Introduction to network analysis

Chen-Hsiang Yeang
Institute of Statistical Science
Academia Sinica

Networks, social networks

- What is a network?
- What is a social network?

What is a network?

- A network is a collection of relations between the entities of interest.
- These relations can include:
  - Physical proximity (examples?).
  - Role participation (examples?).
  - Exchange (examples?).
  - Reference (examples?).

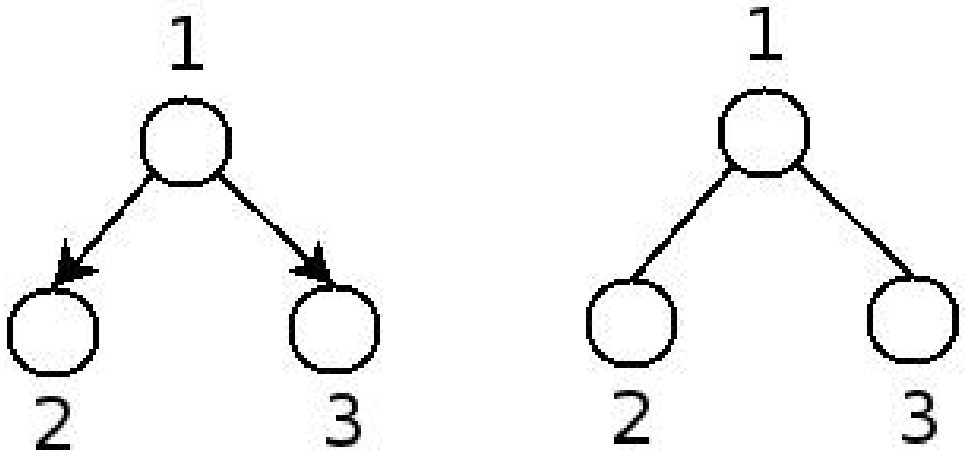
What is a social network?

- A social network is a collection of social relations.
- Some examples include:
  - Friendship.
  - Collaboration.
  - Family bonds.
  - Exchange of money/goods/information.

Mathematical representation of networks

- A network is often represented as a graph.
- Formally, $G = (V, E), V \subseteq N, E \subseteq V \times V$.
- V: nodes or vertices, E: edges, links or ties.
- Nodes can be annotated with various attributes (gene names, gene functions, genders, incomes, etc.).
- Edges can be annotated with various attributes (directions, functions, strength or length, etc.).

Adjacency matrix



- A common and simple representation of a graph is an adjacency matrix.
- An adjacency matrix A is an $|V| \times |V|$ binary matrix.
- $A_{ij} = 1$ if $(i, j) \in E$, 0 otherwise.
- What are the adjacency matrices of the two networks above?

Outline: Four aspects of social networks

- Power-law or scale-free distributions.
- Small-worldness.
- Community structures.
- Network motifs.

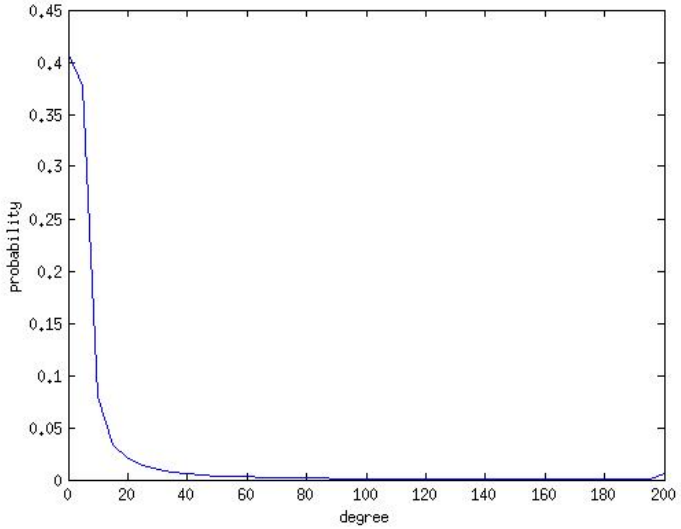
Power-law or scale-free distributions

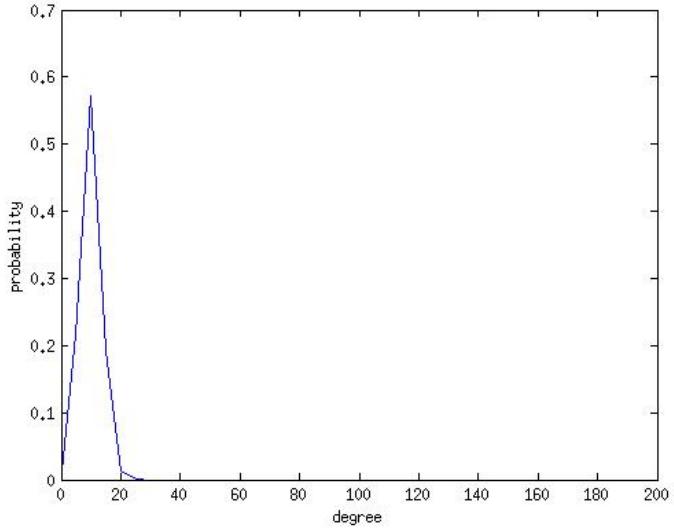
Exercise 1:

- Divide the class into two groups.
- Group 1 is given the edge list of the Enron-Email dataset.
- Group 2 is given the edge list of a randomly generated graph with the same number of nodes (36692) and similar number of edges (183799).
- The degree of a node is the number of its first neighbors in the graph.
- Calculate the degree of each node in your network.
- Plot the histogram of node degrees in your network.

