Social Data Management
Graph Formation Models

Benoît Groz (slides and content are mostly from Silviu Maniu)

November 3rd, 2023

M2 Data Science

Table of contents

Random Networks

We saw that the real networks are **sparse**: is that the only relevant measure?

Objective of graph models: reproduce the complexity of real networks via simple models

Assume we have only two parameters:

- the number of nodes N,
- the probability of an edge existing, p.

What is the graph model, and what properties does it have?

Random Networks: Algorithm

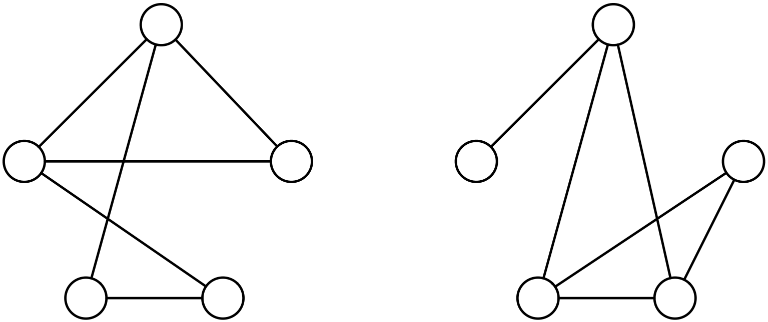
Random Network models, discovered and studied by P. Erdős and A. Rényi.

1. Start with N disconnected nodes.
2. For a node pair, add an edge between them with probability p.
3. Repeat this for all $N(N-1)$ node pairs.

Two possible models:

- $G(N, p)$ model: a graph of N nodes, and each link is connected with a probability p, or
- $G(N, L)$ model: a graph of N nodes, where L links are chosen randomly.

Example: Random Networks ($p = 0.5$)



Random Networks: Basic Measures

Expected number of links:

$$\langle L \rangle = p\frac{N(N-1)}{2}$$

Average degree:

$$\langle k \rangle = p(N-1)$$

Random Networks: Degree Distribution

To compute the probability of a given degree k, we need:

- the probability that exactly k links are present: p^k,
- the probability that the other $N-1-k$ links are not present: $(1-p)^{N-1-k}$, and
- the number of ways one can select k links for the $N-1$ available: $\binom{N-1}{k}$.

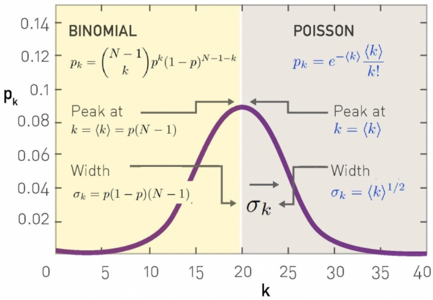
This is exactly a binomial distribution:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

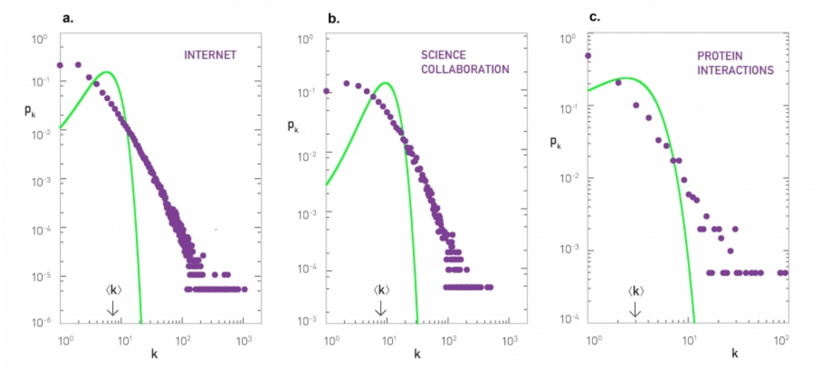
Random Networks: Degree Distribution

For very sparse networks, $\langle k \rangle \ll N$, the degree distribution is also well approximated by the Poisson distribution:

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle}{k!}.$$



Real Networks Do Not Have Poisson Distributions



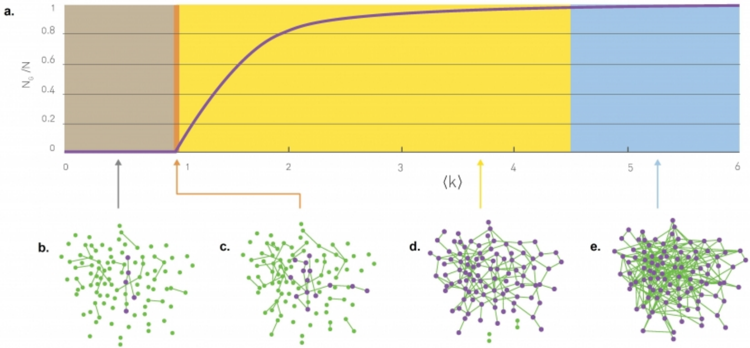
Predicted distribution (green) versus actual one

