7-1- Introduction to Graph Generation

Complex Network Analysis Course
Reza Rezazadegan
Shiraz University, Spring 2025
https://dreamintelligent.com/complex-network-analysis-2025/

Main questions in graph/network generation

- How can the formation/evolution of a network be formulated?
- How can we sample graphs (from a probability distribution on graphs) that resemble real-world graphs?
- How can we learn this probability distribution from a set of networks, or just from the evolution of a single network?
- How can we predict or simulate the future structure of an evolving network? (Similar to predicting a time series.)
- What rules or procedures govern the formation/evolution of different networks?

Applications of Graph Generation

- Designing new molecules/drugs
- Generating graphs for testing or simulating
- Social network modeling
  - example: Jefferson Highschool
- Text generation using graph representation of text

Traditional methods start by making assumptions about network formation:

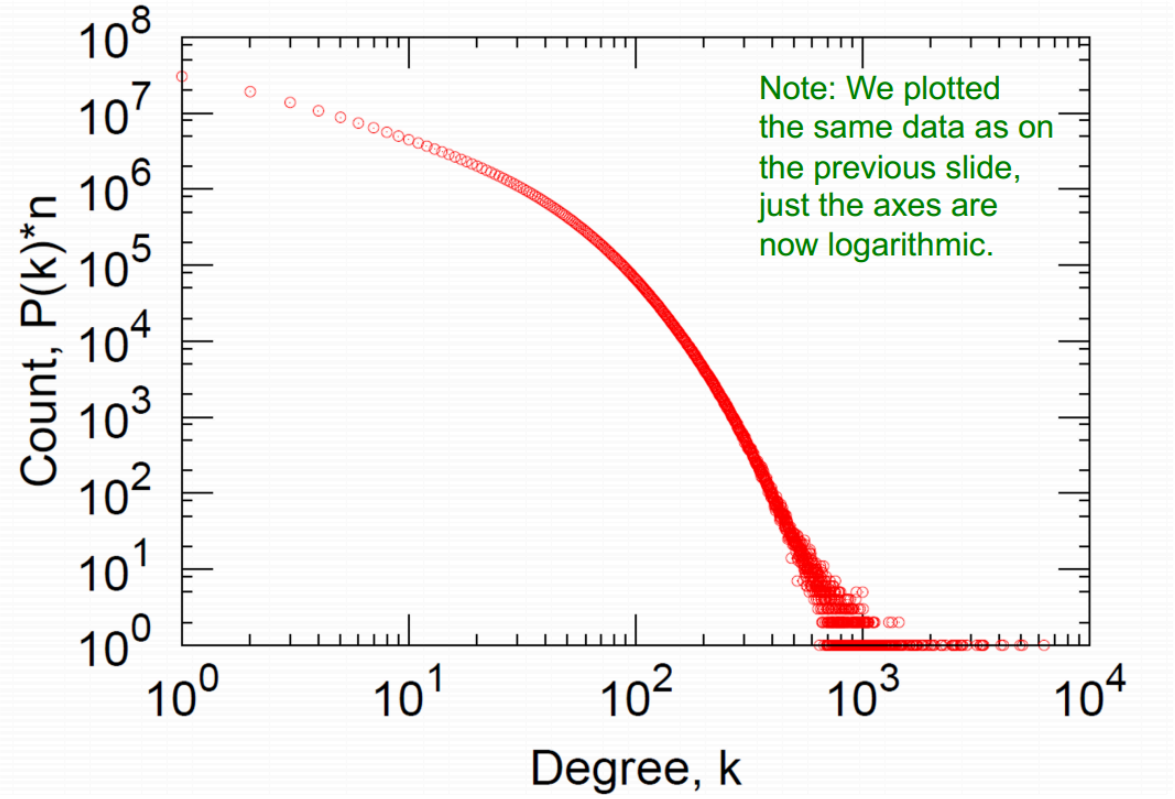
- **Random graph model:** the probability of edge formation between any two nodes is the same.
- **Barabasi-Albert model:** new nodes are more likely to link to existing nodes of higher degree.
- **Bianconi-Barabasi model:** similar to the Barabasi-Albert model but with increasing number of nodes.

