Graph Fundamentals 2

Sarunas Girdzijauskas

ID2211

March 2019

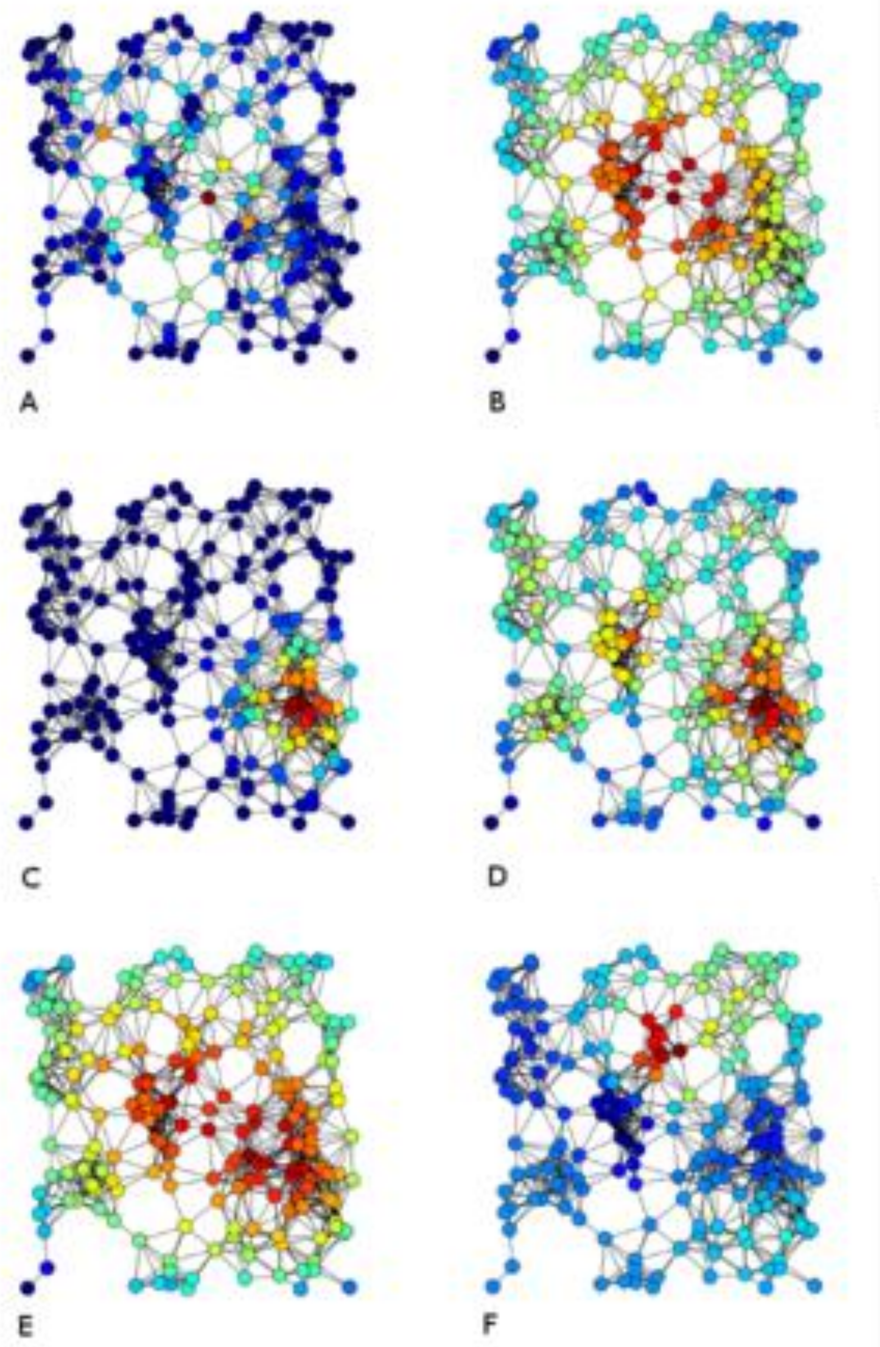
Recap

- Basic Notations
- Type of graphs
- Paths/cycles/Connectivity/Giant component
- Centrality measures (almost finished)

"Importance" Centrality

- Idea:
  - Importance of a node depends on the importance of its neighbors
    - Recursive definition!
    - $v_i \leftarrow \sum_j A_{ij} v_j$
  - E.g, let's start with value "1" at each edge and calculate the importance of all the nodes
    - What happens now?
    - What bad can happen?
    - Divergence! How can we fix it?
    - $v_i \leftarrow \mathbf{1/\lambda} * \sum_j A_{ij} v_j$
    - in matrix terms: $\mathbf{Av} = \lambda \mathbf{v}$
    - Importance of the nodes is described by **v**, which is Principal Eigenvector of **A** ($\lambda = \lambda_1$)
  - I.e., This is **EigenVector Centrality**.
- Similar centrality measures:
  - Katz centrality

normalizing



Examples

- From Wikipedia:
- A) Betweenness centrality,
- B) Closeness centrality,
- C) Eigenvector centrality,
- D) Degree centrality,
- E) Harmonic centrality and
- F) Katz centrality of the same graph.

- Absolute centrality measures might not be that always important. Rank is important (sorted list of centralities)
  - If you want to compare two ranks you can use Kendall tau rank

Kahoot Time!

Metrics Comparison

- Let’s sort all the nodes by a centrality measures and rank them for each metric (betweeneness, eigen, degree, etc)
- Do we get similar rank all the time? How to evaluate?
- E.g., take ranks produced by a quadratic function and a linear function
  - Both of them monotonically increasing and will give the same ranks
  - But simple correlation coefficient will not reveal perfect correlation.

Kendall tau rank

- From information retrieval (e.g., search results by Google and Bing)
  - Counts pairwise agreements (disagreements) between two ranks lists.
  - n_c – number of concordant pairs
  - n_d – number of discordant pairs

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

  - Perfect agreement when tau=1, complete disagreement tau=-1
  - Example Rank1: A B C D E. Rank 2: D C A B E

concordant

discordant

Recap

- Basic Notations
- Type of graphs
- Paths/cycles/Connectivity/Giant component
- Centrality measures
- **Clustering coefficient**

Clustering coefficient

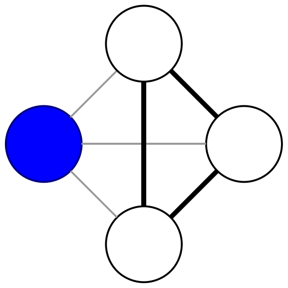
- **Local *clustering coefficient*** $C(v)$ of vertex v is given by the fraction of:

$$C(v) = \frac{e(v)}{deg(v)\,(deg(v)-1)\,/2}$$

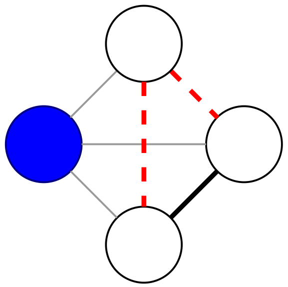
where $e(v)$ denotes the links between the vertices within the neighborhood of v

- **Network average *clustering coefficient*** $\tilde{C}$ is given by the fraction of:

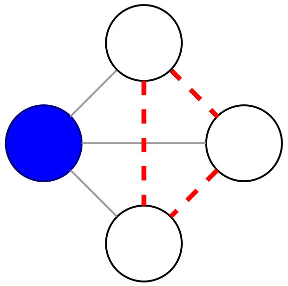
$$\tilde{C} = \frac{1}{N}\sum_{i=1}^{N} C(i)$$



c = 1



c = 1/3



c = 0

How to interpret clustering coef.?

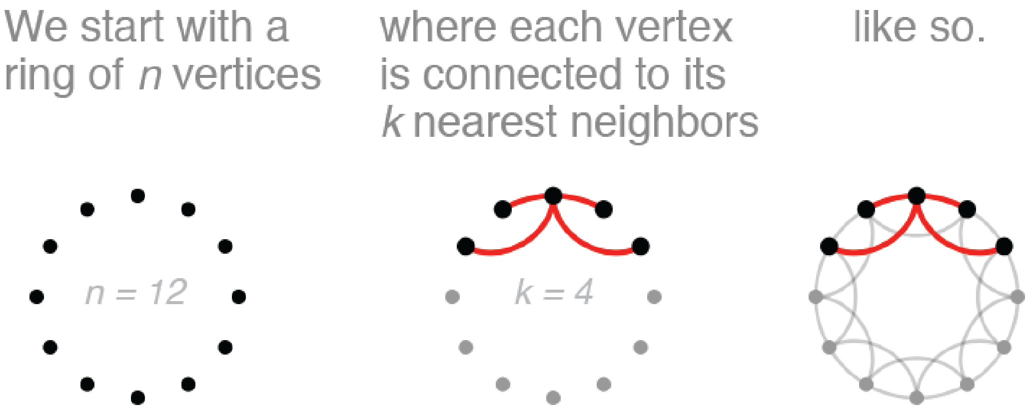
- Clustering coefficient denotes what is the fraction of your neighbors are neighbors themselves
- Compare to a purely random chance that the “triangles” form.
- Edge density of a network:
  - E is total number of edges
  - P is the probability that two nodes are connected in a random graph

$$\mathrm{p} = \frac{E}{0{,}5 * N(N-1)}$$

- If C(G)>>p then we can claim that the **graph is clustered**

Examples

- Regular graph with degree k connected to nearest neighbors



- What's clustering coefficient whe
  - Possible neighbor friendships: 6
  - Actual friendships: 3
- Clustering coef 3/6=0,5
- Compare it with random graph?
- What if the graph has 1000x more nodes?
- What's a clustering coefficient of bipartite graph?

