CAIM: Cerca i Anàlisi d’Informació Massiva
FIB, Grau en Enginyeria Informàtica

Slides by Marta Arias, José Luis Balcázar, Ramon Ferrer-i-Cancho, Ricard Gavaldá
Department of Computer Science, UPC

Fall 2018
http://www.cs.upc.edu/~caim
7. Introduction to Network Analysis
1. Examples of real networks
2. What do real networks look like?
   ▶ real networks exhibit small diameter
     ▶ .. and so does the Erdös-Rényi or random model
   ▶ real networks have high clustering coefficient
     ▶ .. and so does the Watts-Strogatz model
   ▶ real networks’ degree distribution follows a power-law
     ▶ .. and so does the Barabasi-Albert or preferential attachment model
Examples of real networks

- Social networks
- Information networks
- Technological networks
- Biological networks
Social networks

Links denote social “interactions”

- friendship, collaborations, e-mail, etc.
Information networks

Nodes store information, links associate information

- citation networks, the web, p2p networks, etc.
Technological networks

Man-built for the distribution of a commodity

- telephone networks, power grids, transportation networks, etc.
Biological networks

Represent biological systems

- protein-protein interaction networks, gene regulation networks, metabolic pathways, etc.
Representing networks

- Network $\equiv$ Graph
- Networks are just collections of “points” joined by “lines”

<table>
<thead>
<tr>
<th>points</th>
<th>lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>vertices</td>
<td>edges, arcs</td>
</tr>
<tr>
<td>nodes</td>
<td>links</td>
</tr>
<tr>
<td>sites</td>
<td>bonds</td>
</tr>
<tr>
<td>actors</td>
<td>ties, relations</td>
</tr>
</tbody>
</table>

- math
- computer science
- physics
- sociology
Types of networks
From [?]
Small-world phenomenon

- A friend of a friend is also frequently a friend
- Only 6 hops separate any two people in the world
Measuring the small-world phenomenon, I

- Let $d_{ij}$ be the shortest-path distance between nodes $i$ and $j$.
- To check whether “any two nodes are within 6 hops”, we use:
  - The diameter (longest shortest-path distance) as
    \[ d = \max_{i,j} d_{ij} \]
  - The average shortest-path length as
    \[ l = \frac{2}{n(n+1)} \sum_{i>j} d_{ij} \]
  - The harmonic mean shortest-path length as
    \[ l^{-1} = \frac{2}{n(n+1)} \sum_{i>j} d_{ij}^{-1} \]
## Social

<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>$n$</th>
<th>$m$</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449913</td>
<td>25,164,828</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>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</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>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253339</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>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>-</td>
<td>0.45</td>
<td>0.56</td>
<td>0.333</td>
<td>311, 313</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>
</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>2.1</td>
<td>-</td>
<td>0.16</td>
<td>0.16</td>
<td>0.092</td>
<td>321, 316</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>0.17</td>
<td>0.13</td>
<td>0.000</td>
<td>45, 316</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, 316</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, 316</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>
</tr>
</tbody>
</table>

## Information

<table>
<thead>
<tr>
<th>information</th>
<th>type</th>
<th>$n$</th>
<th>$m$</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s.)</th>
</tr>
</thead>
<tbody>
<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>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>2,035,490,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.11</td>
<td>0.29</td>
<td>-0.067</td>
<td>74</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>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>45</td>
<td>351</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>
</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>0.44</td>
<td>0.19</td>
<td>0.189</td>
<td>-0.189</td>
<td>86, 157</td>
</tr>
</tbody>
</table>

## Technology

<table>
<thead>
<tr>
<th>technological</th>
<th>type</th>
<th>$n$</th>
<th>$m$</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</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, 157</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>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>-</td>
<td>0.69</td>
<td>-0.033</td>
<td>366</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>
</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>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</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>
</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>0.011</td>
<td>-0.366</td>
<td>6, 354</td>
</tr>
</tbody>
</table>

## Biological

<table>
<thead>
<tr>
<th>biological</th>
<th>type</th>
<th>$n$</th>
<th>$m$</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s.)</th>
</tr>
</thead>
<tbody>
<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>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>
But...

