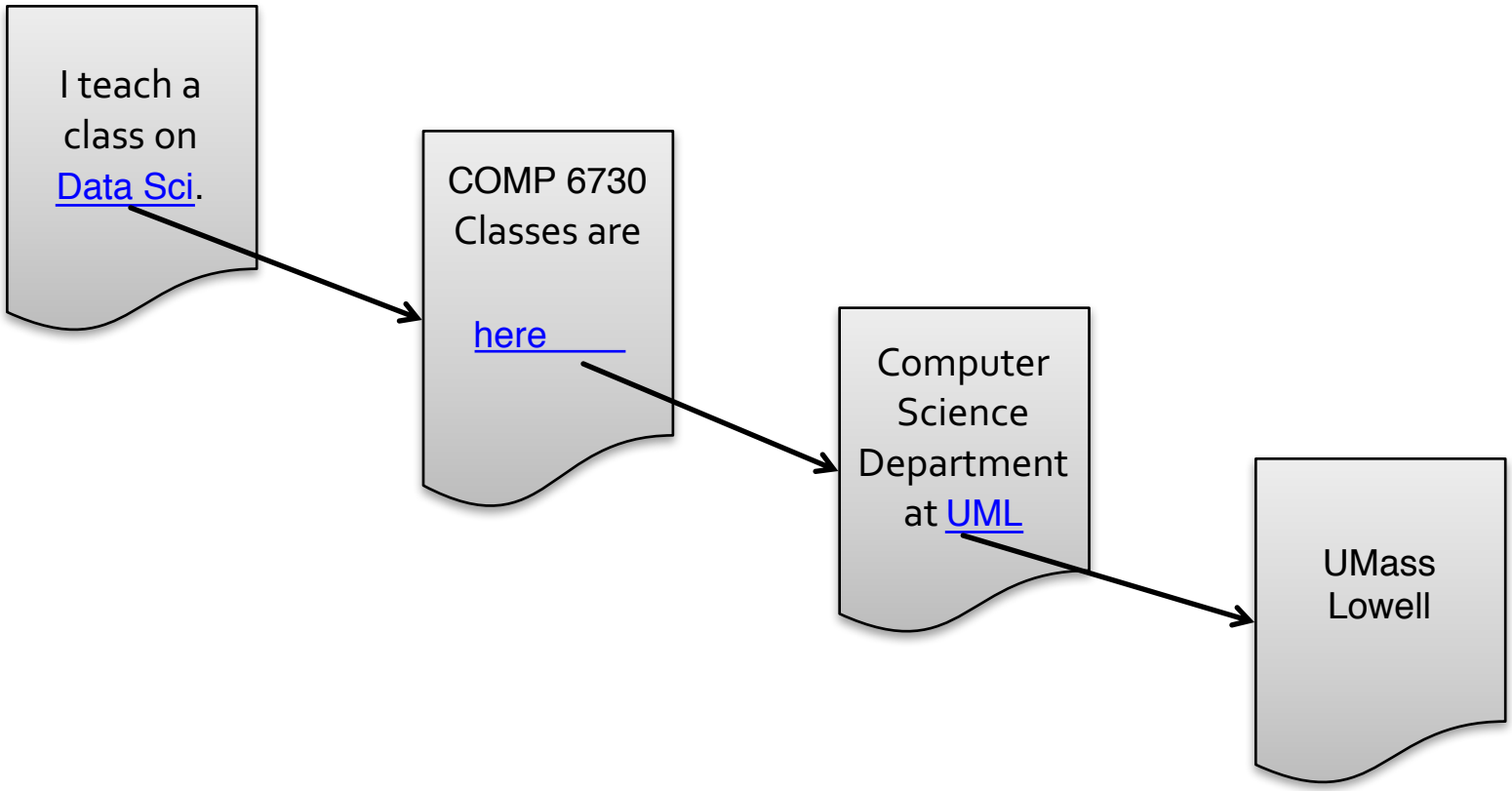


  - **Side issue:** What is a node?
    - Dynamic pages created on the fly
    - "dark matter" – inaccessible database generated pages

The Web as a Graph

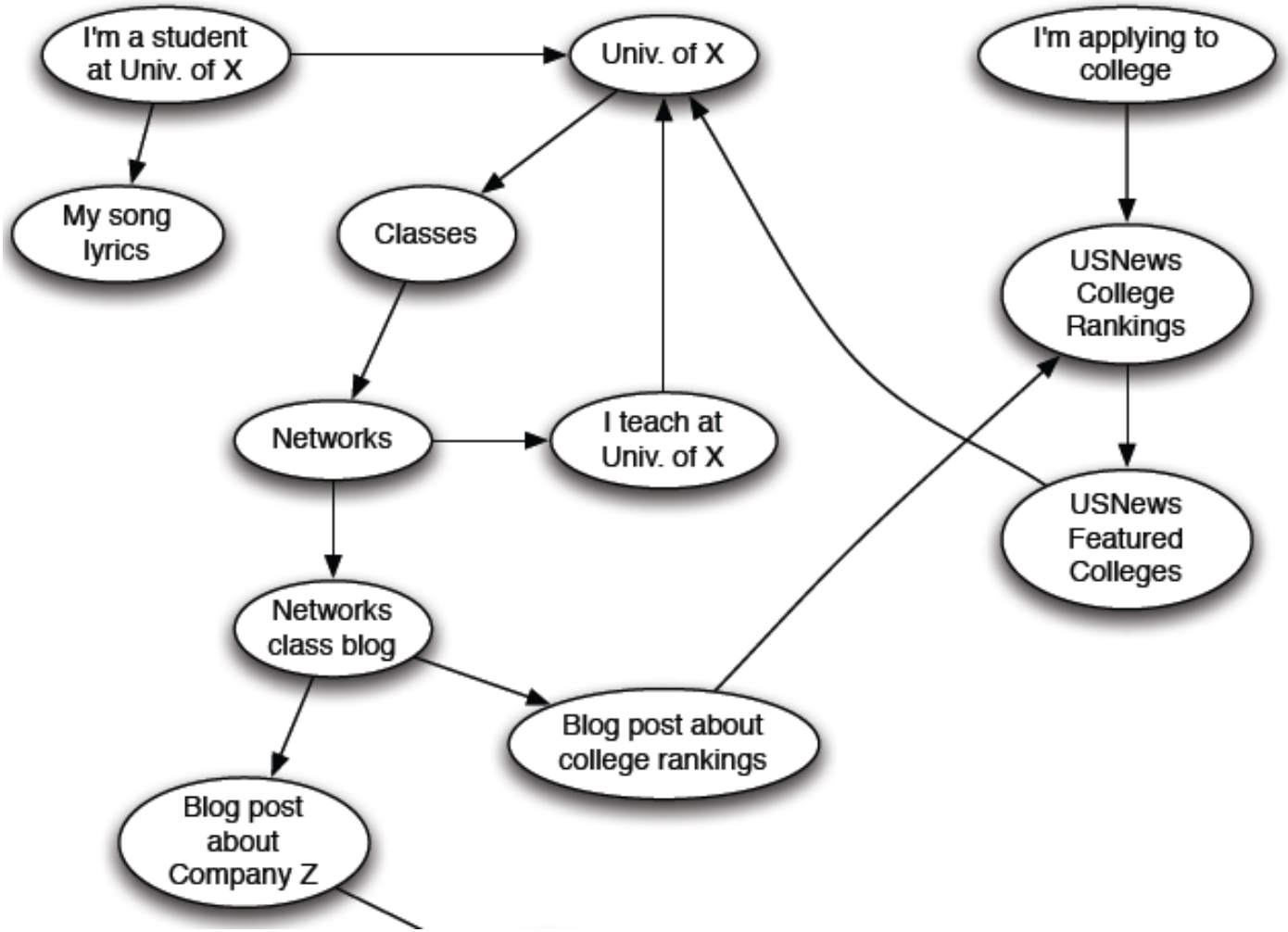


The Web as a Graph

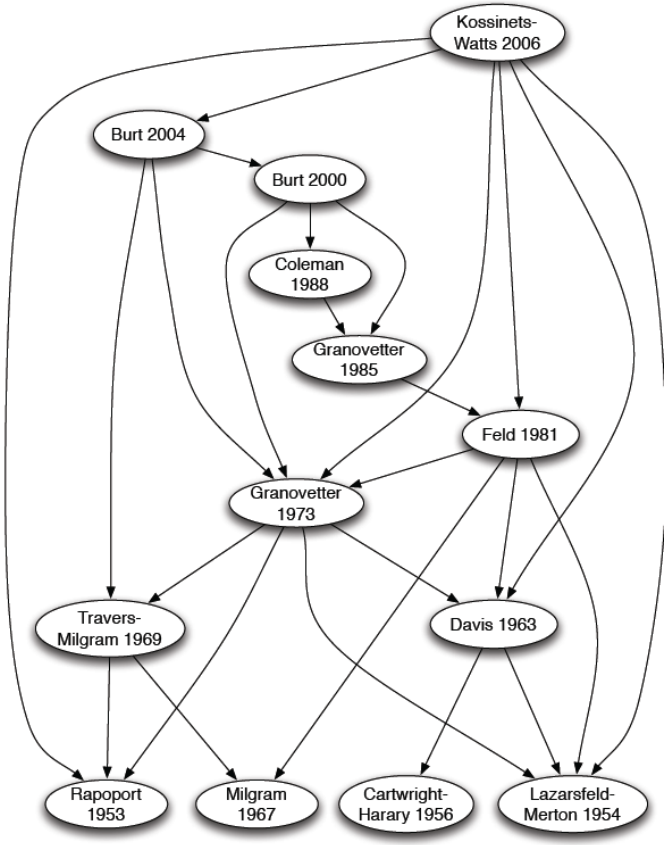


- In early days of the Web links were **navigational**
- Today many links are **transactional** (used not to navigate from page to page, but to post, comment, like, buy, …)