Power-law or scale-free distributions

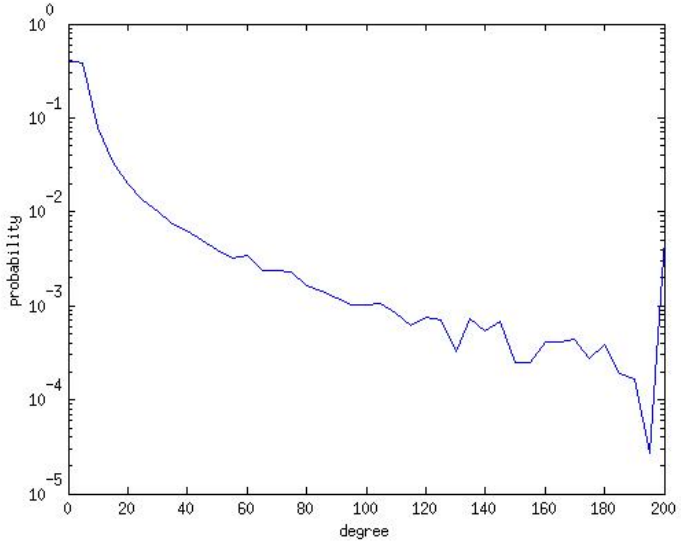
What's the difference between the two distributions?

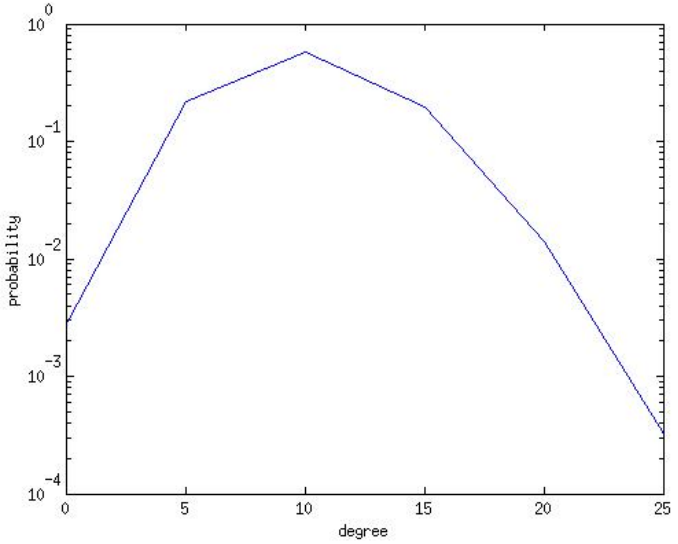




Power-law or scale-free distributions

What's the difference between the their log distributions?





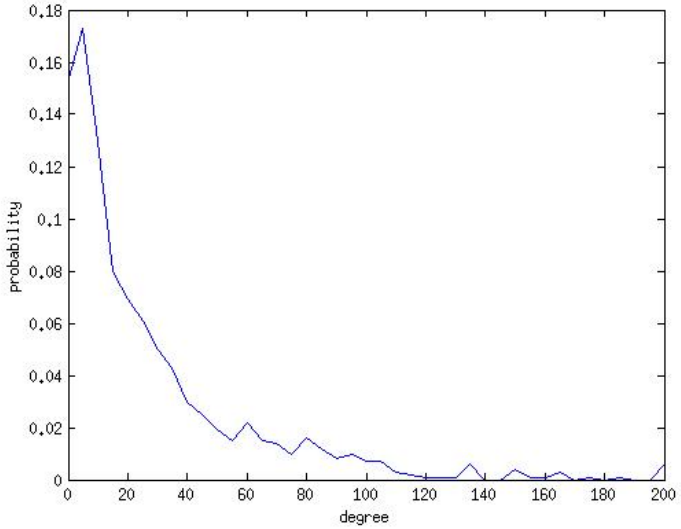
Power-law or scale-free distributions

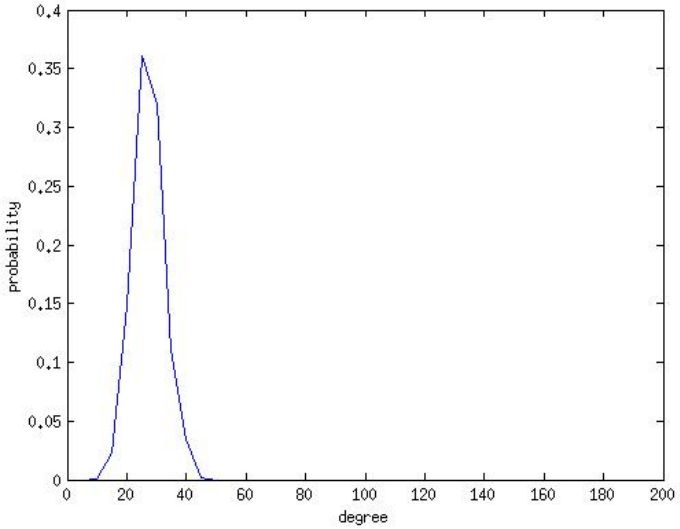
Exercise 2:

- A graph of 1000 nodes subsampled from the Enron data was generated.
- A random graph of the same size and density was also generated.
- Calculate the degree of each node in your network.
- Plot the histogram of node degrees in your network.

