Network Science Overview (2)

By: Ralucca Gera, NPS
Overview

• Section 1: Graph theory:
  – Origins (Eulerian graphs)

• Section 2: Complex networks:
  – Random graphs (Erdos-Renyi)
  – Small world graphs (Watts-Strogatz)
  – Scale free graphs (Barabasi-Albert)
  – The configuration model (Molloy-Reed)
  – Random geometric model
Analysis of Complex Networks:
Watts-Strogatz (WS) small world model
(A) Node degree

(B) The clustering coefficient is

\[ \frac{3}{\text{number of}} \]

Shown here for a central node and its six neighbors. These neighbors maintain 8 out of 15 possible edges, for a clustering coefficient of 0.53 (8 triangles of 15 possible ones).

(C) Each network can be decomposed into subgraphs of motifs. The plot shows two examples of two different classes of three-node motifs.

(D) The distance between two nodes is the length of the shortest path \( d(A, B) = 3 \). The average graph’s path length is the average of all finite distance between any two vertices.

(E) This network forms two modules/communities interconnected by a single hub (hub = highly connected vertex)

Small world

- So *small world* refers to *small average path length* compared to the number of nodes in the network and *high clustering coefficient*.
- It is a property that has been observed in networks in general, regardless of the physical distance between the people the network connects.
- They explained: some nodes are believed to be at the core of a network, when in reality it is an effect of everyone being close to everyone.
<table>
<thead>
<tr>
<th>Network</th>
<th>What is it</th>
<th>Average path length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milgram’s experiment</td>
<td>Paper letters being handed person to person, to get from A to B</td>
<td>6</td>
</tr>
<tr>
<td>WWWWW</td>
<td>Hyperlinks connecting websites</td>
<td>19</td>
</tr>
<tr>
<td>Food web</td>
<td>Predators praying on other species</td>
<td>2</td>
</tr>
<tr>
<td>Molecules in the cell</td>
<td>Chemical reactions</td>
<td>3</td>
</tr>
<tr>
<td>Research collaboration network</td>
<td>Papers published by authors in different fields</td>
<td>4-6 (depending on the field)</td>
</tr>
<tr>
<td>Internet</td>
<td>Routers linked by fiber optic cables and wireless communications</td>
<td>10</td>
</tr>
<tr>
<td>IMDb.com</td>
<td>Actors linked by the movies they played in</td>
<td>3</td>
</tr>
<tr>
<td>Erdos’ number game</td>
<td>Researcher collaborations</td>
<td>2-5 (depends on the field)</td>
</tr>
</tbody>
</table>

19 may look large, but the importance of the number is in the large number of nodes that are connected, namely hundreds of billions of websites.
## Newman’s overview table of networks

