Systems Biology Across Scales: A Personal View
VII. Networks: Models I

Sitabhra Sinha
IMSc Chennai
Theoretical understanding of networks

- Regular lattice or grid (*Physics*)
  - average path length $\sim N$ (no. of nodes)
  - clustering *high*
  - delta function distribution of degree (links/node)

- Random networks (*Graph theory*)
  - Also known as Erdos-Renyi networks
  - average path length $\sim \log N$
  - clustering *low*
  - Poisson distribution of degree
Random networks

Erdos-Renyi model (1959): Two closely related probability-based models for generating random networks

Behavior of random networks is typically studied in the limit where the number of nodes $N \to \infty$

- The $G(N,L)$ model: when any member of a family of all graphs with $N$ nodes and $L$ links is chosen uniformly at random.
  Example: $G(3,2)$ comprises three possible networks of three nodes A, B and C such that each of Graph 1: {AB, BC}, Graph 2: {AC, AB} and Graph 3: {AC, BC} can be picked with probability $1/3$

- The $G(N,p)$ model: when a network is constructed by randomly placing a link between each possible pair of nodes with a probability $p$ (0 < $p$ < 1)
  Example: $G(3,0.5)$ comprises the ensemble of all possible networks of three nodes A, B and C such that each of {AB, BC, AC} are inserted or not, based on tossing a fair coin

As $N \to \infty$, if $p \geq 2\ln(N)/N$ then a network will almost surely be connected.
Percolation & Random networks

A largest connected component (LCC, also referred to as giant component) is a connected component (for directed networks, strongly connected) whose size $N_l$ is a finite fraction of that of the size $N$ of the entire network, even as the network becomes larger and larger, i.e., $\lim_{N \to \infty} N_l / N = c > 0$.

In the $G(N,p)$ random network model, the LCC size is $N_l = 1$ when $p = 0$ (no nodes have any links) and $N_l = N$ when $p = 1$ (clique).

As $p$ is gradually increased from 0 to 1, a phase transition occurs: the fraction $N_l / N$ suddenly increases from 0 to a finite value ($>0$) at the critical value of $p$, $p_c = 1/N$.

This concept is related to the theory of bond percolation.

The Question: Consider a 2-dimensional lattice of $N \times N$ sites in which the links between any two neighboring sites is open with probability $p$ [and hence, closed with prob $(1 - p)$]. What is the probability that a connected path exists from one side of the lattice to the other?

[Percolation theory has origins in understanding the process of transport through porous medium, e.g., of toxic chemical molecules through the filtering agent of a gas mask]
The average path length in the random network is \( \langle L \rangle \sim \log \langle N \rangle / \log \langle k \rangle \)

Intuition:
Locally, a random network \( G(N,p) \) with very small \( p \) – as cycles or closed loops involving only a few nodes are unlikely – will be approximately like a tree

The average number of neighbors located at distance \( d \) away from a node is:
\( N_d = \langle k \rangle^d \)
\[ \Rightarrow N = \langle k \rangle + \langle k \rangle^2 + \langle k \rangle^3 + \ldots + \langle k \rangle^d \sim \langle k \rangle^d \]

The average clustering coefficient in a \( G(N,p) \) random network is approximately \( \langle C \rangle \sim p \approx \langle k \rangle / N \)
This is because if you randomly select a node \( i \) and look at two neighboring nodes \( j \) and \( k \) connected to it, the probability that \( j \) & \( k \) will be connected is just \( p \)
Degree distribution of Random networks

The G(N,p) model:

A given node in the network is connected with independent probability $p$ to each of the $N - 1$ other nodes.

Thus the probability of being connected to $k$ (and only $k$) other nodes is $p^k(1 - p)^{N - 1 - k}$

There are $N - 1 \binom{k}{k}$ ways to choose those $k$ other vertices, and hence the total probability of being connected to exactly $k$ others is $p_k = \frac{N - 1}{k} p^k(1 - p)^{N - 1 - k}$

which is the Binomial distribution having mean $Np$ and variance $Np(1 - p)$

As $N$ becomes large with $p$ being extremely small ($\rightarrow 0$), such that $Np = \langle k \rangle = \lambda$ is finite, this tends to the Poisson distribution

$P(k) = e^{-\lambda} \left( \frac{\lambda^k}{k!} \right)$

Both the mean and variance is given by $\lambda$.

For large values of $\lambda$ this converges to the bell-shaped Gaussian or Normal distribution
Empirical networks are not random – many have certain structural patterns.

Theoretical understanding of networks

- **Regular lattice or grid (Physics)**
  - average path length $\sim N$ (no. of nodes)
  - clustering high
  - delta function distribution of degree (links/node)

- **Random networks (Graph theory)**
  Also known as Erdos-Renyi networks
  - average path length $\sim \log N$
  - clustering low
  - Poisson distribution of degree
Example: *small-world* networks

Regular Network  
\( p = 0 \)  
High clustering, Large path length

“Small-world” Network  
\( 0 < p < 1 \)  
High clustering, Short path length

Random Network  
\( p = 1 \)  
Low clustering, Short path length

Increasing Randomness

\( p: \text{fraction of random, long-range connections} \)

Watts and Strogatz (1998): Many biological, technological and social networks have connection topologies that lie between the two extremes of completely regular and completely random.
It's a small world: The Milgram Experiment

Stanley Milgram (1933-1984), US social psychologist

Arbitrarily selected individuals in Nebraska were asked to generate acquaintance chains (knowing on first name basis) connecting them to a target individual in Boston.