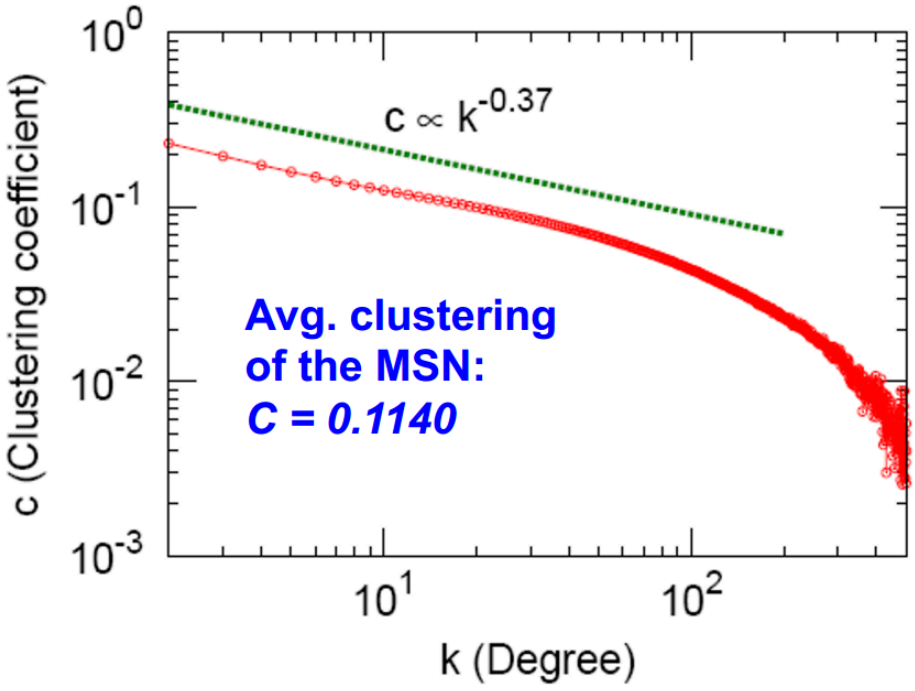
Newer, deep graph generation models *learn* the formation process from the data! (Future chapters.)

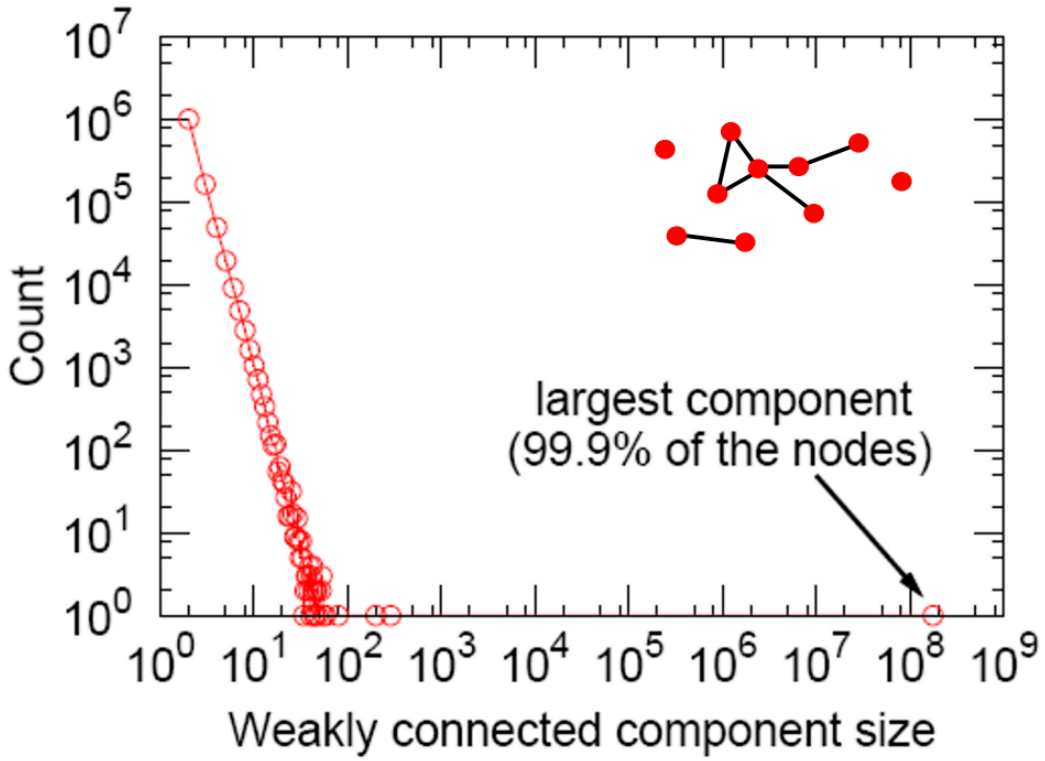
Properties of Natural Networks

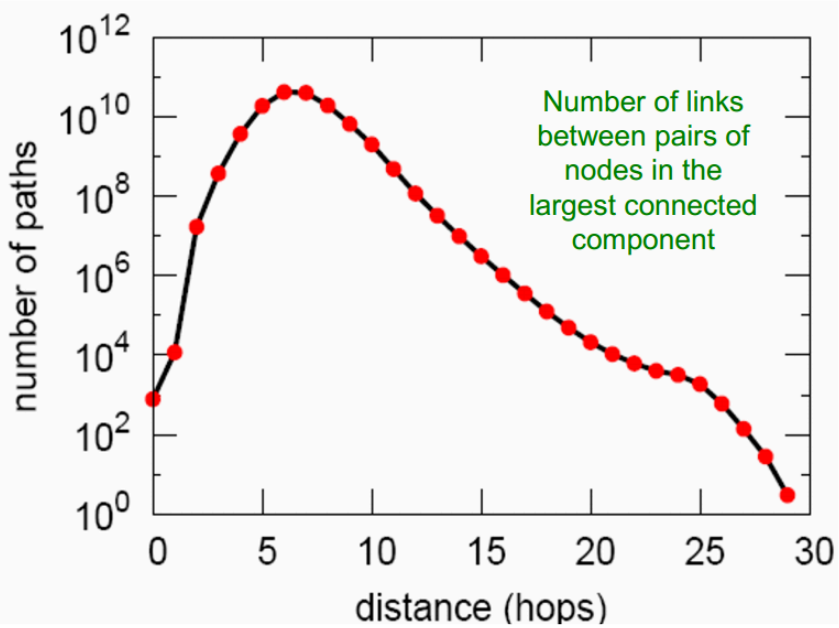
- Degree distribution
- Clustering coefficient
- Connectivity: connected component sizes
- Path length distribution

