Networks and Random Graphs
Network Science and Erdos Legacy

Paris Siminelakis

School of Electrical and Computer Engineering at the National Technical University of Athens

7/2/2011
What are Networks?

- A network is a collection of points (nodes) and their connections (edges).
- It is a simplified representation that reduces a system to an abstract structure capturing only the basics of connection patterns.

**Figure**: A example of a small network with the local neighbourhood of a vertex highlighted
Why do we study them?

- There are many systems of interest that are composed of individual components linked together in some way (*Internet, Computer Networks, Societies, Organisms, Language*).

- Many aspects of these networks are worth of study but almost always crucial to the behaviour of the system is the pattern of connection between components. Exactly these patterns are modelled with networks.

- Through the years scientists have developed a wide variety of tools—mathematical, computational, statistical—for analysing, modelling and understanding networks.

- Thus, due to the ubiquity of networks advancing our knowledge of their properties translates to a deeper understanding of many real world phenomena.
What is Network Science?

Network Science lies in the intersection of Statistics, Computer Science, Physics and Sociology. It aims in:

- Providing **measures** that encapsulate the **structural properties of networks** as well as the **role(importance) of individual components** in the network.

- Explaining and understanding the **evolution pattern** of networks by considering different **Generative Models** of networks (Preferential Attachment, Network Optimization models etc).

- Studying **processes on networks** such as epidemics, rumor spreading, searching, routing, voting and synchronization of dynamical systems.

- Providing **empirical data and evidence**.

Main tools are: **Graph Theory, Random Graphs, Spectral Graph Theory, Probability-Combinatorics**.
**Figure:** Structure of the Internet at the level of autonomous systems. The map was constructed using traceroute data. The vertices are AS and the edges show the routes taken by data travelling between them. Copyright Lumeta Corporation.
Figure: The vertices are "Class C subnets" - groups of computers with similar internet addresses - and the connections between them represent the routes taken by Internet data packets as they hop between subnets. Colours: Asia Pacific, Europe/Middle East/Central Asia/Africa, North America, Latin American and Caribbean. Copyright The Opte Project.
Figure: Characteristics of the yeast proteome. Map of protein protein interactions. The largest cluster, which contains 78% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein (lethal, nonlethal, slow growth, unknown). Copyright Macmillan Publishers Ltd. From Jeong et al. Lethality and Centrality in protein Networks.
Figure: Protein Interactions and signalling network in Liver Cells. Map was constructed using data obtained from the high throughput Luminex technology and optimized by an ILP formulation. From Mitsos et. al Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data
Figure: Network of *romantic adolescent relationships* in Jefferson High. The data are from 800 students in a period of 18 months. From Bearman et. al *Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks*
Early Contributors

- **Karinthy** (1929): Hungarian writer who claims, in a short story entitled *Chains*, that people are increasingly connected to each other via their acquaintances and the dense web of friendship leads to a world where everyone on Earth is at most five acquaintances away from anyone else.

- **Rapoport** (1951): Russian mathematician who worked in mathematical biology, introduced for the first time the notion of a random graph and demonstrates the phenomenon of the *giant component*.

- **Erdos-Renyi** (1959-1968): they are considered the fathers of modern theory of random graphs and during these years they published 8 papers that set the tone for network research for many decades.

- **Pool-Kochen** (1958): political scientist and mathematican, their paper discusses for the first time in scientific terms the *small world* effect. They base their work on the random graph inspired by Rapoport and ask questions that set the mark in sociology for the years to come.

- **Travers-Milgram** (1969): experimental social scientist working in Harvard designed an experiment to test Pool-Kochen ideas and introduced the notion of *six degrees of separation*. 
Recent Contributors

- **Price** (1965, 1976): in his paper *Network of Scientific Papers* which appeared in *Science* price studies the network of scientific citations and observed power laws in both the in and out-degrees which now are known to occur in a variety of different settings and are known as scale-free networks. A decade later Price proposed a possible mechanism for the observed distribution now known as the Rich get Richer Phenomenon or Preferential Attachment mechanism.

