Network Science: Erdős-Rényi Model for Network Formation

Ozalp Babaoglu
Dipartimento di Informatica — Scienza e Ingegneria
Università di Bologna
www.cs.unibo.it/babaoglu/

Why model?

- Simpler representation of possibly very complex structures
- Can gain insight into how networks *form* and how they *grow*
- May allow mathematical derivation of certain properties
- Can serve to “explain” certain properties observed in real networks
- Can predict new properties or outcomes for networks that do not even exist
- Can serve as benchmarks for evaluating real networks

Modeling approaches

- Random models — choices *independent* of current network structure
  - Erdős-Rényi (ER)
  - Watts-Strogatz (clustered)
- Strategic models — choices *depend* on current network structure
  - Barabási-Albert (preferential attachment)
- Limited knowledge models — choices based on local information only
  - Newscast
  - Cyclone

Erdős-Rényi model

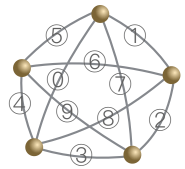
- Network is undirected
- Start with all isolated nodes (no edges) and add edges between pairs of nodes one at a time randomly
- Perhaps the simplest (dumbest) possible model
- Very unlikely that real networks actually form like this (certainly not social networks)
- Yet, can predict a surprising number of interesting properties
- Two possible choices for adding edges randomly:
  - Randomize edge presence or absence
  - Randomize node pairs

Erdős-Rényi model
Randomize edge presence/absence

- Two parameters
  - Number of nodes: n
  - Probability that an edge is present: p
- For each of the $n(n-1)/2$ possible edges in the network, flip a (biased) coin that comes up "heads" with probability p
- If coin flip is "heads", then add the edge to the network
- If coin flip is "tails", then don't add the edge to the network
- Also known as the "$G(n, p)$ model" (graph on n nodes with probability p)

Erdős-Rényi model
Randomize edge presence/absence

- Example: $n=5$, $p=0.6$

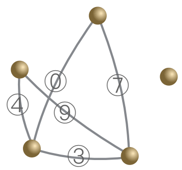


- Number of possible edges: $n(n-1)/2=5\times4/2=10$
- Ten flips of a coin that comes up heads 60%, tails 40%

H T T H H T T H T H
0 3 4 7 9

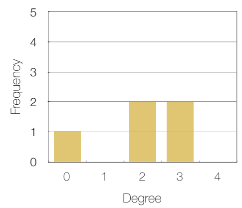
- Add the edges corresponding to the "heads" outcomes

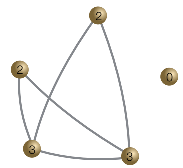
Erdős-Rényi model
Randomize edge presence/absence



- Expected mean node degree: $p(n-1)$
- What about node degree distribution?

Erdős-Rényi model
Degree distribution





- Expected mean node degree: $p(n-1)=0.6\times4=2.4$
- Observed mean node degree: $(3+3+2+2+0)/5=2.0$
- Distribution

Erdős-Rényi model
Degree distribution

- Need to quantify the probability that a node has degree k for all $0 \leq k \leq (n-1)$
- A node has degree zero if all coin flips are "tails"
- A node has degree $(n-1)$ if all coin flips are "heads"
- For a node to have degree k, the $(n-1)$ coin flips must have resulted in k "heads" and $(n-1-k)$ "tails"
- Since the probability of a "heads" is p, the probability of a "tails" is $(1-p)$

Erdős-Rényi model
Degree distribution

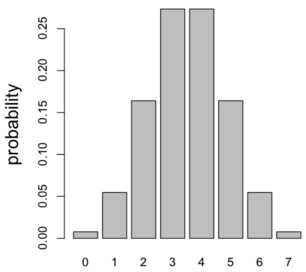
- The outcome "k "heads" and $(n-1-k)$ "tails"" occurs with probability

$$p^k(1-p)^{n-1-k}$$

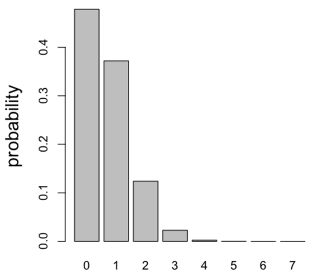
- Since the order of the flip results does not matter, there are several ways for this outcome to occur
- In fact, there are exactly "$(n-1)$ choose k" ways in which this outcome can occur
- Thus, the probability that a given node has degree k is given by the *Binomial distribution*

$$\binom{n-1}{k} p^k(1-p)^{n-1-k}$$

Erdős-Rényi model
Binomial distribution



n=8, p=0.5



n=8, p=0.1

- Mean of the binomial distribution is $\mu=p(n-1)$ (which is also the average node degree we saw earlier)