Recap

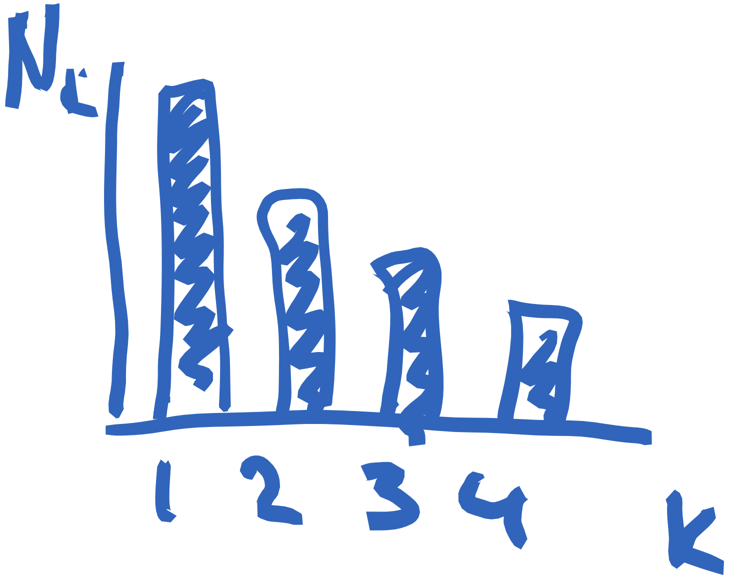
- Type of graphs
- Paths/cycles/Connectivity/Giant component
- Centrality measures
- Clustering coefficient

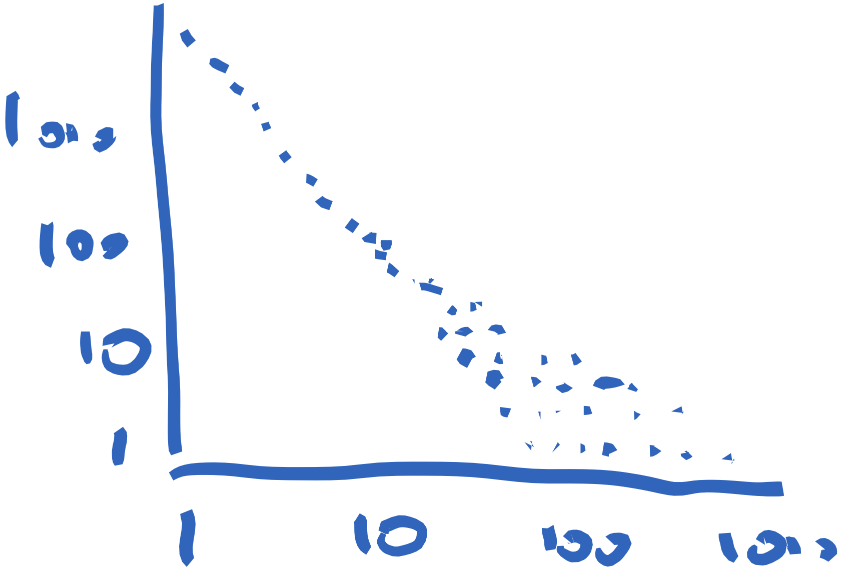
- What’s left?

- **Degree distributions**

Degree Distribution

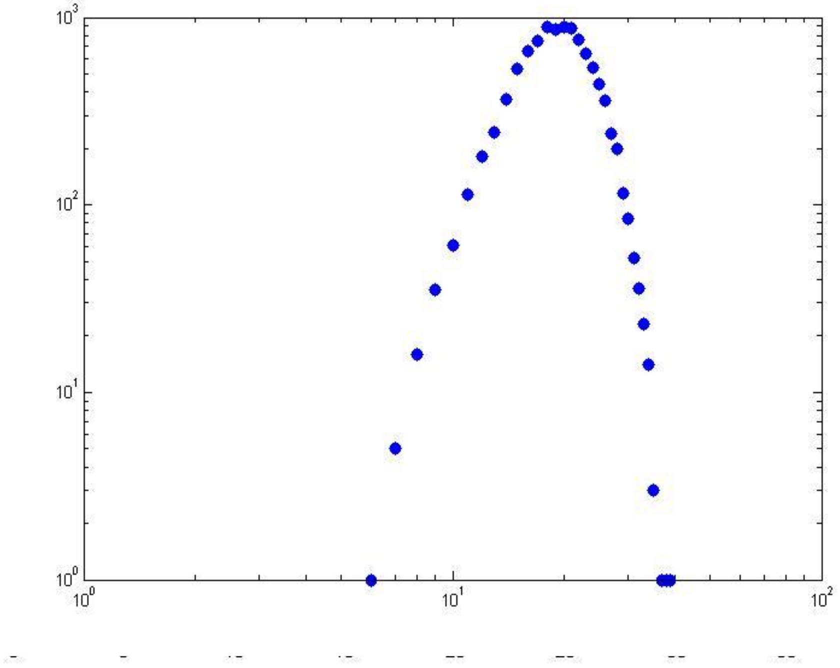
- N_k is the number of nodes with degree k
- $P(k)$ is the probability that a randomly chosen node has degree k.
  - $P(k) = N_k / N$, i.e., normalized
  - Often power-law distributions (linear in loglog scale)



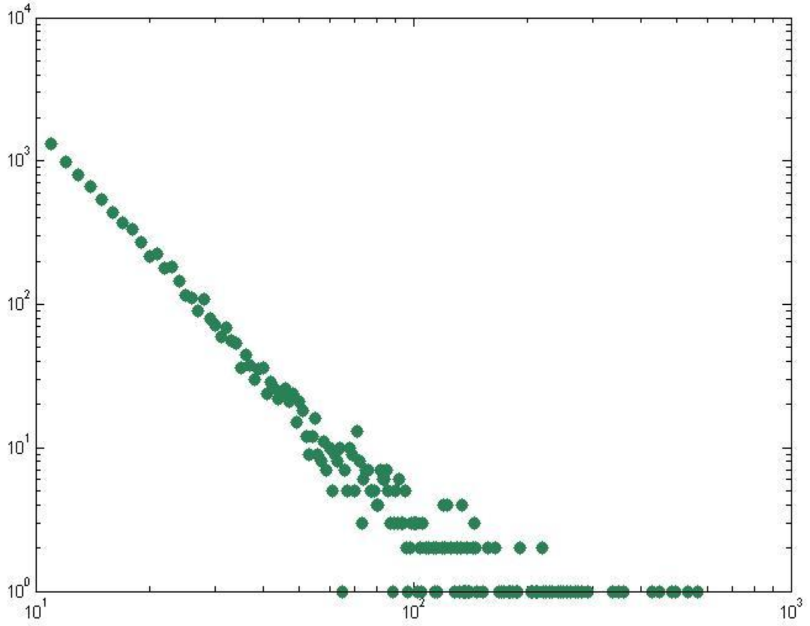


More degree distributions

- Normal vs. power-law distributions
- N=10k nodes, avg d=20,

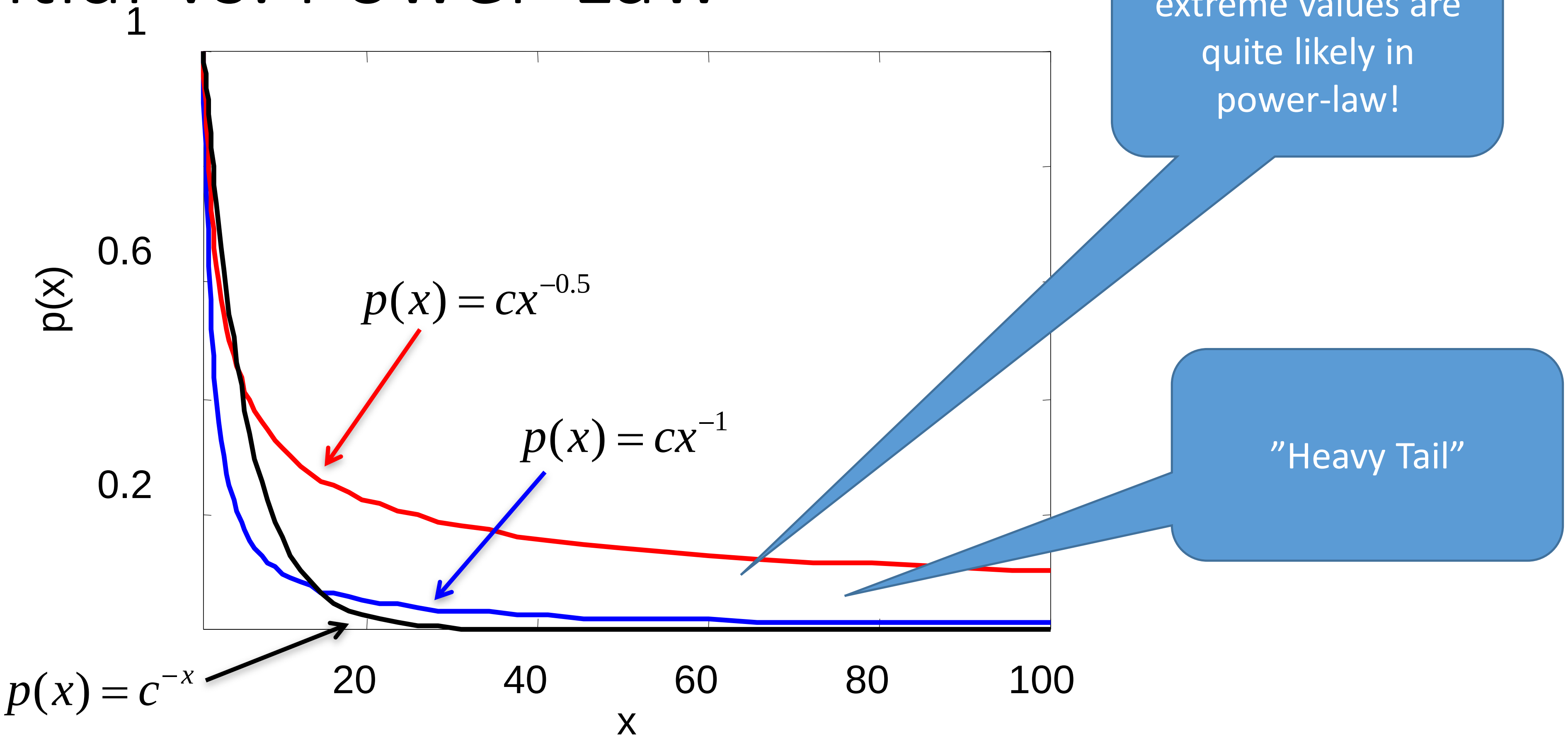


Random Graph



Preferential attachment (power law)