Power-law or scale-free distributions

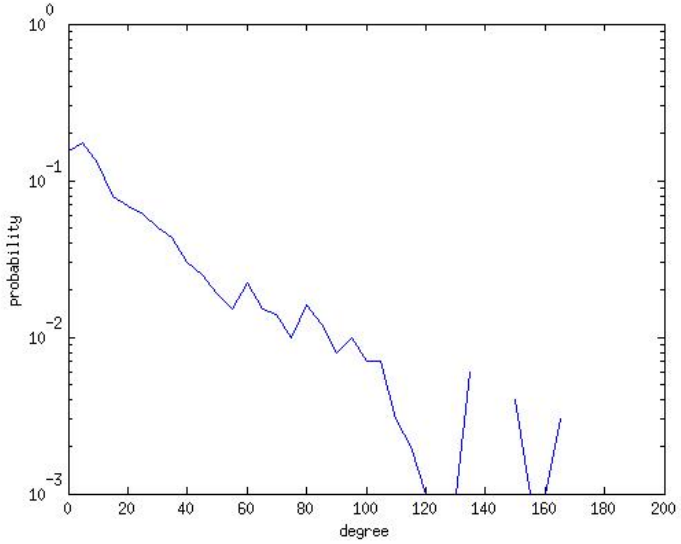
What's the difference between the two distributions?

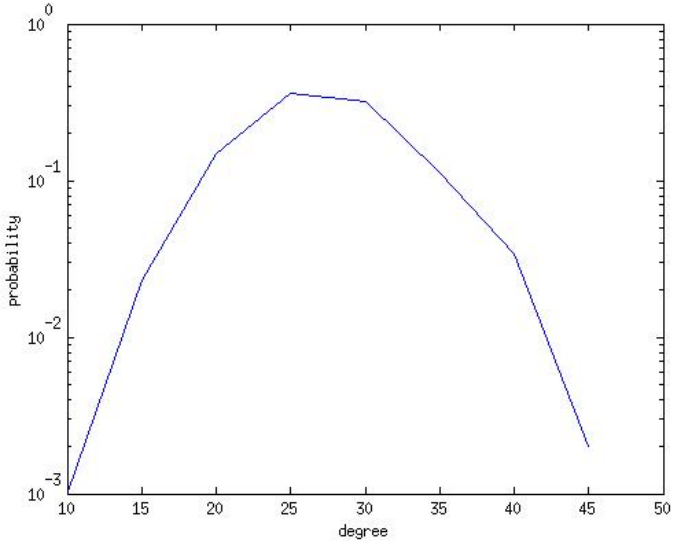




Power-law or scale-free distributions

What's the difference between the their log distributions?





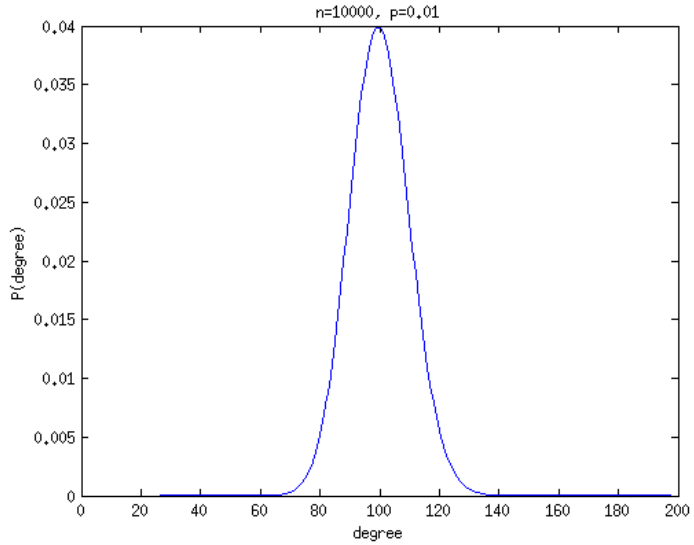
Erdős-Rényi random graphs

Procedures for constructing a random graph $G(n, p)$:

- Fix the number of nodes to n.
- For each pair of nodes, construct an edge connecting them with probability p.