Erdős-Rényi model
Binomial distribution—approximations

Binomial $\qquad \binom{n-1}{k} p^k(1-p)^{n-1-k}$

for p small ↓

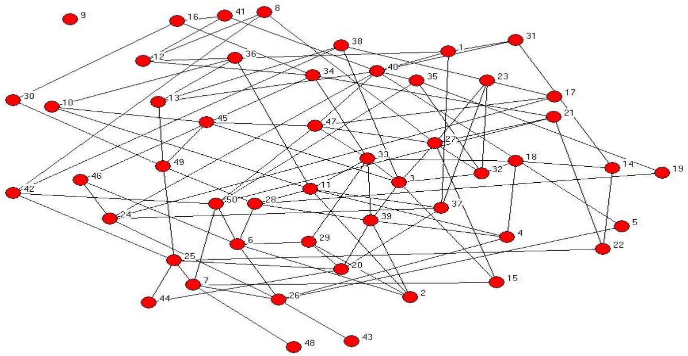
Poisson $\qquad \dfrac{\mu^k e^{-\mu}}{k!}$

for n large ↓

Normal (Gaussian) $\qquad \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(k-\mu)^2}{2\sigma^2}}$

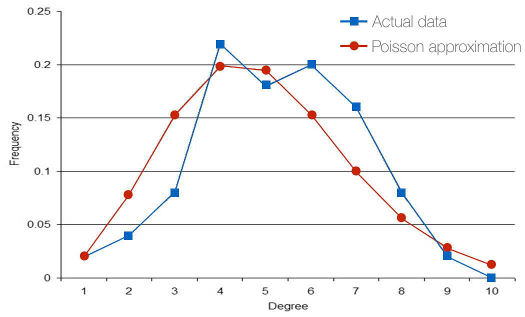
Erdős-Rényi model
Binomial distribution

- Random network with n=50, p=0.08

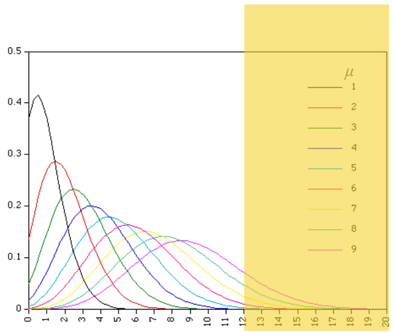


Erdős-Rényi model
Binomial distribution

- Degree distribution of random network with n=50, p=0.08

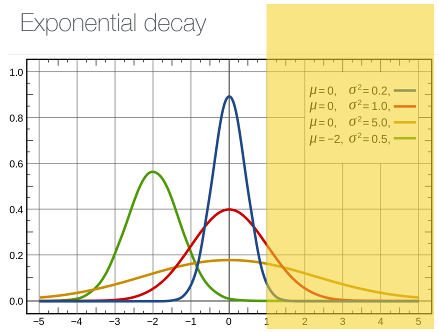


Erdős-Rényi model
Binomial distribution



Poisson distribution with different means

Exponential decay



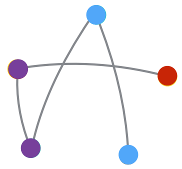
Normal distribution with different means and standard deviations

Erdős-Rényi model
Randomize node pairs

- Alternative method for adding edges randomly
- Two parameters
  - Number of nodes: n
  - Number of edges: m
- Pick a pair of nodes at random among the n nodes and add an edge between them if not already present
- Repeat until exactly m edges have been added
- Also known as the "$G(n, m)$ model" (graph on n nodes with m edges)
- For large n, the two versions of ER are equivalent