- **Watts-Strogatz** (1998): American mathematician known for his work in the synchronization of coupled oscillators along with his student discovered the small world phenomenon by considering regular graphs with some randomness while studying the almost instantaneous synchronization of fireflies.
Recent Contributors

- **Barabasi-Albert** (1999): Hungarian physicist that rediscovered the preferential attachment mechanism and shown that it results in power laws. They have performed many empirical studies concerning the Internet, Biological Networks, Mobile Networks and Human Mobility Patterns.

- **Kleinberg** (2000): American Computer Scientist who considered the algorithmic question that the small world effect implies that there must be a way (algorithm) that individuals search this vast network with very little information. He answered the question by considering regular lattices with some randomness in correspondence with the Watts-Strogatz model.

- **Newman** (2001): introduced the configuration model in the society and explained many aspects of the small world effect along with other contributions in Community Detection, Percolation and Clustering of Complex Networks.

Other important contributors are *Vespignani, Pastor-Satorras, Dorogotsev, Mendez, Krapivsky, Bianconi, Amaral.*
A quantity $X$ follows a power law with exponent $a$ if the fraction of units that have size $x$ is:

$$P(X = x) = C(a) \cdot x^{-a}$$

Power laws have been observed in a wide variety of phenomena such as: the Internet, city populations, river channels, frequency of use of words, book sales, number of citations papers receive, peoples annual income...
The crucial property of such power laws is that they allow the occurrence of extreme events, e.g. a finite number of persons accumulate a finite fraction of the total wealth or a finite number of web-pages accumulate a finite fraction of the links.

Another key aspect is that it is scale invariant. For example the "80-20" rule (Pareto Principle). That 20% have 80% of the wealth, but also that between those 20%, 80% of the total wealth goes to the top 20% of them as well. This is the notion of self-similarity.

Most importantly power laws usually rise when there is a latent feedback process, which is popularly coined as the "Rich-Get-Richer" phenomenon. This effectively generalizes the notion of feedback introduced by Norbert Wiener in his book Cybernetics.
Wealth Distribution

- One of the issues that we are facing are the increasing economic inequalities. Numerous studies have shown that the wealth distribution among individuals follows a power law, allowing for wealth condensation among few individuals. In order to remove these inequalities we must first understand their origin.

- Bouchaud-Mezard (2000) in their work "Wealth condensation in a simple model of economy" considered a simplified model of the economy. In their model every agent can:
  - Invest an amount of his wealth (stock market, housing etc.)
  - Exchange money with other agents (food, cars, clothes etc.).
  - The amount returned by the investment is a gaussian random variable with same parameters for all agents, making them effectively equal in their investment potential.
  - They also assumed a constant rate of exchange $J$ between all agents.
Their model boils down to this stochastic differential equation:

\[
\frac{dW_i(t)}{dt} = \eta_i(t)W_i + J(W - W_i)
\]

The equilibrium long time solution of the above equation exhibits power-law tails for large \( W \) with exponent \( a = 1 + \mu \). They also observed that for \( \mu = 1 \) there is a phase transition where for \( \mu < 1 \) wealth condensation occurs.

Figure: Simulation for \( \mu = 3 \)

Figure: Simulation for \( \mu = 0.5 \)
Reducing Inequalities

The authors also augmented the model with capital and income taxes. They observed that:

- **Favouring exchanges** between individuals allows faster diffusion of Wealth and thus less inequalities.

- **Income taxes** tend to reduce inequalities, even more so if part of this tax is redistributed evenly.

- **Capital taxes** if used simultaneously with incomes taxes and not-adequately redistributed, result in an increase of the inequalities.
Residential Segregation splits communities between the black (ghettos) and whites (rich suburb areas). It is believed to be the main source of poverty among African Americans. It is attributed to White flight, where Whites actively leave. Consequences:

- Money leaves, Poverty Grows, Crime rates jump.
- Businesses leave following the money, Job Shortage.
- Tax base thinning. Less funding for education.

The above render ”black” regions poverty traps that blacks can’t escape from. But is this really why there is segregation?
T.C. Schelling in 1971 (Nobel prize in Economics 2005) conducted a computer experiment simulating the segregation process.