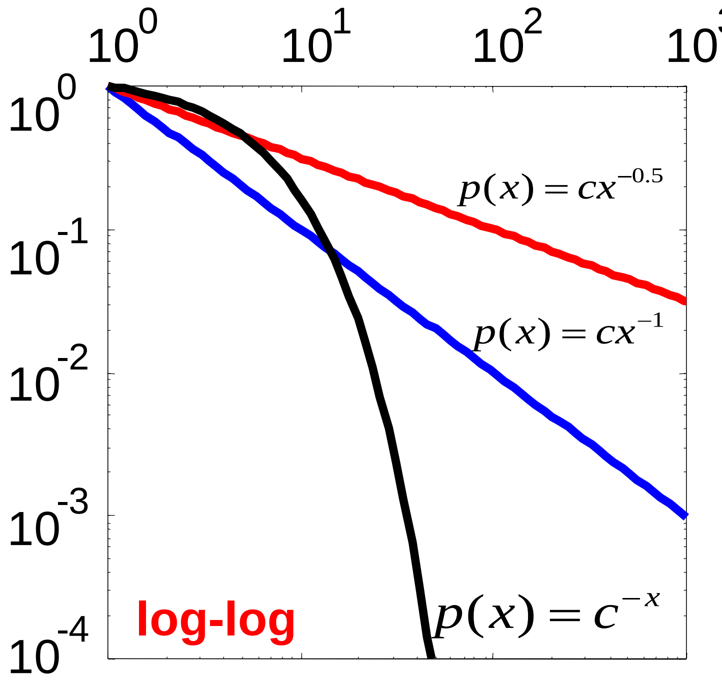
Exponential vs. Power-Law

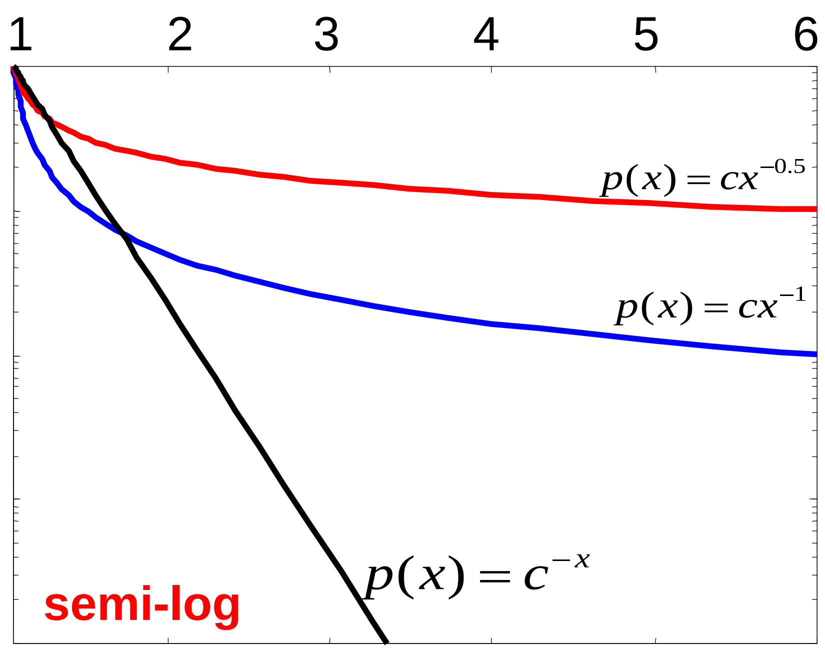


- Above a certain ***x*** value, the power law is always higher than the exponential!

