Small-World Phenomena

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu
How to characterize networks?

- Degree distribution $P(k)$
- Clustering Coefficient $C$
- Diameter (avg. shortest path length) $h$

How to model networks?

- **Erdös-Renyi Random Graph** [Erdös-Renyi, ‘60]

  - $G_{n,p}$: undirected graph on $n$ nodes where each edge $(u,v)$ appears independently with prob. $P$
  - **Degree distribution**: Binomial$(n, p)$
  - **Clustering coefficient**: $C \approx \frac{k}{n}$
  - Diameter: (next)
Random k-Regular Graphs

- Assume each node has $d$ spokes (half-edges):
  - $k=1$: Graph is a set of pairs
  - $k=2$: Graph is a set of cycles
  - $k=3$: Arbitrarily complicated graphs
- Randomly pair them up
Definition: Expansion

- Graph $G(V, E)$ has **expansion** $\alpha$: if $\forall S \subseteq V$:
  
  $\# \text{ of edges leaving } S \geq \alpha \cdot \min(|S|, |V \setminus S|)$

- Or equivalently:

  \[
  \alpha = \min_{S \subseteq V} \frac{\# \text{edges leaving } S}{\min(|S|, |V \setminus S|)}
  \]
Expansion: Intuition

(A big) graph with “good” expansion
Expansion is a measure of robustness:

- To disconnect $l$ nodes, we need to cut $\geq \alpha \cdot l$ edges.

Low expansion:

High expansion:

Social networks:

"Communities"
**Expansion: k-Regular Graphs**

- **k-regular graph** (every node has degree $k$):
  - Expansion is at most $k$ (when $S$ is 1 node)

- Is there a graph on $n$ nodes ($n \to \infty$), of fixed max deg. $k$, so that expansion $\alpha$ remains constant?

**Examples:**

- **n×n grid:** $k=4$: $\alpha = 2n/(n^2/4) \to 0$
  - (S=n/2 × n/2 square in the center)

- **Complete binary tree:**
  - $\alpha \to 0$ for $|S|=(n/2)-1$

- **Fact:** For a random **3-regular graph** on $n$ nodes, there is some const $\alpha$ ($\alpha > 0$, independent. of $n$) such that w.h.p. the expansion of the graph is $\geq \alpha$
Fact: In a graph on $n$ nodes with expansion $\alpha$ for all pairs of nodes $s$ and $t$ there is a path of $O((\log n) / \alpha)$ edges connecting them.

Proof:

- Proof strategy:
  - We want to show that from any node $s$ there is a path of length $O((\log n)/\alpha)$ to any other node $t$.
  - Let $S_j$ be a set of all nodes found within $j$ steps of BFS from $s$.
  - How does $S_j$ increase as a function of $j$?
Proof (continued):

- Let $S_j$ be a set of all nodes found within $j$ steps of BFS from $s$.
- Then:

\[
|S_{j+1}| \geq |S_j| + \frac{\alpha |S_j|}{k} =
\]

Edges can "collide"

\[
= |S_j| \left( 1 + \frac{\alpha}{k} \right) = \left( 1 + \frac{\alpha}{k} \right)^{j+1}
\]
Proof (continued):

In how many steps of BFS we reach \( > n/2 \) nodes?

Need \( j \) so that: \( \left( 1 + \frac{\alpha}{k} \right)^{j} \geq \frac{n}{2} \)

Let’s set: \( j = \frac{k \log_2 n}{\alpha} \)

Then:

\[
\left( 1 + \frac{\alpha}{k} \right)^{\frac{k \log_2 n}{\alpha}} \geq 2^{\log_2 n} = n > \frac{n}{2}
\]

In \( O(2k/\alpha \cdot \log n) \) steps \( |S_j| \) grows to \( \Theta(n) \).
So, the diameter of \( G \) is \( O(\log(n)/\alpha) \)

Note

\[
\left( 1 + \frac{\alpha}{k} \right)^{\frac{k \log_2 n}{\alpha}} \geq 2^{\log_2 n}
\]

Remember \( n > 0, \alpha \leq k \) then:

if \( \alpha = k : (1 + 1)^{\frac{k \log_2 n}{k}} = 2^{\log_2 n} \)

if \( \alpha \to 0 \) then \( \frac{k}{\alpha} = \frac{1}{x} \to \infty \):

and \( \left( 1 + \frac{1}{x} \right)^{\log_2 n} = e^{\log_2 n} > 2^{\log_2 n} \)
Network Properties of $G_{np}$

Degree distribution: 
$$P(k) = \binom{n-1}{k} p^k (1 - p)^{n-1-k}$$

Path length: 
$$O(\log n)$$

Clustering coefficient: 
$$C = p = \bar{k}/n$$
“Evolution” of the $G_{np}$

What happens to $G_{np}$ when we vary $p$?
Remember, expected degree \( E[X_v] = (n - 1) p \)

We want \( E[X_v] \) be independent of \( n \)

So let: \( p = \frac{c}{n-1} \)

Observation: If we build random graph \( G_{np} \) with \( p = \frac{c}{n-1} \) we have many isolated nodes

Why?

\[
P[v \text{ has degree 0}] = (1 - p)^{n-1} = \left(1 - \frac{c}{n-1}\right)^{n-1} \xrightarrow{n \to \infty} e^{-c}
\]

\[
\lim_{n \to \infty} \left(1 - \frac{c}{n-1}\right)^{n-1} = \left(1 - \frac{1}{x}\right)^{-x \cdot c} = \left[ \lim_{x \to \infty} \left(1 - \frac{1}{x}\right)^x \right]^{-c} = e^{-c}
\]

By definition:

\[
e = \lim_{x \to \infty} \left(1 - \frac{1}{x}\right)^x
\]

Use substitution \( \frac{1}{x} = \frac{c}{n-1} \)
How big do we have to make $p$ before we are likely to have no isolated nodes?