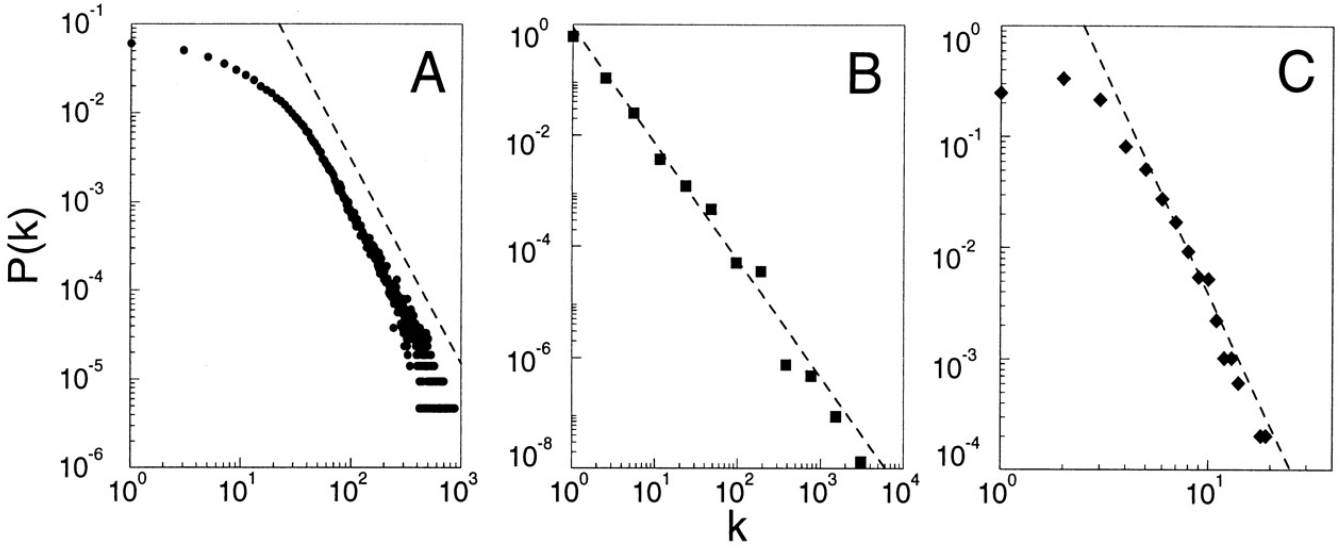
Erdős P. and Rényi A., 1960.

Characteristics of Erdős-Rényi random graphs



- When n is large, the node degree follows a Poisson distribution $P(deg(v) = k) \approx \frac{(np)^k e^{-np}}{k!}$.
- The majority of nodes are adjacent to np neighbors.
- If $p < \frac{(1-\epsilon)logn}{n}$, then $G(n,p)$ will almost surely be disconnected.
- If $p > \frac{(1-\epsilon)logn}{n}$, then $G(n,p)$ will almost surely be connected.
- Thus $\frac{logn}{n}$ is a hard threshold of transitioning from disconnected to connected graphs.

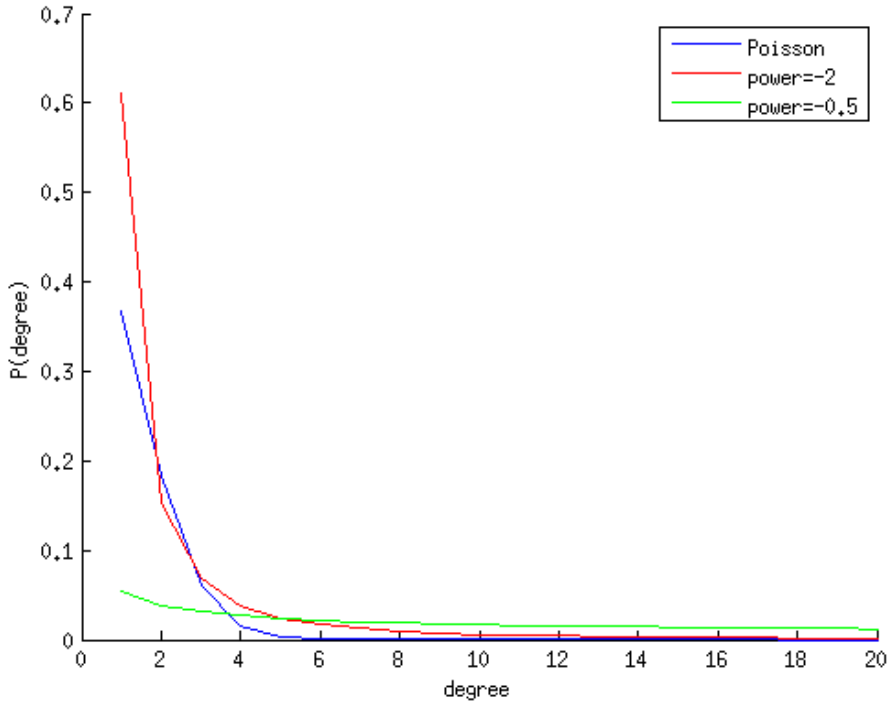
Scale-free networks



- Node degrees follow a power law distribution $P(deg(v) = k) \propto k^{\gamma}$.
- Many real-world networks are scale-free (e.g., social networks, Internet, web documents, protein-protein interaction networks).

Barabási A.L. and Albert R., Science 1999.