<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th></th>
<th>V(G)</th>
<th>n</th>
<th></th>
<th>E(G)</th>
<th>m</th>
<th>Avg deg</th>
<th>Avg path</th>
<th>Exponent</th>
<th>Powerlaw</th>
<th>Clustering coeff</th>
<th>Homophily</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449 913</td>
<td>25 516 482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7 673</td>
<td>55 392</td>
<td>14.44</td>
<td>4.60</td>
<td>-</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253 339</td>
<td>496 489</td>
<td>3.92</td>
<td>7.57</td>
<td>-</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52 909</td>
<td>245 300</td>
<td>9.27</td>
<td>6.19</td>
<td>-</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1 520 251</td>
<td>11 803 064</td>
<td>15.53</td>
<td>4.92</td>
<td>-</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47 000 000</td>
<td>80 000 000</td>
<td>3.16</td>
<td>-</td>
<td>2.1</td>
<td>-</td>
<td>0.005</td>
<td>0.001</td>
<td>0.029</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59 912</td>
<td>86 300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>-</td>
<td>0.13</td>
<td>0.922</td>
<td>136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16 881</td>
<td>57 029</td>
<td>3.38</td>
<td>5.22</td>
<td>-</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>-</td>
<td>0.005</td>
<td>0.001</td>
<td>-0.029</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2 810</td>
<td>3.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>265, 266</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269 504</td>
<td>1 497 135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>-0.067</td>
<td>14, 34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWW AltaVista</td>
<td>directed</td>
<td>203 549 046</td>
<td>2 130 000 000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.074</td>
<td>0.35</td>
<td>0.248</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783 339</td>
<td>6 716 198</td>
<td>8.57</td>
<td>3.0/−</td>
<td>-</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1 022</td>
<td>5 103</td>
<td>4.99</td>
<td>4.87</td>
<td>-</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460 902</td>
<td>17 000 000</td>
<td>70.13</td>
<td>2.7</td>
<td>-</td>
<td>0.44</td>
<td>0.39</td>
<td>-0.189</td>
<td>119, 157</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10 097</td>
<td>31 992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>-0.189</td>
<td>86, 148</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4 941</td>
<td>6 594</td>
<td>2.67</td>
<td>18.99</td>
<td>-</td>
<td>0.10</td>
<td>0.080</td>
<td>-0.003</td>
<td>416</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19 603</td>
<td>66.70</td>
<td>2.16</td>
<td>-</td>
<td>0.69</td>
<td>-0.033</td>
<td>366</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1 439</td>
<td>1 723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>-0.016</td>
<td>318</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1 377</td>
<td>2 213</td>
<td>1.61</td>
<td>1.51</td>
<td>-</td>
<td>0.033</td>
<td>0.012</td>
<td>-0.119</td>
<td>395</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>electronic circuits</td>
<td>directed</td>
<td>24 097</td>
<td>53 248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>-0.154</td>
<td>155</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1 296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.001</td>
<td>-0.366</td>
<td>6, 354</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3 686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>-0.240</td>
<td>214</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>21 15</td>
<td>2 240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>-0.156</td>
<td>212</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>-</td>
<td>0.16</td>
<td>0.23</td>
<td>-0.263</td>
<td>204</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>-</td>
<td>0.20</td>
<td>0.087</td>
<td>-0.326</td>
<td>272</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2 359</td>
<td>7.68</td>
<td>3.97</td>
<td>-</td>
<td>0.18</td>
<td>0.28</td>
<td>-0.226</td>
<td>416, 421</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance l; exponent α of degree distribution if the distribution follows a power law (or “−” if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
• Assume in a network, the nodes’ average degree is $k$, start at some arbitrary node. Then:
  • In one hop: reach $k$ nodes (at most)
  • In 2 hops: reach $k^2$ nodes
  • In $d$ hops: reach $k^d$ nodes

Therefore, for large values of $k$ and large number of nodes $n$, we obtain a very small $d$ (where $d$ is average number of hops):

$$k^d = n \quad \rightarrow \quad d \log k = \log n \quad \rightarrow \quad d \sim \log n$$
Average path as a function of the number of nodes in the network

Logarithm of the number of nodes in the network
(a plot of log x, not measured/observed values)

$log_{10}(n)$, where $n$ is number of nodes
How to construct Small-words?

• Watts offered a modified alternative of the E-R random graphs, that
  – reconciled clustering
  – with radon graphs

• By envisioning that people live on a circle, and they make friends based on proximity to neighbors → not a small world anymore

• So to make it realistic, they add randomness to this world (since we have friends and relatives that could be half way across the world)
How to construct Small-words?

This is a model introduced by Watts-Strogatz: Networks that share properties of both regular and random graphs (Watts and his advisor Strogatz)

<table>
<thead>
<tr>
<th></th>
<th>Regular/lattice</th>
<th>Small world</th>
<th>Random graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>clustering coefficient</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>average path length</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
</tbody>
</table>

$p = \text{probability of rewiring edges of the lattice}$

The WS model turned out to be a wonderful model generating interests among the researchers, and still used today in complex networks analysis, as an ordered world sprinkled with randomness.

Downside: both ER and WS depict an egalitarian world, where edges attach to existing equal nodes based on rolling a dice.

More models need to be introduced.
However, average path length is not enough to study these networks.

Has been augmented by clustering, and even that is not enough.

Generally: degree distribution, average path and clustering, plus possibly more,

And even together they don’t characterize networks

(There are plenty of networks with the same values for all 3, yet they are not the same network)
Overview

• Section 1: Graph theory:
  – Origins (Eulerian graphs)

• Section 2: Complex networks:
  – Random graphs (Erdos-Renyi)
  – Small world graphs (Watts-Strogatz)
  – Scale free graphs (Barabasi-Albert)
  – The configuration model (Molloy-Reed)
  – The random geometric model
Analysis of Complex Networks: Barabasi-Albert (BA) preferential attachment model
The Inception of the World Wide Web

• Tim Berners-Lee in 1980 said “Suppose all the information stored on computers everywhere were linked...All the best information in every computer at CERN (European Organization for Nuclear Research, a French organization) and on the planet would be available to me and everyone else.”

• He wrote a program that allows computers to share information. A year later, this turned into the WWW.