- He considered a flat world and randomly distributed black and white people.
- The residents of the world moved by the simple rule: If more than $R$ (e.g. 50%) percent of my neighbours are different than me then I would prefer a neighbourhood where there at least $R$ of the residents are like me.

Note that there is no negative emotion towards people of different race, rather a slight preference for neighbourhoods with a fair presence of one’s race as well.
Figure: Initial distribution of residents in a urban area. Each triangle corresponds to a person and the colours correspond to different races. Note that the colours are evenly distributed in the residential area.
Figure: Final state of the simulation process when everyone is satisfied by his position. We observe that although neither race has any sort of disliking against the other, due to the slight preference that they have towards people of the same race the city has been partitioned in blocks of whites and blacks. We show results for two homophily percentages.
**Figure:** Final state of the simulation process when everyone is satisfied by his position. We observe that although neither race has any sort of disliking against the other, due to the slight preference that they have towards people of the same race the city has been partitioned in blocks of whites and blacks. We show results for two homophily percentages.
Random Graphs

- In general a random graph is a model network where some specific set of parameters take fixed values, but is random in other respects.
- One of the simplest examples is the networks where we fix only the number of vertices $n$ and the number of edges $m$. Specifically, we choose $m$ pairs of vertices uniformly at random and connect them. This model is referred as $G(n, m)$.

Mathematically a random graph is defined as a probability distribution $P(G)$ over all graphs $G$ that belong to a class $G$ which satisfies the specific constraints - parameters. For example in the case of $G(n, m)$ we have:

$$P(G) = \frac{1}{|G_{n,m}|} = \frac{1}{\binom{n^2}{m}}$$

Since we defined $G(n, m)$ as the class where each (multi)graph with $m$ edges has the same probability(uniform), and there are $\binom{n^2}{m}$ such graphs.
In general a random graph is a model network where some specific set of parameters take fixed values, but is random in other respects.

One of the simplest examples is the networks where we fix only the number of vertices \( n \) and the number of edges \( m \). Specifically, we choose \( m \) pairs of vertices uniformly at random and connect them. This model is referred as \( G(n, m) \).

Mathematically a random graph is defined as a probability distribution \( P(G) \) over all graphs \( G \) that belong to a class \( G \) which satisfies the specific constraints - parameters. For example in the case of \( G(n, m) \) we have:

\[
P(G) = \frac{1}{|G_{n,m}|} = \frac{1}{\binom{n^2}{m}}
\]

Since we defined \( G(n, m) \) as the class where each (multi)graph with \( m \) edges has the same probability(uniform), and there are \( \binom{n^2}{m} \) such graphs.
Erdos-Renyi’s $G_{n,p}$

An other slightly different model that is considerably easier to handle is the $G_{n,p}$. In this model we don’t fix the number of edges but we place each edge (of the complete graph) independently with probability $p$.

- Each graph with $m$ edges appears with probability:

$$\mathbb{P}(G) = p^m (1 - p)^{\binom{n}{2} - m}$$

- The total probability of having a graph with $m$ edges is:

$$\mathbb{P}(m) = \binom{n}{2}^m p^m (1 - p)^{\binom{n}{2} - m}$$

which is the standard binomial distribution.

- The mean value of $m$ is:

$$\langle m \rangle = \sum_{m=0}^{\binom{n}{2}} m \cdot P(m) = \binom{n}{2} p.$$
Degree Distribution

- Using above results we can calculate the mean degree:

\[ c = \langle k \rangle = \sum_{m=0}^{n} \frac{2m}{n} \cdot P(m) = \frac{2}{n} \binom{n}{2} p = (n - 1)p \]

- A given vertex is connected with probability \( p \) to each of the other \( n - 1 \) vertices. Thus the probability of being connected to a particular \( k \) other vertices is \( p^k (1 - p)^{n - 1 - k} \), but there are \( \binom{n-1}{k} \) ways to choose these \( k \) vertices. So, the probability of having exactly \( k \) connections is:

\[ p_k = \binom{n-1}{k} p^k (1 - p)^{n - 1 - k} \]