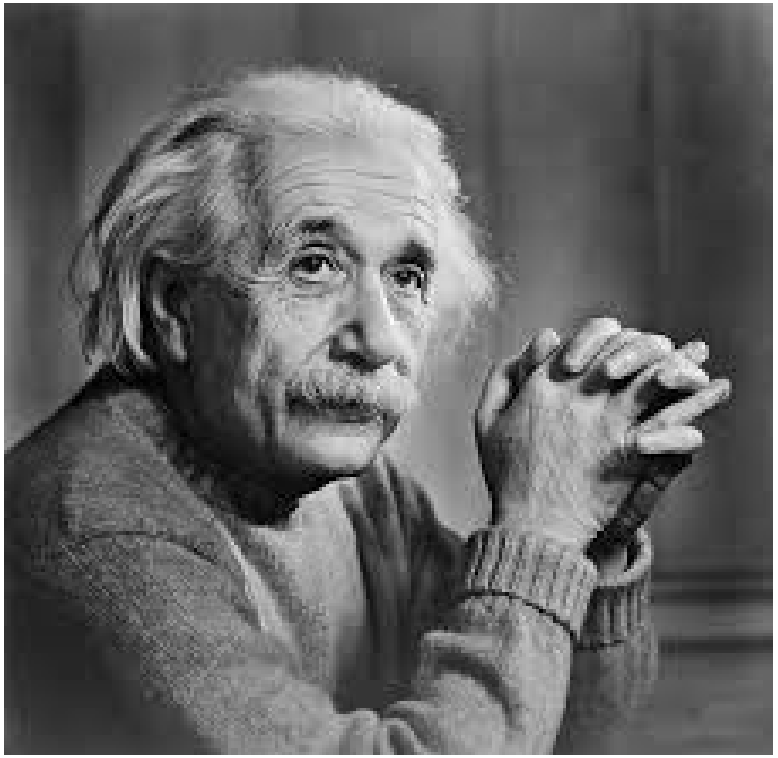
Characteristics of scale-free networks



- Scale-free networks possess heavy tails compared to random graphs.
- The majority of nodes are adjacent to a few neighbors.
- A small number of *hubs* are highly connected.
- The networks are *scale-free* as the shape of the degree distribution is invariant with network scales.

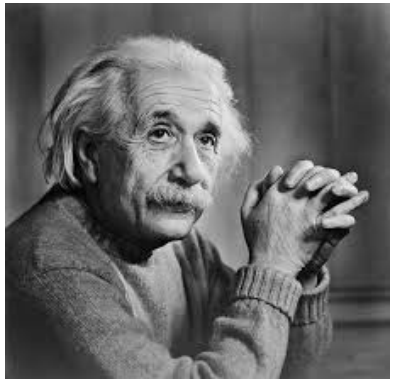
Small-world networks

Six degree of separation: every two persons in the world are connected by paths of less than 6 acquaintance relations.



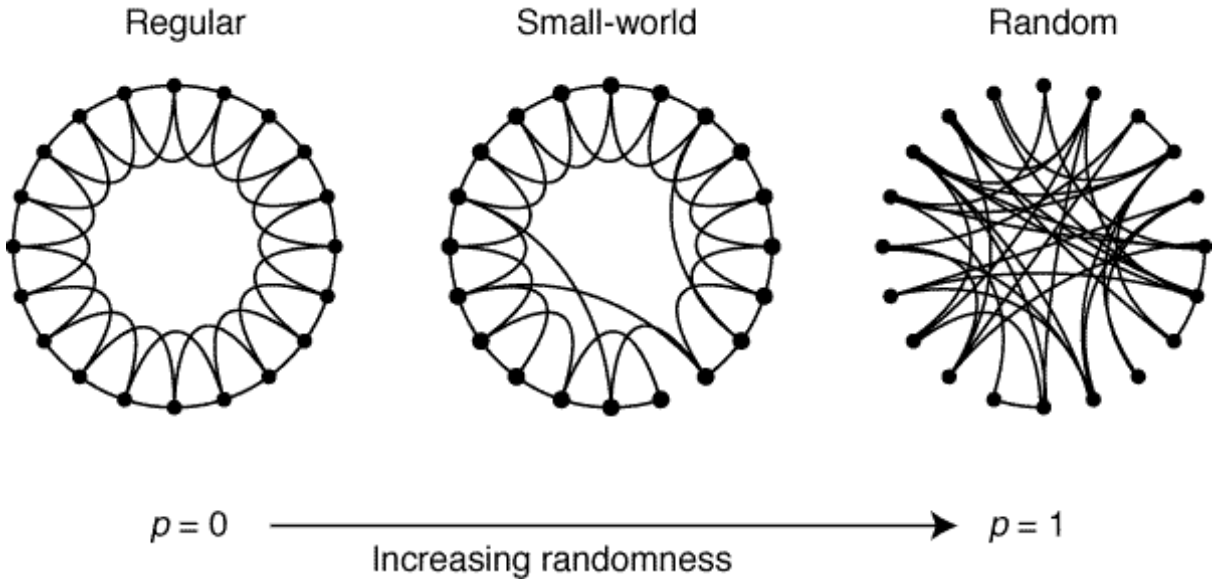
How many degrees separate Einstein from me?

Small-world networks




Albert Einstein $\rightarrow$ Freeman Dyson $\rightarrow$ Arnold Levine $\rightarrow$ Chen-Hsiang Yeang.