Exponential vs. Power-Law

- **Power-law vs. Exponential on log-log and semi-log (log-lin) scales**

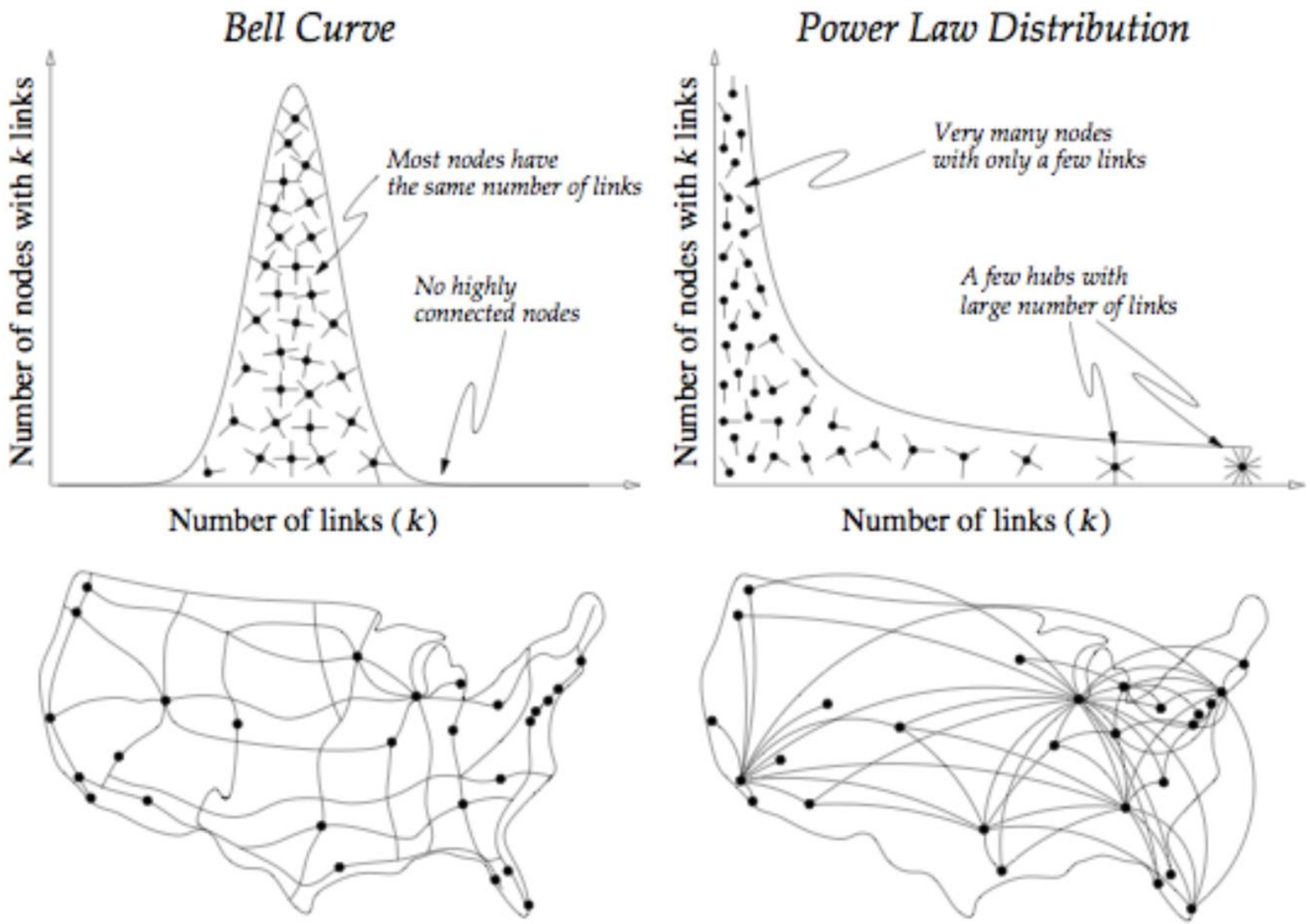




x … logarithmic axis
y … logarithmic axis

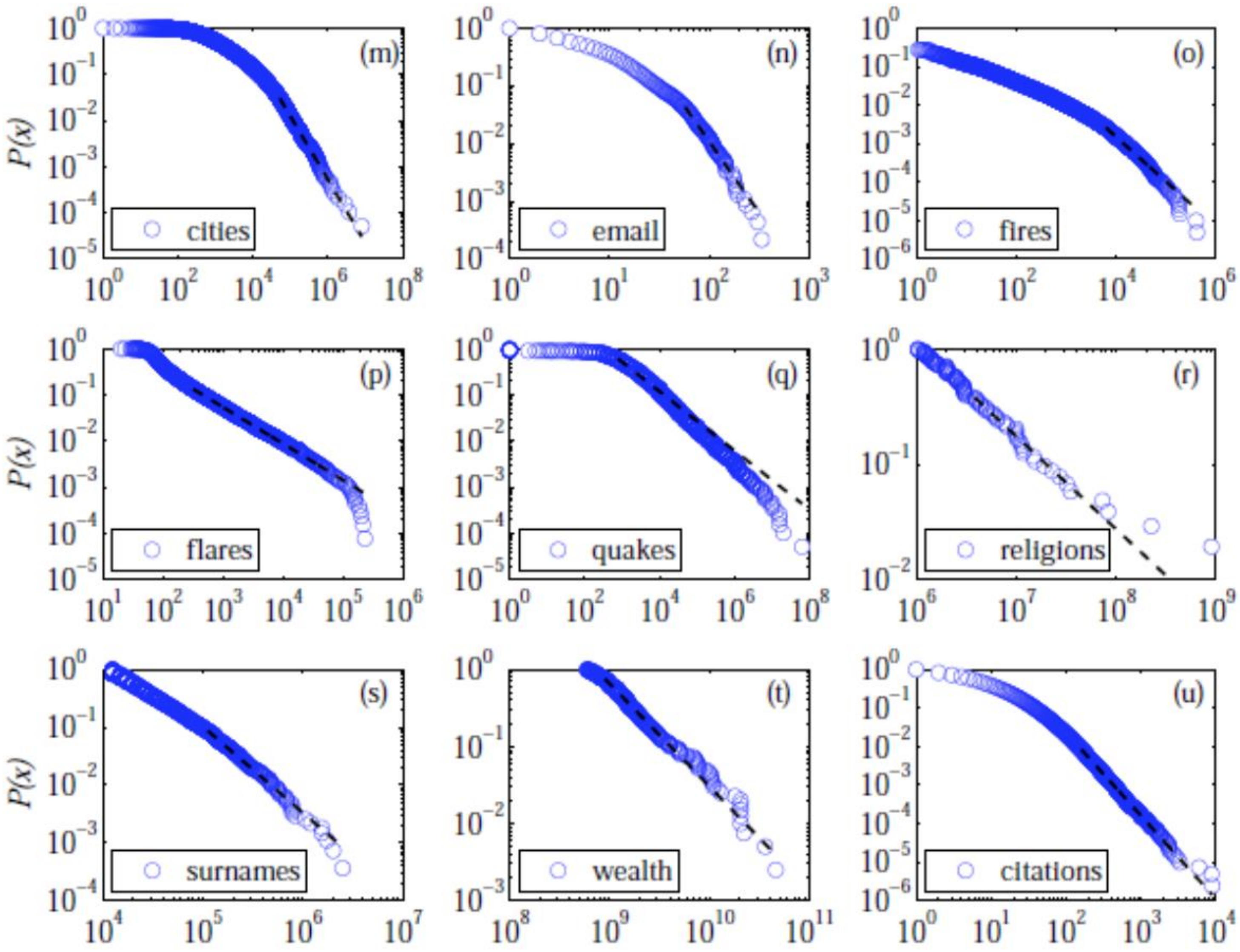
x … linear
y … logarithmic

Binomial Distribution vs Power-Law



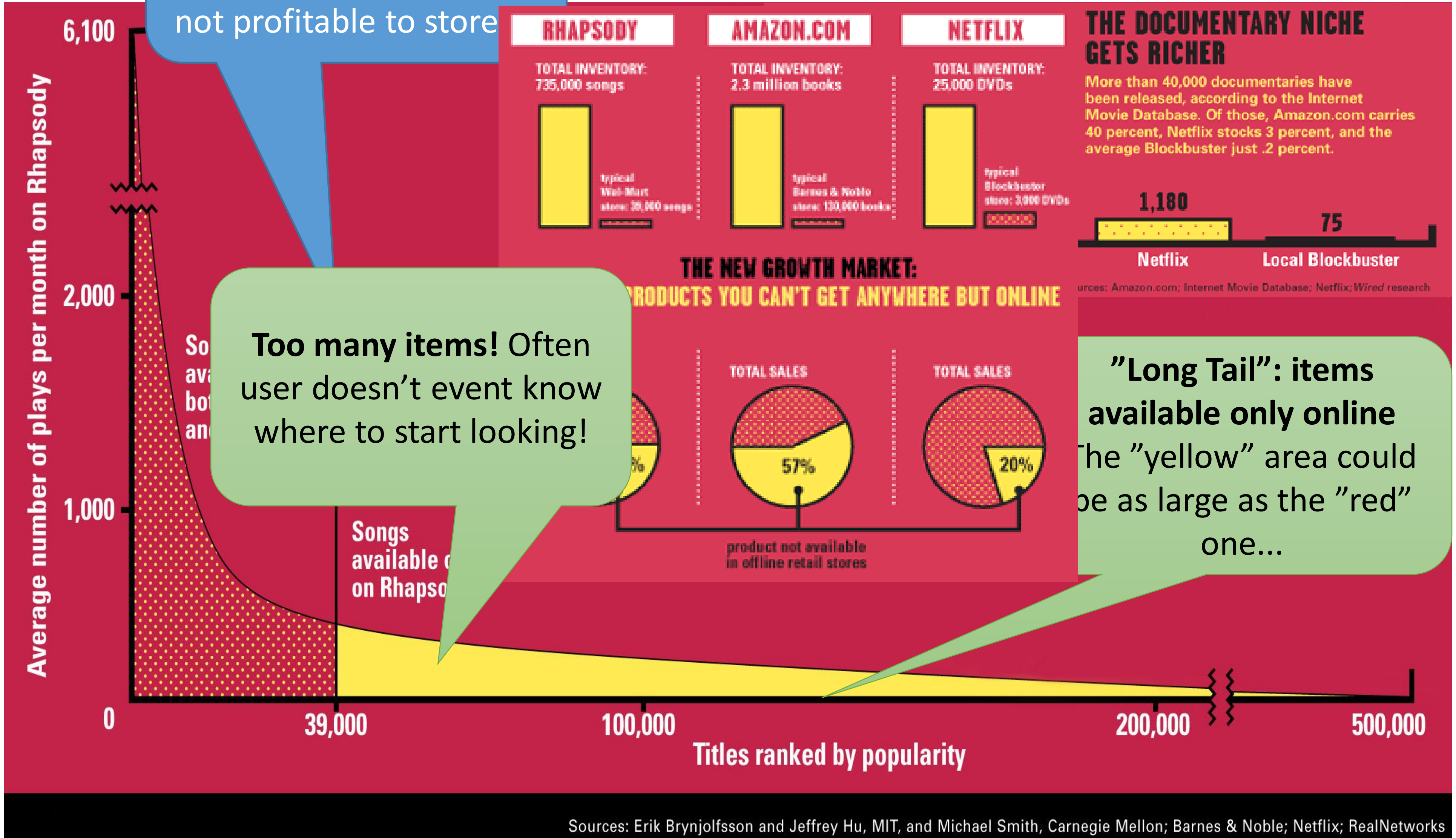
Slide Credit: Jure Leskovec

Natural world is full of power-laws



[Clauset-Shalizi-Newman 2007]

Heavy



”cut-off point”
e.g., Item bought only once a month (becomes not profitable to store

Too many items! Often user doesn’t event know where to start looking!

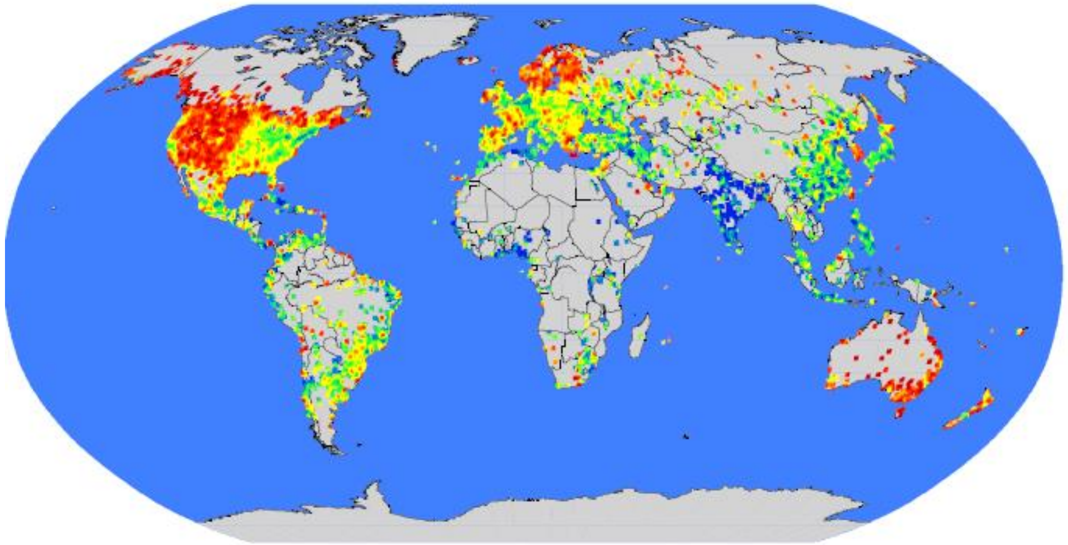
”Long Tail”: items available only online
The ”yellow” area could be as large as the ”red” one...

Source: Chris Anderson (2004)

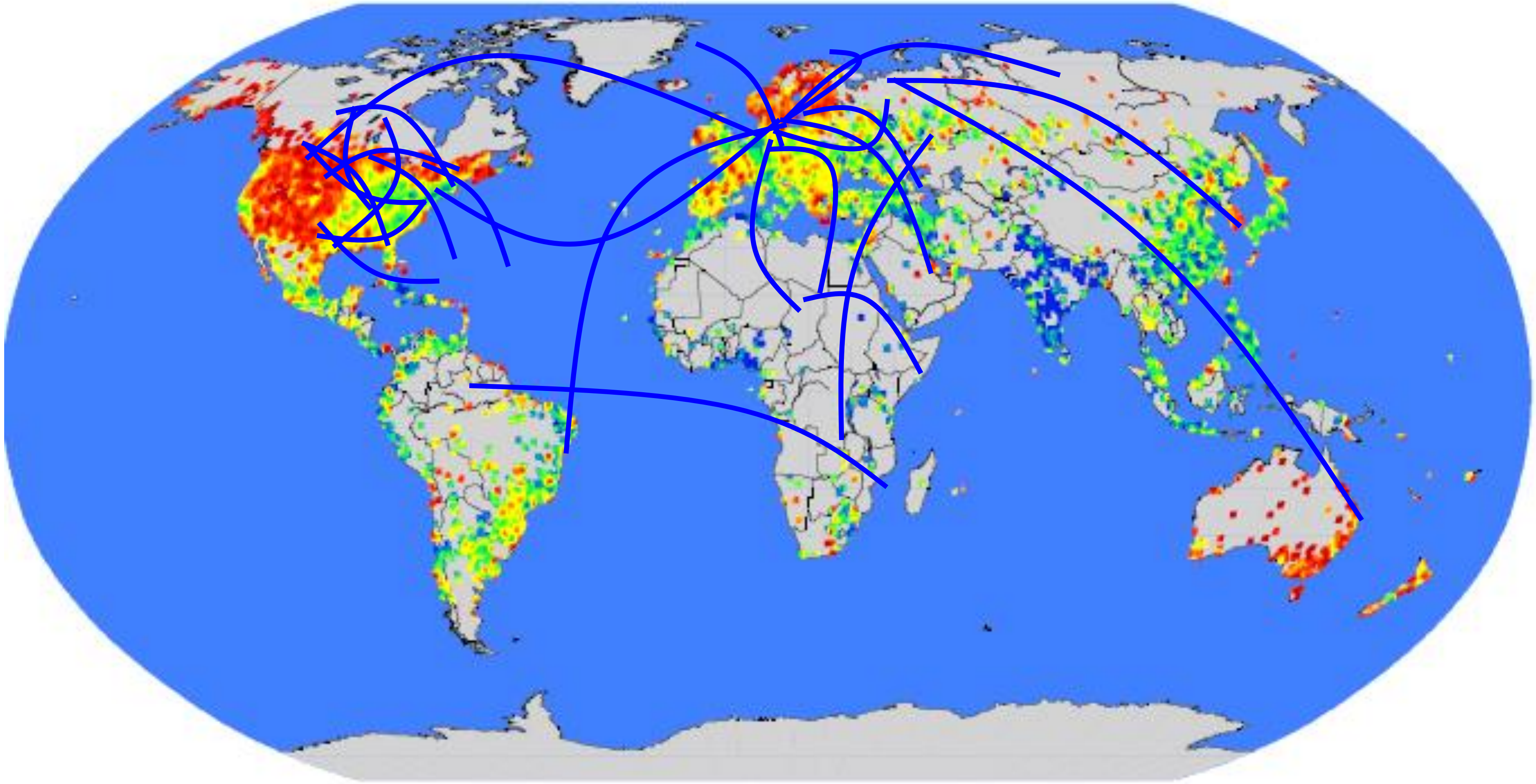
MSN Messenger

J. Leskovec, E. Horvitz. *Worldwide Buzz: Planetary-Scale Views on an Instant-Messaging Network*. Proc. International WWW Conference, 2008.

