Hands-on experience on network topology

Chen-Hsiang Yeang
Institute of Statistical Science
Academia Sinica

Outline

- Network preparation
- Lecture:
  - Network visualization and ERGM
  - Power law distributions
  - Small-worldness
  - Community structures
  - Network motifs
- Exercise:
  - Group 1: Network visualization
  - Group 2: Node degree distributions
  - Group 3: Path lengths and clustering coefficients
  - Group 4: Community detection
  - Group 5: Network motif detection

Network preparation

Please email the TA the following information right away:

- Your name
- Names of the students in the class with whom you have frequent interactions
- Your gender
- Your age
- The school and department you are attending
- The number of years you've been studying at college.
- Your city of residence.

Outline

- Network preparation
- Lecture:
  - Network visualization and ERGM
  - Power law distributions
  - Small-worldness
  - Community structures
  - Network motifs
- Exercise:
  - Group 1: Network visualization
  - Group 2: Node degree distributions
  - Group 3: Path lengths and clustering coefficients
  - Group 4: Community detection
  - Group 5: Network motif detection

Network visualization using Cytoscape

- Launch Cytoscape
- Input files
- Graph layout and manipulation
- Visualize node and edge attributes

Determination of links

What determines the presence of links?

- Homophily – birds of a feather flock together.
- Transitivity – people sharing more common friends tend to be friends.
- Match of physical characteristics – molecular bindings.

Exponential Random Graph Models (ERGM):

- $\mathbf{h}_i$: attributes of node i.
- $Y_{ij} \in \{0, 1\}$: indicator of whether (i, j) forms a link.
- $P(Y_{ij} = 1|\theta) = \frac{1}{Z(\theta)} exp(\theta_{ij} \mathbf{h}_i \mathbf{h}_j)$.
- Probability of an edge existence depends of its node attributes.

Outline

- Network preparation
- Lecture:
  - Network visualization and ERGM
  - Power law distributions
  - Small-worldness
  - Community structures
  - Network motifs
- Exercise:
  - Group 1: Network visualization
  - Group 2: Node degree distributions
  - Group 3: Path lengths and clustering coefficients
  - Group 4: Community detection
  - Group 5: Network motif detection

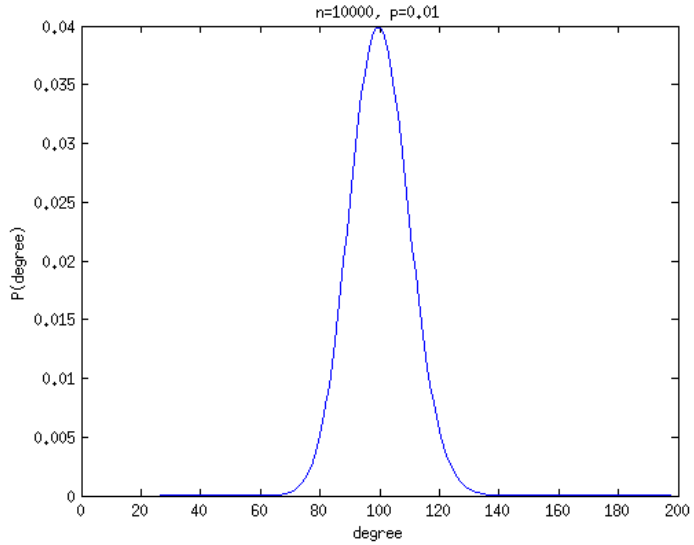
Erdős-Rényi random graphs

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.

