The Beginning of Graph Theory...

Can you take a walk around old Koenigsberg such that you
- Pass through each of the 7 bridges exactly once and
- End up where you started?

- Abstraction with nodes (or vertices) and edges (or arcs)
- The answer is no (Euler 1736) – “A Eulerian cycle does not exist”
Some Favorite Graphs

- Complete graphs or cliques
- Bipartite graphs
- Lattice graphs

Some favorite problems:
- Eulerian/Hamiltonian cycles/paths
- Chromatic number
- Graph/subgraph isomorphism

Some characteristics:
- Small, finite graphs
- Regular structure
- Combinatorial in approach
Small, regular graphs are fine until things get more complex...

How to describe such large (→ infinite), irregular, seemingly random structures?

- Metabolic and protein interaction networks
- Internet and WWW
- Social networks
Random Graphs and the ER Model

Erdös and Rényi first studied random graphs in the late 1950s, using probabilistic methods to derive large-scale, statistical properties of random graphs.

Construction:

- Start with N nodes
- Connect each possible edge with probability p
- And you get a random graph!
Some interesting features to look at...

Consider an ER random graph with \( N \) nodes and connection probability \( p \):

**Degree** = the number of edges (or neighbors) a node has

What's the **average degree** of the graph?

\[
<k> = \frac{2E}{N} = \frac{2{N \choose 2}p}{N} = (N-1)p
\]

What's the probability that **a node has degree** \( k \)?

\[
P_i(k) = \left( \frac{N - 1}{k} \right) p^k (1 - p)^{N - 1 - k}, \quad \text{Binomial}
\]

How many nodes have a given degree \( k \)? (**degree distribution**)

\[
P(k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad \text{where} \quad \lambda = P_i(k) = \left( \frac{N - 1}{k} \right) p^k (1 - p)^{N - 1 - k}, \quad \text{Poisson}
\]
Some more network parameters...

**Degree** = number of neighbors
Average degree and degree distribution

**Clustering Coefficient** = \( m / \binom{k}{2} \)
- Are neighbors more likely to interact? (local density)
- What’s the CC of a random graph?

**Characteristic path length** \( L \):  
- Shortest path between a pair of nodes  
- Average over all pairs  
- \( L \) is short for random graphs \( \sim \ln(N) / \ln(k) \)

**Betweenness** and **Closeness**

**Assortativity** (or degree correlation)

**Intuitive understanding!**
Think of examples!
Random Graphs and the Erdös-Rényi model

- **Construction**
  - Start with $N$ nodes ($>>1$)
  - Connect each pair with probability $p$ ($<<1$)

- **Properties**
  - Node degree $k$ follows Poisson distribution
  - Short average path length
  - Low clustering coefficient ($=p$)

![Graph Example](image)

- $N = 10$
- $p = 0.2$
- $<k> = 1.8$

![Poisson Distribution Graph](image)
Random graphs are useful, but...

- Are real-world complex networks really random?
- What are the organizing principles behind such networks?
- How could such networks have evolved?

If you have two friends, are they more likely to know each other?
High CC, locally dense

How far are you separated from your celebrity of choice on Facebook?
L is short, small-world

Do you have a fixed social circle, or (hopefully!) new people join? Do people ever leave?
Networks grow (or shrink) over time, N is not fixed

Would you rather make friends with someone who is already popular?
Preferential attachment, connection probability \( p \) is not uniform

You and Bill Clinton, whose friends are more likely to know each other?
CC might depend on \( k \)!
“Small-World” Networks

- Start with a regular ring lattice (each vertex connected to its k nearest neighbors)
- Randomly rewire each edge with probability p (in this example stops after 2 circles)

Predict the effect of the first few rewires:
- Big effect on CC? On L?
- Suppose you met your future husband/wife while on vacation abroad...
A few short-cuts are enough to make it “small-world”
Real-World Examples

Table 1 Empirical examples of small-world networks

<table>
<thead>
<tr>
<th>Network</th>
<th>$L_{\text{actual}}$</th>
<th>$L_{\text{random}}$</th>
<th>$C_{\text{actual}}$</th>
<th>$C_{\text{random}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film actors</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
</tr>
<tr>
<td>Power grid</td>
<td>18.7</td>
<td>12.4</td>
<td>0.080</td>
<td>0.005</td>
</tr>
<tr>
<td>C. elegans</td>
<td>2.65</td>
<td>2.25</td>
<td>0.28</td>
<td>0.05</td>
</tr>
</tbody>
</table>

$L \sim L_{\text{ran}}, C_{\text{ran}} \gg C_{\text{ran}}$

Effect of small-world

Spread of infectious disease (figures familiar?!)
“Small-world” focuses on L (and to a lesser extent CC): The effect of long-range short-cuts

Now we look at another topological parameter: Node degree and degree distribution

Some historical perspectives:
- Most complex networks emerged only recently (Internet, WWW, genomics, etc.)
- Even for "older" networks (e.g. social), data collection became possible only recently
- Complex networks had been modeled on random graphs – for lack of data!