- Which is the binomial distribution. Using a first order approximation of \( \ln(1 + x) \) and approximating \( \binom{n-1}{k} \) by \( \frac{(n-1)^k}{k!} \) we get in the limit of large \( n \):

\[ p_k = \frac{(n - 1)^k}{k!} p^k e^{-c} = \frac{(n - 1)^k}{k!} \left( \frac{c}{n - 1} \right)^k e^{-c} = e^{-c} \frac{c^k}{k!} \]

- Thus, \( G(n, p) \) in the limit of large \( n \) has Poisson Degree Distribution. This is why sometime \( G(n, p) \) is referred to as Poisson Random Graph.
Giant Component

- Consider the case of $G(n, p)$ when $p = 0$ and $p = 1$. In the first case there is no edge in the network and all vertices are disconnected. In the later case all edges are present in the network. So, when $p = 0$ the size of the largest component is 1, whereas when $p = 1$ it is $n$.

- The distinction between the two cases is an important one as in the first case the size of the largest component is independent of the size of the network, whereas in the second case it is proportional to $n$.

- In many applications of networks it is crucial that there is a component that fills most of the network. For example the Internet wouldn’t serve its purpose if the network didn’t connect most computers together.

- It is interesting to see how the transition between the two extremes occurs?
A network component whose size grows in proportion to $n$ we call a **Giant Component**. We will calculate exactly the *size of the giant component* in the limit of large $n$:

- Let $u$ be the average fraction of vertices in the random graph that don’t belong to the giant component.
- For a vertex not to belong to the giant component, its neighbours must also not belong to the giant component.
- Due to the uniformity of $G(n, p)$, $u$ is also the probability that any vertex doesn’t belong to the giant component. We have:

$$u = (1 - p + pu)^{n-1} = \left[1 - \frac{c}{n-1}(1 - u)\right]^{n-1} = e^{-c(1-u)}$$
But if $u$ is the fraction of vertices that doesn’t belong to the giant component, then $S = 1 - u$ is the fraction that belongs to the giant component. Then $S$ must satisfy:

$$S = 1 - e^{-cS}$$

This equation describes the size of the giant component as a fraction of the size of the network in the limit of large $n$ for any given value of the mean degree $c$. Although quite simple it doesn’t have analytic solution, thus we will solve it numerically.

Observe that $S = 0$ will always be a solution to the equation so we are wondering whether there is one non-trivial solution. Note that the function $f(S) = 1 - e^{-cS}$ is an increasing, concave function (it’s derivative is decreasing).
Because both lines $y = S, f(S)$ start from $(0, 0)$ we will have a second solution if $f'(0) > 1$, since $f$ is bound to intersect the line $y = S$. Thus, the transition between the two regimes takes place when:

$$f'(S)|_{S=0} = 1 \Rightarrow ce^{-cS}|_{S=0} = 1 \Rightarrow c = 1$$
Technically, we have shown that no giant component exists if $c < 1$ but not that there has to be a giant component if $c > 1!$. A heuristic argument for the existence of the giant component is:

- Imagine that somewhere in the graph there is a small set of connected vertices. We will divide the set into its core (vertices in the set) and its periphery $\mathcal{P}$ (vertices that connect to the core).

- Now, imagine adding progressively vertices from the periphery in the core. If we have $s$ vertices (core + periphery) then each time the new periphery $\mathcal{P}_{k+1}$ will be:

$$\mathcal{P}_{k+1}(n - s) = p(n - s) \cdot \mathcal{P}_k = c \frac{n - s}{n - 1} \mathcal{P}_k \approx c \cdot \mathcal{P}_k$$

- If we repeat this argument we see that the size of the periphery will increase exponentially if $c > 1$ or decrease if $c < 1$. If it grows exponentially, at some point the size of the component will occupy a significant fraction of the network and the Giant Component will rise.
Uniqueness of the Giant Component

We have shown that a giant component must rise if $c > 1$ that covers a significant fraction of the network, but what happens to the remaining fraction? **Can there be another giant component?**

- Assume that there are **two giant components** with $S_1n$, $S_2n$ number of vertices respectively. We will show that they **intersect with high probability**.