(Barabasi, “Linked”)
• The WWW is a network
  – Nodes: website (web pages) containing all the information publicly available
  – Links: hyperlinks

• The strength of this network is in the links: the links turn a set of documents into a powerful resource available to us at our fingertips

• In a decade (in 1999) WWW had a billion documents

• What does WWW look like? Is it a small world?
Measuring the World Wide Web

- **What** does WWW look like? Why would we care if the WWW is a small word?
  - Are the WWW and social networks similar? Or like the brain? Or other networks..
  - Is there hope to learn from one network and transfer to the other if they are similar?
- We don’t really know how to measure this similarity: are we saying that if two networks have the same property (small world) then they are similar?
- If they are different, then why? Finding the “why” behind the observations is what we search for in Complex Networks.
• **How to infer the WWW?**

• Barabasi and Albert (1999) created a crawler that randomly walked the Notre Dame nd.edu domain and gathered information that created a map of this domain:
  – 300,000 nodes
  – Average number of hops = 11

• They asked:
  – How does this information transfer to the WWW that was 3,000 times larger in 1999?
  – Is the average path length of 3,000 x 11 = 33,000?
• Since nobody has a true map of the WWW (called ground truth), researchers turned to statistical mechanics approach.
  – Statistical mechanics is a field of physics that studies random systems and unpredictable outcomes
Estimating the average path length in 1999
(the numbers used in the plot are estimated from
samples using statistical mechanics,
and not the ground truth values)

In 1999 for 800 million pages, the estimated average path length was 19

“Through analyzing the data of BGP tables in recent 5 years, we find that the average path length of the Internet is descending and the descending rate is about 0.00025 which is much different from the result induced from the theories.”

Form “Does the Average Path Length Grow in the Internet?”
By Jinjing Zhao, Peidong Zhu, Xicheng Lu, Lei Xuan
• Since the average degree for WWW is 7, it follows that in 1 click we can discover 7 pages, in 2 clicks we can discover $7^2$ pages, and so on, and in 19 clicks we can discover $7^{19}$ pages which is 10 million more pages than the pages available on the Web.

• Why?

• Some of them are repeated.

• We don’t generally follow random links, we follow the meaning of the words in the links, so we don’t have to click that many times.
• In crawling the Web, Barabasi and Albert discovered hubs, very large degree nodes that connect large parts of the networks.

• There is a need for a new model to account for the hubs

• Why? Do you believe that everybody’s website has an equal chance of being discovered in a random path?

• Thinking of the topology (the layout and connections), what does this discovery depend on?
• Will the information you post be discovered with the same probability as your department’s website?
  – One study: If you look at the Web as directed (in the direction of hyperlinks, and some could be bidirectional) then 82% of the pages had less than 3 incoming hyperlinks (only 42 pages had more than 1000 incoming hyperlinks).

  (from Barabasi, “Linked”)

• What about vulnerability of networks?
• Or virus propagation?
Random (Poisson degree distribution) vs. Observed emerging networks (power-law)

Scale free

National highway network

Airline traffic network

FIGURE 1 Random and scale-free networks. The degree distribution of a random network follows a Poisson distribution close in shape to the bell curve, telling us that most nodes have the same number of links, and that nodes with a large number of links don’t exist (a). Thus, a random network is similar to a national highway network in which the nodes are the cities and the links are the major highways connecting them. Indeed, most cities are served by roughly the same number of highways (c). In contrast, the power-law degree distribution of a scale-free network predicts that most nodes have only a few links held together by a few highly connected hubs (b). Such a network is similar to the air traffic system, in which a large number of small airports are connected to each other by means of a few major hubs (d). After [1].
The same way that WS networks challenged the belief that ER is the model (they were missing the small world short average path),

now a new challenge is present for the small worlds: hubs are not present in ER nor WS, so it looks like we need another model.

Nature generates random distributions (distribution of heights among people) and power law distributions (i.e. scale-free networks)

A transition from randomness to order?
Scale-free Networks

• A **scale-free network** is a **growing** network whose degree distribution is power-law
• Generally **emerges under preferential attachment**, but it doesn’t have to.

**Scale free property:** The ratio of hubs to the number of nodes in the rest of the network remains **constant** as the network changes in size.

