Winter School on Complex Networks

SophiaTech campus
12-16 January 2015
General information

- Website
  - www-sop.inria.fr/members/Giovanni.Neglia/complexnetworks15/

- Organization of the school
- Spirit
- Presence
- Exam
- For any question: giovanni.neglia@inria.fr
Winter School on Complex Networks

Lecture 1:
Introduction to Complex Networks

Giovanni Neglia
INRIA – EPI Maestro
12 January 2015
Which network?
Which network?
Network Science

1. Common properties to many existing networks
   - Social nets, transportation nets, electrical power grids, Internet AS net, P2P nets, gene regulatory net,
   - These are the "complex networks" that exhibit "non-trivial topological features—features that do not occur in simple networks such as lattices or random graphs but often occur in real graphs" [confusing wikipedia’s definition]

2. Important dynamic processes on these networks show the same properties
Contagion

Physical Contacts

- w/o disease
- w/ disease
Contagion

Potential infection of a disease: w/o disease and w/ disease.
Contagion

- FB friendship
- w/o rumour
- w/ rumour
Contagion

[Diagram of network with nodes and P2P overlay link]
Contagion

Graph showing contact patterns with and without packets.
Take Home Lesson

If we understand how topological properties influence contagion

• We can speed-up or slow-down contagion
• We can use these lessons to engineer new protocols (overlay topologies, replication mechanisms, …)
Properties of Complex Networks

- Small diameter
- High Clustering
- Hubs and heavy tails

Physical causes

What is Network Science?

- Is it really a new science? Different from graph theory?
Milgram’s experiment (1967)
6 degrees of separation

Six degrees of separation is the idea that everyone is on average approximately six steps away, by way of introduction, from any other person in the world, so that a chain of "a friend of a friend" statements can be made, on average, to connect any two people in six steps or fewer.

J. Vaucher, Small World Networks, April 2005
Small Diameter, more formally

- A linear network has diameter N-1 and average distance $\Theta(N)$
  - How to calculate it?
- A square grid has diameter and average distance $\Theta(\sqrt{N})$
- Small Diameter: diameter $O((\log(N))^a)$, $a>0$
- Lessons from model: a few long distance random connections are enough
Erdös-Rényi graph

- A ER graph $G(N,q)$ is a stochastic process
  - $N$ nodes and edges are selected with prob. $q$
- Purpose: abstract from the details of a given graph and reach conclusions depending on its average features
A ER graph $G(N,q)$ is a stochastic process

- $N$ nodes and edges are selected with prob. $q$
- Degree distribution: $P(d) = \binom{d}{N-1} q^d (1-q)^{N-1-d}$
  - Average degree: $<d> = q (N-1)$
  - For $N \to \infty$ and $Nq$ constant: $P(d) = e^{-<d>}<d>^d/d!$
    - $<d^2> = <d>(1+<d>)$
- Average distance: $<l> \approx \log N / \log<d>$
  - Small diameter
Clustering

- "The friends of my friends are my friends"
- Local clustering coefficient of node i
  - \( \frac{\text{(# of closed triplets with i at the center)}}{\text{(total # of triplets)}} = \frac{\text{(links among i's neighbors of node i)}}{\text{(potential links among i's neighbors)}} \)

- Global clustering coefficient
  - \( \frac{\text{(total # of closed triplets)}}{\text{(total # of triplets)}} \)
    - # of closed triplets = 3 # of triangles
    - Or \( 1/N \sum_i C_i \)