Evolution of Random Networks [Erdös, Rényi]

Depending on p (we compare p to $1/N$ or $\langle k \rangle$ to 1) we have several regimes of: **random networks**: (N_G= Nb of nodes of largest cluster)

1. **Subcritical regime**, $p < 1/N$: numerous tiny connected components. $N_G \in O(\ln N)$, clusters are trees.
2. **Critical point**, $p = 1/N$: the shifting point: a "large" component and small components $N_G \sim N^{2/3}$, large cluster may have loops, small are mostly trees.
3. **Supercritical regime**, $p > 1/N$: one connected component that dominates other small ones. $N_G \sim (p - p_C)N$, giant has loops, others are mostly trees, where $p_C \sim 1/N$ is critical point.
4. **Connected regime**, $p > \ln N/N$: one single connected component. $N_G \sim N$, has loops.

Evolution of Random Networks



Real Networks are Supercritical

name	$\lvert V\rvert$	$\lvert E\rvert$	$\langle k\rangle$	$\ln N$
LiveJournal	4,847,571	68,993,773	14.23	15.39
WikiTalk	2,394,385	5,021,410	2.09	14.68
Enron	36,692	183,831	4.99	10.51
CondMat	23,133	93,497	4.04	10.04
RoadCA	1,965,206	2,766,607	1.40	14.49
Web	875,713	5,105,039	5.82	13.68

However, the random network predicts multiple connected components in the supercritical regime – this does not occur in real networks.

Random Networks: Clustering Coefficient

We need to estimate the **expected number of links L_i** of a node i's k_i neighbors:

$$\langle L_i \rangle = p \frac{k_i(k_i - 1)}{2}.$$

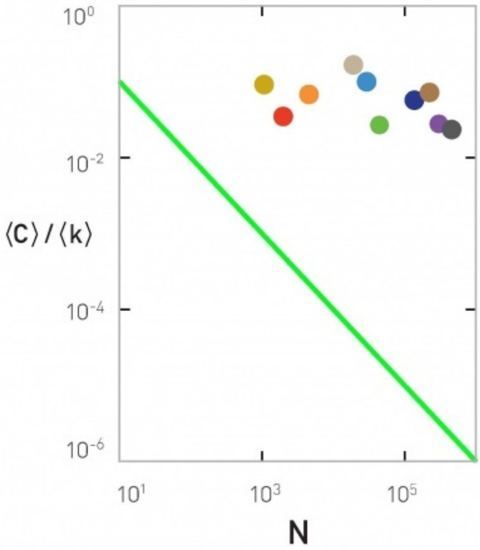
Then, the **clustering coefficient C_i** is:

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

Two interpretations:

1. for a constant $\langle k \rangle$, the larger the network the smaller a node's clustering coefficient, and
2. the clustering coefficient for a node is independent of the degree.

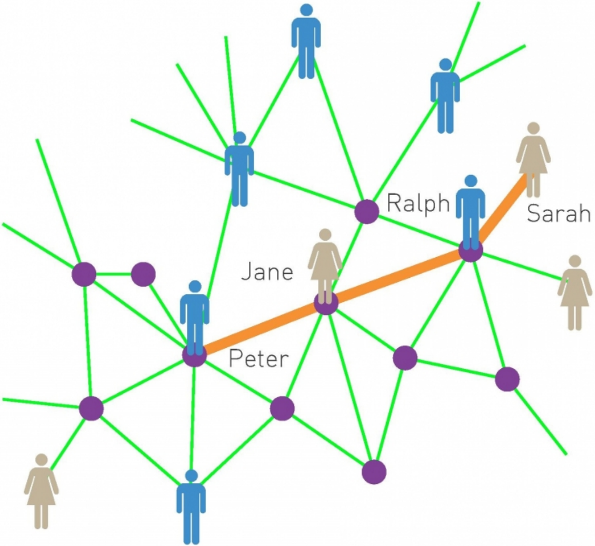
Random Networks Do Not Capture Clustering Coefficients



Predicted clustering coefficient (green) versus actual one

Table of contents

Six Degrees of Separation



Six Degrees of Separation

Small-world phenomenon (or **six degrees of separation**): choosing any two persons, one can find a path of few acquaintances between them.