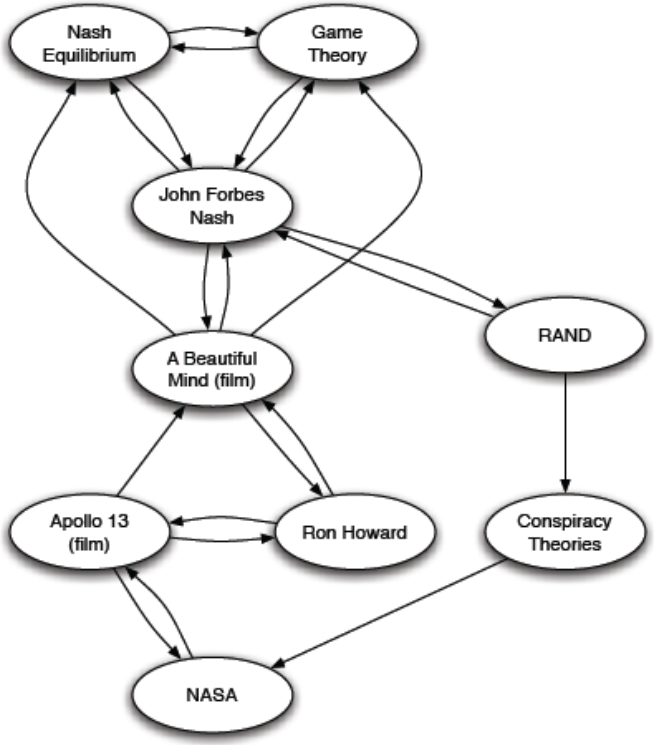
The Web as a Directed Graph



Other Information Networks



Citations



References in an Encyclopedia

Structure of the Web

- **Broder et al.: Altavista web crawl (Oct '99)**
  - Web crawl is based on a large set of starting points accumulated over time from various sources, including voluntary submissions.
  - 203 million URLS and 1.5 billion links
- **Computer:** Server with 12GB of memory

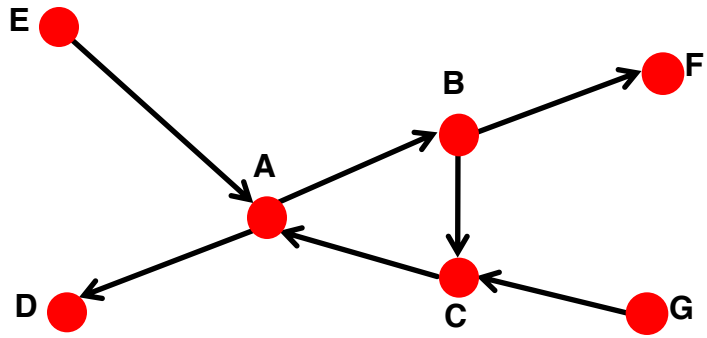
Tomkins, Broder, and Kumar

What Does the Web Look Like?

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

Web as a directed graph [Broder et al. 2000]:

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



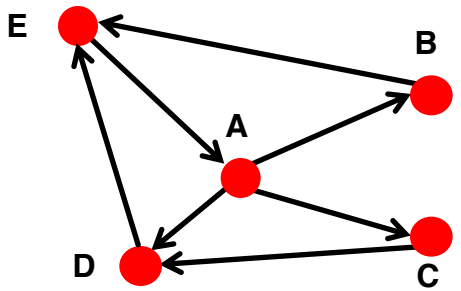
$$In(v) = \{w \mid w \text{ can reach } v\}$$
$$Out(v) = \{w \mid v \text{ can reach } w\}$$

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

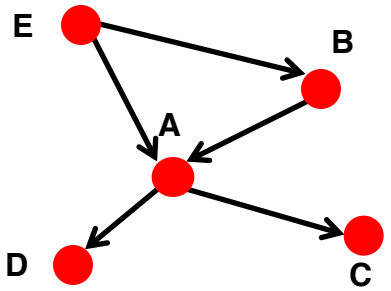
Reasoning about Directed Graphs

- **Two types of directed graphs:**
  - **Strongly connected:**
    - Any node can reach any node via a directed path

      $In(A)=Out(A)=\{A,B,C,D,E\}$



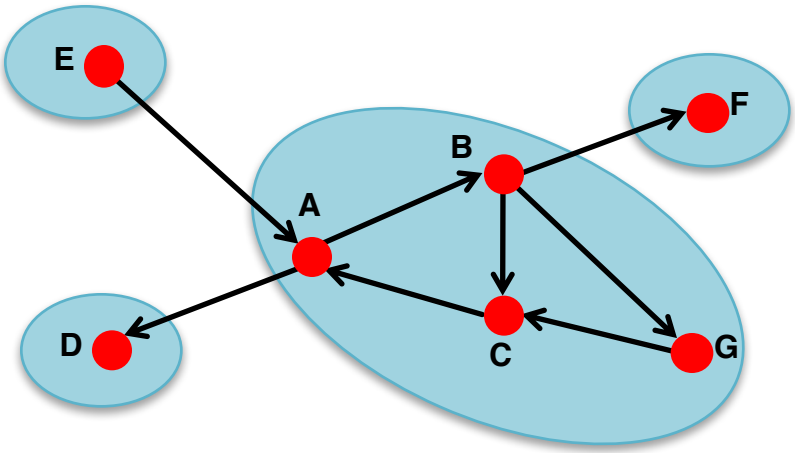
  - **Directed Acyclic Graph (DAG):**
    - Has no cycles: if u can reach v, then v cannot reach u



- **Any directed graph (the Web) can be expressed in terms of these two types!**
  - Is the Web a big strongly connected graph or a DAG?

Strongly Connected Component

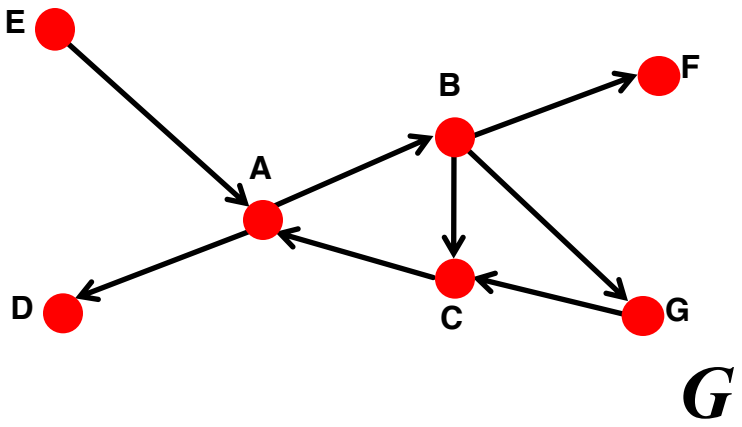
- **A Strongly Connected Component (SCC)** is a set of nodes S so that:
  - Every pair of nodes in S can reach each other
  - There is no larger set containing S with this property



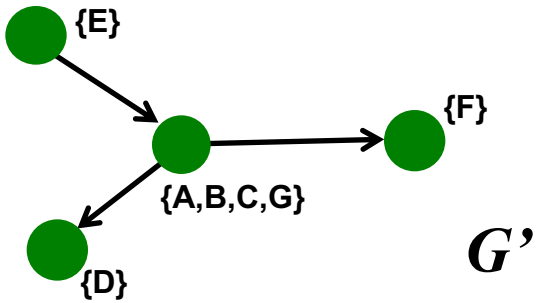
Strongly connected components of the graph: {A,B,C,G}, {D}, {E}, {F}

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

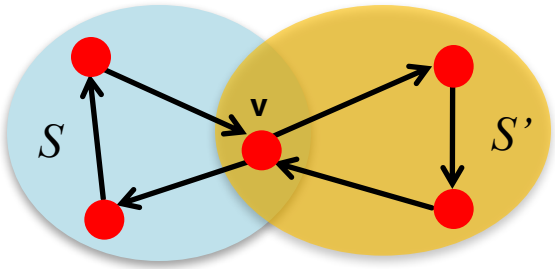


(1) Strongly connected components of graph G: {A,B,C,G}, {D}, {E}, {F}
(2) G' is a DAG:



Proof of (1)

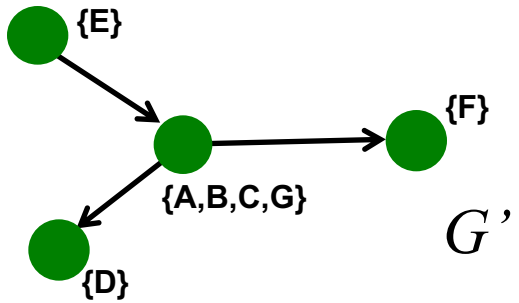
- **Claim: SCCs partition nodes of G.**
  - This means: Each node is member of exactly 1 SCC
- Proof by contradiction:
  - Suppose there exists a node v which is a member of two SCCs S and S'

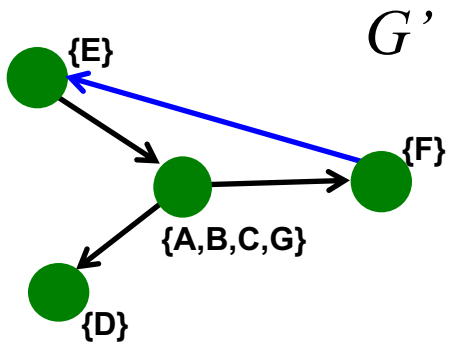


  - But then $S \cup S'$ is one large SCC!
    - **Contradiction:** By definition SCC is a maximal set with the SCC property, so S and S' were not two SCCs.