We know: $P[v \text{ has degree 0}] = e^{-c}$

Event we are asking about is:

- $I = \text{some node is isolated}$
- $I = \bigcup_{v \in N} I_v$, where $I_v$ is the event that $v$ is isolated

We have:

$$P(I) = P\left(\bigcup_{v \in N} I_v\right) \leq \sum_{v \in N} P(I_v) = ne^{-c}$$
We just learned: \( P(I) = n \ e^{-c} \)

Let’s try:

- \( c = \ln n \) \thinspace \text{then:} \quad n \ e^{-c} = n \ e^{-\ln n} = n \cdot 1/n = 1 \)
- \( c = 2 \ln n \) \thinspace \text{then:} \quad n \ e^{-2 \ln n} = n \cdot 1/n^2 = 1/n \)

So if:

- \( p = \ln n \) \thinspace \text{then:} \quad P(I) = 1 \)
- \( p = 2 \ln n \) \thinspace \text{then:} \quad P(I) = 1/n \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty \)
Graph structure of $G_{np}$ as $p$ changes:

- **Emergence of a Giant Component:**
  - avg. degree $k=2E/n$ or $p=k/(n-1)$
    - $k=1-\varepsilon$: all components are of size $\Omega(ln n)$
    - $k=1+\varepsilon$: 1 component of size $\Omega(n)$, others have size $\Omega(ln n)$
$G_{np}$ Simulation Experiment

- $G_{np}$, $n=100k$, $p(n-1) = 0.5 \ldots 3$
How well does $G_{np}$ correspond to real networks?
Data statistics: Total activity

- Activity in June 2006:
  - 245 million users logged in
  - 180 million users engaged in conversations
  - More than 30 billion conversations
  - More than 255 billion exchanged messages
Fraction of country’s population on MSN:
• Iceland: 35%
• Spain: 28%
• Netherlands, Canada, Sweden, Norway: 26%
• France, UK: 18%
• USA, Brazil: 8%
Communication graph
- Edge \((u, v)\) if users \(u\) and \(v\) exchanged at least 1 msg
- \(N=180\) million people
- \(E=1.3\) billion edges
MSN Network: Connectivity

- **Count** vs. **Weakly connected component size**
- The largest component includes 99.9% of the nodes.
Degree distribution of the MSN looks nothing like the $G_{np}$.
We plot the same data as on the previous slide, just the axes are logarithmic.
**MSN: Clustering**

 Avg. clustering of the MSN: 
 \[ C = 0.1140 \]

 Avg. clustering of corresponding \( G_{np} \): 
 \[ C = \frac{k}{n} \approx 8 \cdot 10^{-8} \]

\[ C_k: \text{average } C_i \text{ of nodes of degree } k \]

\[ C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i \]
MSN: Diameter

Avg. path length 6.6
90% of the people can be reached in < 8 hops
Are real networks like random graphs?

- Average path length: 😊
- Clustering Coefficient: 😞
- Degree Distribution: 😞

Problems with the random network model:

- Degree distribution differs from that of real networks
- Giant component in most real network does NOT emerge through a phase transition
- No local structure – clustering coefficient is too low

Most important: Are real networks random?

- The answer is simply: NO
If $G_{np}$ is wrong, why did we spend time on it?

- It is the reference model for the rest of the class.
- It will help us calculate many quantities, that can then be compared to the real data.
- It will help us understand to what degree is a particular property the result of some random process.

So, while $G_{np}$ is WRONG, it will turn out to be extremely USEFUL!
The Small-World
Six Degrees of Kevin Bacon

Origins of a small-world idea:

- **Bacon number:**
  - Create a network of Hollywood actors
  - Connect two actors if they co-appeared in the movie
  - **Bacon number:** number of steps to Kevin Bacon

- As of Dec 2007, the highest (finite) Bacon number reported is 8
- Only approx. 12% of all actors cannot be linked to Bacon
Erdős numbers are small!
What is the typical shortest path length between any two people?