- **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

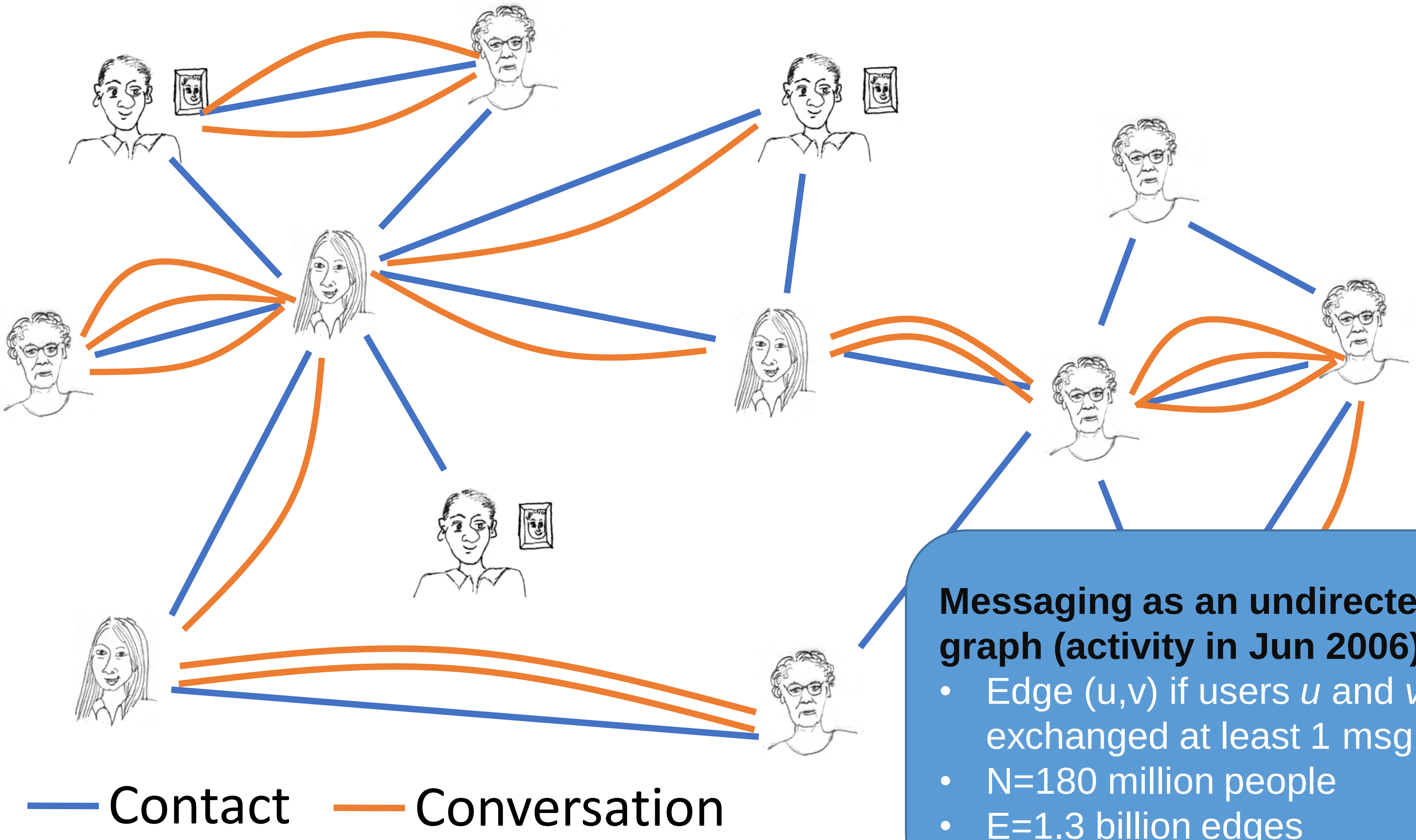


How do we connect?



Network: 180M people, 1.3B edges

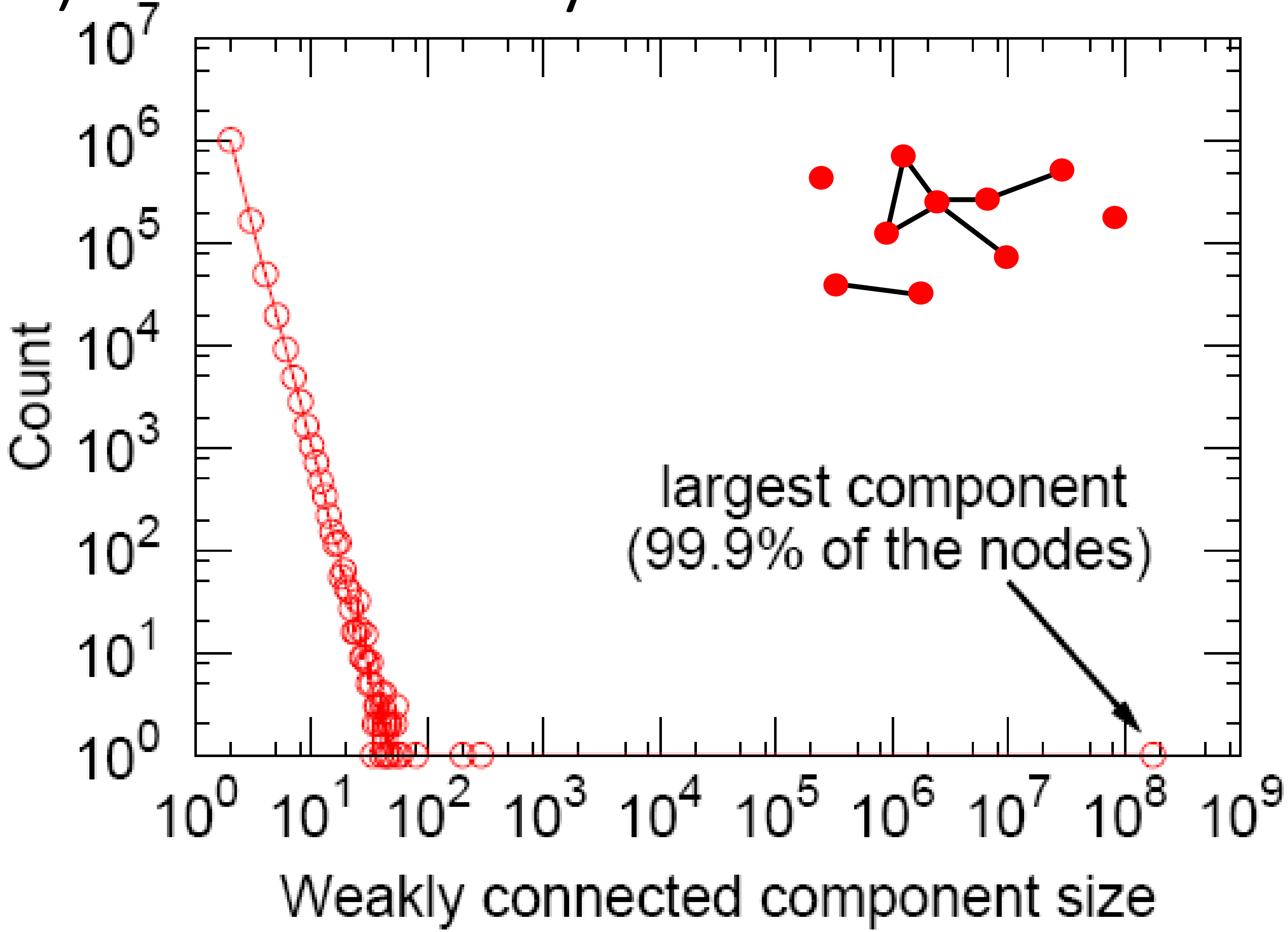
Messaging as a Multigraph



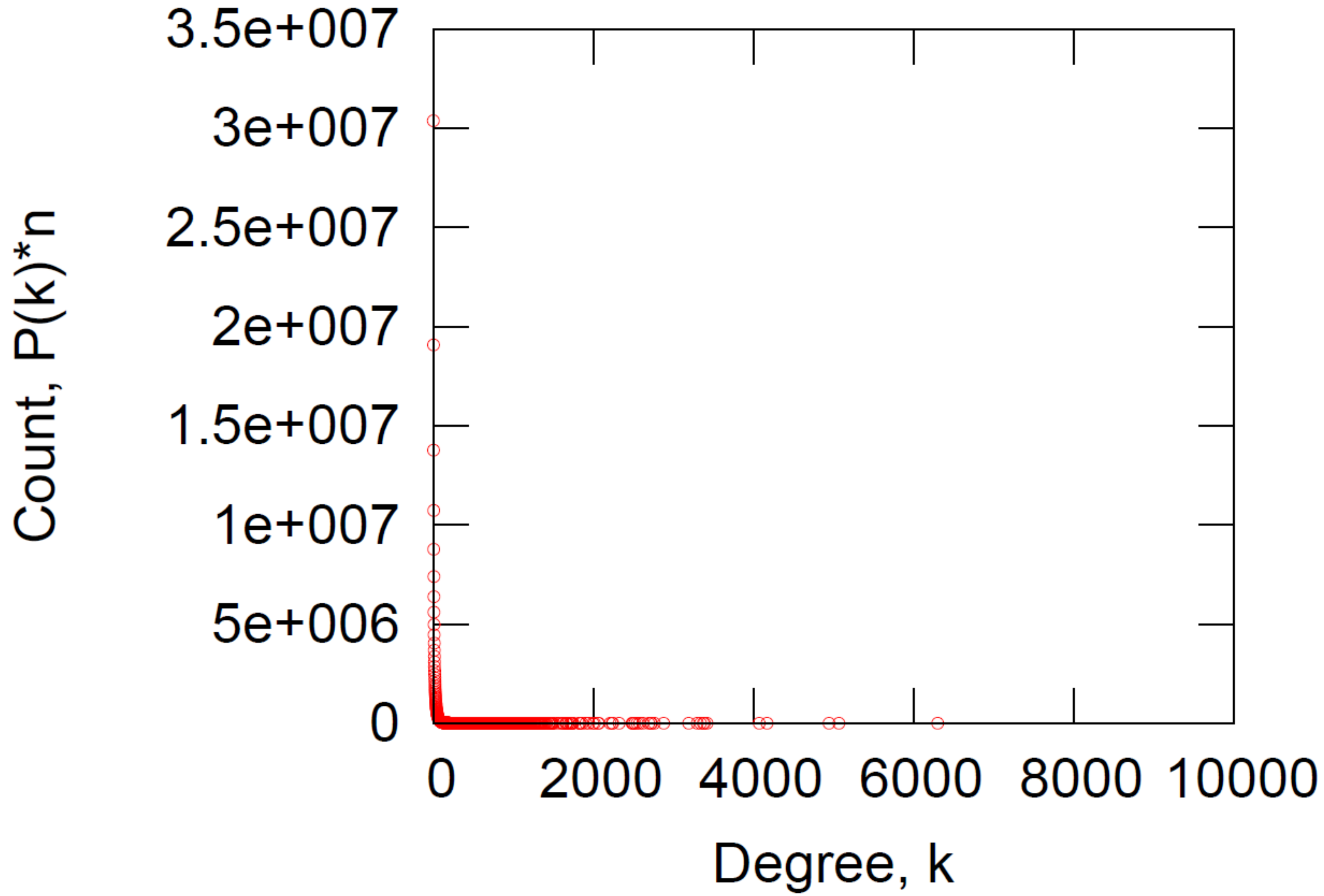
Messaging as an undirected graph (activity in Jun 2006)