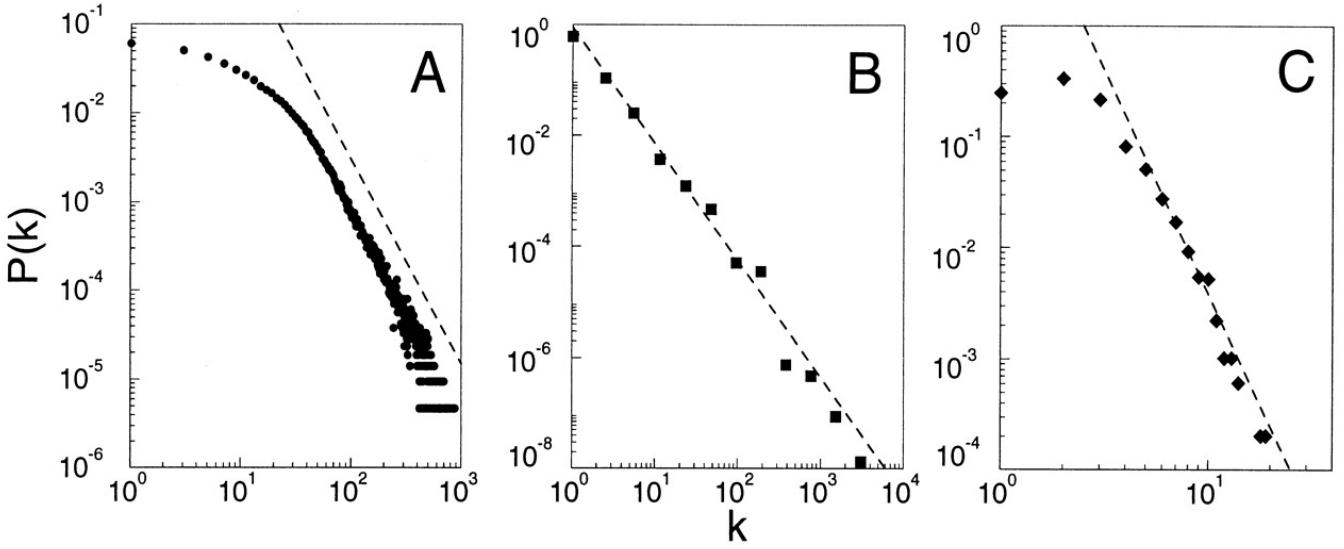
Erdős P. and Rényi A., 1960.

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.