Proof of (2)

- **Claim: G' (graph of SCCs) is a DAG.**
  - This means: G' has no cycles
- Proof by contradiction:
  - Assume G' is not 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!

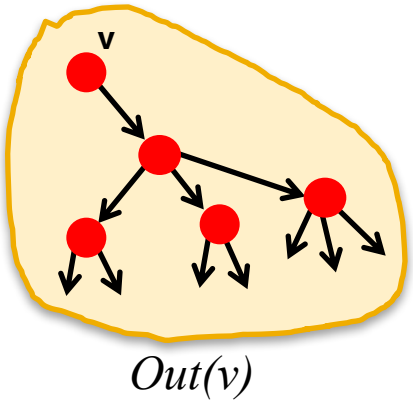
Back to…

How is the Web linked?

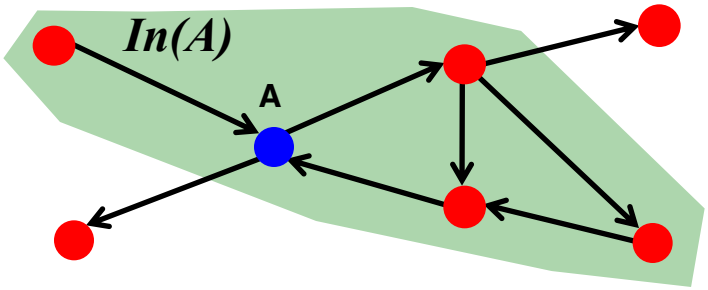
Goal: Take a large snapshot of the Web and try to understand how its SCCs "fit together" as a DAG

Graph Structure of the Web

- **Computational issue:**
  - Want to find a SCC containing node v?

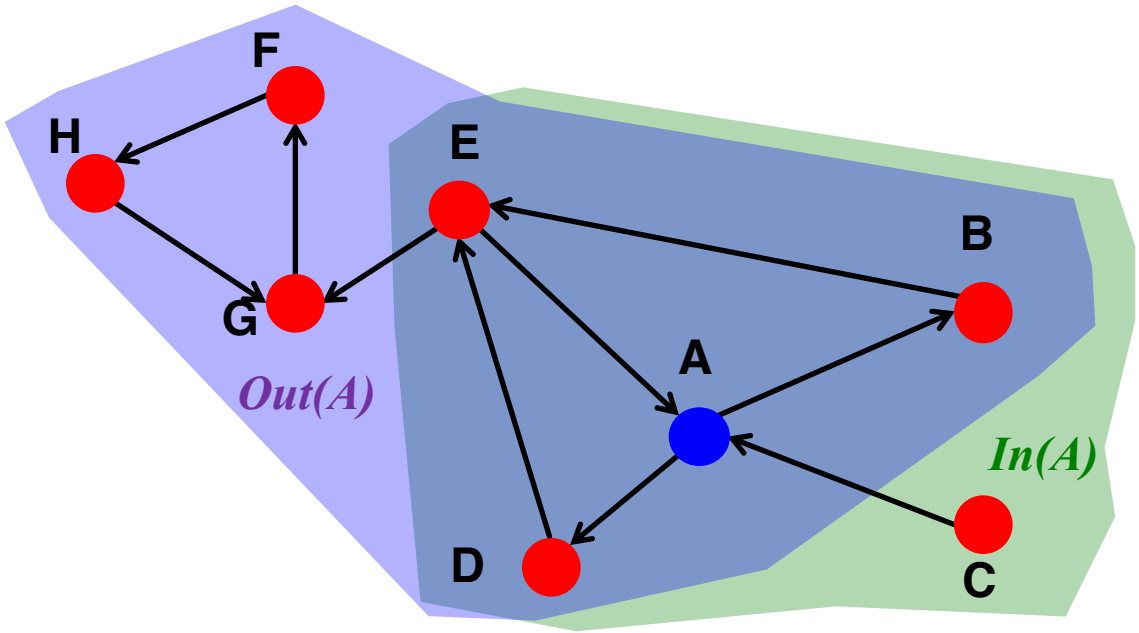


- **Observation:**
  - ***Out(v)*** … nodes that can be reached from v
  - **SCC containing v is:** $Out(v) \cap In(v)$
    $= Out(v,G) \cap Out(v,G')$, where G' is G with all edge directions flipped



Out(A) ∩ In(A) = SCC

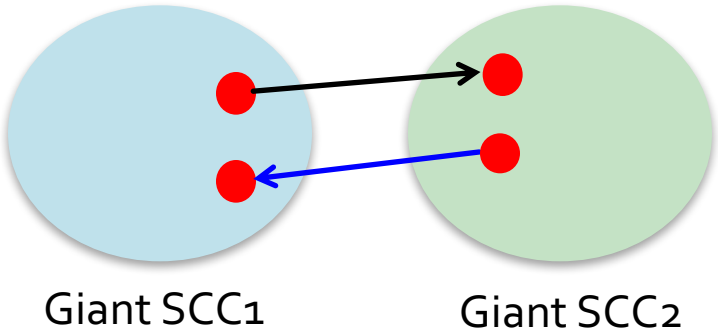
- **Example:**



  - 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
- **Why only 1 big SCC? Heuristic argument:**
  - Assume two equally big SCCs.
  - It just takes 1 page from one SCC to link to the other SCC.
  - If the two SCCs have millions of pages the likelihood of this not happening is very very small.

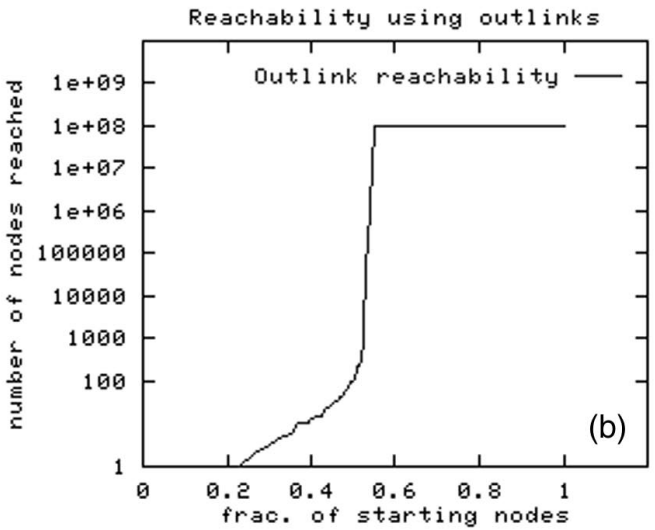


Structure of the Web

- **Directed version of the Web graph:**
  - Altavista crawl from October 1999
    - 203 million URLs, 1.5 billion links

Computation:

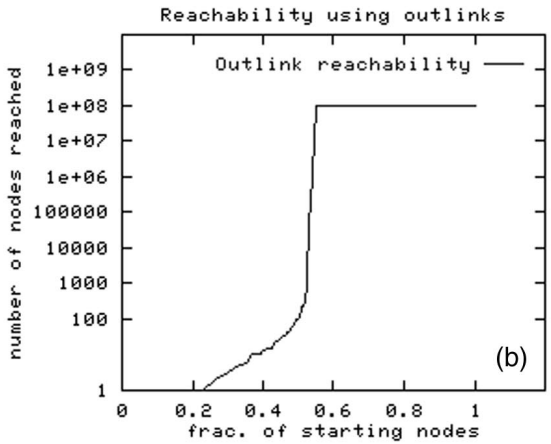
- Compute IN(v) and OUT(v) by starting at random nodes.
- **Observation:** The BFS either visits many nodes or gets quickly stuck.



Structure of the Web

Result: Based on IN and OUT of a random node *v*:

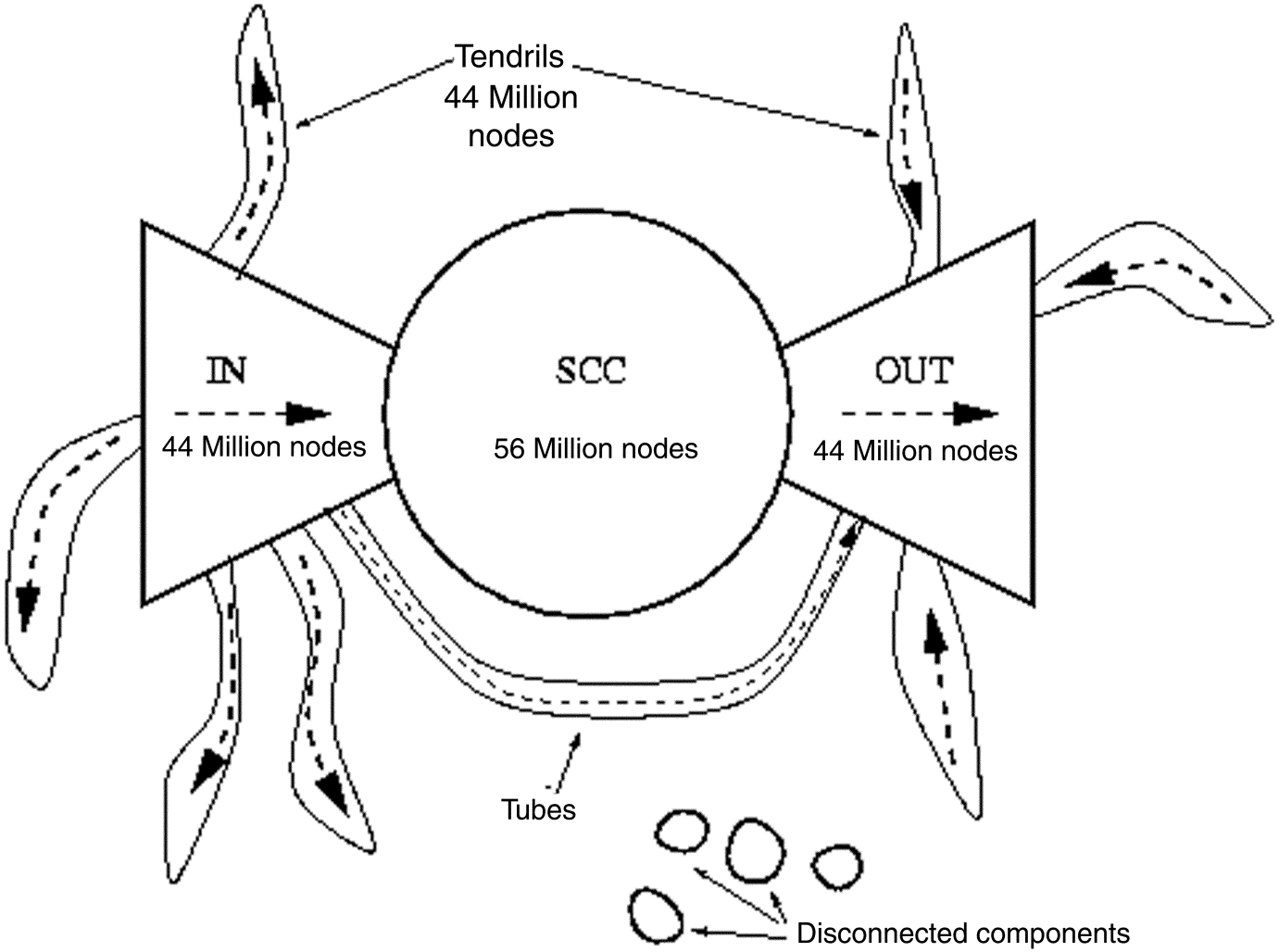
- **Out(v)** ≈ 100 million (**50%** nodes)
- **In(v)** ≈ 100 million (**50%** nodes)
- **Largest SCC:** 56 million (**28%** nodes)



x-axis: rank
y-axis: number of reached nodes

- **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

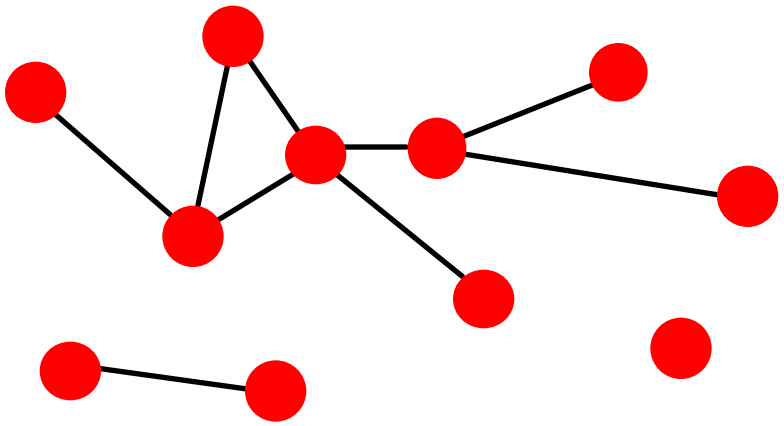
(1) Degree Distribution

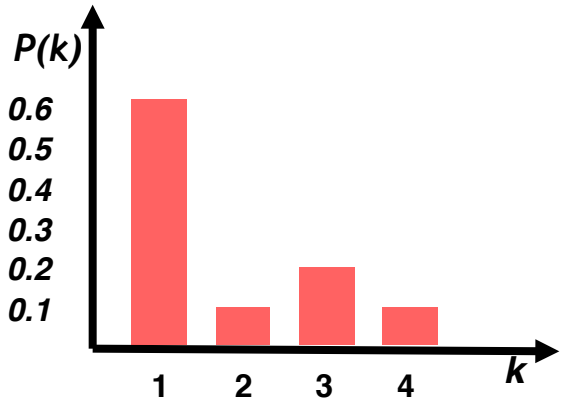
- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

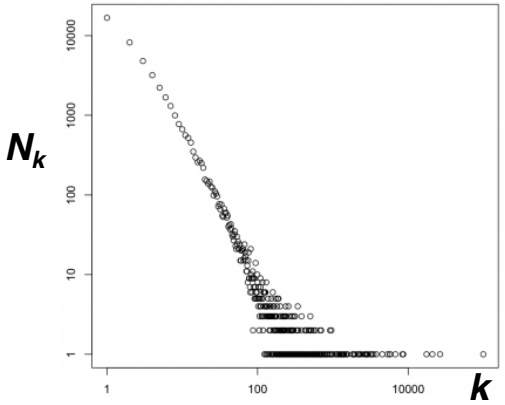
  N_k = # nodes with degree k

- Normalized histogram:

  $P(k) = N_k / N$ → **plot**





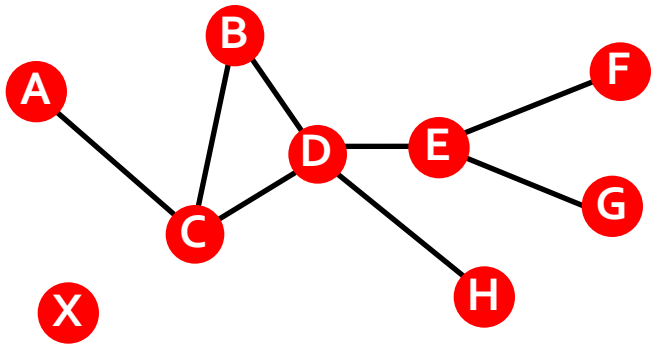


(2) Paths in a Graph

- 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 $A_{uk}A_{kv}=0$

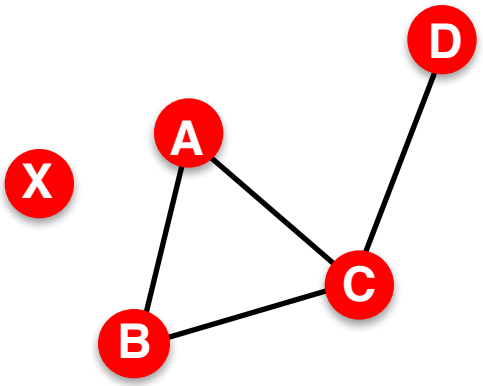
$$H_{uv}^{(2)} = \sum_{k=1}^{N} A_{uk}A_{kv} = [A^2]_{uv}$$

  - **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

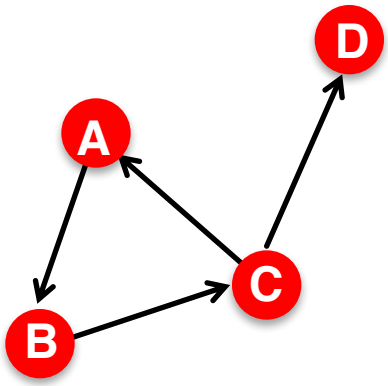
$$H_{uv}^{(h)} = [A^h]_{uv}$$

(holds for both directed and undirected graphs)

Distance in a Graph



$h_{B,D} = 2$
$h_{A,X} = \infty$



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

- **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 not connected, 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

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

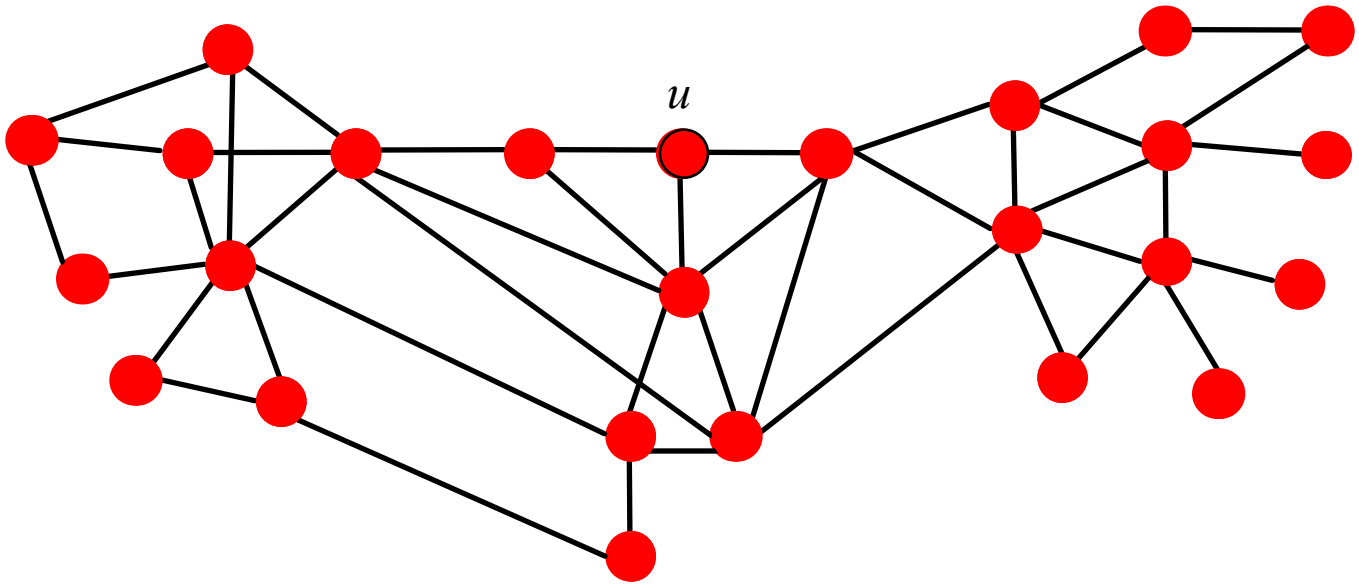
where h_{ij} is the distance from node i to node j

