Performance Evaluation

Lecture 1: Complex Networks

Giovanni Neglia
INRIA – EPI Maestro
10 December 2012
Outline

- Properties of Complex Networks
  - Small diameter
  - High Clustering
  - Hubs and heavy tails
- Physical causes
- Consequences
Small Diameter (after Milgram’s experiment)

Six degrees - the science of a connected age, 2003

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 World: diameter $O((\log(N))^a)$, $a>0$
- Lessons from model: long distance random connections are enough
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 \( N \to \infty \) and \( Nq \) constant: \( P(d) = e^{-<d><d>^d/d!} \)
    - \( <d^2> = <d>(1+<d>) \)
  - Average distance: \( <l> \approx \log N / \log <d> \)
    - Small world
Clustering

"The friends of my friends are my friends"

Local clustering coefficient of node i

- $(\# \text{ of closed triplets with } i \text{ at the center}) / (\# \text{ of triplets with node } i \text{ at the center}) = (\text{links among } i\text{'s neighbors of node } i) / (\text{potential links among } i\text{'s neighbors})$

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

Global clustering coefficient

- $(\text{total } \# \text{ of closed triplets}) / (\text{total } \# \text{ of triplets})$
  - $\# \text{ of triplets} = 3 \# \text{ of triangles}$
- $1/N \sum_i C_i$
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>$\ell_{\text{rand}}$</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>Yook et al., 2001a, Pastor-Satorras et al., 2001</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>Watts and Strogatz, 1998</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>Newman, 2001a, 2001b, 2001c</td>
<td>4</td>
</tr>
<tr>
<td>MEDLINE co-authorship</td>
<td>152 25</td>
<td>11.1</td>
<td>8.6</td>
<td>7.3</td>
<td>0.56</td>
<td>3×10^{-5}</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>11 994</td>
<td>6.59</td>
<td>9.7</td>
<td>7.4</td>
<td>0.49</td>
<td>3×10^{-4}</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>4941</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
Kleinberg's result

Model: Each node has
- Short-range connections
- 1 long-range connection, up to distance $r$ with probability prop. to $r^{-\alpha}$
- For $\alpha=0$ it is similar to Watts-Strogatz model: there are short-paths
Kleinberg’s result

- If $\alpha = 2$ the greedy algorithm (forward the packet to the neighbor with position closest to the destination) achieves avg path length $O((\log(N))^2)$
Kleinberg’s result

- If $\alpha < 2$ no local information algorithm can take advantage of small world properties
  - avg path length $\Omega(N^{\beta/2})$
    - where $\beta = (2 - \alpha)/3$ for $0 \leq \alpha \leq 2$, $\beta = (\alpha - 2)/(\alpha - 1)$, for $\alpha > 2$
Kleinberg’s result

Conclusions

- The larger $\alpha$, the less distant long-range contacts move the message, but the more nodes can take advantage of their “geographic structure”
- $\alpha=2$ achieved the best trade-off
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:
\[ P(d) \sim d^{-\alpha} \]

PDF

ER
Power law

CCDF
Power law degree distributions

Wikipedia Out-Degree Distribution

Wikipedia In-Degree Distribution

Total/In/Out Degree Distribution

Degree (n)

Frequency (C(n))

coauthorship
... and more

Deaths in terroristic attacks

Pr(X ≥ x)

Severity (deaths), x

Pr(≥ x)
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-Barabási’s model

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

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

- Then degree distribution at time $t$ is:

$$P(k_i(t) < k) = P(t_i > t \frac{m^{1/\beta}}{k^{1/\beta}})$$
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 \( (k_i+a)^{-1} \)
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

\[ \pi_i = \sum_{j \in N_i} \frac{1}{k_j} \pi_i \]

\[ \pi_i = \frac{k_i}{N} \sum_{j=1}^{i} k_j = \frac{k_i}{2M} \]

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 \( \frac{kp_k}{<k>} \)

- Prob. to arrive to a given node of degree \( k \):
  \[
  \frac{kp_k}{p_kN <k>} = \frac{k}{2M}
  \]

- Avg. time to arrive to this node \( 2M/k \)
Distributed navigation (speed up random walks)

- Every node knows its neighbors
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\}
We discover $i$ when we sample one of the links of $i$'s neighbors.

Avg # of these links:

$$k_i \sum_k (k-1) \frac{kp_k}{\langle k \rangle} = k_i \left( \frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right)$$

Prob. to arrive at one of them:

$$\frac{k_i}{2M} \left( \frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right)$$
Distributed navigation reasoning 2

- Prob that a node of degree $k$ is neighbor of node $i$
  \[ 1 - \left(1 - \frac{k_i}{2M}\right)^{k-1} \approx \frac{k_i(k-1)}{2M} \]

- 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# \[ \frac{2M}{k_i} \left( \frac{<k^2>}{<k>} - <k> \right) \]

  - 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 \( \alpha \to 2 \ldots \)
Distributed navigation

- Application example:
  - File search in unstructured P2P networks through RWs
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