Erdős-Rényi model
Randomize node pairs

- Example: $n=5,\ m=4$



- The two versions of the model are related through the equation for the number of edges: $m=pn(n-1)/2$
- In the first case we pick p, and m is established by the model
- In the second case we pick m, and p is established by the model
- The above example corresponds to the second case where $p=2m/n(n-1)=2\times4/(5\times4)=0.4$

Erdős-Rényi model vs real networks
Degree distribution

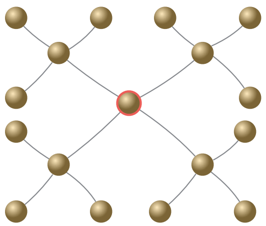
- The ER model is a poor predictor of *degree distribution* compared to real networks
- The ER model results in Poisson degree distributions that have exponential decay
- Whereas most real networks exhibit power-law degree distributions that decay much slower than exponential

Erdős-Rényi diameter

- Recall that the d*iameter* of a network is the longest shortest path between pairs of nodes
- Equivalently, the average distance between two randomly selected nodes
- In a connected network with n nodes, the diameter is in the range 1 (completely connected) to $n-1$ (linear chain)
- For a given n as we vary the model parameter p from 0 to 1, at some critical value of p, the diameter becomes finite (network becomes connected) and continues to decrease, becoming 1 when $p=1$
- What is the relation between the diameter and p in the region where the network is connected?

Erdős-Rényi diameter

- Suppose the model results in a tree-structured network of nodes with identical degrees, all equal to the mean $z=p(n-1)$
- Starting from a given node, how many nodes can we reach in ℓ steps?



At step 1, reach z nodes

then, reach $z(z-1)$ new nodes

then, reach $z(z-1)^2$ new nodes

…

the number of new nodes reached grows exponentially with steps

Erdős-Rényi diameter

- After ℓ steps, we have reached a total of

  $z + z(z-1) + z(z-1)^2 + \ldots + z(z-1)^{\ell-1}$
- nodes, which is

  $z((z-1)^{\ell} - 1) / (z-2)$
- which is roughly $(z-1)^{\ell}$
- How many steps are required to reach $(n-1)$ nodes?

  $(z-1)^{\ell} = (n-1)$
- Solving for ℓ we conclude has to be on the order of $\log(n)/\log(z)$

Erdős-Rényi diameter

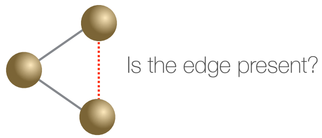
- The diameter will be roughly twice $\log(n)/\log(z)$
- Confirms the empirical data we observed in real networks
- Can be shown to hold for the general ER model without the strong assumptions
  - In reality, not all nodes have the same degree
  - In reality, not tree-structured (there could be backwards edges)
- Proof based on a weaker set of conditions
  - n large
  - $z \geq (1-\varepsilon)\log(n)$ for some $\varepsilon > 0$ (connected)
  - $z/n \to 0$ (but not too connected)

Erdős-Rényi model vs real networks Diameter

- The ER model is a good predictor of *diameter* and *average path length* compared to real networks
- The model results in networks with small diameters, capturing very well the "small-world" property observed in many real networks

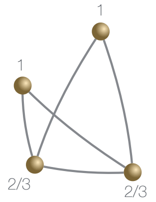
Erdős-Rényi clustering coefficient

- Recall *clustering coefficient* of a node: probability that two randomly selected *friends* of it are friends themselves



- In the ER model, an edge between any two nodes is present with probability p (independent of their context)
- So, the *clustering coefficient* of the ER random network is equal to p

Erdős-Rényi clustering coefficient



- Example: n=5, p=0.6
- CC=(0+1+1+2/3+2/3)/5=0.6667
- Compare with p which is 0.6

Erdős-Rényi clustering coefficient

- Recall *edge density* of a network: actual number of edges in proportion to the maximum possible number of edges
- In the ER model, on average, $pn(n-1)/2$ edges are added, thus $m=pn(n-1)/2$
- Edge density of ER network:

$$\rho = \frac{2m}{n(n-1)} = \frac{2(pn(n-1)/2)}{n(n-1)} = p$$

- Since the edge density is exactly equal to the background probability of triangles being closed, the networks produced by the ER model *cannot* be considered highly clustered

