PhD course on Network Science

Giovanni Neglia
Inria – EPI Maestro

University of Pisa
23-27 March 2015
Institut National de Recherche en Informatique et Automatique
Général De Gaulle in 67
8 research centers in France
4200 persons including
1300 permanent researchers
1200 PhD students
250 postdocs and R&D engineers
900 non-French researchers from 78 countries
General information

- Website
  - www-sop.inria.fr/members/Giovanni.Neglia/complexnetworks15pisa/

- Who are the audience?

- What can you expect from me?

- At home

- Evaluation

- For any question: giovanni.neglia@inria.fr

- First homework
  - send me an email with your name
PhD course on Network Science

Module 1:
Introduction to Network Science
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

![Network diagram with nodes and edges labeled as Facebook friendship and states of w/o rumour and w/ rumour.](image)
Contagion

P2P overlay link

w/o file

w/ file
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)
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: $<d>=q(N-1)$
    - For $N->\infty$ 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 $i$
  - $\frac{\text{(# of closed triplets with } i \text{ at the center})}{\text{(# of triplets with node } i \text{ at the center})} = \frac{\text{(links among } i\text{'s neighbors of node } i\text{)}}{\text{(potential links among } i\text{'s neighbors)}}$

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

- Global clustering coefficient
  - $\frac{\text{(total # of closed triplets)}}{\text{(total # of triplets)}}$
    - # of closed triplets = 3 # of triangles
  - $\frac{1}{N} \sum_i C_i$
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>$\ell_{rand}$</th>
<th>$C$</th>
<th>$C_{rand}$</th>
<th>Reference</th>
<th>Nr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW, site level, undir.</td>
<td>153127</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>Yook et al., 2001a,</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pastor-Satorras et al., 2001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Watts and Strogatz, 1998</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Barabási et al., 2001</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Barabási et al., 2001</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Wagner and Fell, 2000</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Wagner and Fell, 2000</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Montoya and Solé, 2000</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Montoya and Solé, 2000</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ferrer i Cancho and Solé, 2001</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yook et al., 2001b</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Watts and Strogatz, 1998</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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?

“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
$L = n/(2k)$
Highly clustered $C=3/4$

Random Graph (k=4)
Short path length
$L = \log_k N$
Almost no clustering $C=k/n$

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

Power law:
\[ P(d) \sim d^{-\alpha} \]

PDF

ER
Power law

CCDF
Power law degree distributions

coauthorship
null
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-Barabási’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?

We will

- study more formally these properties
- evaluate the effect of such properties on a specific dynamic process (infection)
- learn about software tools we can use to study complex networks
- see what a specific complex network (Twitter) looks like
What after?

We will

☑️ study more formally these properties
☑️ evaluate the effect of such properties on a specific dynamic process (infection)
☑️ learn about software tools we can use to study complex networks
☑️ see what a specific complex network (Twitter) looks like

...if we have enough time!
2nd Homework

- Watch “How Kevin Bacon Cured Cancer”
  - http://tiny.cc/kevinbacon
- Send me an email with 2 questions rising from the movie you would like the course to address