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

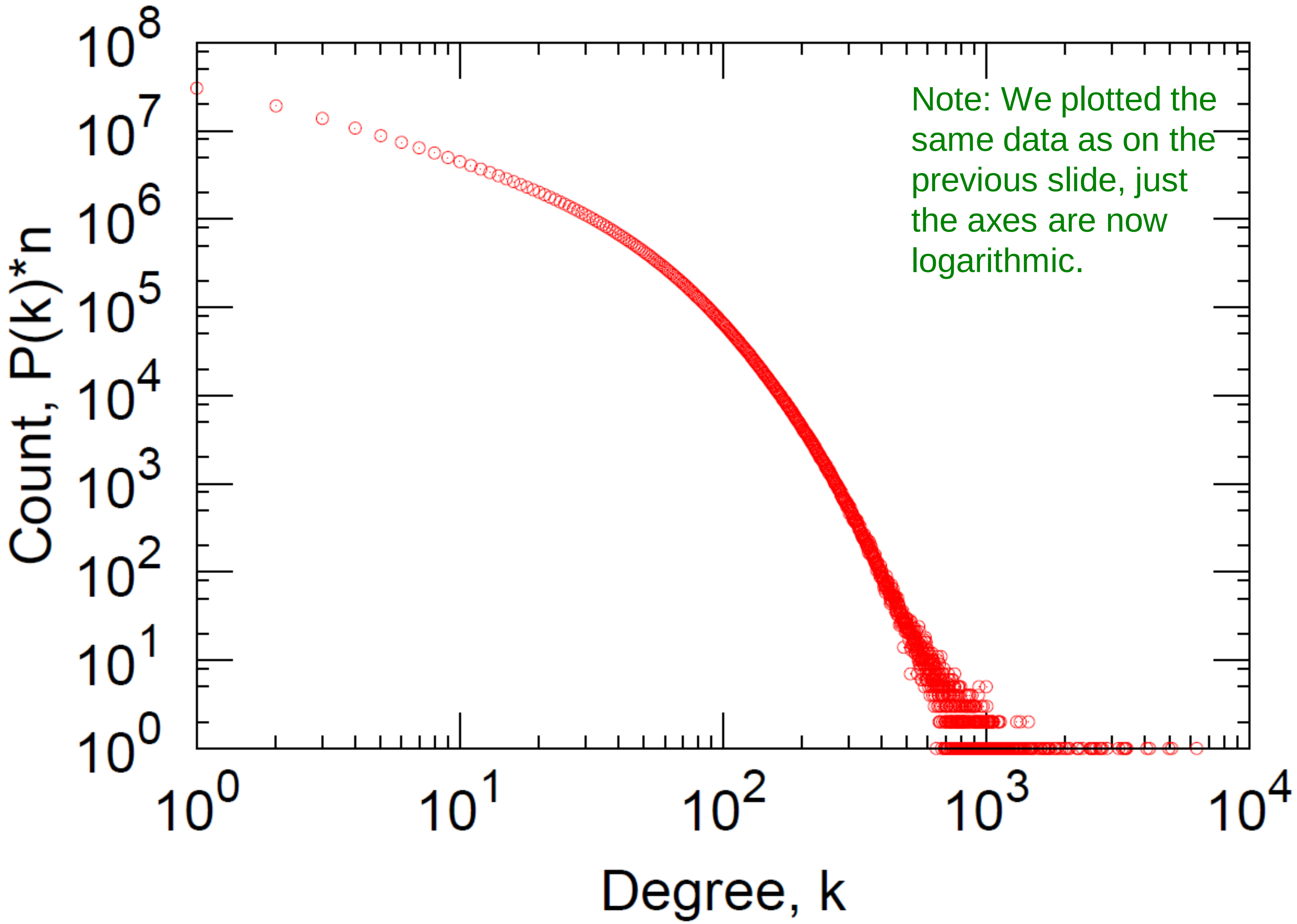
MSN: (1) Connectivity



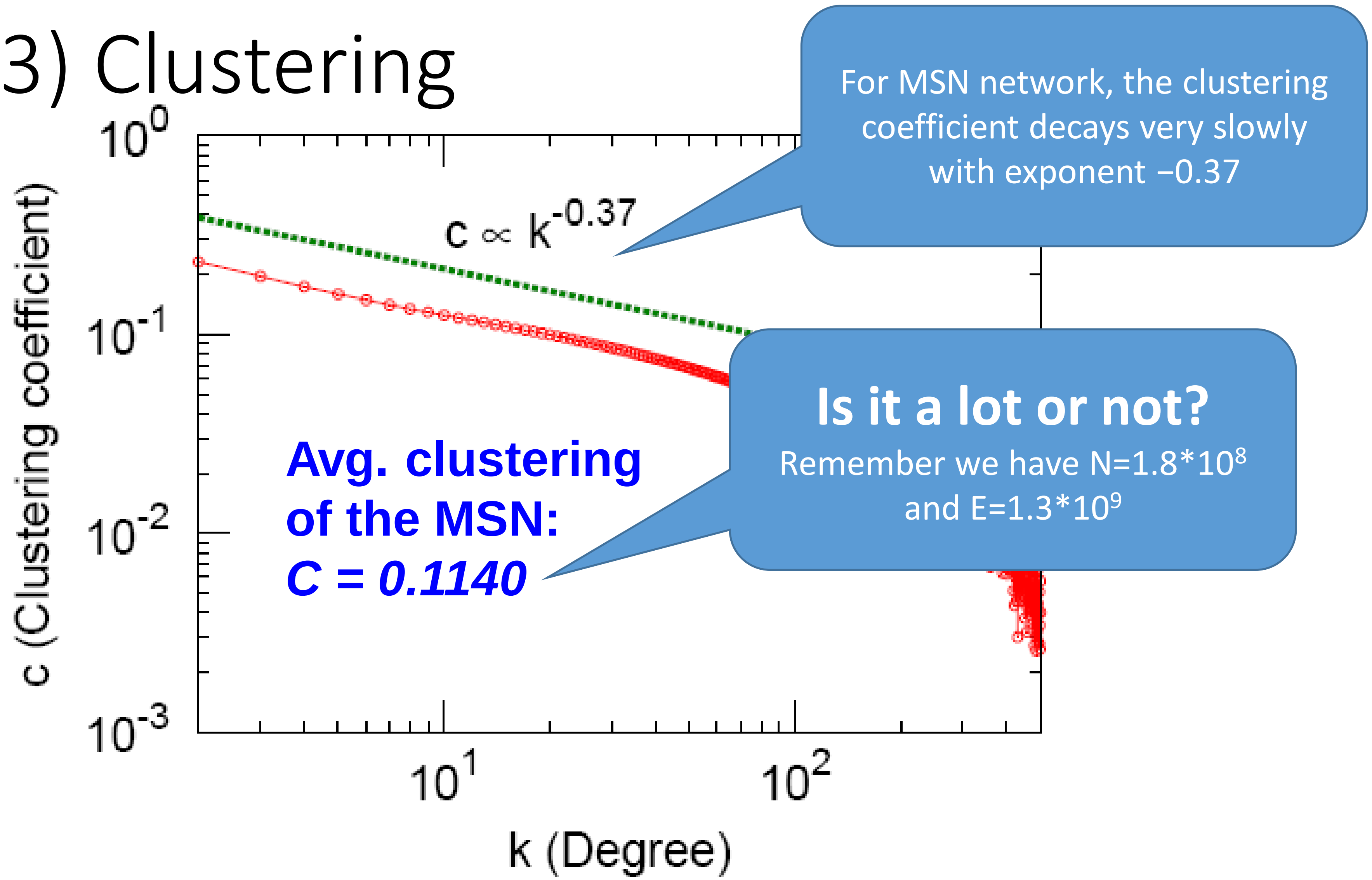
MSN: (2) Degree Distribution



MSN: Log-Log Degree Distribution

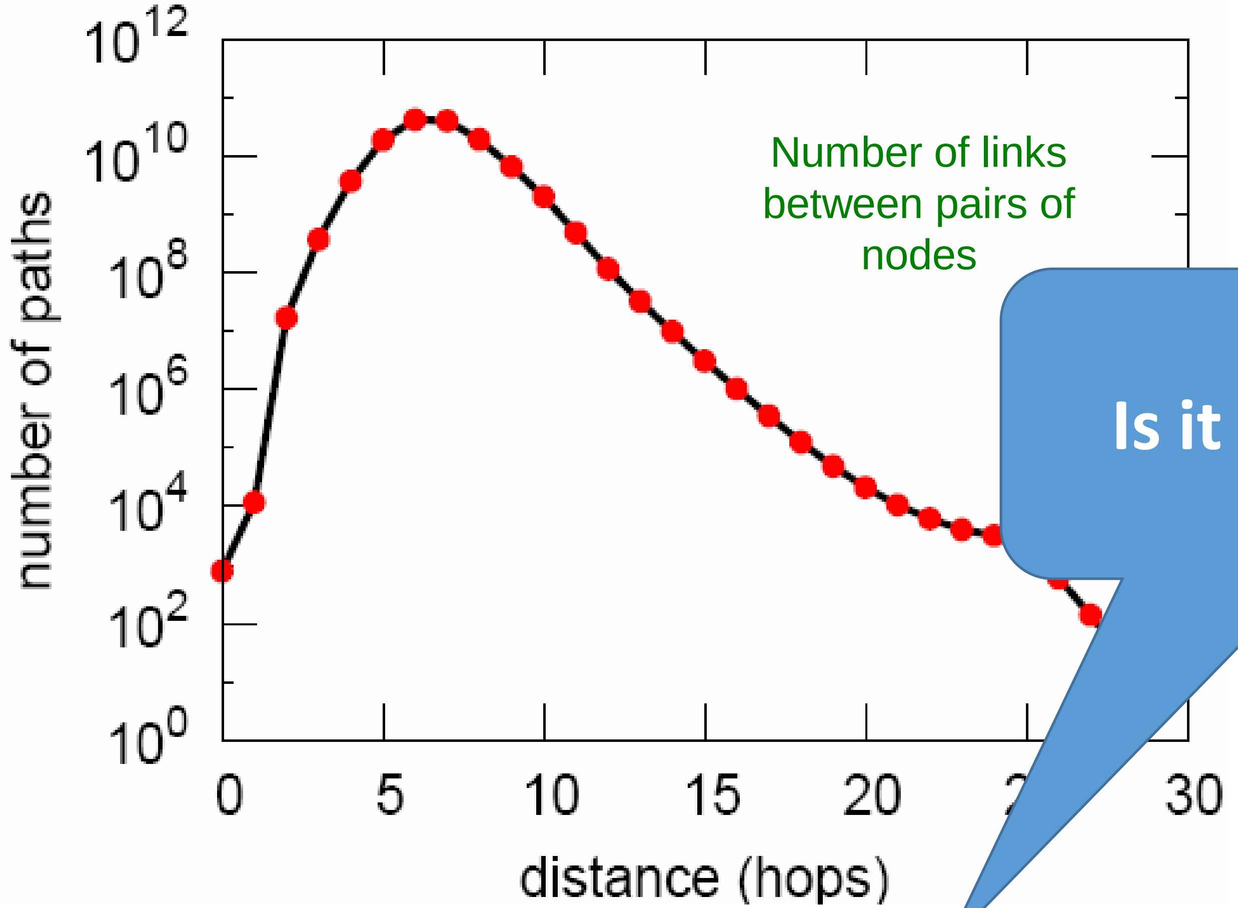


MSN: (3) Clustering



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



Is it a lot or not?