For many complex networks:
- Most nodes have few links
- A few nodes have many links (so-called “hubs”) – think of the above examples!
- But how abundant are those hubs?

More precisely, what's the probability $P(k)$ that a node has $k$ neighbors?
- Both the ER (random) and WS (small-world) models predict exponential decay: You basically don't see any hubs!
- Is this true? Think of the above examples.
Instead of exponential decay, we have power-law decay!
Such networks have been termed scale-free

Fig. 1. The distribution function of connectivities for various large networks. (A) Actor collaboration graph with $N = 212,250$ vertices and average connectivity $\langle k \rangle = 28.78$. (B) WWW, $N = 325,729$, $\langle k \rangle = 5.46$ (6). (C) Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{\text{actor}} = 2.3$, (B) $\gamma_{\text{WWW}} = 2.1$ and (C) $\gamma_{\text{power}} = 4$.

Collection of data is the huge first step!
After observation comes modeling

ER and WS fail to predict power-law degree distribution:

- What's missing in those models?
- Do real networks come out of nowhere?
- No, they grow gradually. → ER and WS start with a fixed number of nodes
- How do they grow? Each edge with equal probability? Rewiring?

Key features to incorporate into a new model:

- Growth (continuous addition of new nodes)
- Preferential attachment (new nodes more likely to connect to existing hubs)

Again, think of those real-world examples!

Once you have a model, it's time to

- Run simulations – do they produce the desired outcome (power-law)?
- Fine-tune your models – are current features sufficient/necessary/improvable?
- Analyze your model (i.e. math!)
Simulation steps:
- Start with some initial nodes (m₀)
- At every time step add a new node with m edges (m ≤ m₀)
- For each of those m new edges, an existing node’s probability of receiving that edge corresponds to its own degree (as a fraction of the total degree) before this time step

Model produces power-law degree distribution
Both “growth” and “preferential attachment” are necessary features
P(k) does not depend on time or system size (hence “scale-free”)

Fig. 2. (A) The power-law connectivity distribution at t = 150,000 (⊙) and t = 200,000 (□) as obtained from the model, using m₀ = m = 5. The slope of the dashed line is γ = 2.9. (B) The exponential connectivity distribution for model A, in the case of m₀ = m = 1 (⊙), m₀ = m = 3 (□), m₀ = m = 5 (◇), and m₀ = m = 7 (△). (C) Time evolution of the connectivity for two vertices added to the system at t₁ = 5 and t₂ = 95. The dashed line has slope 0.5.
Consequences of the model – “rich gets richer”

Math of the model – you can actually solve for the power coefficient!

Let $k_i(t)$ be the degree of node $i$ at time $t$. Then the rate of change of $k_i$ is

$$\frac{\partial k_i}{\partial t} = m \Pi(k_i) = m \sum_i k_i = m \frac{k_i}{2mt} = k_i \frac{1}{2t}$$

Suppose node $i$ was added at time $t_i$, so $k_i(t_i) = m$. This is the initial condition for the above first-order ODE.

$$k_i(t) = m \sqrt{\frac{t}{t_i}}$$

To calculate $P(k)$, we have

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{\partial}{\partial k} \left( m \sqrt{\frac{t}{t_i}} < k \right) = \frac{\partial}{\partial k} \left( t_i > \frac{m^2 t}{k^2} \right) = 1 - \frac{\partial}{\partial k} \left( t_i \leq \frac{m^2 t}{k^2} \right)$$

$P(t_i)$ follows the uniform distribution with height $1 / (m_0 + t)$. Thus

$$P \left( t_i \leq \frac{m^2 t}{k^2} \right) = \frac{m^2 t}{k^2 (m_0 + t)}$$

Combining the two, we obtain

$$P(k) = \frac{2m^2 t}{m_0 + t} k^{-3}$$

For large $t$, $t / (m_0+t) \to 1$, so $P(k) = 2m^2 / k^3$, the power coefficient being 3.
Scale-free implies hubs are common, but why do hubs matter?

Lethality and Centrality
Most biological networks known to date are small-world and scale-free

Interactomes:

Yeast (Nature 2000)
Fly (Science 2003)
Worm (Science 2004)
Human (Nature 2005)