Winter School on Complex Networks

SophiaTech campus
16-20 January 2017
General information

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

- Organization of the school
- Spirit
- Presence
- Exam

For any question: giovanni.neglia@inria.fr
Lecture 1:
Introduction to Complex Networks

Giovanni Neglia
INRIA – EPI Maestro
25 January 2017
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

Physical Contacts

w/o disease

w/ disease
Contagion

Contagion w/o rumour

Contagion w/ rumour

FB friendship

w/o rumour

w/ rumour
Contagion

[Diagram showing a network with nodes labeled 'w/o file' and 'w/ file' connected by 'P2P overlay link']
Contagion

- DTN contact
- w/o packet
- w/ packet
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, …)
Outline

- Properties of Complex Networks (high-level view)
  - 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.
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
Erdös-Rényi graph

- A ER graph $G(N,q)$ is a stochastic process
  - $N$ nodes and edges are selected with prob. $q$
  - Degree distribution: $P(d)=C^d_{N-1}q^d(1-q)^{N-1-d}$
    - Average degree: $\langle d \rangle = q(N-1)$
    - For $N->\infty$ and $Nq$ constant: $P(d)=e^{-\langle d \rangle}\langle d \rangle^d/d!$
      - $\langle d^2 \rangle = \langle d \rangle (1+\langle d \rangle)$
  - Average distance: $\langle l \rangle \approx \log N/\log \langle d \rangle$
    - Small diameter
Clustering

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

\[
C_i = \frac{2}{(4\times3/2)} = \frac{1}{3}
\]

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

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

- In real networks

### Table: Clustering Coefficients in Various Networks

<table>
<thead>
<tr>
<th>Network</th>
<th>Size</th>
<th>$\langle k \rangle$</th>
<th>$l/l_{\text{rand}}$</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>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>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>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>Watts and Strogatz, 1998</td>
<td>4</td>
</tr>
<tr>
<td>MEDLINE co-authorship</td>
<td>152,25</td>
<td>16</td>
<td>16</td>
<td>1.87</td>
<td>0.97</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>Newman, 2001a, 2001b, 2001c</td>
<td>6</td>
</tr>
<tr>
<td>NCSTRL co-authorship</td>
<td>41,994</td>
<td>3.59</td>
<td>9.7</td>
<td>7.41</td>
<td>0.49</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>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>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>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>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>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>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>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>Yook et al., 2001b</td>
<td>15</td>
</tr>
<tr>
<td>Power grid</td>
<td>4941</td>
<td>2.67</td>
<td>18.7</td>
<td>12.4</td>
<td>0.08</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>Watts and Strogatz, 1998</td>
<td>17</td>
</tr>
</tbody>
</table>

**Remarks:**
- Good matching for avg distance.
- Bad matching for clustering coefficient.
How to model real networks?

“Geometric” 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

The 2D case
The 2D case
Small World

- to denote
  1. Small diameter
  2. Small diameter + high clustering
  3. Small diameter + navigability

- **Cause**
  - Nodes are embedded in some multidimensional space (e.g. geography, jobs, hobbies)
  - There are some random far-away links
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 a later module
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 $N=1000$, $\langle d \rangle = 5$, $P(d) \approx e^{-\langle d \rangle} \frac{\langle d \rangle^d}{d!}$
  - #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

Wikipedia Out-Degree Distribution

Wikipedia In-Degree Distribution

Total/In/Out Degree Distribution

coauthorship
... and more
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

\[
\Pi(k_i) = \frac{k_i}{\sum_{j=1,N} k_j}
\]

- It generates power-law

The rich becomes richer
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...
What after?

- Albert–Barabasi’s model
- how complex nets properties affect a specific dynamic process (infection)
- software tools to study complex networks (F. Huet)
- learn complex nets properties through random walks (K. Avrachenkov)
What after?

- studying mobility through complex nets (T. Spyropoulos)
- how to navigate in complex nets? (N. Nisse)
- What Twitter looks like (A. Legout)
- What R&D networks look like (M. Napoletano)
- how to describe and query the semantic web graph (C. Faron Zucker)
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
      \[
      \Pi(k_i) = \frac{k_i}{\sum_{j=1,N} k_j}
      \]
      The rich becomes richer
Albert-Barabasi’s model

- Node $i$ arrives at time $t_i$, its degree keeps increasing
- With a continuum approximation:

\[
\frac{\partial k_i}{\partial t} = \frac{m k_i}{2tm} = \frac{m k_i}{2t} = \frac{k_i}{2t} \rightarrow k_i(t) = m \left( \frac{t}{t_i} \right)^\beta, \beta = \frac{1}{2}
\]

- 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 \((k_i + a)^{-1}\)