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
	08
	07
	10
	45
	16
	40
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 … ndom node

MSN: Key Network Properties

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

Path length: *6.6*

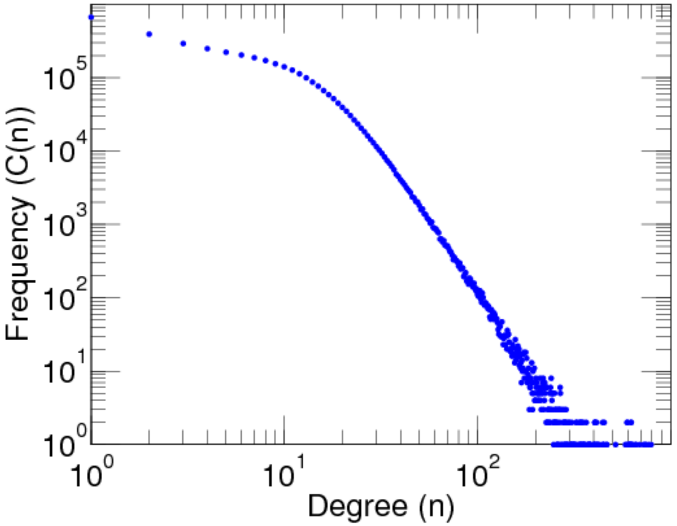
Clustering coefficient: *0.11*

Connectivity: *Giant Component*

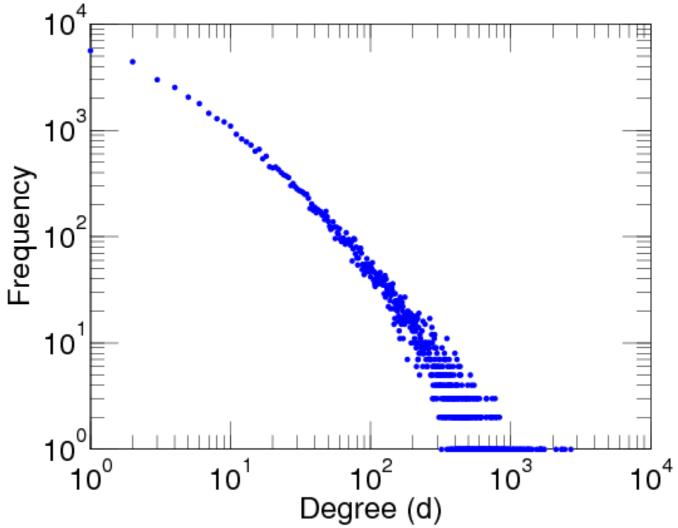
Are these values "expected"?
Are they "surprising"?

Some real world examples (from http://konect.uni-koblenz.de)

- US patents
  - Patent-patent citation
  - N=3774768
  - E=16522438
  - 90-percentile effective diameter 9,79
  - Diameter (longest shortest path) 22

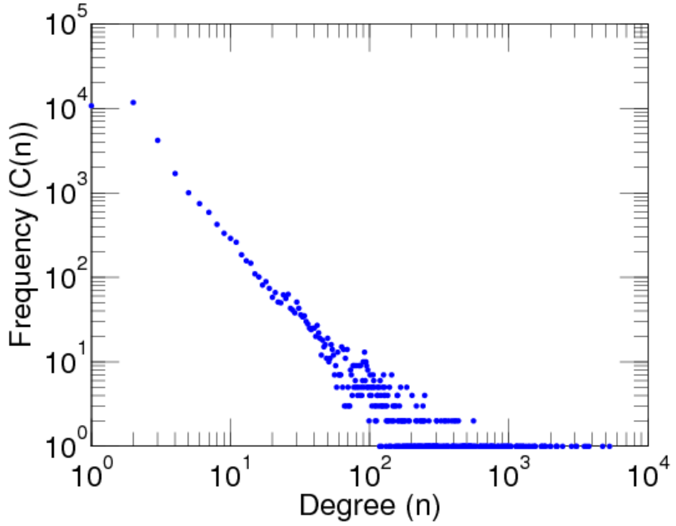


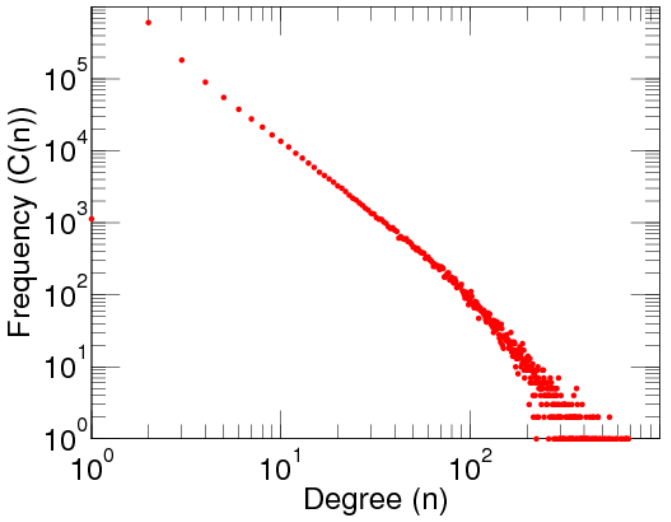
- Facebook (user-user wall posts)
  - Directed
  - N=63891, E=876993
  - CC=19,1% Effective diameter=7,25



Some real world examples (cont.)

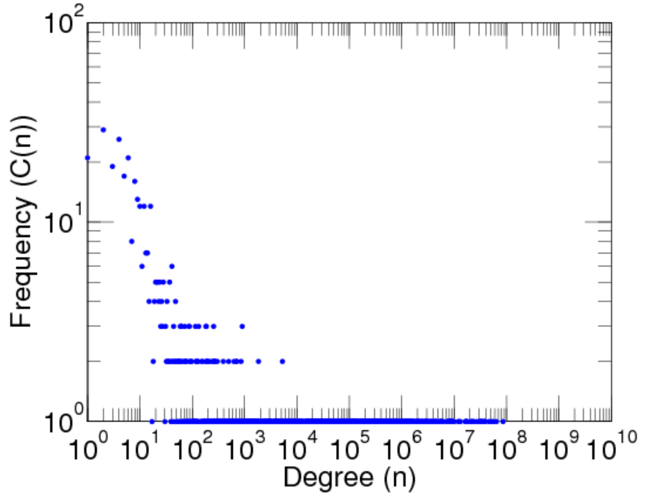
- Internet Topology
  - AS-AS connection
  - N=34761, E=171403, CC=4,851%,
  - Mean shortest path 3,77
- DBLP
  - Author-publication authorship
  - Bipartate
  - N=4337293
  - E=6651968
  - Effective diameter 15,3



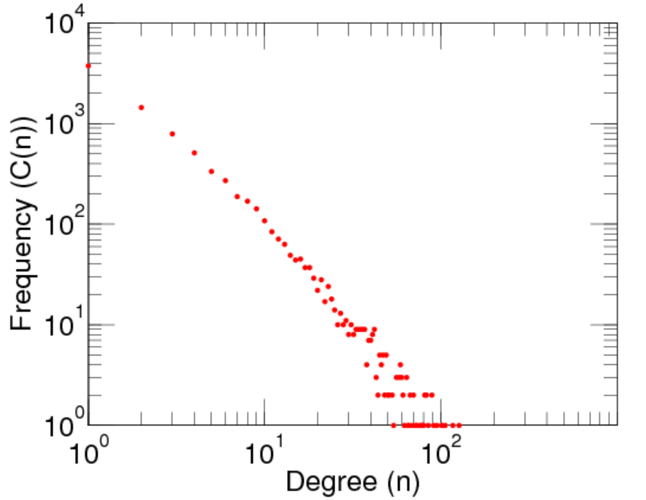


Some real world examples (cont.)

- USA airports (airport-airport flights)
  - N=1574, E=28236, CC=38,4%, dim=8



- Sexual escorts (Buyer–escort contact)
  - bipartate graph
  - N=16730, E=50632, dim=6,05



What do we see?

- ***Sparse*** networks, with ***low diameter***, ***hubs*** and ***non-random clustering***.

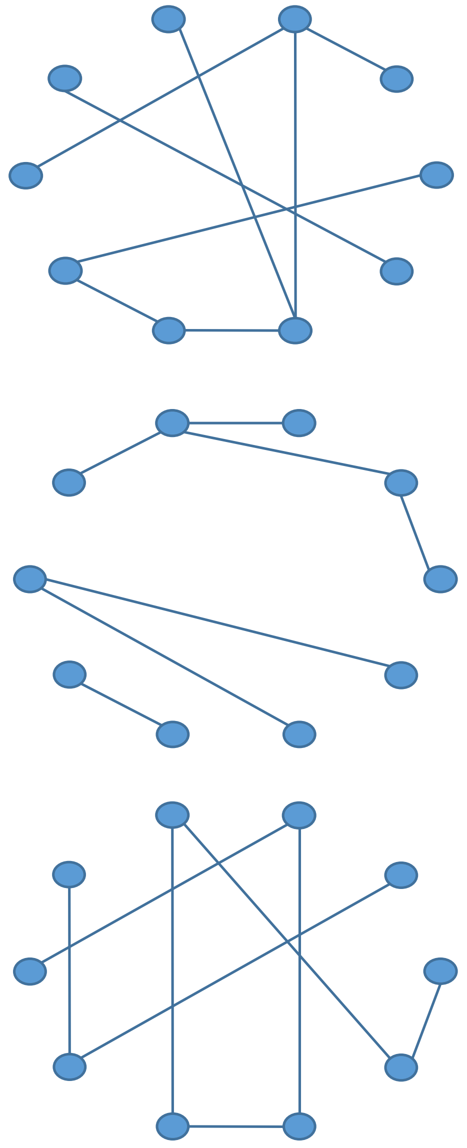
How do we Approach Networks

- **Observations**
  - Structure, properties, patterns, evolution of the networks
  - Our empirical observations often find:
    - Sparse networks
    - Small diameter
    - Large clustering coefficients
    - Power law degree distributions
    - One giant component
  - How to explain them?