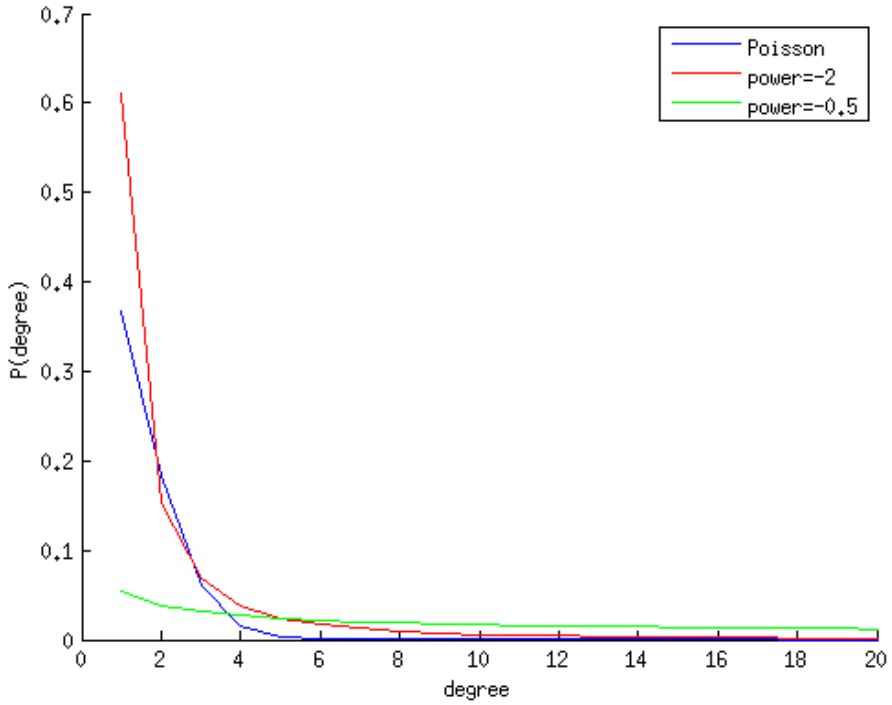
Scale-free networks



- 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



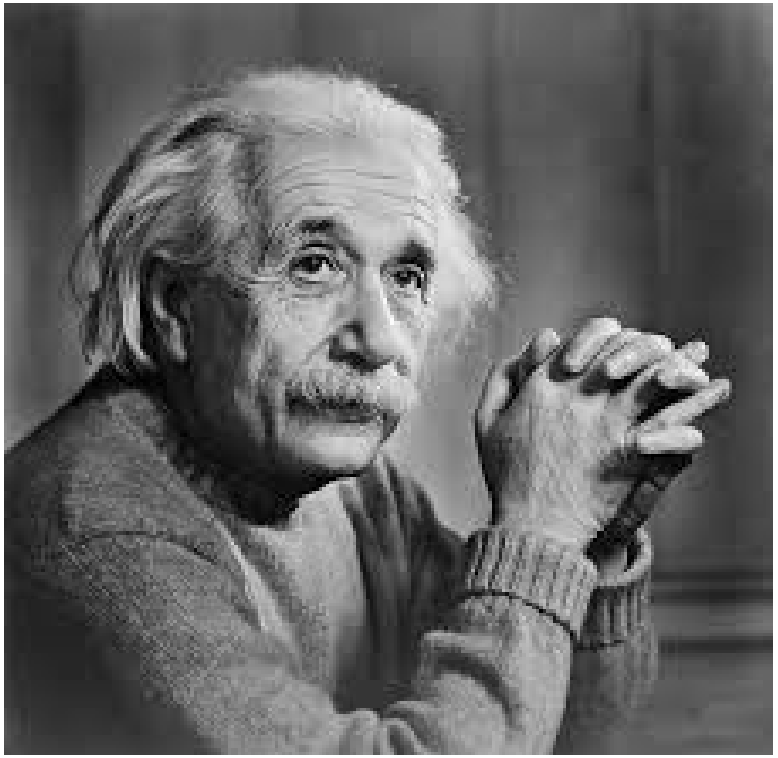
- Scale-free networks possess heavy tails compared to random graphs.
- The majority of nodes are adjacent to a few neighbors.
- A small number of *hubs* are highly connected.
- The networks are *scale-free* as the shape of the degree distribution is invariant with network scales.

Outline

- Network preparation
- Lecture:
  - Network visualization and ERGM
  - Power law distributions
  - Small-worldness
  - Community structures
  - Network motifs
- Exercise:
  - Group 1: Network visualization
  - Group 2: Node degree distributions
  - Group 3: Path lengths and clustering coefficients
  - Group 4: Community detection
  - Group 5: Network motif detection