Example: Microsoft Messenger Network




c (Clustering coefficient)
10^0
10^{-1}
10^{-2}
10^{-3}
$c \propto k^{-0.37}$
Avg. clustering
of the MSN:
C = 0.1140
10^1
10^2
k (Degree)





Case Study: The Network of Romantic Relationships in Jefferson High

The researchers examine three different hypotheses for the formation of the network.

The first one is that relationships are formed at random. However this hypothesis does not account for the "spanning tree" structure of the network.

The second hypothesis is that relationships are formed based on homophily. The researchers used QAP to evaluate the difference in attribute means between *actual romantic pairs* and the *randomly simulated partnerships*.
Quadratic Assignment Procedure (QAP) tests whether there is a significant association between two matrices.

For each feature F we have two matrices, **X** and **Y**, which might represent:

- X: **attribute distance matrix** ($M_{i,j} = |F_i - F_i|$) or the **attribute matching matrix** ($M_{i,j} = 1$ if nodes i, j have the same attribute and 0 otherwise.)
- Y: romantic relation matrix (same as adjacency matrix)

You want to test whether X and Y are significantly correlated.

1. **Compute the observed correlation** (or regression coefficient) between corresponding elements of X and Y.
2. **Permute the rows and columns** of one matrix (say, Y) the same way (to preserve structure) many times (e.g., 1,000 permutations).
3. For each permutation, **recalculate the correlation**.
4. Compare the original observed value to the distribution of permuted values to get a **p-value**.

	QAP Mean Difference[a]	
Variable	Full Network	Cross-Sex Only
Family SES	.299***	.295***
Grade	.331***	.367***
GPA	.096**	.102***
Expect to graduate college	.202***	.222***
School attachment	.118***	.132***
Trouble in school	.029	.019
Gets drunk	.180***	.195***
Delinquency[b]	−.058	−.070
Hours watching TV	−.149	−.027
Religiosity (praying)	−.006	−.012
Popularity (in-degree)	−.377*	−.211
Self-esteem	.004	.008
Autonomy	.008	.002
Expect to get HIV	.003	−.007
Expect to marry by 25	.025	.020
Attractiveness	.013	.047
Vocabulary (AH_PVT)	1.508***	1.671***
Religion	−.034*	−.043*
Sexually active	−.100***	−.124***
Smoking	−.087***	−.110***
School suspension	−.028	−.066**
Tattoo	−.003	−.016

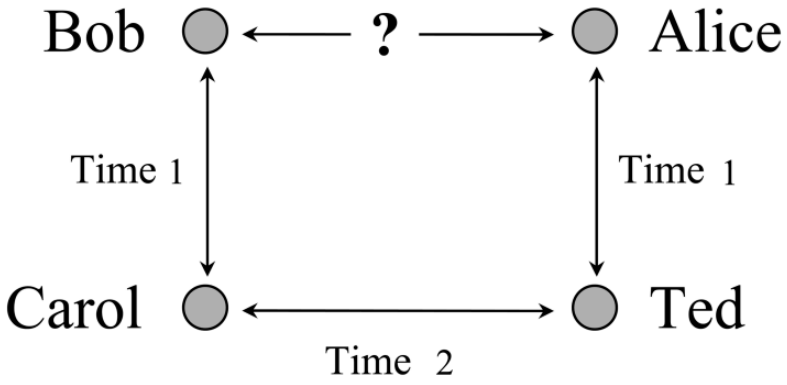
Source: Peter Bearman, James Moody, and Katherine Stovel, American Journal of Sociology, 110(1):44–99, 2004.

Even though there is strong evidence for homophily, it does not extend to all characteristics, e.g. age.

The researchers then considered the effect of random rewiring of the network. To rewire the empirically observed graph, they select 5% of the relationships at random and reassign them conditional only on the degree distribution of the original graph.

The rewired graphs are quite similar to the observed network. But the rewired networks have almost twice as many cycles as are observed in
Jefferson.

The third and last hypothesis is a type of exclusion rule that persons do not date the former (or current) partner of their former (or current) partner by prohibiting all cycles of length 4.



Source: Peter Bearman, James Moody, and Katherine Stovel, American Journal of Sociology, 110(1):44–99, 2004.

The third hypothesis generated networks quite similar to the observed network, in terms of degree distribution and the number and length of cycles.