- Experiment on the global friendship network
  - Can’t measure, need to probe explicitly

Small-world experiment [Milgram ’67]

- Picked 300 people in Omaha, Nebraska and Wichita, Kansas
- Ask them to get a letter to a stockbroker in Boston by passing it through friends

How many steps did it take?
The Small-World Experiment

- **64 chains completed:** (i.e., 64 letters reached the target)
  - It took 6.2 steps on the average, thus “6 degrees of separation”
- **Further observations:**
  - People what owned stock had shortest paths to the stockbroker than random people: 5.4 vs. 5.7
  - People from the Boston area have even closer paths: 4.4
Milgram: Further Observations

- **Boston vs. occupation networks:**

- **Criticism:**
  - **Funneling:**
    - 31 of 64 chains passed through 1 of 3 people as their final step → Not all links/nodes are equal
  - Starting points and the target were non-random
  - People refused to participate (25% for Milgram)
  - **Some sort of social search:** People in the experiment follow some strategy (e.g., geographic routing) instead of forwarding the letter to everyone. They are not finding the shortest path!
  - There are not many samples (only 64)
  - People might have used extra information resources
In 2003 Dodds, Muhamad and Watts performed the experiment using email:

- 18 targets of various backgrounds
- 24,000 first steps (~1,500 per target)
- 65% dropout per step
- 384 chains completed (1.5%)

Avg. chain length = 4.01

**Problem:** People stop participating

**Correction factor:**

\[ n^*(h) = \frac{n(h)}{\prod_{i=0}^{h-1} (1 - r_i)} \]

\( r_i \) .... drop-out rate at hop \( i \)
After the correction:
- Typical path length $L=7$

Some not well understood phenomena in social networks:
- Funneling effect: Some target’s friends are more likely to be the final step.
  - Conjecture: High reputation/authority
- Effects of target’s characteristics: Structurally why are high-status target easier to find
  - Conjecture: Core-periphery net structure
Two Questions

1. What is the structure of a social network?
2. Which mechanisms do people use to route and find the target?
Assume each human is connected to 100 other people. Then:

- Step 1: reach 100 people
- Step 2: reach $100 \times 100 = 10,000$ people
- Step 3: reach $100 \times 100 \times 100 = 1,000,000$ people
- Step 4: reach $100 \times 100 \times 100 \times 100 = 100M$ people

In 5 steps we can reach 10 billion people.

What’s wrong here?

- 92% of new FB friendships are to a friend-of-a-friend.
Scientific Collaborations
Clustering Implies Edge Locality

- **MSN network has 7 orders of magnitude larger clustering than the corresponding $G_{np}$!**
- **Other examples:**
  
<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Average degree $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor Collaborations (IMDB)</td>
<td>225,226</td>
<td>61</td>
</tr>
<tr>
<td>Electrical power grid</td>
<td>4,941</td>
<td>2.67</td>
</tr>
<tr>
<td>Network of neurons</td>
<td>282</td>
<td>14</td>
</tr>
</tbody>
</table>

**Table 1 Empirical examples of small-world networks**

<table>
<thead>
<tr>
<th></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>

$L$ ... Average shortest path length  
$C$ ... Average clustering coefficient
Consequence of expansion:

- Short paths: $O(\log n)$
  - This is the “best” we can do if the graph has constant degree and $n$ nodes

But networks have local structure:

- Triadic closure:
  - Friend of a friend is my friend

How can we have both?
Where should we place social networks?

Clustered?

Random?
Could a network with high clustering be at the same time a small world?

- How can we at the same time have high clustering and small diameter?

- Clustering implies edge “locality”
- Randomness enables “shortcuts”
Small-world Model [Watts-Strogatz ‘98]:
2 components to the model:

(1) Start with a low-dimensional regular lattice
   - Has high clustering coefficient

(2) Rewire:
   - Add/remove edges to create shortcuts to join remote parts of the lattice
   - For each edge with prob. $p$ move the other end to a random node
The Small-World Model

Rewiring allows us to interpolate between regular lattice and a random graph

\[ h = \frac{N}{2k} \quad C = \frac{3}{4} \]

\[ h = \frac{\log N}{\log \alpha} \quad C = \frac{\bar{k}}{N} \]
The Small-World Model

It takes a lot of randomness to ruin the clustering, but a very small amount to overcome locality.

Parameter region of high clustering and low diameter

Clustering coefficient, $C = 1/n \sum C_i$

mean vertex-vertex distance

clustering coefficient
Diameter of the Watts-Strogatz

Alternative formulation of the model:

- Start with a square grid
- Each node has 1 random long-range edge
  - Each node has 1 spoke. Then randomly connect them.

\[ C_i \geq 2 \times 12/(8 \times 7) \geq 0.43 \]

What’s the diameter?
It is \( \log(n) \)
Why?
Proof:

- Consider a graph where we contract 2x2 subgraphs into supernodes.
- Now we have 4 edges sticking out of each supernode.
  - 4-regular random graph!
- From Thm. we have short paths between supernodes.
- We can turn this into a path in a real graph by adding at most 2 steps per hop.

⇒ Diameter of the model is \( O(2 \log n) \)
Could a network with high clustering be at the same time a small world?

- Yes. You don’t need more than a few random links.

The Watts Strogatz Model:

- Provides insight on the interplay between clustering and the small-world
- Captures the structure of many realistic networks
- Accounts for the high clustering of real networks
- Does not lead to the correct degree distribution
- Does not enable navigation (next lecture)
(1) What is the structure of a social network?
(2) Which mechanisms do people use to route and find the target?