Enables us to study a scaled version of large networks.
Robustness: Random vs. Targeted Attack

Random Network, Accidental Node Failure

Before

After

Scale-Free Network, Accidental Node Failure

Before

After

Scale-Free Network, Attack on Hubs

Before

After
number of nodes found

Power-law graph
Poisson distribution

number of nodes found

93
19
15
11
15
7
3
1

WWW.NPS.EDU
Power-law networks are robust to random breakdown
But are especially vulnerable to targeted attack

- Targeting and removing hubs can quickly break up the network
In social networks, it’s nice to be a hub
But it depends on what you’re sharing...
From the Synthetic networks slide deck (didn’t have a chance to go over them at that time)
Generating Barabasi-Albert

Barabasi-Albert Scale Free model A (uniform attachment...)

- N – number of nodes in generated network: 100
- m0 – number of nodes at the start time: 20
- M – number of edges coming with every new node: 6

Results:
Average Degree: 13.400

Degree Distribution

Count

Value
Generating Barabasi-Albert

Barabasi-Albert Scale Free model B (no growth)

N – number of nodes in generated network: 100
M – number of edges in generated network: 225

N > 0
M > 0
M <= N = (N - 1) / 2
Generating Barabasi-Albert networks

```python
import networkx as nx
import matplotlib.pyplot as plt
import numpy as np

# Close all figure windows, in case there are open
plt.close('all')

# Define a graph
# barabasi_albert_graph(n, m, seed)
# n = Number of nodes, m = Number of edges per newly added node at each time step, seed = Random seed
G1 = nx.barabasi_albert_graph(50, 3)

#plot the graph
plt.figure()
#choose the format of the display
nx.draw_spring(G1)

# Show plots (wait until end to avoid closing)
plt.show()
```

http://networkx.lanl.gov/reference/generated/networkx.generators.random_graphs.barabasi_albert_graph.html#networkx.generators.random_graphs.barabasi_albert_graph
• Many modifications of this model exists, based on:
  – Nodes “retiring” and losing their status
  – Nodes disappearing (such as website going down)
  – Links appearing or disappearing between the existing nodes (called internal links)
  – Fitness of nodes (modeling newcomers like Google)
• Most researchers still use the standard BA model when studying new phenomena and metrics.
• Why? It is a simple model, and it was the first model that brought in growth (as well as preferential attachment)
• Section 1: Graph theory:
  – Origins (Eulerian graphs)
• Section 2: Complex networks:
  – Random graphs (Erdos-Renyi)
  – Small world graphs (Watts-Strogatz)
  – Scale free graphs (Barabasi-Albert)
  – The configuration model (Molloy-Reed)
  – The random geometric model
THE MALLOY REED CONFIGURATION MODEL
The configuration model

• A random graph model created based on Degree sequence of choice (can be scale free)

• Maybe more than degree sequence is needed to be controlled in order to create realistic models
Overview

• Section 1: Graph theory:
  – Origins (Eulerian graphs)

• Section 2: Complex networks:
  – Random graphs (Erdos-Renyi)
  – Small world graphs (Watts-Strogatz)
  – Scale free graphs (Barabasi-Albert)
  – The configuration model (Molloy-Reed)
  – The random geometric model
THE RANDOM GEOMETRIC MODEL
Random Geometric Model

• Again the connections are created at random, but based on proximity rather than preferential attachment (such as *ad hoc networks*).
• Recall that BA was introduced based on the data obtained from the Web, where physical proximity is irrelevant.
• But if one would want to model something like the Internet, then proximity is relevant.
• There is no perfect model for the world around us, not even for specific types of networks.
• No model has been introduced for the Internet.
An example of a random geometric

https://www.youtube.com/watch?v=NUisb1-INIE
A zoo of complex networks
Scale Free networks:
1. High degree heterogeneity
2. Various levels of modularity
3. Various levels of randomness

FIG. 3 A zoo of complex networks. In this qualitative space, three relevant characteristics are included: randomness, heterogeneity and modularity. The first introduces the amount of randomness involved in the process of network’s building. The second measures how diverse is the link distribution and the third would measure how modular is the architecture. The position of different examples are only a visual guide. The domain of highly heterogeneous, random hierarchical networks appears much more occupied than others. Scale-free like networks belong to this domain.

Various types of Networks, Reproduced from Sole and Valverde (2004).

http://noduslabs.com/radar/types-networks-random-small-world-scale-free/
We tend to characterize networks by their degree distributions:

- Random graphs iff Poisson degree distribution
- Scale free iff power-law degree distribution.

But they are not! Rather:
- If G is a random graphs, then G has Poisson degree distribution
- If G is scale free, then G most probably has a power-law degree distribution.
- If G was constructed using preferential attachment, then G has a power-law degree distribution.
• References to the classes that exist in python:
http://networkx.lanl.gov/reference/generators.html