- Can we mimic this phenomenon in simulated networks ("models")?
- The answer is YES!
The (basic) random graph model
a.k.a. ER model

Basic $G_{n,p}$ Erdös-Rényi random graph model:

- parameter $n$ is the number of vertices
- parameter $p$ is s.t. $0 \leq p \leq 1$
- Generate and edge $(i, j)$ independently at random with probability $p$
Want to show that the diameter in ER networks is small

- Let the average degree be $z$
- At distance $l$, can reach $z^l$ nodes
- At distance $\frac{\log n}{\log z}$, reach all $n$ nodes
- So, diameter is (roughly) $O(\log n)$
ER networks have small diameter

As shown by the following simulation
To check whether “the friend of a friend is also frequently a friend”, we use:

- The transitivity or clustering coefficient, which basically measures the probability that two of my friends are also friends
Global clustering coefficient

\[ C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}} \]

\[ C = \frac{3 \times 1}{8} = 0.375 \]
Local clustering coefficient

- For each vertex $i$, let $n_i$ be the number of neighbors of $i$
- Let $C_i$ be the fraction of pairs of neighbors that are connected within each other

$$C_i = \frac{\text{nr. of connections between } i\text{'s neighbors}}{\frac{1}{2} n_i (n_i - 1)}$$

- Finally, average $C_i$ over all nodes $i$ in the network

$$C = \frac{1}{n} \sum_i C_i$$
Local clustering coefficient example

- $C_1 = C_2 = \frac{1}{1}$
- $C_3 = \frac{1}{6}$
- $C_4 = C_5 = 0$
- $C = \frac{1}{5}(1 + 1 + \frac{1}{6}) = \frac{13}{30} = 0.433$
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>$n$</th>
<th>$m$</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,16,482</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>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</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>
</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>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>529,909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.303</td>
<td>311, 313</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>
</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>2.1</td>
<td>–</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</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>0.016</td>
<td>0.13</td>
<td>0.092</td>
<td>321</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>
</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>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td></td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265, 266</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>
</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></td>
<td></td>
<td></td>
<td>74</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>78,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>351</td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5103</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>
</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></td>
<td></td>
<td>119, 157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</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>
</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>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>–0.033</td>
<td></td>
<td>366</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>
</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>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</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>
</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>0.011</td>
<td>–0.366</td>
<td>6, 354</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>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>
ER networks do not show transitivity

- $C = p$, since edges are added independently
- Given a graph with $n$ nodes and $e$ edges, we can “estimate” $p$ as
  $$\hat{p} = \frac{e}{1/2 \ n \ (n - 1)}$$
- We say that clustering is high if $C \gg \hat{p}$
  - Hence, ER networks do not have high clustering coefficient since for them $C \approx \hat{p}$
ER networks do not show transitivity

Table 1: Clustering coefficients, $C$, for a number of different networks; $n$ is the number of node, $z$ is the mean degree. Taken from [146].