- For the two components to be separate we require that there is no edge between them. That none of the $S_1n \times S_2n = S_1 S_1n^2$ edges between them occur. The probability of that event is **exponentially small**:

$$\mathbb{P}(S_1 \neq S_2) = \left(1 - \frac{c}{n-1}\right)^{S_1 S_1n^2}$$

We have thus proven that there is a **unique giant component**.
Rigorous Proof

We will prove the existence of the phase transition for the rise of the giant component in a more rigorous way.

- Consider an arbitrary node in the network, e.g node 1. We will gradually explore the graph. The number of neighbours of node 1 follow Binomial distribution: $B(n - 1, \frac{c}{n-1})$

- Likewise, the number of unexplored neighbours of a node after $s$ vertices have been explored follows: $B(n - s, \frac{c}{n-1})$

- We will approximate the Erdos-Renyi process with a Branching Process, where the number of new vertices follow $B(n, \frac{c}{n})$...(Proof on board)
Sub-critical phase $c < 1$
Super-critical phase $c > 1$
Small Components
Tree-Like Structure

We have seen that there is a unique giant component with \( S \) fraction of the vertices. **What about the other \( 1 - S \) percent?** Since there is only one giant component the remaining vertices must be distributed in small components.

- Let \( \pi_s \) be the fraction of vertices that belong to a component with size \( s \), then: \( \sum_{s=0}^{\infty} \pi_s = 1 - S \)

- We want to calculate the component distribution \( \pi_s \). The **crucial insight** is that the small components are tree-like, that is the probability of a cycle is very small.

- Consider a component of size \( s \), there must be at least \( s - 1 \) edges to form the component. **If we add another edge a cycle will form.** There are \( \binom{s}{2} - (s - 1) = \frac{1}{2} (s - 1)(s - 2) \) possible edges.

- Assuming that \( s \) increases more slowly than \( \sqrt{n} \) then the probability tends to zero:

\[
P_s(Cycle) = p \cdot \frac{1}{2} (s - 1)(s - 2) \sim c \frac{o(n)}{n-1} \to 0
\]
The Subgraph Argument

We will calculate the probability $\pi_s$ as follows. Consider a vertex $i$ in a small component. Each of $i$’s edges leads to a separate subgraph. These subgraphs are not connected to one another, other via the vertex $i$, because then the whole component wouldn’t be a tree.
The Subgraph Argument

We will calculate the probability $\pi_s$ as follows. Consider a vertex $i$ in a small component. Each of $i$’s edges leads to a separate subgraph. These subgraphs are not connected to one another, other via the vertex $i$, because then the whole component wouldn’t be a tree.

Thus the size of the component is the sum of the sizes of the subgraphs reachable by $i$ plus 1: $s = \sum_{j=1}^{k} n_j + 1$.
The Subgraph Argument

We will calculate the probability $\pi_s$ as follows. Consider a vertex $i$ in a small component. Each of $i$’s edges leads to a separate subgraph. These subgraphs are not connected to one another, other via the vertex $i$, because then the whole component wouldn’t be a tree.

Thus the size of the component is the sum of the sizes of the subgraphs reachable by $i$ plus 1: $s = \sum_{j=1}^{k} n_j + 1$

Now if we consider a modified network, where $i$ has been removed, the new network has the same probability $p$ and $n - 1$ vertices, thus in the limit of large $n$ all average properties, such as component sizes will be indistinguishable.
The Subgraph Argument

We will calculate the probability $\pi_s$ as follows. Consider a vertex $i$ in a small component. Each of $i$’s edges leads to a separate subgraph. These subgraphs are not connected to one another, other via the vertex $i$, because then the whole component wouldn’t be a tree.

Thus the size of the component is the sum of the sizes of the subgraphs reachable by $i$ plus 1: $s = \sum_{j=1}^{k} n_j + 1$

Now if we consider a modified network, where $i$ has been removed, the new network has the same probability $p$ and $n - 1$ vertices, thus in the limit of large $n$ all average properties, such as component sizes will be indistinguishable.

In this modified network the subgraphs adjacent to $i$ will be now small separate components, and since all average properties are the same the probability that a neighbour of $i$ belongs to a small component with size $s_1$ is itself $\pi_{s_1}$. 
Generating Functions

- if $X$ is a discrete random variable taking values in the non-negative integers $\{0, 1, \ldots\}$ then the probability-generating function of $X$ is defined as:

$$g(z) = \mathbb{E}(z^X) = \sum_{x=0}^{\infty} p(x)z^x = x_0 + x_1 z + x_2 z^2 \ldots$$

where $p(x)$ is the probability mass function of $X$.