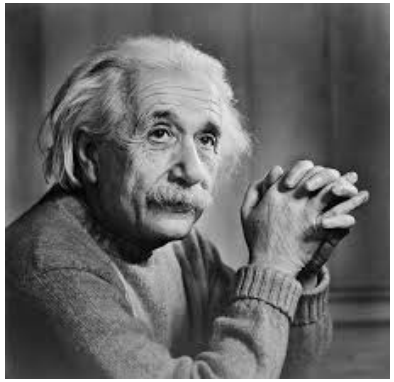
Small-world networks

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



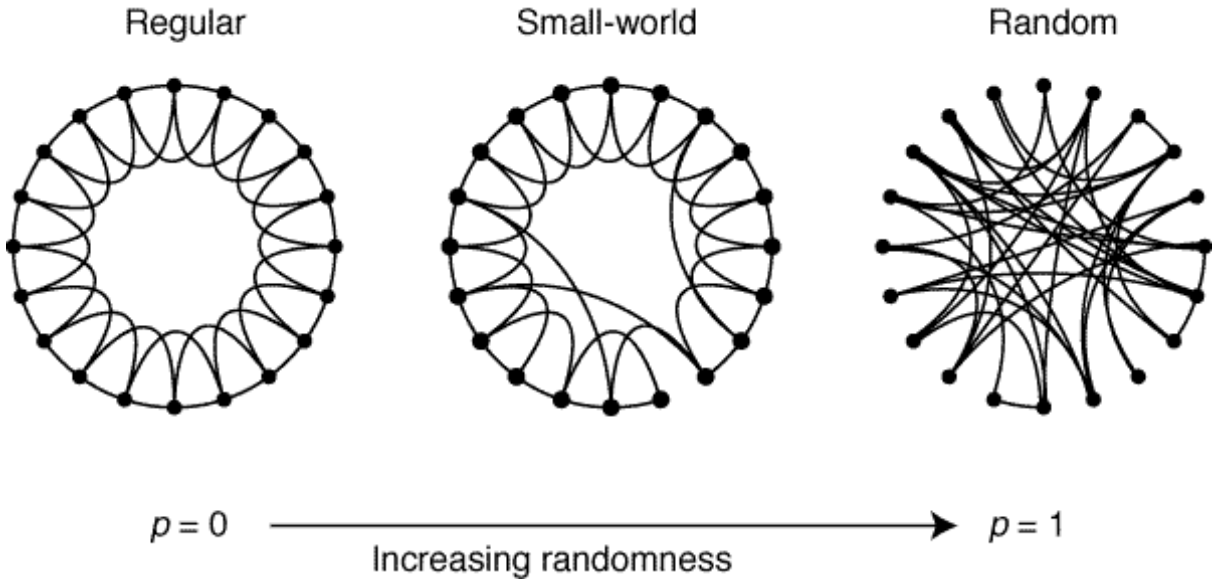
How many degrees separate Einstein from me?