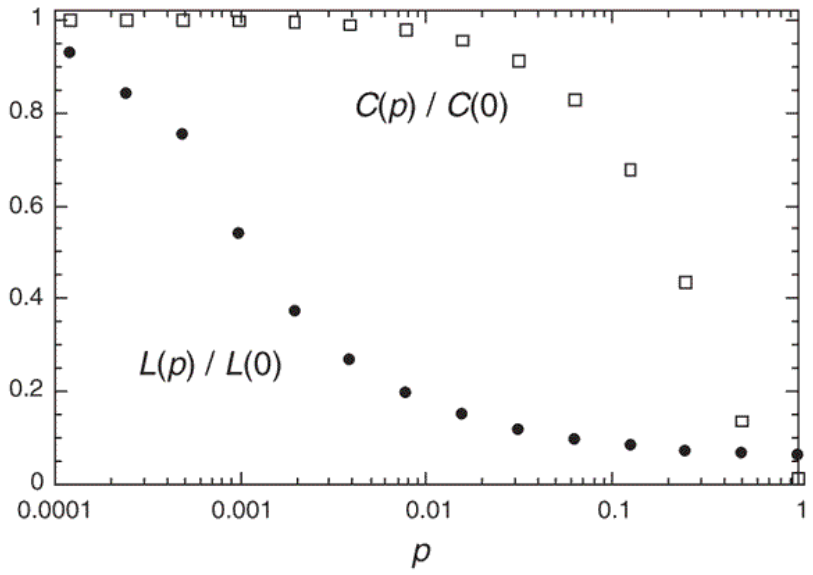
Small-world networks



- Many networks (such as social networks) are highly clustered but also have short characteristic path lengths.
- Six degrees of separation.
- Procedures for constructing small-world networks:
  - Start with a regular graph $R(n, k)$ with n nodes, each nodes are adjacent to k neighbors.
  - Rewire each edge to randomly selected nodes with probability p.

Watts D.J. and Strogatz S.H., Nature 1998.

Characteristics of small-world networks



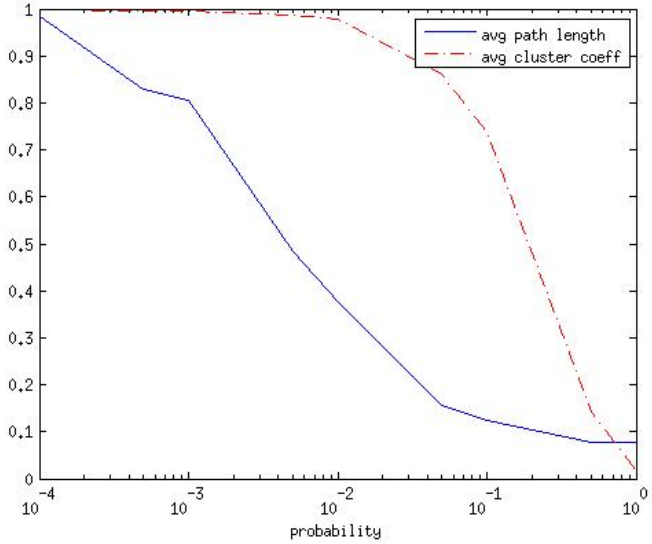
- Vary p from 0 (regular graph) to 1 (random graph).
- Normalized characteristic path length drops quickly with increasing p.
- Normalized clustering coefficient is robust against p.
- Thus graphs within mid range p values satisfy small-world properties.

Small-world networks

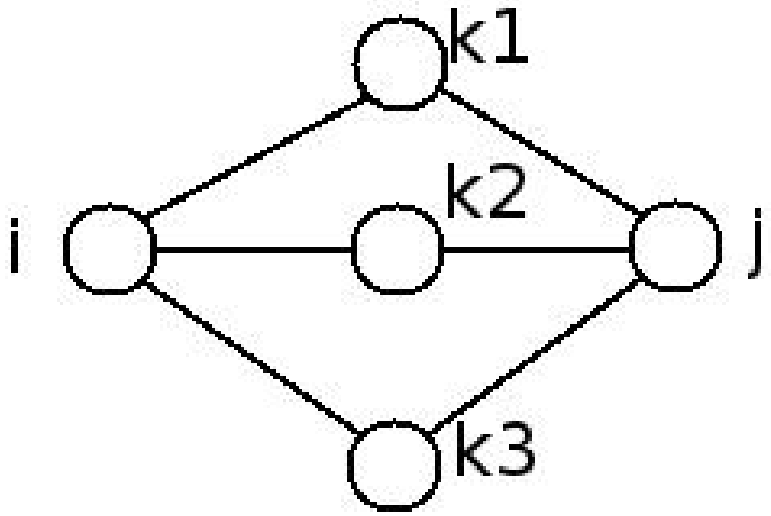
Exercise 3:

- Each group is given a perturbed network.
- smallworldnet1.csv: $p = 0.001$. smallworldnet2.csv: $p = 0.1$.
- Calculate the clustering coefficient of each node in the network.
- Clustering coefficient of node i: #(edges in the subgraph spanned by first neighbors of i)/#(all node pairs in the same subgraph).
- Calculate the length of the shortest path connecting each pair of nodes in the network.

Small-world networks

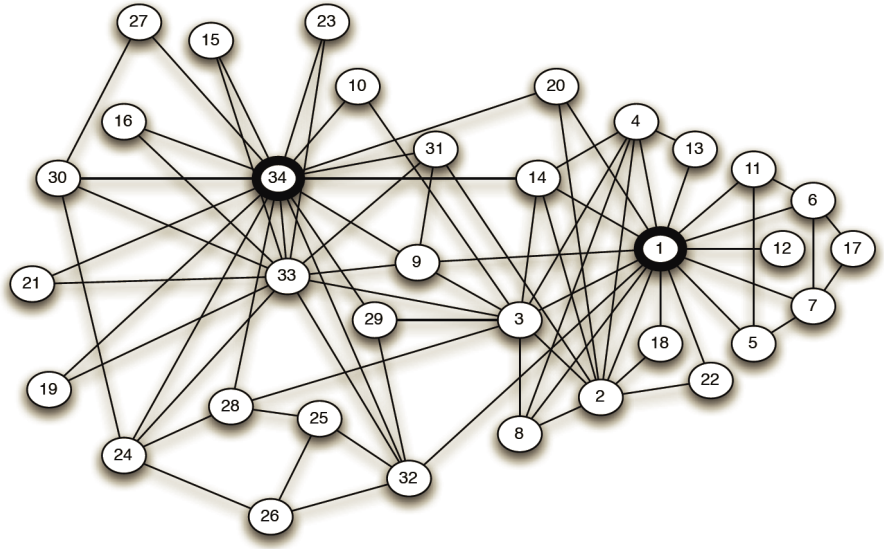


How to evaluate the shortest path lengths?