<table>
<thead>
<tr>
<th>Network</th>
<th>$n$</th>
<th>$z$</th>
<th>$C$ measured</th>
<th>$C$ for random graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet [153]</td>
<td>6,374</td>
<td>3.8</td>
<td>0.24</td>
<td>0.00060</td>
</tr>
<tr>
<td>World Wide Web (sites) [2]</td>
<td>153,127</td>
<td>35.2</td>
<td>0.11</td>
<td>0.00023</td>
</tr>
<tr>
<td>power grid [192]</td>
<td>4,941</td>
<td>2.7</td>
<td>0.080</td>
<td>0.00054</td>
</tr>
<tr>
<td>biology collaborations [140]</td>
<td>1,520,251</td>
<td>15.5</td>
<td>0.081</td>
<td>0.000010</td>
</tr>
<tr>
<td>mathematics collaborations [141]</td>
<td>253,339</td>
<td>3.9</td>
<td>0.15</td>
<td>0.000015</td>
</tr>
<tr>
<td>film actor collaborations [149]</td>
<td>449,913</td>
<td>113.4</td>
<td>0.20</td>
<td>0.00025</td>
</tr>
<tr>
<td>company directors [149]</td>
<td>7,673</td>
<td>14.4</td>
<td>0.59</td>
<td>0.0019</td>
</tr>
<tr>
<td>word co-occurrence [90]</td>
<td>460,902</td>
<td>70.1</td>
<td>0.44</td>
<td>0.00015</td>
</tr>
<tr>
<td>neural network [192]</td>
<td>282</td>
<td>14.0</td>
<td>0.28</td>
<td>0.049</td>
</tr>
<tr>
<td>metabolic network [69]</td>
<td>315</td>
<td>28.3</td>
<td>0.59</td>
<td>0.090</td>
</tr>
<tr>
<td>food web [138]</td>
<td>134</td>
<td>8.7</td>
<td>0.22</td>
<td>0.065</td>
</tr>
</tbody>
</table>
So ER networks do not have high clustering, but..

- Can we mimic this phenomenon in simulated networks ("models"), while keeping the diameter small?
- The answer is YES!
Reconciling two observations from real networks:

- **High clustering**: my friend’s friends are also my friends
- **small diameter**
The Watts-Strogatz model, II

- Start with all $n$ vertices arranged on a ring
- Each vertex has initially 4 connections to their closest nodes
  - mimics local or geographical connectivity
- With probability $p$, rewire each local connection to a random vertex
  - $p = 0$ high clustering, high diameter
  - $p = 1$ low clustering, low diameter (ER model)
- What happens in between?
  - As we increase $p$ from 0 to 1
    - Fast decrease of mean distance
    - Slow decrease in clustering
For an appropriate value of $p \approx 0.01$ (1 %), we observe that the model achieves high clustering and small diameter.
Degree distribution

Histogram of nr of nodes having a particular degree

\[ f_k = \text{fraction of nodes of degree } k \]
The degree distribution of most real-world networks follows a **power-law** distribution

\[ f_k = ck^{-\alpha} \]

- “heavy-tail” distribution, implies existence of **hubs**
- hubs are nodes with very high degree
Random networks are not scale-free!

For random networks, the degree distribution follows the binomial distribution (or Poisson if $n$ is large)

$$f_k = \binom{n}{k} p^k (1 - p)^{n-k} \approx \frac{z^k e^{-z}}{k!}$$

- Where $z = p(n - 1)$ is the mean degree
- Probability of nodes with very large degree becomes exponentially small
  - so no hubs
So ER networks are not scale-free, but..

- Can we obtained scale-free simulated networks?
- The answer is YES!
Preferential attachment

“Rich get richer” dynamics
  The more someone has, the more she is likely to have

Examples
  the more friends you have, the easier it is to make new ones
  the more business a firm has, the easier it is to win more
  the more people there are at a restaurant, the more who want to go
Barabási-Albert model
From [?]

- “Growth” model
  - The model controls how a network grows over time
  - Uses preferential attachment as a guide to grow the network
    - new nodes prefer to attach to well-connected nodes
- (Simplified) process:
  - the process starts with some initial subgraph
  - each new node comes in with $m$ edges
  - probability of connecting to existing node $i$ is proportional to $i$’s degree
  - results in a power-law degree distribution with exponent $\alpha = 3$
ER vs. BA

Experiment with 1000 nodes, 999 edges \((m_0 = 1\) in BA model).
In summary..