Small-world networks




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

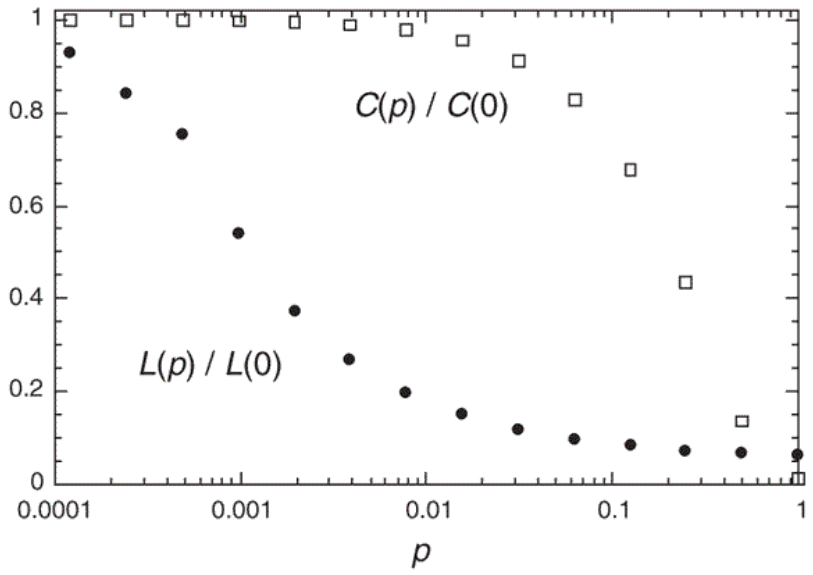
Small-world networks



- 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

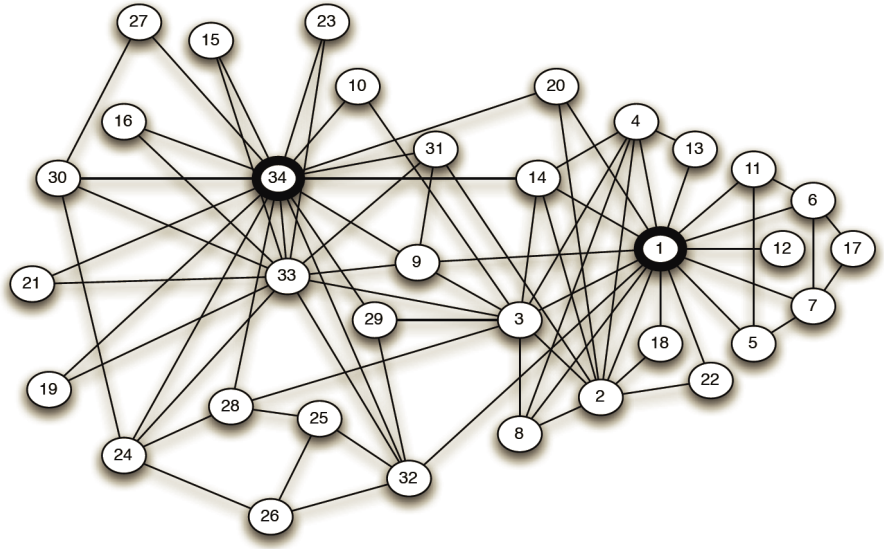


- Vary p from 0 (regular graph) to 1 (random graph).
- Normalized characteristic path length drops quickly with increasing p.
- Normalized clustering coefficient is robust against p.
- Thus graphs within mid range p values satisfy small-world properties.

Outline

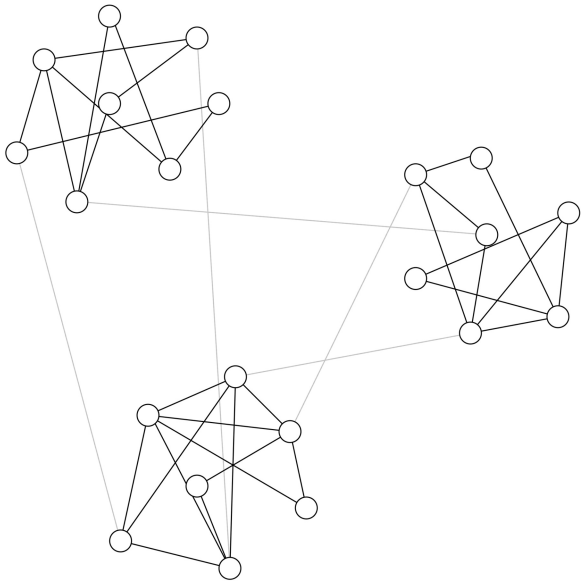
- Network preparation
- Lecture:
  - Network visualization and ERGM
  - Power law distributions
  - Small-worldness
  - Community structures
  - Network motifs
- Exercise:
  - Group 1: Network visualization
  - Group 2: Node degree distributions
  - Group 3: Path lengths and clustering coefficients
  - Group 4: Community detection
  - Group 5: Network motif detection

Community structures



- A society typically consists of multiple factions.
- In the small and famous social network of a karate club, members belong to two communities.

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.

Outline

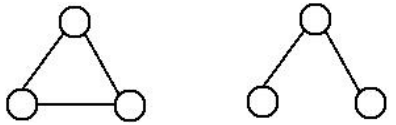
- Network preparation
- Lecture:
  - Network visualization and ERGM
  - Power law distributions
  - Small-worldness
  - Community structures
  - Network motifs
- Exercise:
  - Group 1: Network visualization
  - Group 2: Node degree distributions
  - Group 3: Path lengths and clustering coefficients
  - Group 4: Community detection
  - Group 5: Network motif detection

Network motifs

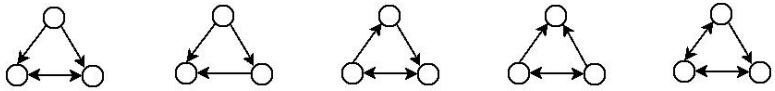
- Motifs are recurrent patterns in the data.
- Network motifs are recurrent topological structures in large-scale networks.
- Two motif instances are illustrated above.