- $A_{ij}^2 = \sum_k A_{ik} A_{kj}$.
- In other words, A_{ij}^2 is the number of 2-link paths connecting i and j.
- In extension, A_{ij}^l counts the number of l-link paths connecting i and j.
- If the shortest path connecting i and j is l, then $A_{ij}^{l-1} = 0$ and $A_{ij}^{l-1} > 0$.
- Can calculate the path length of each node pair accordingly.

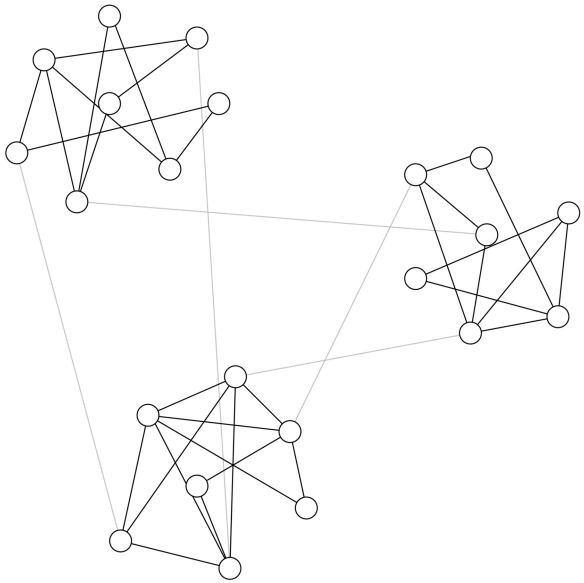
Community structures



Exercise 4: Consider the social network of a karate club above.

- Propose an algorithm to divide the network into communities.
- What are the communities in the karate club network according to this method?

Existence and detection of communities



- Many real-world networks consist of communities.
- There are dense *intra-community* connections and sparse *inter-community* connections.
- Define *betweenness* of an edge as the number of shortest paths traversing the edge.
- Edges of high betweenness are bridges between communities.
- Iteratively remove edges of high betweenness and recalculate the betweenness of the remaining edges.

Girvan M. and Newmann M.E.J., PNAS 2002.

Network motifs

- Motifs are recurrent patterns in the data.
- Network motifs are recurrent topological structures in large-scale networks.
- Two motif instances are illustrated above.

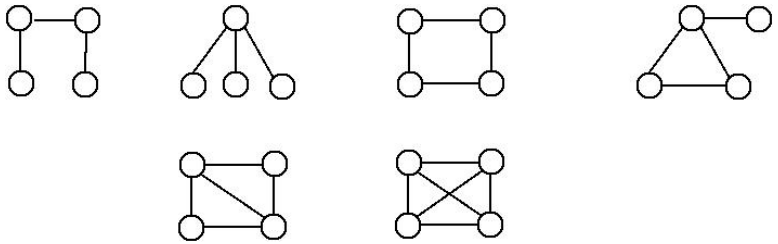
Network motifs

Exercise 5:

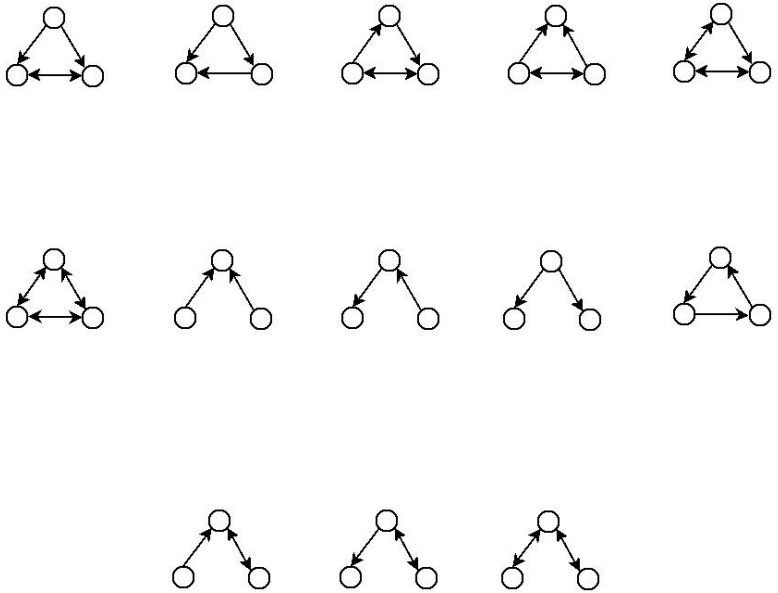
- Group 1: List all the 4-node undirected motifs.
- Group 2: List all the 3-node directed motifs.

Network motifs

Undirected 4-motifs:



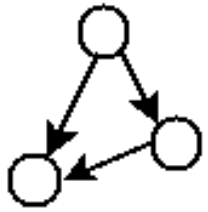
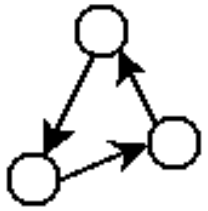
Directed 3-motifs:



Enrichment of network motifs

- Over-represented motifs are detected by comparing the frequencies of small structures in the real networks versus properly constructed randomized networks.