Erdős-Rényi model vs real networks Clustering coefficient

- The ER model is a poor predictor of *clustering* compared to real networks
- The model results in clustering coefficients that are too small and too close to the edge density
- Whereas most real networks are often highly clustered with clustering coefficients that are much greater (sometimes several orders of magnitude) than their edge densities

Erdős-Rényi giant component

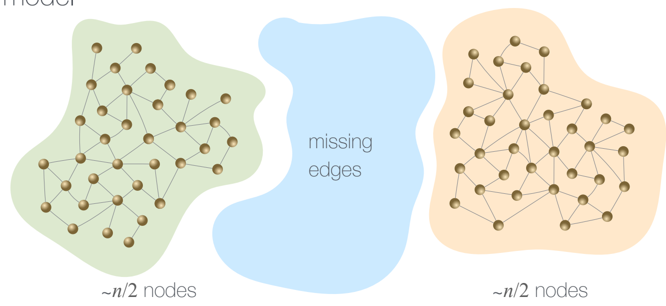
- Suppose we add edges randomly with probability p
- If p=0, no edges added, so edge density of the network is 0
- As p tends towards 1, the edge density tends towards 1
- In fact, for the ER model, edge density follows the edge probability exactly
- What structural properties are likely at a given density ρ?
- When do certain structures emerge as a function of ρ?
  - Many interesting properties occur at small densities
  - And they occur very suddenly (tipping points)

Erdős-Rényi giant component
Tipping point

- Note that at edge density ρ, the expected node degree is $\rho(n-1) \sim \rho n$ for large n
- Run the NetLogo *Library/Networks/GiantComponent* simulation
- In the ER model, giant components start forming at very low values of edge density
- For large n, we can show that
  - If $\rho < 1/n$, the probability of a giant component tends to 0
  - If $\rho > 1/n$, the probability of a giant component tends to 1 and all other components have size at most $\log(n)$
- At the tipping point $\rho = 1/n$, the average node degree is $\rho n = 1$
- Network is very sparse but ER uses edges very efficiently

Erdős-Rényi giant component
Tipping point

- Why is it very unlikely that two large components form?
- Run the NetLogo *ErdosRenyiTwoComponents* simulation
- Suppose two large components containing roughly half the nodes each do form in the ER model



Erdős-Rényi giant component
Tipping point

- How many potential edges are missing?
- The number of cross component edges is $\sim n/2 \times n/2 = n^2/4$
- Compare to the total number of possible edges: $n(n-1)/2$
- In other words, more than half of all possible edges are missing
- Selecting a new edge to add that is *not* one of the missing “cross edges” becomes increasingly more unlikely
- Imagine enrolling 10,000 friends to Facebook asking them to keep their friendships strictly among themselves
- Impossible to maintain since all it takes is just one of the 10,000 to make one external friendship

Erdős-Rényi giant component
Tipping point

- In those rare cases where two giant components have co-existed for a long time, their merger is sudden and often dramatic
- Imagine the arrival of the first Europeans in the Americas some 500 years ago
- Until then, the global socio-economic-technological network likely consisted of two giant components — one for the Americas, another for Europe-Asia
- In the two components, not only technology, but also human diseases developed independently
- When they came in contact, the results were disastrous

Erdős-Rényi diameter
Tipping point

- In the ER model, emergence of small diameter is also sudden and has a tipping point
- For large n, we can show that
  - If $\rho < n^{-5/6}$, the probability of the network having diameter 6 or less tends to 0
  - If $\rho > n^{-5/6}$, the probability of the network having diameter 6 or less tends to 1
- For the US, n=300M and the tipping point is $\rho n \sim 25.8$
- For the world, n=7B and the tipping point is $\rho n \sim 43.7$

Erdős-Rényi
Other tipping points

- In fact, we can prove a much more general result
- In the ER model, any *monotone* property of the network has a tipping point
- In networks, a property is *monotone* if it continues to hold as we add more edges to the network
- Examples of monotone properties:
  - The network has a giant component
  - The diameter of the network is at most k
  - The network contains a cycle of length at most k
  - The network contains at most k isolated nodes
  - The network contains at least k triangles

Erdős-Rényi
Summary

- The ER model *is* able explain
  - Small diameter, path lengths
  - Giant components
- The ER model *is not* able explain
  - Degree distributions
  - Clustering