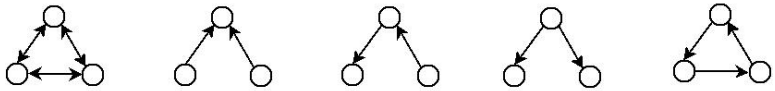
Network motifs

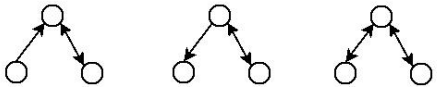
Undirected 3-motifs:



Directed 3-motifs:



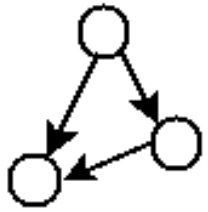
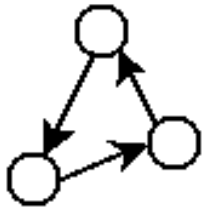




Enrichment of network motifs

- 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.

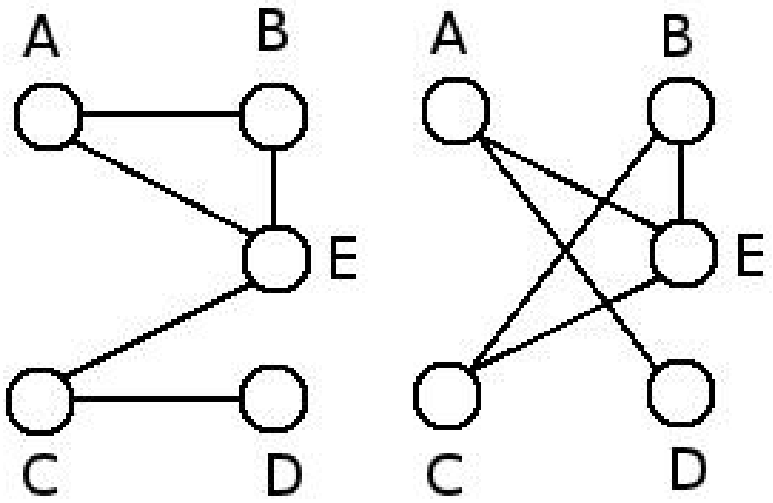
	feed-forward loop	3-node cycle
		
random graphs	1.7 +/- 1.3	0.6 +/- 0.8
E. coli regulatory network	42	0

Evaluating statistical significance of network motif frequencies

- To quantify statistical significance we need a proper null model.
- A null model generates randomized networks resembling the base network in the following aspects:
  - Have the same size as the base network.
  - Have the same degree distribution as the base network.
  - Have the same frequencies of lower-order motifs.