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

Finding Shortest Paths

Extra

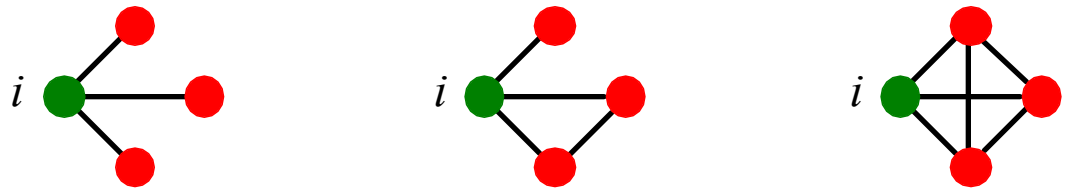
- **Breadth First Search:**
  - Start with node u, mark it to be at distance $h_u(u)=0$, add u to the queue
  - While the queue not empty:
    - Take node v off the queue, put its unmarked neighbors w into the queue and mark $h_u(w)=h_u(v)+1$



(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



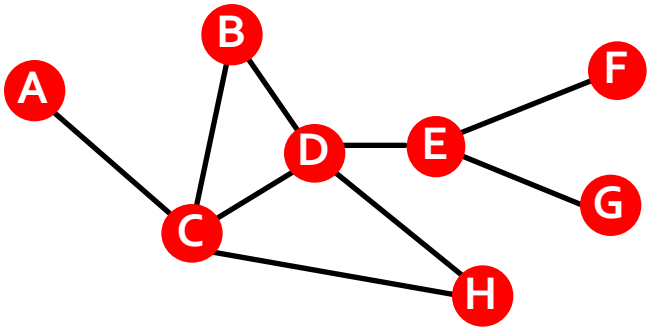
C_i=0 C_i=1/3 C_i=1

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

Clustering Coefficient

- **Clustering coefficient:**
  - What portion of ***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