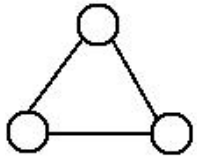
- Instance: feed-forward loops versus 3-node cycles.

	feed-forward loop	3-node cycle
		
random graphs	1.7 +/- 1.3	0.6 +/- 0.8
E. coli regulatory network	42	0

Undirected motifs

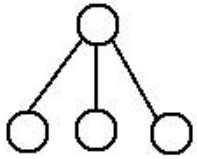
Undirected motifs enriched in the datasets.

motif 1



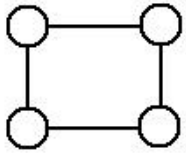
CA-HepTh
CA-GrQc
CA-CondMat
netscience
lesmis
dolphins
football
polbooks
Hsapi201010CR
Hsapi201010
Scere201010
PTT

motif 0



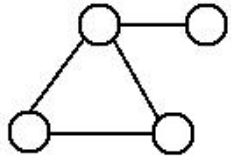
CA-GrQc
adjnoun
lesmis
football
polbooks
Hspi201010CR

motif 3



karate
football
polbooks
Hsapi20101010CR
Hsapi20201010
PTT

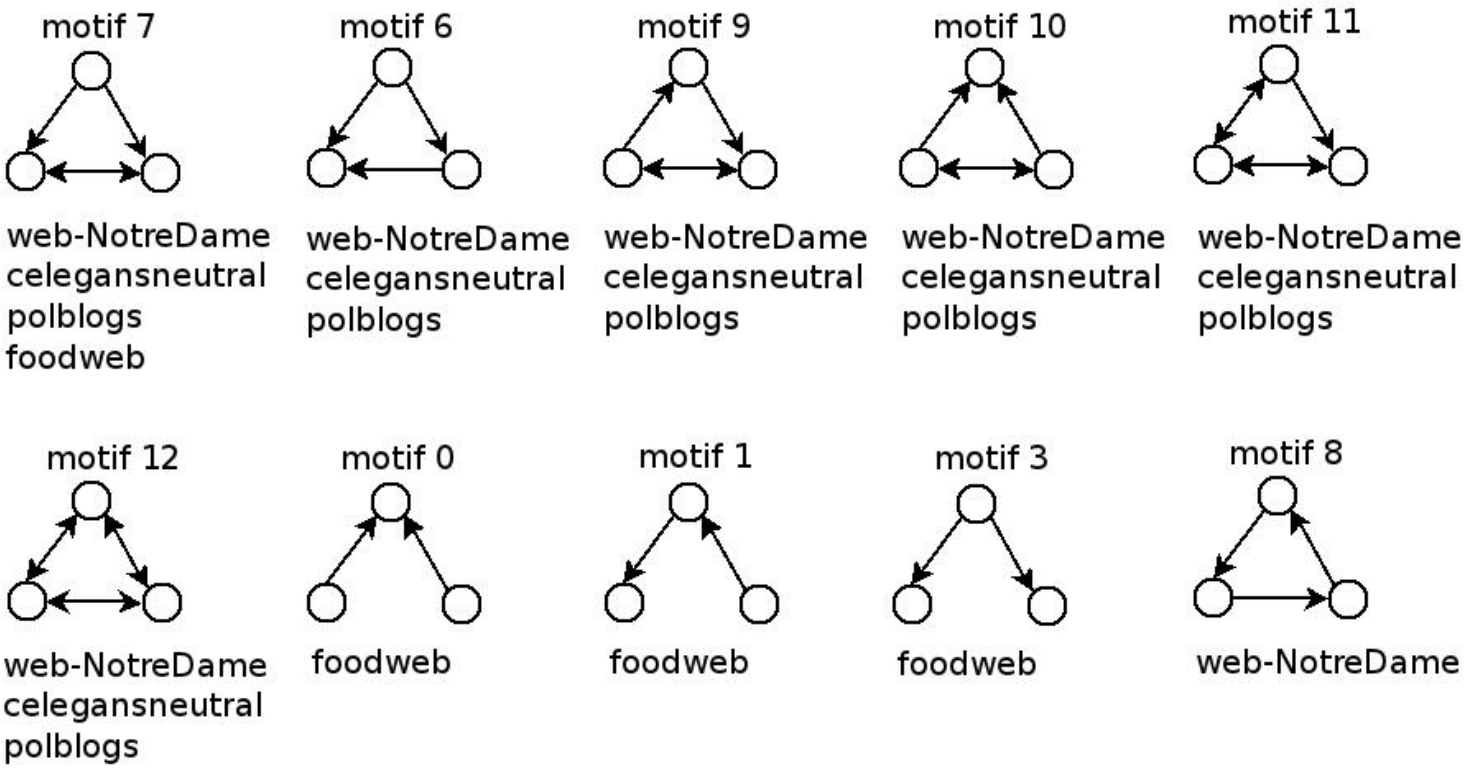
motif 2



football

Directed motifs

Three-node directed motifs enriched in the datasets.



Conclusion

- Networks are collections of relations.
- Many networks observed in the real world possess distinct characteristics from random networks.
- Network analysis is relevant in a wide range of application domains (e.g., biology, sociology, economics, etc.).
- You will learn more about networks in subsequent classes.