\[ C_i = \frac{2}{(4 \times 3/2)} = 1/3 \]
Clustering

- In ER
  - $C \approx q \approx \langle d \rangle / N$
Clustering

- In real networks

<table>
<thead>
<tr>
<th>Network</th>
<th>Size</th>
<th>$\langle k \rangle$</th>
<th>$\ell$</th>
<th>$\ell_{\text{rand}}$</th>
<th>$C$</th>
<th>$C_{\text{rand}}$</th>
<th>Reference</th>
<th>Nr.</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>
<td>Adamic, 1999</td>
<td>1</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>
<td>Watts and Strogatz, 1998</td>
<td>2</td>
</tr>
<tr>
<td>Movie actors</td>
<td>225 226</td>
<td>61</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
<td>Yook et al., 2001a, Pastor-Satorras et al., 2001</td>
<td>3</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>
<td>Watts and Strogatz, 1998</td>
<td>4</td>
</tr>
<tr>
<td>MEDLINE co-authorship</td>
<td>153 25</td>
<td>11.1</td>
<td>7.66</td>
<td>5.72</td>
<td>0.72</td>
<td>0.003</td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>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>0.72</td>
<td>0.003</td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>6</td>
</tr>
<tr>
<td>NCSTRL co-authorship</td>
<td>14 994</td>
<td>15.5</td>
<td>9.7</td>
<td>7.64</td>
<td>0.49</td>
<td>3×10^{-2}</td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>7</td>
</tr>
<tr>
<td>Math. co-authorship</td>
<td>70 975</td>
<td>32</td>
<td>9.5</td>
<td>8.2</td>
<td>0.59</td>
<td>5.4×10^{-5}</td>
<td>Barabási et al., 2001</td>
<td>8</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>
<td>Barabási et al., 2001</td>
<td>9</td>
</tr>
<tr>
<td>E. coli, 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>
<td>Wagner and Fell, 2000</td>
<td>10</td>
</tr>
<tr>
<td>E. coli, 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>
<td>Wagner and Fell, 2000</td>
<td>11</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>
<td>Montoya and Solé, 2000</td>
<td>12</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>
<td>Montoya and Solé, 2000</td>
<td>13</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>
<td>Ferrer i Cancho and Solé, 2001</td>
<td>14</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>
<td>Yook et al., 2001b</td>
<td>15</td>
</tr>
<tr>
<td>Power grid</td>
<td>49 41</td>
<td>2.67</td>
<td>18.7</td>
<td>12.4</td>
<td>0.08</td>
<td>0.005</td>
<td>Watts and Strogatz, 1998</td>
<td>16</td>
</tr>
<tr>
<td>C. Elegans</td>
<td>282</td>
<td>14</td>
<td>2.65</td>
<td>2.25</td>
<td>0.28</td>
<td>0.05</td>
<td>Watts and Strogatz, 1998</td>
<td>17</td>
</tr>
</tbody>
</table>

Good matching for avg distance,
Bad matching for clustering coefficient
How to model real networks?

Regular Graphs have a high clustering coefficient but also a high diameter.

Random Graphs have a low diameter but a low clustering coefficient.

--> Combine both to model real networks: the Watts and Strogatz model.

Regular Graph (k=4)
- Long paths
  \[ L = \frac{n}{2k} \]
- Highly clustered
  \[ C = \frac{3}{4} \]

Random Graph (k=4)
- Short path length
  \[ L = \log_k N \]
- Almost no clustering
  \[ C = \frac{k}{n} \]

Regular ring lattice

Watts and Strogatz model

Random rewiring of regular graph
With probability $p$ rewire each link in a regular graph to a randomly selected node
Resulting graph has properties both of regular and random graphs
--> High clustering and short path length

Small World

- Usually to denote
  - small diameter + high clustering
Intermezzo: navigation

- In Small world nets there are short paths $O((\log(N))^a)$
- But can we find them?
  - Milgram's experiment suggests nodes can find them using only local information
  - Standard routing algorithms require $O(N)$ information!
  - The answer will arrive in Nicolas Nisse's lecture on "Navigation in Small Worlds"
Kleinberg's result

- Networks with underlying geographical structure

**Model:** Each node has

- Short-range connections
- 1 long-range connection, up to distance $r$ with probability proportional to $r^{-\alpha}$
- For $\alpha = 0$ it is similar to Watts-Strogatz model: there are short-paths
Hubs

- 80/20 rule
  - few nodes with degree much higher than the average
  - a lot of nodes with degree smaller than the average
  - (imagine Bill Clinton enters this room, how representative is the avg income)

- ER with $N=1000$, $\langle d \rangle = 5$, $P(d) \approx e^{-\langle d \rangle} \langle d \rangle^d / d!$
  - #nodes with $d=10$: $N \times P(10) \approx 18$
  - #nodes with $d=20$: $N \times P(20) \approx 2.6 \times 10^{-4}$
Hubs

Power law: $P(d) \sim d^{-\alpha}$
Power law degree distributions

coauthorship
... and more

Deaths in terroristic attacks

Pr(X ≥ x)

Severity (deaths), x

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10
Power Law

- Where does it come from?
  - Albert-Barabasi’s growth model
  - Highly Optimized Model
  - And other models
    - See Michael Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions
Albert-Barabasi's model

- Two elements
  - Growth
    - $m_0$ initial nodes, every time unit we add a new node with $m$ links to existing nodes
  - Preferential attachment
    - The new node links to a node with degree $k_i$ with probability

\[
\Pi(k_i) = \frac{k_i}{\sum_{j=1}^{N} k_j}
\]
Albert-Barabasi’s model

- Node $i$ arrives at time $t_i$, its degree keeps increasing
- With a continuum approximation:
  \[
  \frac{\partial k_i}{\partial t} = \frac{mk_i}{\sum_{j=1,N} k_j} = \frac{mk_i}{2tm} = \frac{k_i}{2t} \quad \rightarrow \quad k_i(t) = m\left(\frac{t}{t_i}\right)^\beta, \beta = \frac{1}{2}
  \]
- Then degree distribution at time $t$ is:
  \[
  P(k_i(t) < k) = P(t_i > t \frac{m^{1/\beta}}{k^{1/\beta}})
  \]
Albert-Barabasi's model