Or: *distance between any two nodes in a network is short*

How can we justify this?

Average and Maximum Distance

Given a graph with average degree $\langle k \rangle$ a node has on average $\langle k \rangle^d$ nodes at distance d.

Number of nodes upto distance d is:

$$N(d) \approx 1 + \langle k \rangle + \langle k \rangle^2 + \cdots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1}.$$

Setting $N(d_{\max}) \approx N$ and assuming $\langle k \rangle \gg 1$:

$$\langle k \rangle^{d_{\max}} \approx N,$$

and hence:

$$d_{\max} = \frac{\ln N}{\ln \langle k \rangle}.$$

Small Worlds in Random Networks

For most networks, the previous equation offers a better approximation for the average distance:

$$\langle d \rangle = \frac{\ln N}{\ln \langle k \rangle}$$

Generally $\ln N \ll N$, implying that distances are orders of magnitudes smaller than the size of the graph.

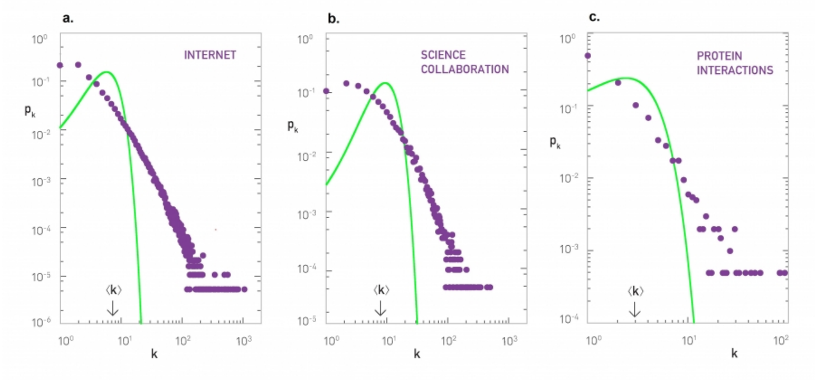
The $1/\ln\langle k \rangle$ terms implies that the denser the network, the smaller the distances are.

This estimator works also for real-world networks, with some small corrections.

Table of contents

Degree Distributions in Real Networks

Let us look again at the degree distribution:



Degree Distributions in Real Networks

In real networks, we have **hubs**: a few extremely well-connected nodes, pointing to many links.

These are effectively **forbidden by random networks**

Degree Distributions in Real Networks

The degrees seem to approximately follow a power law distribution, roughly of the form:

$$p_k \sim k^{-\gamma}.$$

Scale-free network: a network whose degree distribution follows a power law.

Power-laws have long tails, such as the hubs in the real networks.

Degree Distributions in Real Networks

The degree distribution is of the form:

$$p_k = Ck^{-\gamma}.$$

Remember that $\sum p_k = 1$ so we need to set:

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\zeta(\gamma)},$$

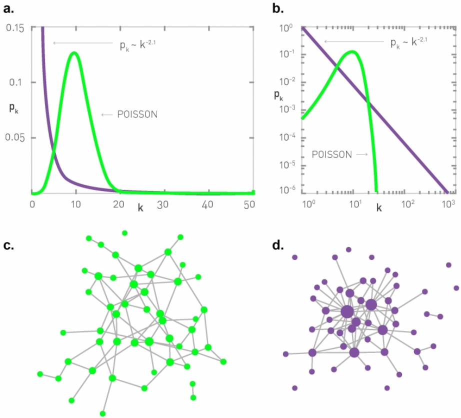
where ζ is Riemann-zeta function.

The final form is then:

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}.$$

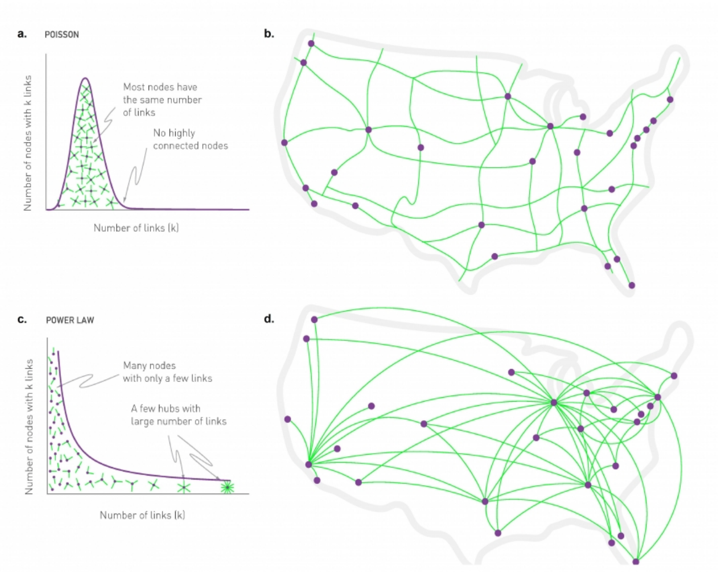
.

