Social and Technological Network Data Analytics

Lecture 2: Small World and Weak Ties

Prof. Cecilia Mascolo
• We will compare random networks with real networks
• We will introduce the concept of small world networks
• We will introduce the concept of weak ties and illustrate their importance
Clustering Coefficient of Real Networks

• From [Watts and Strogatz, 1998]
• Characteristic path length and clustering coefficient for some real networks and for random networks with same number of nodes and average number of edges per node.

• Aim is to check if random graphs can model real networks.
Real Networks vs Random Networks

- Film Actors: actors in movies together
- Power grid: the network of the electricity generators
- C. elegans: network of neurons of a worm
- L is comparable while C is very different

<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>

Random Net: same size and average degree
Small World Model

• Watts & Strogatz built a model which was able to capture these characteristics.
• Start with regular lattice
  – Increase a probability $p$ of “rewiring” a node to another node.
  – When $p$ very high the lattice would become a random graph.
Small World Model (2)
How are L and C in this model?

- There is a zone where C is high and L is low
- These are small world networks
### Other Real Networks Examples

<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>1 520 251</td>
<td>18.1</td>
<td>4.6</td>
<td>4.91</td>
<td>0.066</td>
<td>1.1×10^{-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>0.726</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>3.59</td>
<td>9.7</td>
<td>7.34</td>
<td>0.496</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>3.9</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>
Analysis of Messenger Network

• [Leskovec and Horvitz 2008] analyzed a large dataset of the Microsoft Messenger.
  • Communication Network contained 180 million users and 1.3 billion conversations in 1 month.
  • Buddy Network contained 240 million users.
  • 99.9% users belonged to a connected component.
Analysis of a Messenger Network

- Average shortest path is 6.6 (confirming Milgram’s study).
- Although some longer paths up to 29.
- Average clustering coefficient is quite high: 0.137.
Again on Clustering Coefficient

- We have introduced the clustering coefficient. This indicates:
  - The number of triangles including node A.
  - How connected the friends of A are.
- **Triadic closure**: if C and B are connected to A there is an increased likelihood that they will be connected among themselves in future.
Granovetter interviewed people about how they discovered their jobs
  – Most people did so through personal contacts
  – Often the personal contacts described as acquaintances and not close friends

Basic intuition on this is: close friends are part of triad closures and would know what you know and would know others who would know what you know

We will explain this more formally...
Bridges

- Edge between A and B is a **bridge** if, when deleted, it would make A and B lie in 2 different components
Local Bridges

• An edge is a local bridge if its endpoints have no friends in common
  – If deleting the edge would increase the distance of the endpoints to a value more than 2.
Strong Triadic Closure Property (STPC)

• Links between nodes have different “value”: strong and weak ties
  – E.g: Friendship vs acquaintances

• **Strong Triadic Closure Property (Granovetter):** If a node A has two strong links (to B and C) then a link (strong or weak) must exist between B and C.
Local Bridges and Weak Ties

- If node A satisfies the STCP and is involved in at least two strong ties then any local bridge it is involved in must be a weak tie. (Proof by contradiction)

(assuming STCP) If there are enough strong ties in the network then local bridges must be weak ties.
Real Data Validation

• Granovetter’s theory about the importance of weak ties remained not validated for years for large social networks due to the lack of data.
• [Onnela et al ‘07] tested it over a large cell-phone network (4 millions users):
  – Edge between two users if they called each other within the 18 months period.
  – Data exhibits a giant component (84%).
  – Edge weight: time spent in conversation.
• Extending the definition of local bridge

• Given:

• Neighbourhood overlap:
  \[
  \frac{\text{Number of nodes who are neighbours of both A \& B}}{\text{Number of nodes who are neighbours of at least A or B}}
  \]

• When the numerator is 0 the quantity is 0.
  – Numerator is 0 when AB is a local bridge

• The definition finds “almost local bridges” (~0)
Neighbourhood overlap

\[ O_{ij} = 0 \]

\[ O_{ij} = \frac{1}{3} \]

\[ O_{ij} = \frac{2}{3} \]

\[ O_{ij} = 1 \]
Relationship of Overlap with Tie Strength

- Red: random shuffled weights over links.
- Blue: real ones. Correlation with tie strength.
Real tie weights in a portion of the graph (around a random node)

A = Real
B = Randomly shuffled
Effect of edge removal
Overlap based link removal

![Diagram showing the relationship between the size of the largest component and the fraction of removed links, with two curves: one for low overlap and another for high overlap.](image-url)
Weak ties matter!

• We have just seen that weak ties matter and if they are removed, they lead to a breakdown in the network.

• If strong ties are removed they lead to a smooth degrading of the network
Difference of importance of weak ties in social and other networks

- The importance of weak ties is specific to social networks
- In biological and spatial networks:
  - Deleting an important road [strong tie] damages the network more
  - A central vein in a leaf is more important than smaller veins
Tie strength matters: Facebook Example

• Facebook data analysis of one month of data
• Four networks:
  – Declared friendship
  – Reciprocal communication (messages)
  – One way communication
  – Maintained relationship: clicking on content on news feed from other friend or visiting profile more than once.
What does it look like?
(one random user)
Active Network Size: number of links

Declared friends

News feed effect
Another study on FB shows the impact of ties over information dissemination

- 3 months of FB data
- 253 million users (profile and location)

- Measuring effect of tie strength on sharing
How did they measure tie strength?

- Private interactions
- Public interactions (comments)
- Coappearance in pictures
- Involvement in the same post with comments
Strong ties are more influential

probability of sharing

photo coincidences

thread coincidences

condition

feed

no feed

comments received

messages received
Strong ties are more influential but their effect is not large enough to compensate the abundance of weak ties...
Twitter Analysis

- Huberman at al. have analyzed strong and weak ties in Twitter.
- The “followers” graph in Twitter is directed
  - Someone can follow someone else who does not follow him
- Messages of 140 chars can be posted
- Messages can be addressed to specific users (although they stay readable to all)
- **Weak ties**: users followed
- **Strong ties**: users to whom the user sent at least 2 messages in the observation period
Twitter

Number of strong ties stays below ~50
Summary

• Small world network models are able to capture a good quantity of real networks
  – They have characteristic path length comparable to random networks.
  – But much higher clustering coefficient.

• We have introduced weak and strong ties and shown example of application on real networks
• Kleinberg’s book: Chapter 3 and 20.


• **Social networks that matter: Twitter under the microscope.** Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. First Monday, 14(1), January 2009.