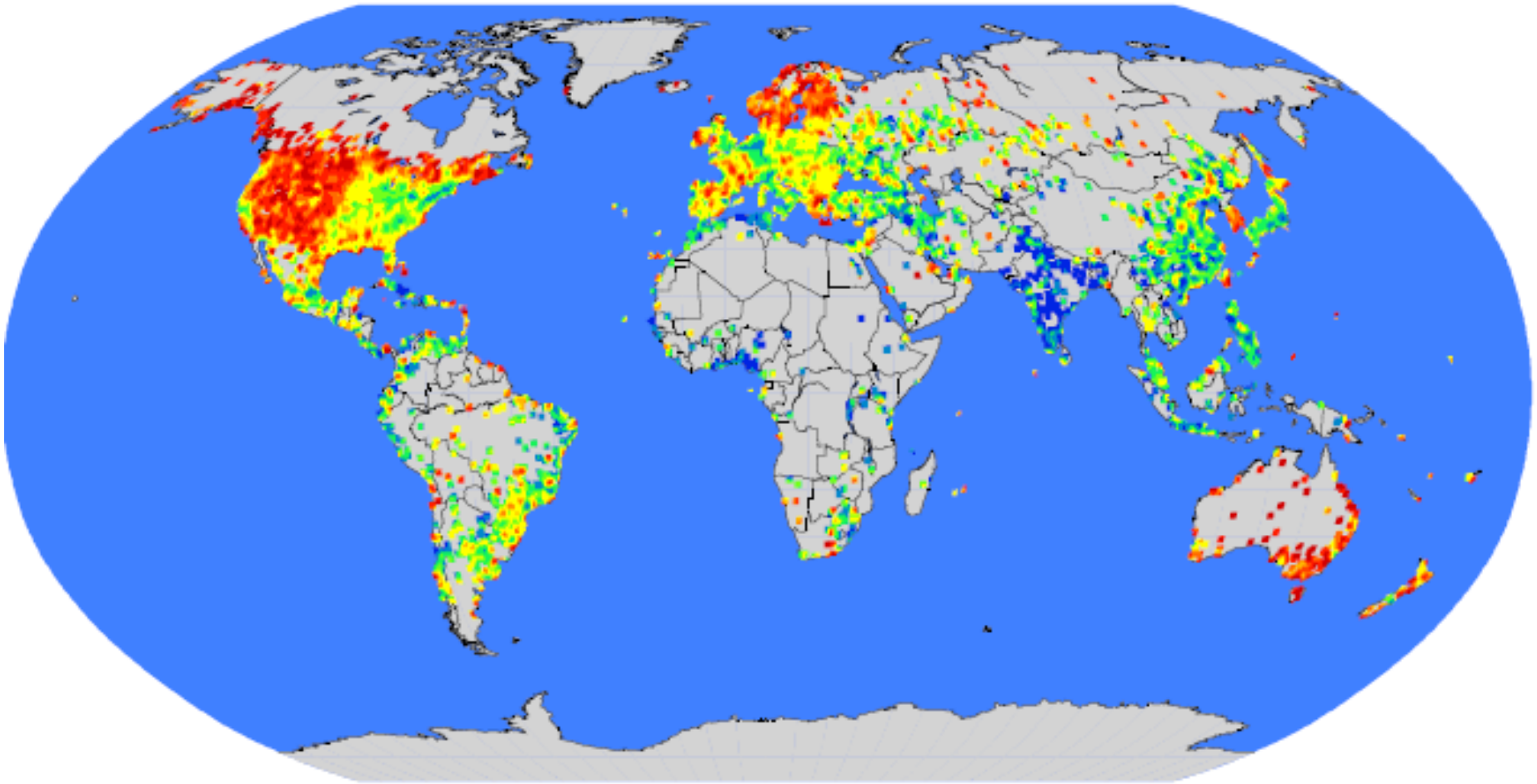
Clustering coefficient: C

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

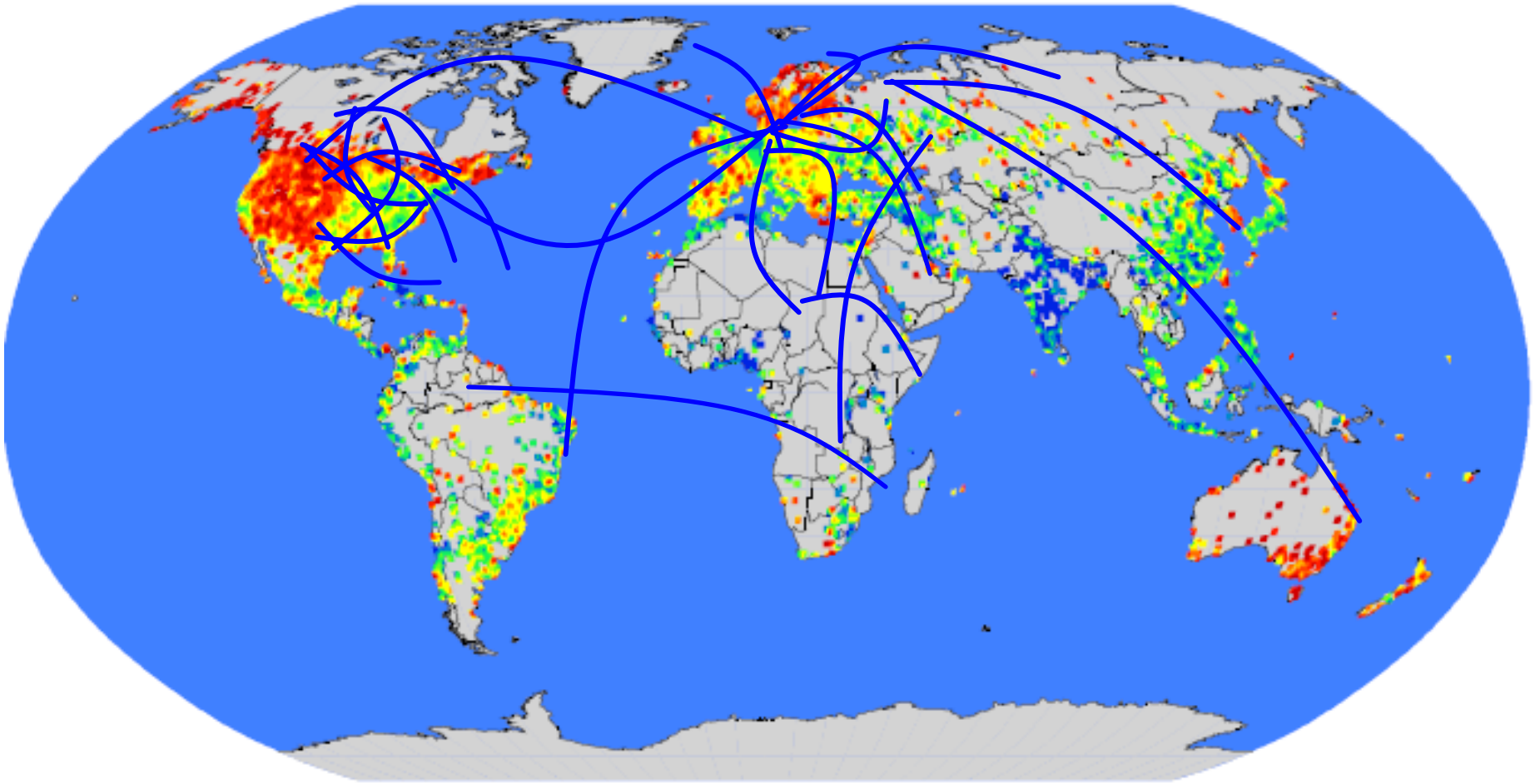
MSN Messenger

- **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

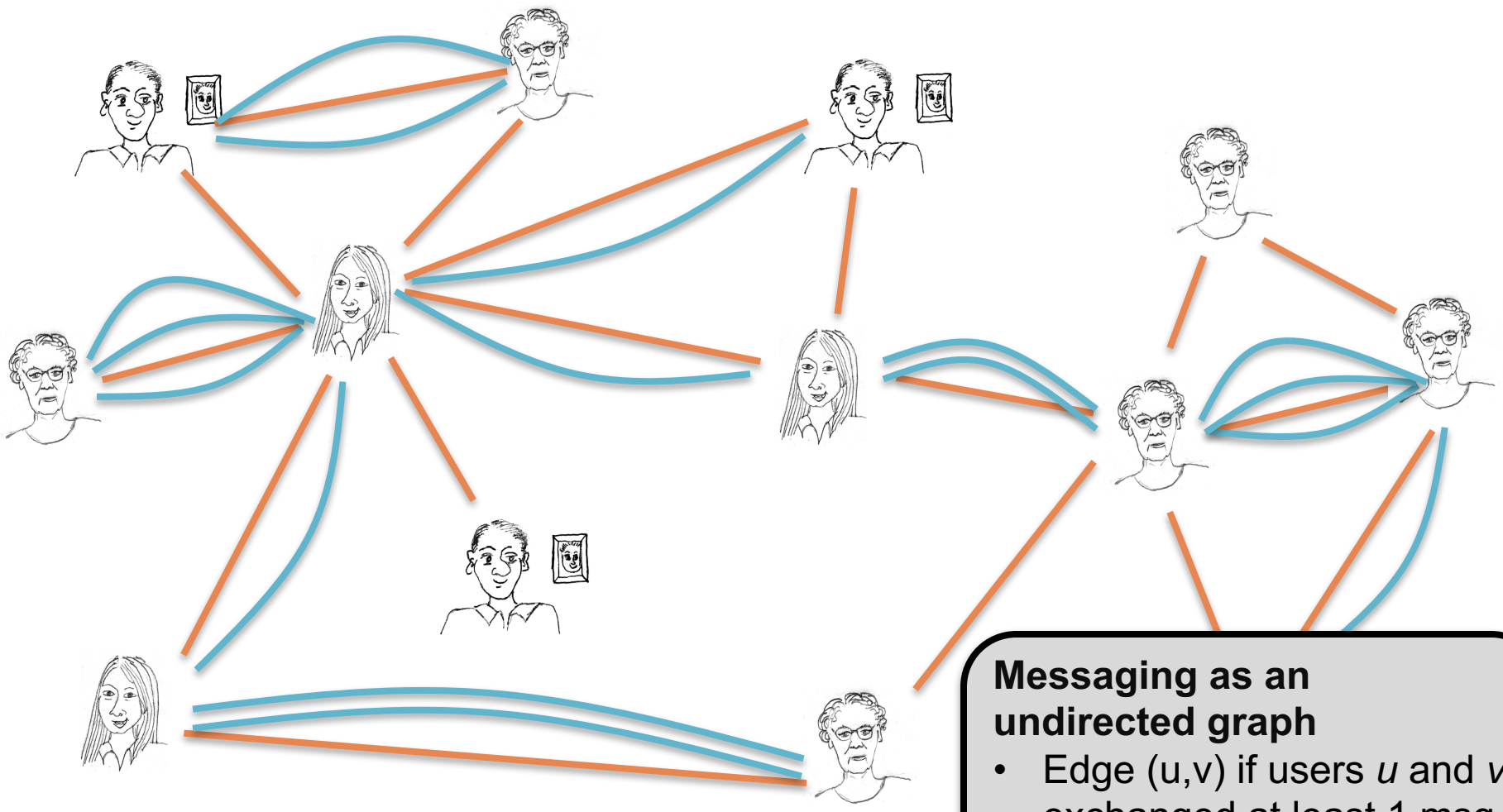


Communication Network



Network: 180M people, 1.3B edges

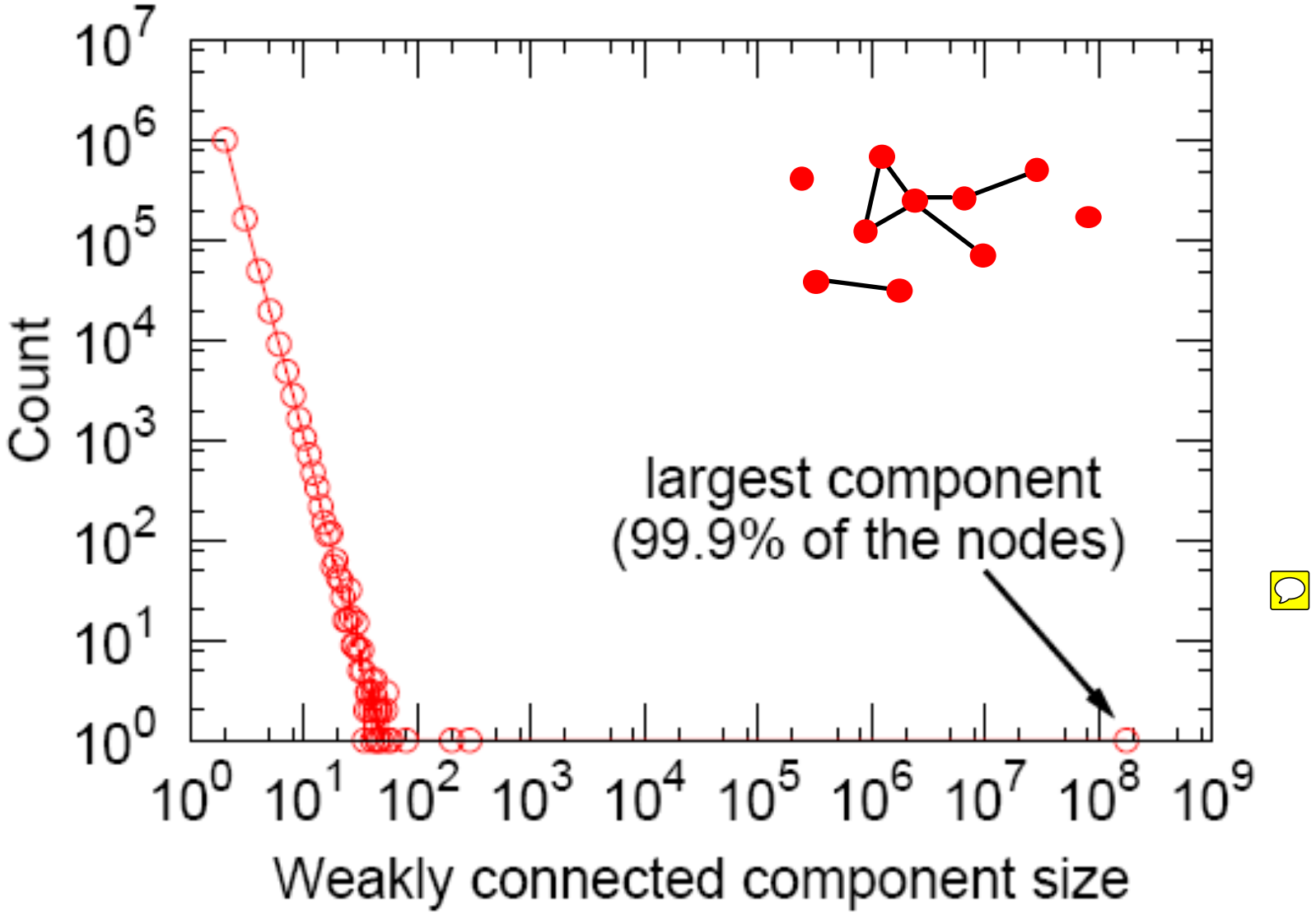
Messaging as a Multigraph



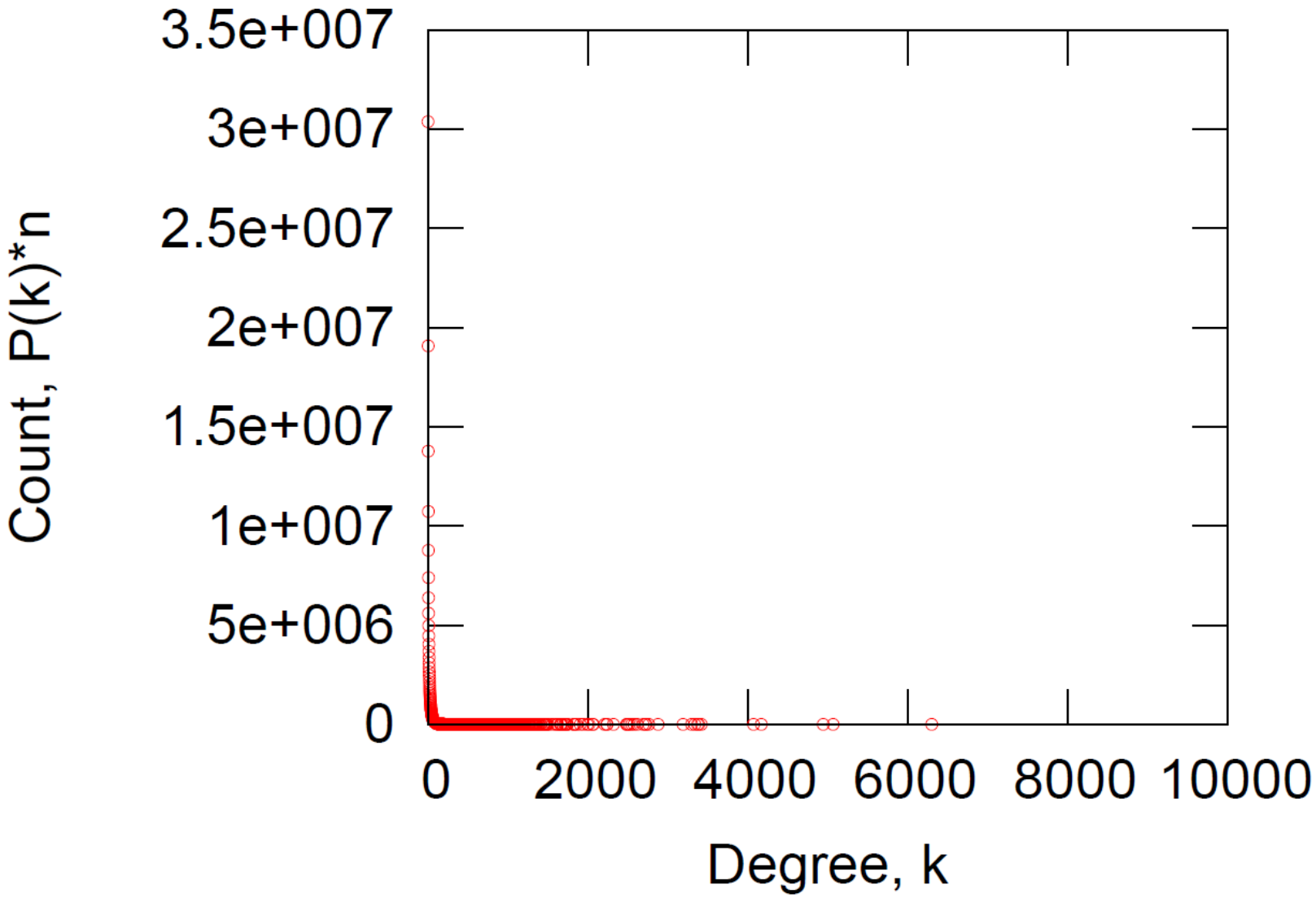
Messaging as an undirected graph

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

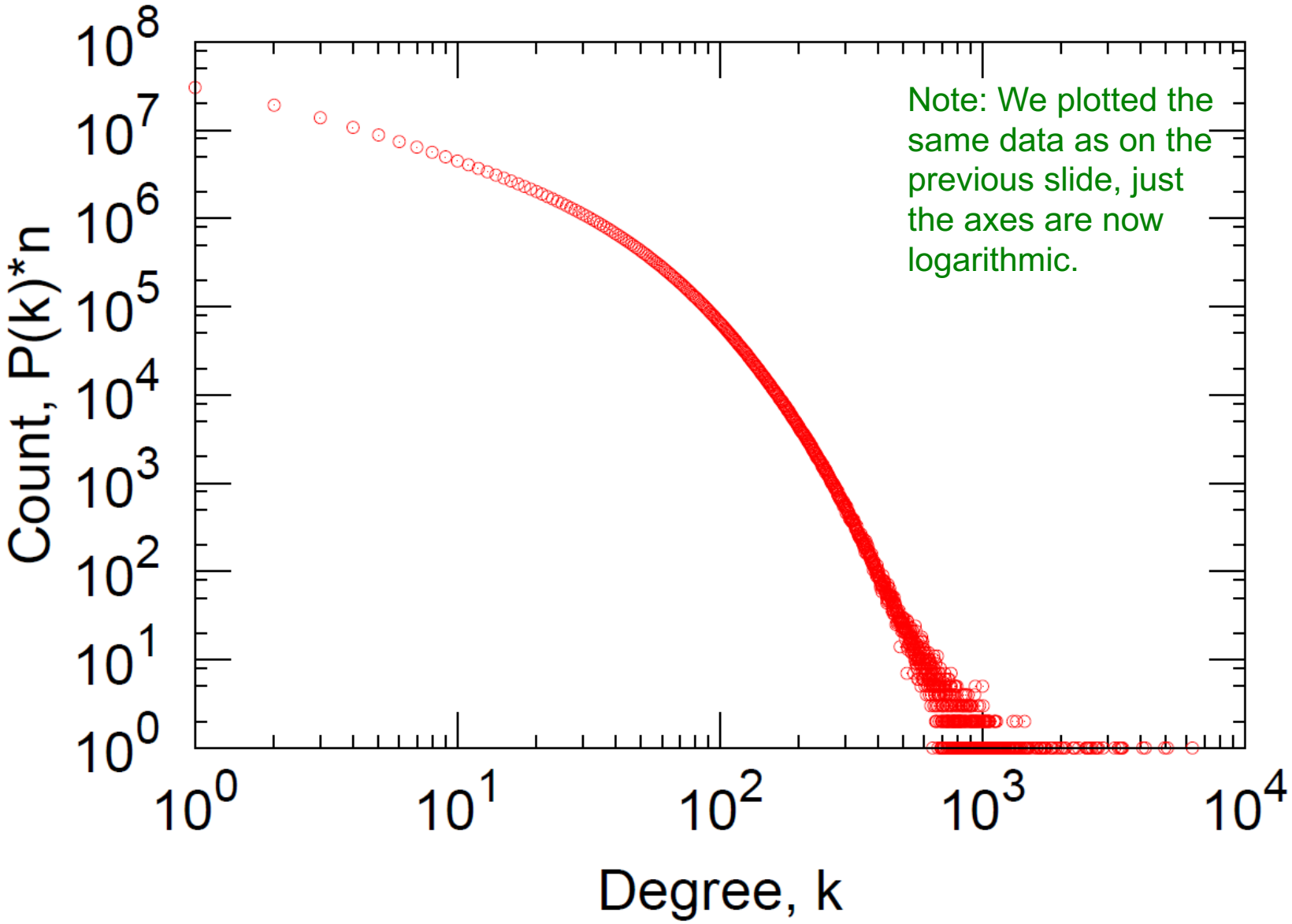
MSN: (1) Connectivity



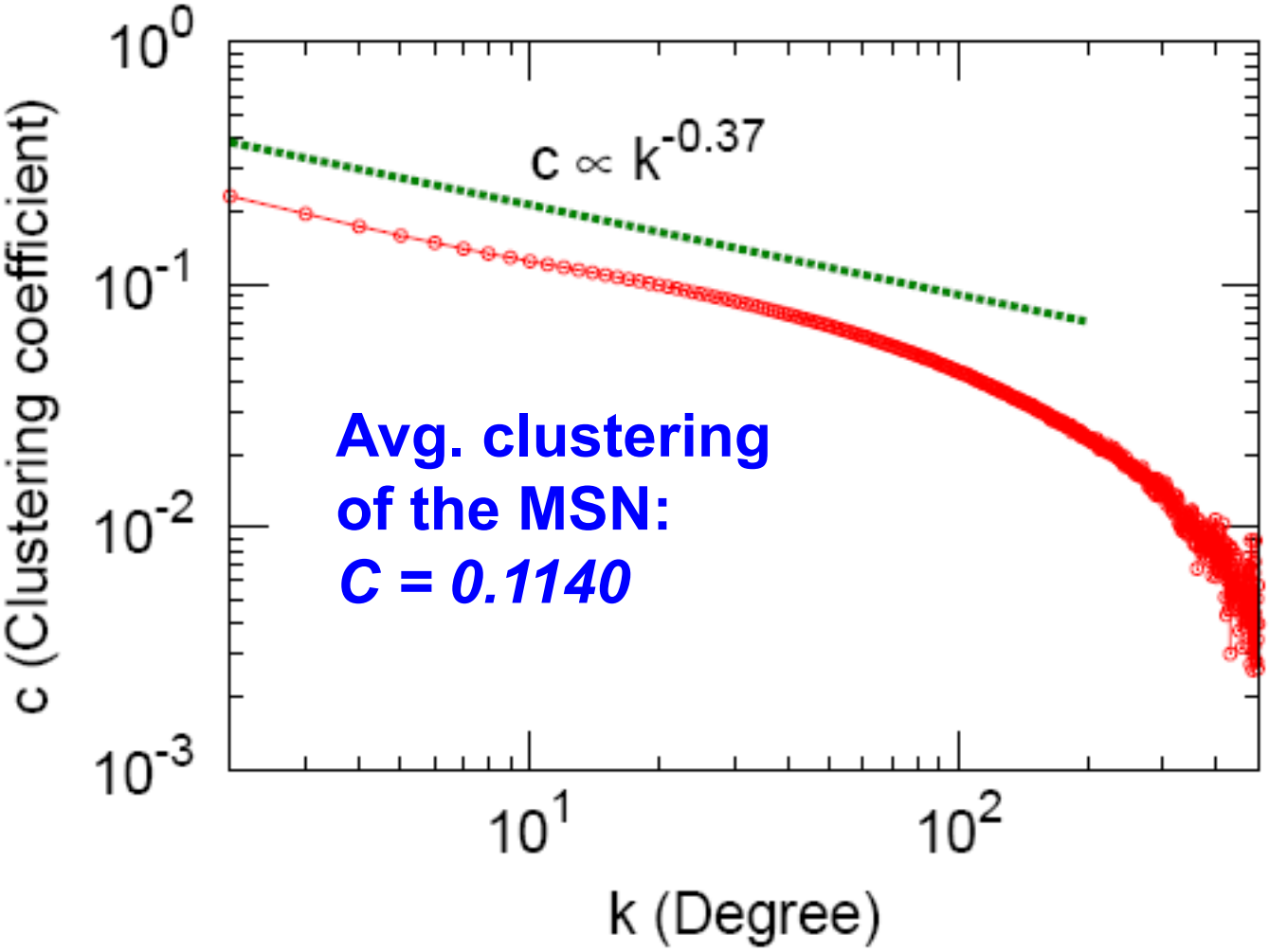
MSN: (2) Degree Distribution



MSN: Log-Log Degree Distribution

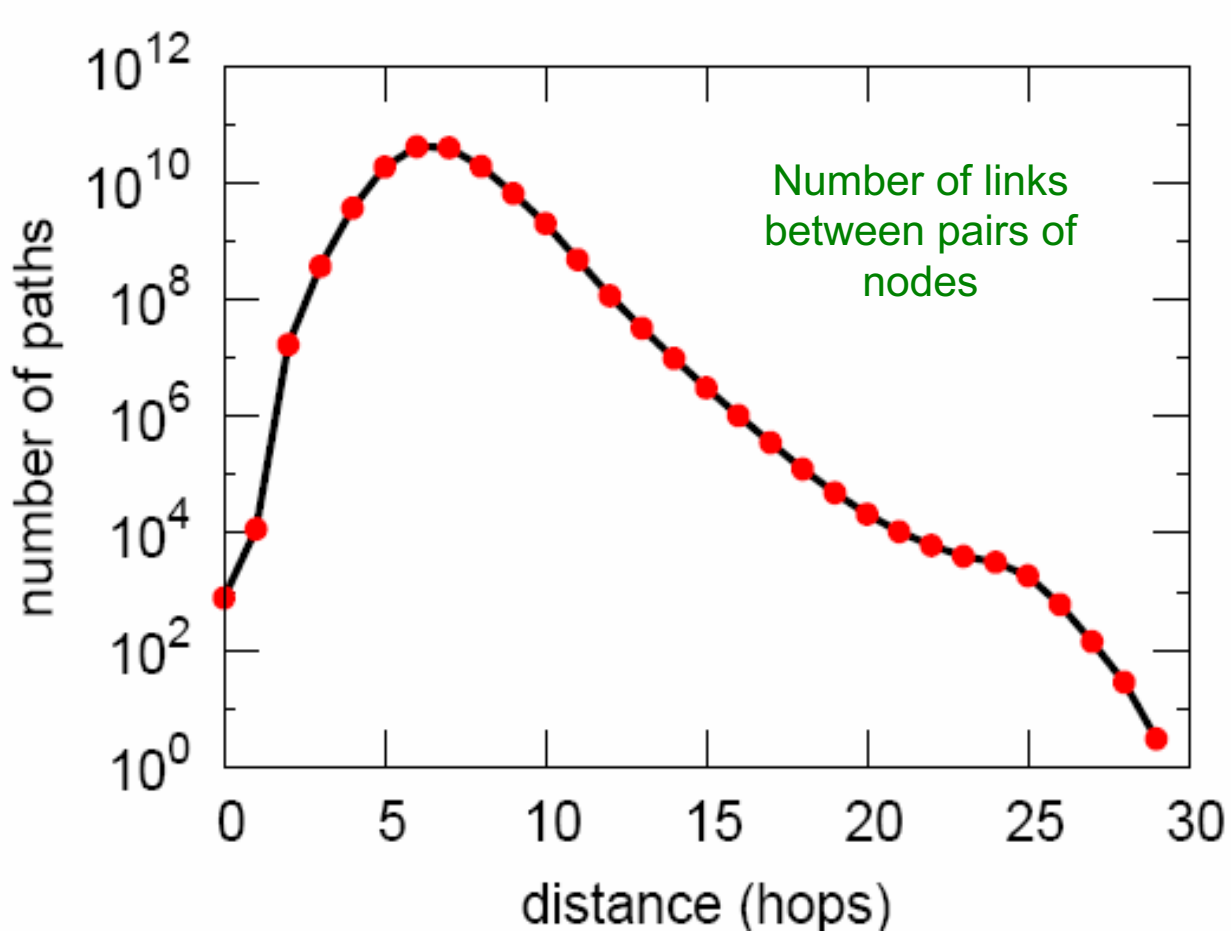


MSN: (3) Clustering



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