At time $t$ there are $m_0 + t$ nodes, if we consider that the $t$ nodes are added uniformly at random in $[0, t]$, then

$$P(t_i > x) = \frac{t - x}{t + m_0}$$

$$P(k_i(t) < k) = \frac{t}{t + m_0} \left(1 - \frac{m^{1/\beta}}{k^{1/\beta}}\right)$$
Albert-Barabasi’s model

- The PDF is

\[ P(k_i(t) = k) = \frac{\partial P(k_i(t) \leq k)}{\partial k} = \frac{t}{t + m_0} \frac{1}{\beta} \frac{m^{1/\beta}}{k^{1/\beta+1}} \]

- For \( t \to \infty \)

\[ P(k_i(t) = k) \xrightarrow{t \to \infty} \frac{1}{\beta} \frac{m^{1/\beta}}{k^{1/\beta+1}} \propto k^{-\gamma}, \quad \gamma = 3 \]
Albert-Barabasi's model

- **If** \( \Pi(k_i) \propto a + k_i \), \( P(k) \propto k^{-\gamma} \), \( \gamma = 3 + \frac{a}{m} \)

- **Other variants:**
  - With fitness \( \Pi(k) = \frac{\eta_i k_i}{\sum_{j=1,N} \eta_j k_j} \)
  - With rewiring (a prob. \( p \) to rewire an existing connection)
  - Uniform attaching with "aging": A vertex is deactivated with a prob. proportional to \( (k_i+a)^{-1} \)
Configuration model

- A family of random graphs with given degree distribution
Configuration model

- A family of random graphs with given degree distribution
  - Uniform random matching of stubs
Configuration model

- A family of random graphs with given degree distribution
  - Uniform random matching of stubs
Back to Navigation: Random Walks

- What can we do in networks without a geographical structure?
  - Random walks
How much time is needed in order to reach a given node?
Random Walks: stationary distribution

- $\pi_i = \sum_{j \in N_i} \frac{1}{k_j} \pi_j$
- $\pi_i = \frac{k_i}{N} \sum_{i=1}^{\sum k_j} = \frac{k_i}{2M}$

- avg time to come back to node $i$ starting from node $i$: $\frac{1}{\pi_i} = \frac{2M}{k_i}$
- Avg time to reach node $i$
  - intuitively $\approx \Theta(M/k_i)$
Another justification

- Random walk as random edge sampling
  - Prob. to pick an edge (and a direction) leading to a node of degree $k$ is $\frac{kp_k}{\langle k \rangle}$
  - Prob. to arrive to a given node of degree $k$: 
    \[
    \frac{kp_k}{p_kN \langle k \rangle} = \frac{k}{2M}
    \]
  - Avg. time to arrive to this node $2M/k$
- …equivalent to a RW where at each step we sample a configuration model
Distributed navigation
(speed up random walks)

- Every node knows its neighbors

{a,b,c,d}
Distributed navigation (speed up random walks)

- Every node knows its neighbors
- If a random walk looking for \( i \) arrives in \( a \) the message is directly forwarded to \( i \)

\( \{a, b, c, d\} \)
Distributed navigation reasoning 1

- We discover \( i \) when we sample one of the links of \( i \)'s neighbors
- Avg # of these links: \( k_i \sum_k ((k-1) \frac{k p_k}{<k>}) = k_i \left( \frac{<k^2>}{<k>} - 1 \right) \)
- Prob. to arrive at one of them: \( \frac{k_i}{2M} \left( \frac{<k^2>}{<k>} - 1 \right) \)
Prob that a node of degree $k$ is neighbor of node $i$ given that RW arrives to this node from a node different from $i$:

$$1 - \left(1 - \frac{k_i}{2M}\right)^{k-1} \approx \frac{k_i(k-1)}{2M}$$

Prob that the next edge brings to a node that is neighbor of node $i$:

$$\sum_k \frac{k_i(k-1)}{2M} \frac{kp_k}{<k>} = \frac{k_i}{2M} \left(\frac{<k^2>}{<k>} - 1\right)$$
Distributed navigation

- **Avg. Hop#**
  \[ \frac{2M}{k_i} \frac{<k>}{<k^2> - <k>} \]
  
  - Regular graph with degree \(d\):
    \[ \frac{2M}{d(d-1)} \]
  
  - ER with \(<k>\):
    \[ \frac{2M}{k_i(<k>-1)} \]

- Pareto distribution
  \[ P(k) \approx \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \]
  
  \[ \approx \frac{2M}{k_i} \frac{(\alpha-2)(\alpha-1)}{x_m-(\alpha-2)(\alpha-1)} \]

If \(\alpha \rightarrow 2\)...
Distributed navigation

- Application example:
  - File search in unstructured P2P networks through RWs
What is Network Science?

- A *natural science*
  - The focus is on existing networks (not graphs in general)
  - Understand observed phenomena
- An interdisciplinary approach, it draws on many different theories and methods
  - graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science, inferential modeling from statistics, social structure from sociology...