- **Models**
  - How do we model edge attachments, epidemics, communities etc?
  - We will start from “simple” to more “sophisticated”.

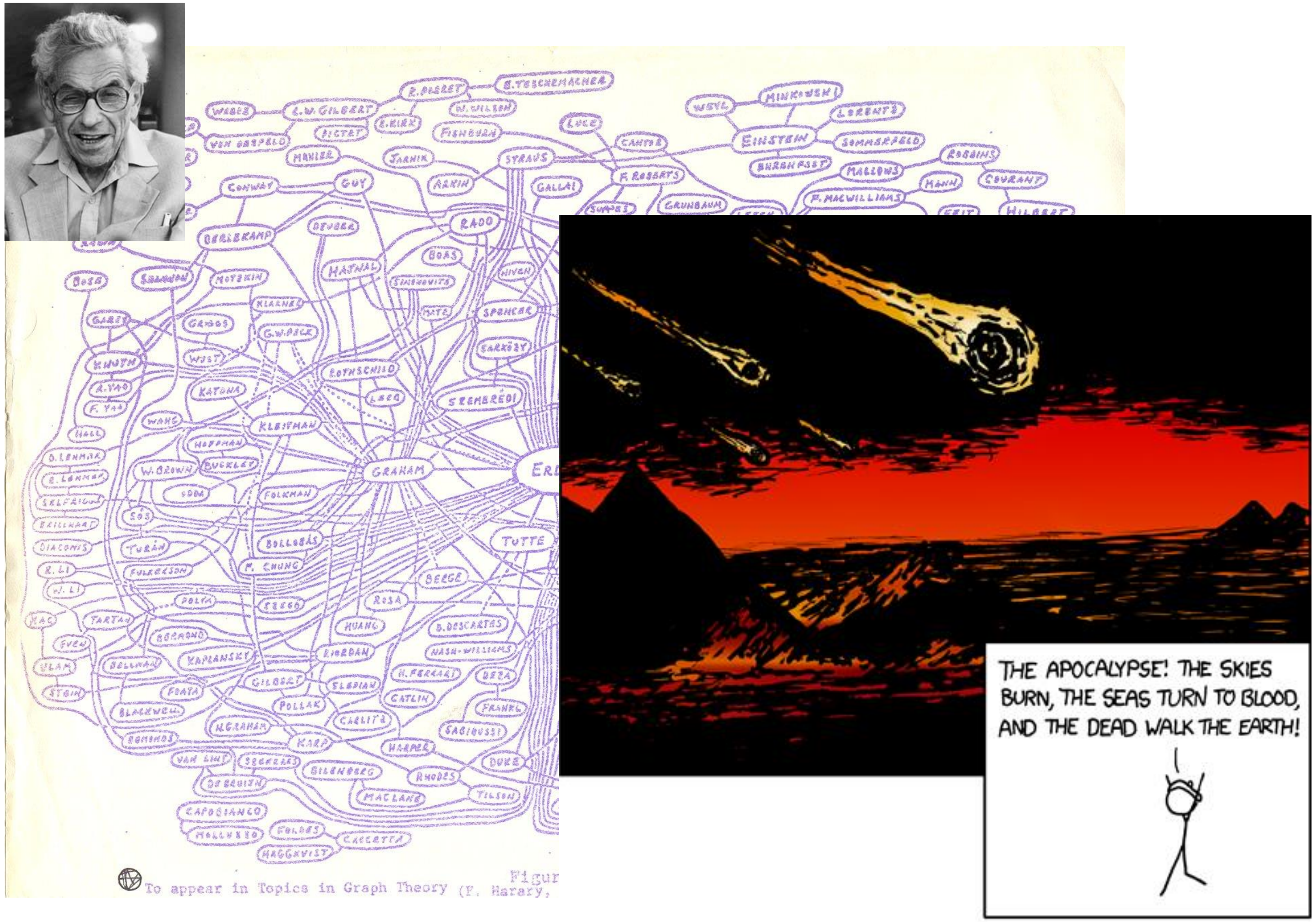
Models of Graphs: Random Graphs

- **G(n,m) model**
  - Start with n isolated vertices
  - Place m edges among them at random.
  - G(n,m) defines a family of graphs (not a particular graph)
- **G(n, p) model (Erdos-Renyi random graph)**
  - Start with n isolated vertices
  - We place and edge between each distinct vertex pair with probability p.
  - n and p do not uniquely determine the graph! It is stochastic!
- **Q1: What's a degree of the above networks?**
- **Q2: Which family is bigger?**

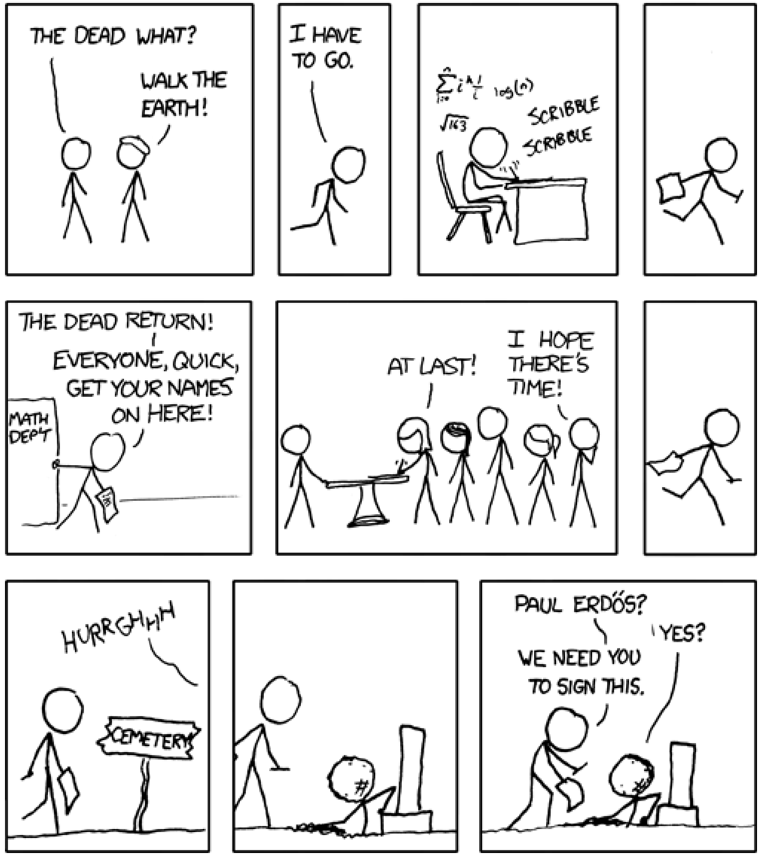


n = 10
p= 1/6

Erdos Number

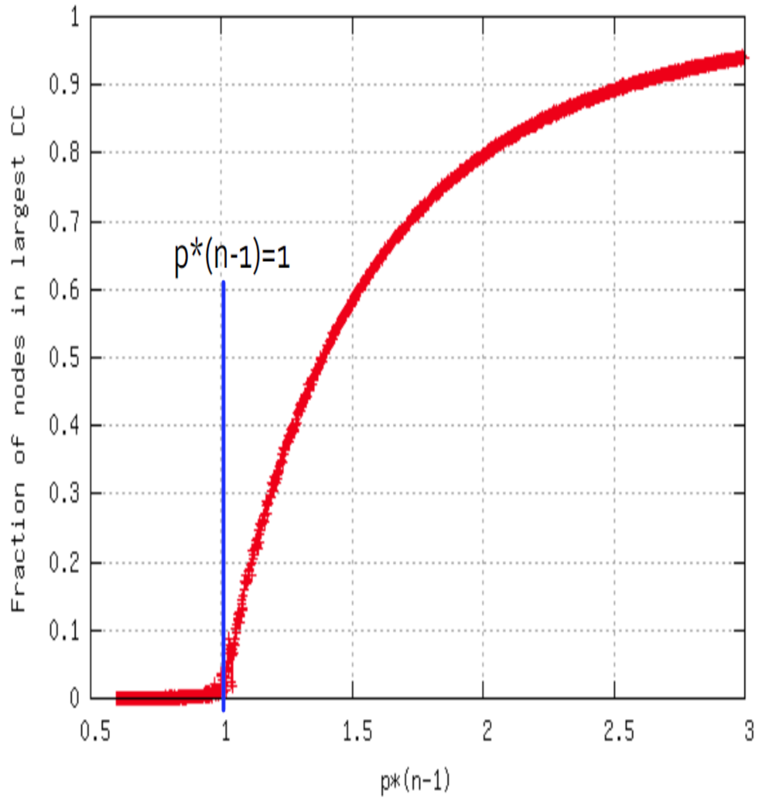


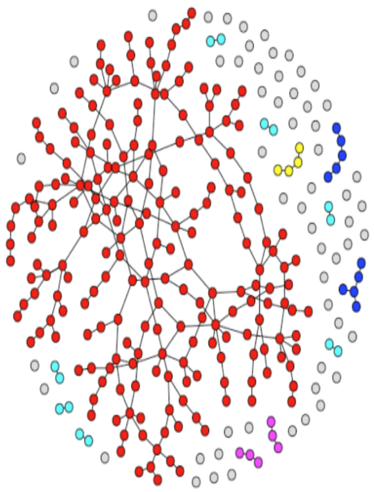
XKCD: What do you do when dead start walking the earth?



Erdos-Renyi random graph

- What can you say about the graph when we move p from 0 to 1?
  - Diameter?
  - How big is a giant component when p=0 and when p=1?
  - How does the giant component grow inbetween those p values?
    - Network undergoes "***phase transition***"





Fraction of nodes in the largest component