Poisson versus Power-law



For small k, the power law is above the Poisson function, i.e., large number of small-degree nodes.
For k around $\langle k \rangle$ the Poisson distribution is above the power law, indicating that a random network has many nodes around the mean.
For large k the power law is again above the Poisson, indicating the presence of hubs.

Poisson versus Power-law



Why Scale-Free?

Scale-free: comes from an area of physics called *phase transitions*, studying power-laws.

To understand it, we use the **moments of a distribution** $\langle k^n \rangle = \sum k^n p_k$, e.g.:

1. $\langle k \rangle$ is the mean of the distribution
2. $\langle k^2 \rangle$ allows to compute the **variance** $\sigma_k^2 = \langle k^2 \rangle - \langle k \rangle^2$, i.e., the spread of the degrees
3. $\langle k^3 \rangle$ measures the **skewness** of the distribution, i.e., how symmetric p_k is

Why Scale-Free?

There are major differences between random networks and scale-free networks:

- **Random networks have a scale**: $\sigma_k = \langle k \rangle^{1/2} < \langle k \rangle$. This means that the nodes in a random networks have comparable degrees.
- **Scale-free networks do not have a scale**: assuming $\gamma < 3$, $\langle k \rangle$ is finite, but $\langle k^2 \rangle$ is *infinite*. That means that node degrees can be arbitrarily tiny or arbitrarily large.

Ultra Small-World Property

Average distance $\langle d \rangle$ depends on N and the exponent γ:

1. **Anomalous Regime** ($\gamma = 2$). The degree of the biggest hub grows linearly with N, so $\langle d \rangle \sim$ constant (**hub-and-spoke**).
2. **Ultra-Small World** ($2 < \gamma < 3$). $\langle d \rangle \sim \ln \ln N$, slower growth than random networks. *This is where most real networks are.*
3. **Critical Point** ($\gamma = 3$). The moment when $\langle k^2 \rangle$ does not diverge any more, i.e., the moment between scale-free and random regime. Here $\langle d \rangle \sim \frac{\ln N}{\ln \ln N}$.
4. **Small World** ($\gamma > 3$). This is the random network regime, when $\langle d \rangle \sim \ln N$.

The Role of the Degree Exponent

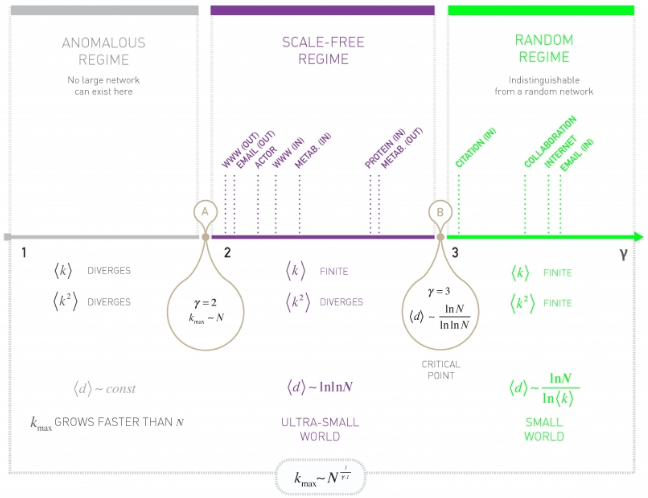


Table of contents

Growth of Real Networks

The random network model assumes we have a *fixed* number of nodes, whereas in real networks the graph **grows continually**.

Moreover, **new nodes prefer to link to more connected nodes**, e.g., following people on Twitter, books, movies, etc.

Growth of Real Networks

We need two ingredients:

1. **Growth**: The model should allow adding nodes, and not only a fixed number of nodes.
2. **Preferential Attachment**: New nodes should tend to link to more connected nodes.

Barabási–Albert Model

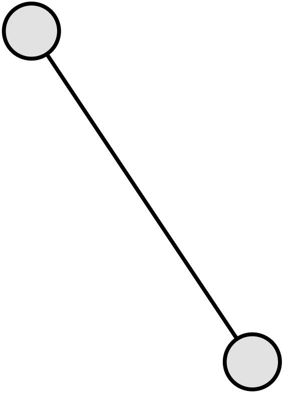
A model that generates **scale-free networks**. It takes a single parameter, m.