<table>
<thead>
<tr>
<th>phenomenon</th>
<th>real networks</th>
<th>ER</th>
<th>WS</th>
<th>BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>small diameter</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>high clustering</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes(^1)</td>
</tr>
<tr>
<td>scale-free</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

\(^1\)clustering coefficient is higher than in random networks, but not as high as for example in WS networks
Network Analysis, Part II

Today’s contents

1. Centrality
   ▶ Degree centrality
   ▶ Closeness centrality
   ▶ Betweenness centrality

2. Community finding algorithms
   ▶ Hierarchical clustering
     ▶ Agglomerative
     ▶ Girvan-Newman
   ▶ Modularity maximization: Louvain method
Centrality in Networks

Centrality is a node’s measure w.r.t. others

- A central node is *important* and/or *powerful*
- A central node has an *influential position in the network*
- A central node has an *advantageous position in the network*
Degree centrality

Power through connections

\[ \text{degree}_\text{centrality}(i) \overset{\text{def}}{=} k(i) \]
Degree centrality
Power through connections

\[ \text{in}_\text{degree}_\text{centrality}(i) \overset{\text{def}}{=} k_{in}(i) \]
Degree centrality
Power through connections

\[ \text{out\_degree\_centrality}(i) \overset{\text{def}}{=} k_{out}(i) \]
By the way, there is a *normalized* version which divides the centrality of each degree by the maximum centrality value possible, i.e. $n - 1$ (so values are all between 0 and 1).

But look at these examples, does degree centrality look OK to you?
Closeness centrality
Power through proximity to others

\[
closeness\_centrality(i) \overset{\text{def}}{=} \left( \frac{\sum_{j \neq i} d(i, j)}{n - 1} \right)^{-1} = \frac{n - 1}{\sum_{j \neq i} d(i, j)}
\]

Here, what matters is to be close to everybody else, i.e., to be easily reachable or have the power to quickly reach others.
Betweenness centrality
Power through brokerage

A node is important if it lies in many shortest-paths

- so it is essential in passing information through the network
Betweenness centrality
Power through brokerage

\[ betweenness_{centrality}(i) \overset{def}{=} \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \]

Where
- \(g_{jk}\) is the number of shortest-paths between \(j\) and \(k\), and
- \(g_{jk}(i)\) is the number of shortest-paths through \(i\)

Oftentimes it is normalized:

\[ norm_{betweenness_{centrality}}(i) \overset{def}{=} \frac{betweenness_{centrality}(i)}{\binom{n-1}{2}} \]
Betweenness centrality

Examples (non-normalized)
What is community structure?
Why is community structure important?
.. but don’t trust visual perception
it is best to use objective algorithms
Main idea
A community is dense in the inside but sparse w.r.t. the outside

No universal definition! But some ideas are:

- A community should be densely connected
- A community should be well-separated from the rest of the network
- Members of a community should be more similar among themselves than with the rest

Most common..

nr. of intra-cluster edges > nr. of inter-cluster edges
Some definitions

Let $G = (V, E)$ be a network with $|V| = n$ nodes and $|E| = m$ edges. Let $C$ be a subset of nodes in the network (a “cluster” or “community”) of size $|C| = n_c$. Then

- **intra-cluster density:**

$$\delta_{int}(C) = \frac{\text{nr. internal edges of } C}{n_c(n_c - 1)/2}$$

- **inter-cluster density:**

$$\delta_{ext}(C) = \frac{\text{nr. inter-cluster edges of } C}{n_c(n - n_c)}$$

A community should have $\delta_{int}(C) > \delta(G)$, where $\delta(G)$ is the average edge density of the whole graph $G$, i.e.

$$\delta(G) = \frac{\text{nr. edges in } G}{n(n - 1)/2}$$
Most algorithms search for tradeoffs between large $\delta_{int}(C)$ and small $\delta_{ext}(C)$

- e.g. optimizing $\sum_C \delta_{int}(C) - \delta_{ext}(C)$ over all communities $C$

Define further:

- $m_c = \text{nr. edges within cluster } C = |\{(u, v)|u, v \in C\}|$
- $f_c = \text{nr. edges in the frontier of } C = |\{(u, v)|u \in C, v \notin C\}|$

- $n_{c1} = 4, m_{c1} = 5, f_{c1} = 2$
- $n_{c2} = 3, m_{c2} = 3, f_{c2} = 2$
- $n_{c3} = 5, m_{c3} = 8, f_{c3} = 2$
Community quality criteria

- **conductance**: fraction of edges leaving the cluster \( \frac{f_c}{2m_c + f_c} \)
- **expansion**: nr of edges per node leaving the cluster \( \frac{f_c}{n_c} \)
- **internal density**: a.k.a. “intra-cluster density” \( \frac{m_c}{n_c(n_c-1)/2} \)
- **cut ratio**: a.k.a. “inter-cluster density” \( \frac{f_c}{n_c(n-n_c)} \)
- **modularity**: difference between nr. of edges in \( C \) and the expected nr. of edges \( E[m_c] \) of a random graph with the same degree distribution

\[
\frac{1}{4m}(m_c - E[m_c])
\]
Methods we will cover

- Hierarchical clustering
  - Agglomerative
  - Divisive (Girvan-Newman algorithm)
- Modularity maximization algorithms
  - Louvain method
Hierarchical clustering
From hairball to *dendogram*
Suitable if input network has hierarchical structure
Agglomerative hierarchical clustering [?]  

Ingredients

- Similarity measure between nodes
- Similarity measure between *sets of nodes*

Pseudocode

1. Assign each node to its own cluster
2. Find the cluster pair with highest similarity and join them together into a cluster
3. Compute new similarities between new joined cluster and others
4. Go to step 2 until all nodes form a single cluster
Example

Data

D. Blei

Clustering 02
iteration 002

Example

iteration 002

V1

V2

0 20 40 60 80

-20 0 20 40 60 80

D. Blei
Example
Example
Example
Example
Example

iteration 010
Example

iteration 011

D. Blei Clustering 02
Example
iteration 013
Example
Example

iteration 015

V1

V2
Example
Example

iteration 018

D. Blei
Clustering 02
Example
Example
Example

iteration 022

V1
V2

D. Blei Clustering 02 5 / 21
Example

iteration 024

V1

V2
Similarity measures $w_{ij}$ for nodes $I$

Let $A$ be the adjacency matrix of the network, i.e. $A_{ij} = 1$ if $(i, j) \in E$ and 0 otherwise.

- **Jaccard index:**

  $$w_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}$$

  where $\Gamma(i)$ is the set of neighbors of node $i$

- **Cosine similarity:**

  $$w_{ij} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2} \sqrt{\sum_k A_{jk}^2}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

  where:

  - $n_{ij} = |\Gamma(i) \cap \Gamma(j)| = \sum_k A_{ik} A_{kj}$, and
  - $k_i = \sum_k A_{ik}$ is the degree of node $i$
Similarity measures $w_{ij}$ for nodes II

- **Euclidean distance:** (or rather Hamming distance since $A$ is binary)

$$d_{ij} = \sum_k (A_{ik} - A_{jk})^2$$

- **Normalized Euclidean distance:**

$$d_{ij} = \frac{\sum_k (A_{ik} - A_{jk})^2}{k_i + k_j} = 1 - 2\frac{n_{ij}}{k_i + k_j}$$

- **Pearson correlation coefficient**

$$r_{ij} = \frac{\text{cov}(A_i, A_j)}{\sigma_i \sigma_j} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i \sigma_j}$$

where $\mu_i = \frac{1}{n} \sum_k A_{ik}$ and $\sigma_i = \sqrt{\frac{1}{n} \sum_k (A_{ik} - \mu_i)^2}$

---

2 From the equation $xy = |x||y| \cos \theta$

3 Uses the idea that the maximum value of $d_{ij}$ is when there are no common neighbors and then $d_{ij} = k_i + k_j$
Similarity measures for sets of nodes

- Single linkage: $s_{XY} = \max_{x \in X, y \in Y} s_{xy}$
- Complete linkage: $s_{XY} = \min_{x \in X, y \in Y} s_{xy}$
- Average linkage: $s_{XY} = \frac{\sum_{x \in X, y \in Y} s_{xy}}{|X| \times |Y|}$
Agglomerative hierarchical clustering on Zachary’s network

Using average linkage
The Girvan-Newman algorithm
A *divisive* hierarchical algorithm [?]