In one experiment, 64 of the 296 chains initiated eventually reached the target – the mean number of intermediaries between source and target being slightly larger than 5.

⇒ Six degrees of separation

Milgram, 1967
“Small world”: Local properties of regular networks but global properties of random networks.

Small-world networks can be highly clustered (like regular networks), yet have small characteristic path lengths (as in random networks).

Characteristic path length:

\[ L = \frac{1}{2n(n+1)} \sum_{i \geq j} d_{ij} \]

where \( d_{ij} \) is the shortest distance between nodes i and j.

Alternatively:

\[ \ell^{-1} = \frac{1}{2n(n+1)} \sum_{i \geq j} d_{ij}^{-1} \]

Clustering coefficient:

\[ C = \frac{1}{n} \sum_{i} C_i \]

where

\[ C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i} \]

Alternatively:

\[ C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \]
Epidemics on “Small world”

Dynamical process:
- Time $t = 0$: single infected individual present.
- Each infected agent can infect any of its neighbours with probability $r$.
- Infected individuals removed (by immunity or death) after unit period of sickness.

Key Results:
- Critical infectiousness $r_{\text{half}}$, at which the disease infects half the population, decreases with $p$.
- Time required for a maximally infectious disease ($r = 1$) to spread throughout the entire population $T(p)$ has same form as characteristic path length $L(p)$.
  $\Rightarrow$ rewiring only a few links in the original lattice causes global infection to occur almost as fast as in random network.

Implication:
“Control the truck-drivers”
### Do small-world networks occur in real life?

<table>
<thead>
<tr>
<th>Network</th>
<th>Size</th>
<th>⟨k⟩</th>
<th>ℓ</th>
<th>ℓ_{rand}</th>
<th>C</th>
<th>C_{rand}</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW, site level, undir.</td>
<td>153 127</td>
<td>35.21</td>
<td>3.1</td>
<td>3.35</td>
<td>0.1078</td>
<td>0.00023</td>
</tr>
<tr>
<td>Internet, domain level</td>
<td>3015–6209</td>
<td>3.52–4.11</td>
<td>3.7–3.76</td>
<td>6.36–6.18</td>
<td>0.18–0.3</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Movie actors</strong></td>
<td>225 226</td>
<td>61</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
</tr>
<tr>
<td>LANL co-authorship</td>
<td>52 909</td>
<td>9.7</td>
<td>5.9</td>
<td>4.79</td>
<td>0.43</td>
<td>1.8×10^{-4}</td>
</tr>
<tr>
<td>MEDLINE co-authorship</td>
<td>1 520 251</td>
<td>18.1</td>
<td>4.6</td>
<td>4.91</td>
<td>0.066</td>
<td>1.1×10^{-5}</td>
</tr>
<tr>
<td>SPIRES co-authorship</td>
<td>56 627</td>
<td>173</td>
<td>4.0</td>
<td>2.12</td>
<td>0.726</td>
<td>0.003</td>
</tr>
<tr>
<td>NCSTRL co-authorship</td>
<td>11 994</td>
<td>3.59</td>
<td>9.7</td>
<td>7.34</td>
<td>0.496</td>
<td>3×10^{-4}</td>
</tr>
<tr>
<td>Math. co-authorship</td>
<td>70 975</td>
<td>3.9</td>
<td>9.5</td>
<td>8.2</td>
<td>0.59</td>
<td>5.4×10^{-5}</td>
</tr>
<tr>
<td>Neurosci. co-authorship</td>
<td>209 293</td>
<td>11.5</td>
<td>6</td>
<td>5.01</td>
<td>0.76</td>
<td>5.5×10^{-5}</td>
</tr>
<tr>
<td><em>E. coli</em>, substrate graph</td>
<td>282</td>
<td>7.35</td>
<td>2.9</td>
<td>3.04</td>
<td>0.32</td>
<td>0.026</td>
</tr>
<tr>
<td><em>E. coli</em>, reaction graph</td>
<td>315</td>
<td>28.3</td>
<td>2.62</td>
<td>1.98</td>
<td>0.59</td>
<td>0.09</td>
</tr>
<tr>
<td>Ythan estuary food web</td>
<td>134</td>
<td>8.7</td>
<td>2.43</td>
<td>2.26</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td>Silwood Park food web</td>
<td>154</td>
<td>4.75</td>
<td>3.40</td>
<td>3.23</td>
<td>0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>Words, co-occurrence</td>
<td>460.902</td>
<td>70.13</td>
<td>2.67</td>
<td>3.03</td>
<td>0.437</td>
<td>0.0001</td>
</tr>
<tr>
<td>Words, synonyms</td>
<td>22 311</td>
<td>13.48</td>
<td>4.5</td>
<td>3.84</td>
<td>0.7</td>
<td>0.0006</td>
</tr>
<tr>
<td><strong>Power grid</strong></td>
<td>4941</td>
<td>2.67</td>
<td>18.7</td>
<td>12.4</td>
<td>0.08</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>C. Elegans</strong></td>
<td>282</td>
<td>14</td>
<td>2.65</td>
<td>2.25</td>
<td>0.28</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Albert & Barabasi, 2003