- The generating function encapsulates all of the information about the probability distribution in a single function. Given $G(z)$ we can recover the probabilities by differentiating:

$$x_s = \frac{1}{s!} \frac{d^s G(0)}{dz^s}$$

- Thus if we can calculate $G(z)$ explicitly or its derivatives at zero, we can recover the probabilities $x_s$. 

Generating Functions Approach

- We will utilize the power of Generating Functions to calculate the probability $\pi_s$ that a randomly chosen vertex belongs to a component of size $s$. Remember the previous setting before vertex $i$ is removed.

- The probability $P(s|k)$ that vertex $i$ belongs to a small component with size $s$, given its degree is $k$, is the probability that its $k$ neighbours belong to small components of size $s_1, s_2, \ldots s_k$ and that those sizes add up to $s - 1$:

$$P(s|k) = \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[ \prod_{j=1}^{k} \pi_{s_j} \right] \delta(s - 1, \sum_j s_j)$$

where $\delta(m, n)$ is the Kronecker delta.
Computing $g(z)$

Now in order to have an expression for $\pi_s$ we just average $P(s|k)$ over the degree distribution $p_k$:

$$
\pi_s = \sum_{k=0}^{\infty} p_k \cdot P(s|k) = e^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \sum_{s_k=1}^{\infty} \left[ \prod_{j=1}^{k} \pi_{s_j} \right] \delta(s - 1, \sum_j s_j)
$$

Substituting the above equation in the definition of $g(z)$ and after (some...) algebra we get:

$$
g(z) = z e^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \left[ \sum_{s=1}^{\infty} \pi_s z^s \right]^k = z e^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} [g(z)]^k
$$

$$
= z \cdot \exp \left[ c \cdot (g(z) - 1) \right]. \quad (1)
$$
We have managed to have a self-consistent equation for $h(z)$. There is a Lemma known as Lagrange Inversion Formula from Complex Analysis and the Theory of Generating functions that states:

**Lemma (Lagrange Inversion Formula)**

If $f(z) = z \cdot \phi(f(z))$ where $\phi$ is a known function which at $f = 0$ is finite, non-zero and differentiable then:

$$\pi_s = \frac{1}{s!} \left[ \frac{d^{s-1}}{df} \left[ \phi(f) \right]^s \right]_{f=0}$$

Using the above lemma with $f(z) \rightarrow g(z)$ and $\phi(f) \rightarrow e^{c(g-1)}$ then we get:

$$\pi_s = \frac{1}{s!} \left[ \frac{d^{s-1}}{dg} \left[ e^{c(g-1)} \right]^s \right]_{g=0} = \frac{e^{-sc}(sc)^{s-1}}{s!}$$
We have managed to calculate the complete distribution of the small components in the random graph. We can see that as the average degree $c$ rises all the probabilities drop exponentially.

One could argue that as the average degree rises, the component distribution would become increasingly right-skewed as now the components would be in aggregate larger. The truth is that this is not the case. The reason for that is that the extra connections that are now available in average to every node are likely to be with nodes in the giant component and thus the small components vanish.
Path Lengths

- We would like to know what is the longest distance in the Poisson Random Graph. Recall that we have seen that the average number of vertices $s$ steps away from a randomly chosen vertex in a random graph is $c^s$. If $c > 1$ this number grows exponentially, so we expect that for $s \approx \ln n$ every vertex will be within $s$ steps away from our starting point.

- The above argument is only approximate, as when the number of explored vertices becomes a non-zero fraction of the network then the number of vertices explored at the next step are smaller than $c$. Thus, we must ensure that the number of explored vertices are $o(n)$. 
The Intersection Argument

- One natural way to resolve this issue is to consider two different vertices \( i \) and \( j \). The average number of vertices \( s \) and \( t \) steps away from them respectively will be \( c^s \) and \( c^t \), as long as both these numbers are \( o(n) \).