We start with m_0 nodes with links chosen arbitrarily.

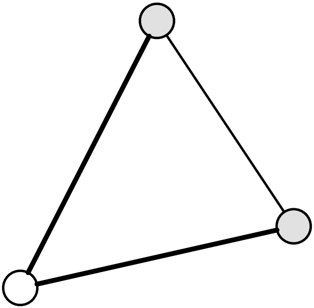
Then, the process goes in two steps:

1. **Growth** At each step, we add a nodes to the network, with m links to connect to other nodes.
2. **Preferential Attachment** Each of the m links can connect to node i with probability $\mathrm{P}(i) = \frac{k_i}{\sum_j k_j}$.

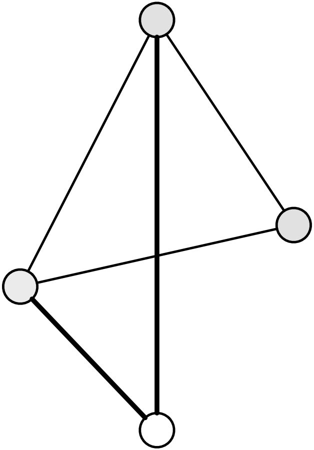
Example: Barabási–Albert Model ($m_0 = 2$, $m = 2$)



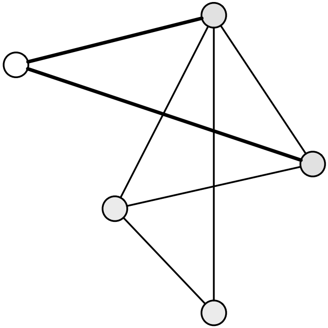
Example: Barabási–Albert Model ($m_0 = 2$, $m = 2$)



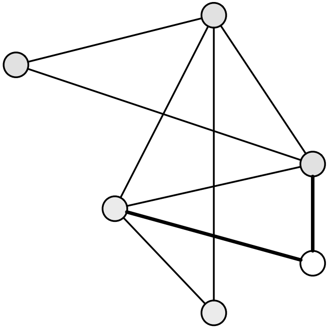
Example: Barabási–Albert Model ($m_0 = 2$, $m = 2$)



Example: Barabási–Albert Model ($m_0 = 2$, $m = 2$)



Example: Barabási–Albert Model ($m_0 = 2$, $m = 2$)



Degree Dynamics

We study next the time-dependent degree of a node i:

$$\frac{\partial k_i}{\partial t} = m\mathrm{P}(i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j}.$$

By using the fact that $\sum_{j=1}^{N-1} k_j = m(2t-1)$ and by integrating, we obtain:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta},$$

where $\beta = 1/2$ is called the dynamical exponent.

Degree Distribution

The previous result leads us to the degree distribution:

$$p_k \approx 2m^{\frac{1}{\beta}} k^{-\gamma},$$

where $\gamma = \frac{1}{\beta} + 1 = 3$.

Interpretation:

1. for large k, $p_k \approx k^{-3}$, resulting in a scale-free network,
2. the degree exponent γ is independent of m, in line with real results, and
3. the model predicts the emergence of stationaly scale-free network.

Other Measures

Average distance:

$$\langle d \rangle \sim \frac{\ln N}{\ln \ln N},$$

i.e., the distances grow slower than in random networks, hence closer to the real network prediction.

Clustering coefficient:

$$\langle C \rangle \sim \frac{(\ln N)^2}{N},$$

meaning that the model predicts a network that is more locally clustered than a random network.

Shortcomings of the Model

1. The model predicts $\gamma = 3$ while the exponent in real networks ranges from 2 to 5.
2. It only works for undirected networks.
3. Linking between already existing nodes and disappearance of nodes is not modeled.
4. It does not allow to distinguish between nodes of different characteristics.

Acknowledgments

Figures in slides 8, 9, 11, 14, 16, 21, 25, and 26 taken from the book "Network Science" by A.-L. Barabási. The contents is partly inspired by the flow of Chapters 3, 4, and 5 of the same book.

`http://barabasi.com/networksciencebook/`

References i

Barabási, A.-L. (2016).
Network Science.
Cambridge University Press.

Easley, D. and Kleinberg, J. (2010).
Networks, Crowds, and Markets: Reasoning about a Highly Connected World.
Cambridge University Press.

Newman, M. (2010).
Networks: An Introduction.
Oxford University Press.