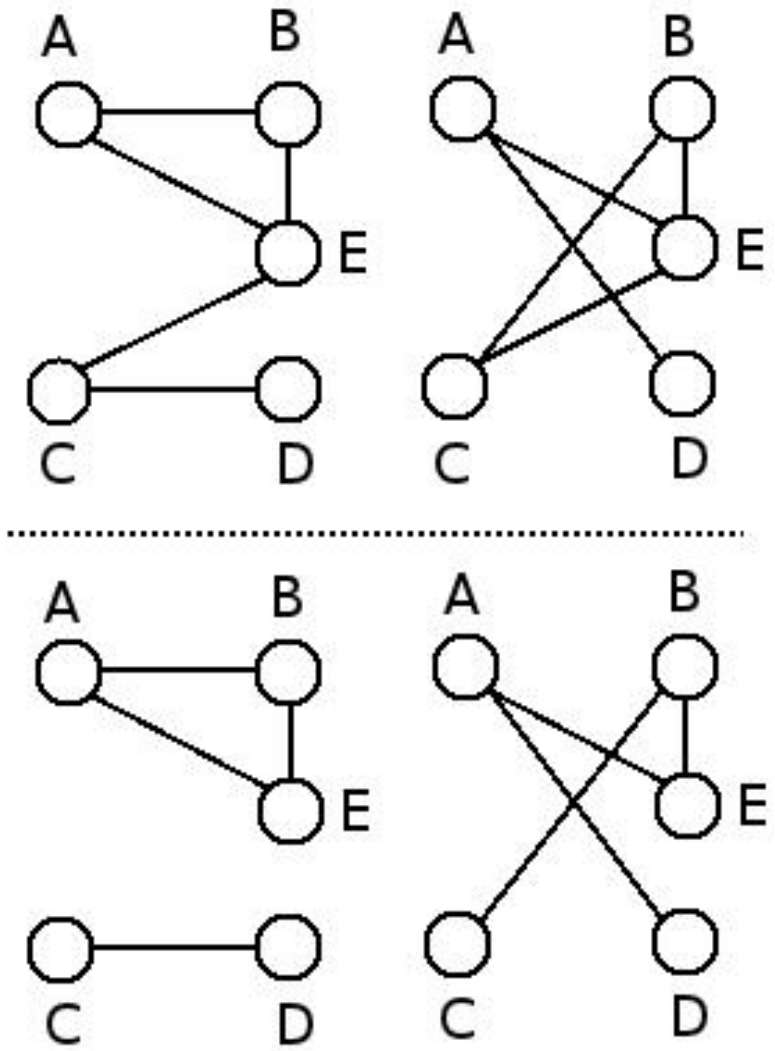
Evaluating statistical significance of network motif frequencies

Edge swaps can satisfy properties 1 and 2.



Evaluating statistical significance of network motif frequencies

However, edge swaps do not necessarily satisfy property 3.



Detection of network motifs

- Inputs: A network G (directed or undirected), a collection of network motifs $M_1, \cdots, M_K$.

- Outputs: The over-represented motifs in the network.

  1. For each motif M_i, evaluate its statistical significance of over-representation in G:

     (a) Count the occurrence f_i of M_i in G.

     (b) Generate random graphs $rG_1, \cdots, rG_n$ by swapping edges that preserve statistics of low-order motifs.

     (c) Count the occurrences rf_{ij} of M_i in rG_j.

     (d) The p-value of M_i is $\frac{1}{n}\sum_{j=1}^{n} I(rf_{ij} \geq f_i)$.

Outline

- Network preparation
- Lecture:
  - Network visualization and ERGM
  - Power law distributions
  - Small-worldness
  - Community structures
  - Network motifs
- Exercise:
  - Group 1: Network visualization
  - Group 2: Node degree distributions
  - Group 3: Path lengths and clustering coefficients
  - Group 4: Community detection
  - Group 5: Network motif detection

Rules of the game

- Divide the class into five groups.
- Each group is given the undirected graph of students acquaintances and attributes.
- Each group has one hour to perform the assigned tasks.
- Finally, each group will solicit one representative to present the results.

Group 1 task: Network visualization

- Visualize the network using Cytoscape.
- Find out the node attributes that are likely to explain the presence of links.
- Visualize your findings using Cytoscape.

Group 2 task: Node degree distributions

- Calculate and present the node degree distribution of the network.
- Generate 1000 Erdős-Rényi random graphs of the same number of nodes and expected number of edges.
- Calculate and present the node degree distribution of the random graphs.

Group 3 task: Path lengths and clustering coefficients

- Calculate the shortest path lengths between all node pairs in the graph.
- Discard when two nodes are not connected.
- Calculate the average shortest path length.
- 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).

Group 4 task: Community detection

- Calculate the betweenness of each edge in the network.
- Report the edges of high betweenness.
- Apply the Girvan-Newmann algorithm to detect communities.

Group 5 task: Network motif detection

- Count the number of 3-node motifs (triangles and chains) in the network.
- Generate 1000 random graphs by edge swaps.
- Calculate the enrichment p-values of motif occurrences.

It's your turn now