- Consider the situation where we are \( s \) steps away from vertex \( i \) and \( t \) steps away from vertex \( j \). Consider a ball around each vertex \( i, j \) consisting of all the vertices within distance \( s \) and \( t \) respectively. Let the most distant nodes be the "surface" of the neighbourhood.

- Up until now the neighbourhoods where separate, that means there is no edge from inner layers to any vertex of the other ball, including the "surface". Thus, there may be only edges between the surfaces.
The Intersection Argument

- So, if there is no edge between the surfaces, that implies also that the shortest path between $i$ and $j$ must be longer than $s + t + 1$. Thus, the absence of an edge between the surfaces $S_i, S_j$ of $i$ and $j$ respectively is a necessary and sufficient condition for the shortest path to be greater than $\ell = s + t + 1$. Therefore: $P(d_{ij} > \ell) = P(e_p, q \notin E, \forall p \in S_i, q \in S_j)$.

- But there are on average $c^s \times c^t$ pairs of vertices between the surfaces. Each pair is connected with probability $p = c/(n - 1) \simeq c/n$, hence we get in the limit of large $n$:

$$P(d_{ij} > \ell) = \left(1 - \frac{c}{n}\right)^{c^{\ell-1}} \simeq \exp\left(-\frac{c^\ell}{n}\right)$$
The Diameter

- The diameter of the network is the smallest value of $\ell$ such that $P(d_{ij} > \ell) \to 0$. This will happen if: $c^\ell$ grows faster than $n$. The smallest value of $\ell$ that achieves such guarantee is:

$$c^\ell = an^{1+\epsilon} \Rightarrow \ell = \frac{\ln a}{\ln c} + \lim_{\epsilon \to 0} \frac{(1 + \epsilon)\ln n}{\ln c} = A + \frac{\ln n}{\ln c}$$

- Note that we can achieve this value of $c^\ell$ while requiring both neighbourhoods to have a $o(n)$ vertices and thus our estimates for the ”surface” are asymptotically correct.

- The Clustering Coefficient $C$ is the probability that two neighbours of a vertex are also neighbours of each other and is a measure of transitivity of the network. In $G_{n,p}$:

$$C = p = \frac{c}{n-1}$$
The idea of *six degrees of separation* says that you can connect any individual in the world with any other by a string of acquaintances. There was a quite popular game on the internet were you would try to connect every actor to Kevin Bacon with a shortest path of co-actors.

Mathematically *small worlds* in the language of networks mean that, the diameter of the network is proportional (or smaller than) to the logarithm of the number of nodes in the network:

\[ D \propto \log N \]
The Complete Picture

\[ \pi_s \]

\[ S \]

\[ d \approx \ln n \]

\[ p_k \]

\[ C > 1 ? \]
Erdos Legacy

- **Randomness ≠ Chaos** rather Randomness $\rightarrow$ Structure
- **Simple local** rules $\rightarrow$ Rich Global Behaviour
- **Okam’s razor** principle:
  
  *The simplest possible explanation is the right one.*

- **$G_{n,p}$ isn’t a realistic model** for real world networks because:
  - Poisson Degree Distribution $\neq$ Power Law
  - Low Transitivity and Clustering (small Clustering Coefficient)
  - Degrees uncorrelated

- **Need for ”simple”, analytically tractable models:**
  - Configuration Model
  - Preferential Attachment
  - Exponential Random Graphs
Thank You!

psiminelakis@gmail.com

”How can I link, with three, four, or at most five links of the chain, trivial, everyday things of life. How can I link one phenomenon to another? How can I join the relative and the ephemeral with steady, permanent things - how can I tie up the part with the whole”

Frigyes Karinthy 1929, Chains.
Further Reading

M.E.J Newman
*Mathematics: An Introduction.*

M.E.J Newman, A.L. Barabasi, D.J. Watts
*Structure and Dynamics of Networks*

Fan Chung, Linyuan Lu
*Complex Graphs and Networks*
CBMS Regional Conference Series in Mathematics, 107

Steven Strogatz
*Sync*
Hyperion Publishing

Albert-Laszlo Barabasi
*Linked*
Plume Publishing

Documentary from ABC about Network Science.