nodes as we do BFS out of a random node

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

MSN: Key Network Properties

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

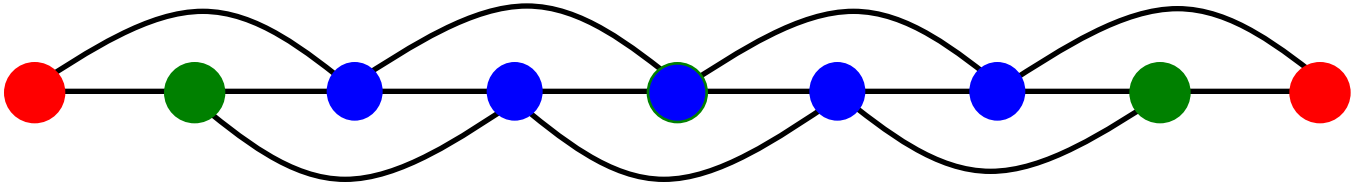
Path length: *6.6*

Clustering coefficient: *0.11*

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

To answer this we need a null-model!

Is MSN Network like a "chain"?



- $P(k) = \delta(k\text{-}4)$ $k_i = 4$ for all nodes
- $C = \frac{1}{N}\left((N-4)\frac{1}{2} + 2 \cdot 1 + 2\frac{2}{3}\right) = ½$ as $N \to \infty$
- Path length: $h_{max} = \frac{N-1}{2} = O(N)$

- So, we have: **Constant degree,**

  Constant avg. clustering coeff.

  Linear avg. path-length

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

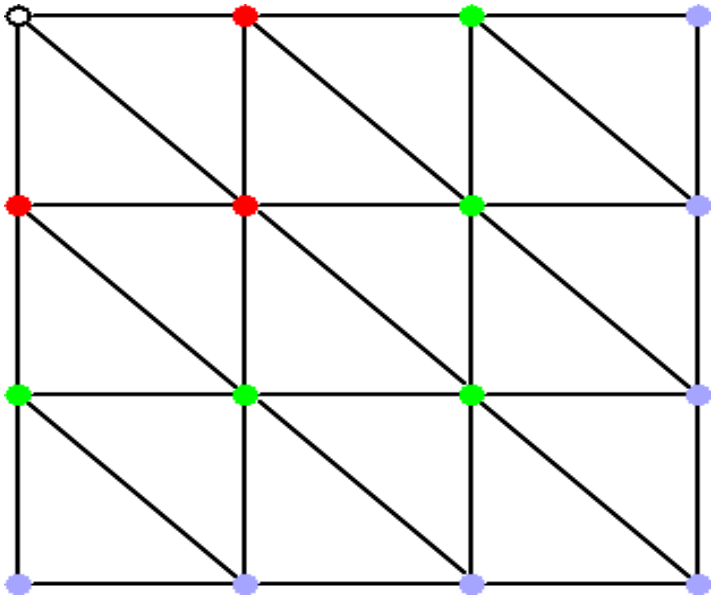
Note about calculations:
We are interested in quantities as graphs get large ($N \to \infty$)

We will use big-O:
$f(x) = O(g(x))$ as $x \to \infty$
if $f(x) < g(x)*c$ for all $x > x_0$
and some constant c.

Is MSN Network like a "grid"?

- $P(k) = \delta(k\text{-}6)$
  - $k = 6$ for each inside node
- $C = 2/5$ for inside nodes
- **Path length:**

$$h_{max} = O(\sqrt{N})$$



- **In general, for lattices:**
  - Average path-length is $\bar{h} \approx N^{1/D}$ (D… lattice dimensionality)
  - Constant degree, constant clustering coefficient

What did we learn so far?

MSN Network is neither a chain nor a grid

Erdös-Renyi Random Graph Model

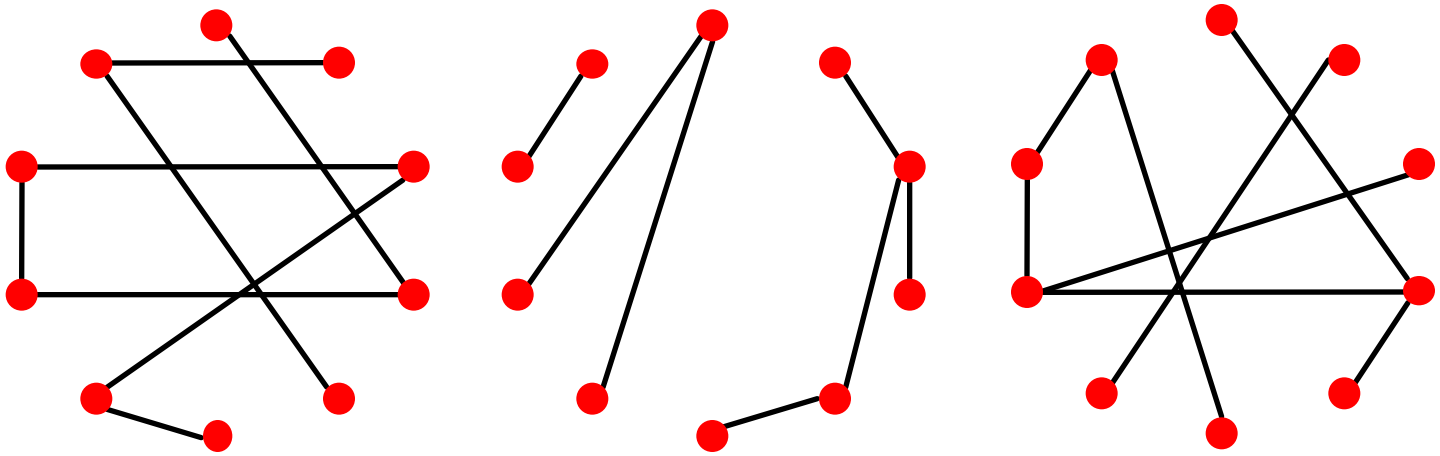
Simplest Model of Graphs

- **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 kind of networks do such models 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