**Edge betweenness**

The betweenness of an edge is the nr. of shortest-paths in the network that pass through that edge.

It uses the idea that “bridges” between communities must have high edge betweenness.

![Network diagram](image)
The Girvan-Newman algorithm

Pseudocode

1. Compute betweenness for all edges in the network
2. Remove the edge with highest betweenness
3. Go to step 1 until no edges left

Result is a dendogram
Definition of modularity [?]

Using a null model

Random graphs are not expected to have community structure, so we will use them as null models.

\[ Q = (\text{nr. of intra-cluster communities}) - (\text{expected nr of edges}) \]

In particular:

\[
Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)
\]

where \( P_{ij} \) is the expected number of edges between nodes \( i \) and \( j \) under the null model, \( C_i \) is the community of vertex \( i \), and \( \delta(C_i, C_j) = 1 \) if \( C_i = C_j \) and 0 otherwise.
How do we compute $P_{ij}$?

Using the “configuration” null model

The “configuration” random graph model choses a graph with the same degree distribution as the original graph uniformly at random.

➤ Let us compute $P_{ij}$

➤ There are $2m$ stubs or half-edges available in the configuration model

➤ Let $p_i$ be the probability of picking at random a stub incident with $i$

$$p_i = \frac{k_i}{2m}$$

➤ The probability of connecting $i$ to $j$ is then $p_ip_j = \frac{k_i k_j}{4m^2}$

➤ And so $P_{ij} = 2mp_ip_j = \frac{k_i k_j}{2m}$
Properties of modularity

\[ Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) \]

- \( Q \) depends on nodes in the same clusters only
- Larger modularity means better communities (better than random intra-cluster density)
- \( Q \leq \frac{1}{2m} \sum_{ij} A_{ij} \delta(C_i, C_j) \leq \frac{1}{2m} \sum_{ij} A_{ij} \leq 1 
- \( Q \) may take negative values
  - partitions with large negative \( Q \) implies existence of cluster with small internal edge density and large inter-community edges
The Louvain method [?]  
Considered state-of-the-art

Pseudocode

1. Repeat until local optimum reached
   1.1 Phase 1: partition network greedily using modularity
   1.2 Phase 2: agglomerate found clusters into new nodes
The Louvain method
Phase 1: optimizing modularity

Pseudocode for phase 1

1. Assign a different community to each node
2. For each node $i$
   - For each neighbor $j$ of $i$, consider removing $i$ from its community and placing it to $j$’s community
   - Greedily chose to place $i$ into community of neighbor that leads to highest modularity gain
3. Repeat until no improvement can be done
The Louvain method
Phase 2: agglomerating clusters to form new network

Pseudocode for phase 2

1. Let each community \( C_i \) form a new node \( i \)
2. Let the edges between new nodes \( i \) and \( j \) be the sum of edges between nodes in \( C_i \) and \( C_j \) in the previous graph (notice there are self-loops)
The Louvain method

Observations

- The output is also a hierarchy
- Works for weighted graphs, and so modularity has to be generalized to

\[ Q^w = \frac{1}{2W} \sum_{ij} \left( W_{ij} - \frac{s_is_j}{2W} \right) \delta(C_i, C_j) \]

where \( W_{ij} \) is the weight of undirected edge \((i, j)\),
\( W = \sum_{ij} W_{ij} \) and \( s_i = \sum_k W_{ik} \).
References I


References II
