Intro to Complex and Social Networks

Argimiro Arratia & R. Ferrer-i-Cancho

Universitat Politècnica de Catalunya

Version 0.4
Complex and Social Networks (2019-2020)
Master in Innovation and Research in Informatics (MIRI)
Instructors

- Ramon Ferrer i Cancho
  - rferredicancho@cs.upc.edu
  - Omega S124, 93 413 4028

- Argimiro Arratia
  - argimiro@cs.upc.edu
  - Omega 323,
Please go to http://www.cs.upc.edu/~csn for all course’s material, schedule, lab work, etc.
Class Logistics

- Friday, 10:00 – 12:00, A6106
  - Theory lectures.

- Tuesday, 12:00 – 14:00, every two weeks, A5S108.
  - Guided lab activities; expected to be complemented with an average estimate of 4-6 additional hours per session of autonomous lab activities.
  - Lab sessions will require handing in a short written report; these count towards the evaluation of the course.
  - Start on the 17th of September
Lab work - important rules

- Lab reports in teams of 2, submission by one member.
- Work with a different partner each lab.
- Do not exchange information other than general ideas with others: that will be considered plagiarism
Evaluation

There will be no exam in this course. Grading is done entirely through reports on various tasks throughout the course.

▶ You are expected to hand in 7 lab work reports
  ▶ The best 5 count for 50% of the final grade
  ▶ Lab reports not handed in penalize, so please complete all of them

▶ You have to do a final course project
  ▶ Project ideas given by instructors towards the end of the course
  ▶ Students pick a project or can suggest their own
  ▶ 50% of the final grade
Contents

As planned today – may go through unpredictable changes

1. Characterization of networks (*how can we describe them*)
   - Lectures 1–7
   - Includes: small-world, degree distribution, finding communities, and other advanced metrics

2. Dynamics of growing networks (*how do networks grow*)
   - Lectures 8–9
   - Includes: random growth, preferential attachment, and other growth models

3. Processing networks and processes on networks (*how can we process large networks and how are processes over networks affected by their topology*)
   - Lectures 10–13
   - Includes: sampling, epidemic models of diffusion, rumor spreading, search, percolation, etc.
So, let’s start! Today, we’ll see:

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
- Financial 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.
Financial networks

Nodes = financial assets, links = associated value or information

- Forex network I: Nodes = currencies, links = exchange value

- Forex network II: Nodes = currencies, links = nominal dollar value of all transactions between those two currencies (volume of trading)

Financial networks

The Forex network (2015): Nodes = currencies, links = exchange value
Representing networks

- Network ≡ Graph
- Networks are just collections of “points” joined by “lines”

<table>
<thead>
<tr>
<th>points</th>
<th>lines</th>
<th>disciplines</th>
</tr>
</thead>
<tbody>
<tr>
<td>vertices</td>
<td>edges, arcs</td>
<td>math</td>
</tr>
<tr>
<td>nodes</td>
<td>links</td>
<td>computer science</td>
</tr>
<tr>
<td>sites</td>
<td>bonds</td>
<td>physics</td>
</tr>
<tr>
<td>actors</td>
<td>ties, relations</td>
<td>sociology</td>
</tr>
</tbody>
</table>
Types of networks
From [Newman, 2003]

(a) unweighted, undirected
(b) discrete vertex and edge types, undirected
(c) varying vertex and edge weights, undirected
(d) directed
Descriptive measures of networks

- real networks exhibit small **diameter**
- real networks have high **clustering coefficient** (or transitivity)
- real networks’ **degree distribution** follows a power-law (i.e. are **scale free**)

Argimiro Arratia & R. Ferrer-i-Cancho  
Intro to Complex and Social Networks
**Presentation and course logistics**

**Intro to Network Analysis**

**Examples of real networks**

**Measuring and modeling networks**

---

**From [Newman, 2003]**

<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>25516482</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>55392</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>496489</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>245300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.383</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1520251</td>
<td>11803064</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>47000000</td>
<td>80000000</td>
<td>3.16</td>
<td>2.1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>265, 266</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59912</td>
<td>86300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>–</td>
<td>–</td>
<td>136</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16881</td>
<td>57029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.17</td>
<td>0.13</td>
<td>0.002</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>–</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2810</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>3.2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>265, 266</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>209504</td>
<td>1497135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.007</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203549046</td>
<td>213000000</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>783339</td>
<td>6716198</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>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1022</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>460002</td>
<td>17000000</td>
<td>70.13</td>
<td>2.7</td>
<td>–</td>
<td>0.41</td>
<td>–</td>
<td>–</td>
<td>119, 157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10697</td>
<td>31992</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>4941</td>
<td>6594</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>19603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.09</td>
<td>0.033</td>
<td>–0.033</td>
<td>366</td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1439</td>
<td>1723</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>1377</td>
<td>2213</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>24097</td>
<td>53248</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>1296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.306</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3686</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>2115</td>
<td>2240</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>2.4</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>2.0</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>2359</td>
<td>7.68</td>
<td>3.97</td>
<td>2.2</td>
<td>0.18</td>
<td>0.28</td>
<td>–0.226</td>
<td>415, 421</td>
</tr>
</tbody>
</table>

$z$ mean deg; $\ell$ mean distance; $\alpha$ exponent of deg. distrib. if power law; $C$ clustering coef.

Argimiro Arratia & R. Ferrer-i-Cancho

Intro to Complex and Social Networks
Small-world phenomenon
Low diameter and high transitivity

- Only 6 hops separate any two people in the world
- A friend of a friend is also frequently a friend
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
    $$l^{-1} = \frac{2}{n(n-1)} \sum_{i>j} d_{ij}^{-1}$$
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 $p$
Measuring the diameter in ER networks

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) $O(1)$

(Show that $z = (n - 1)p$)
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

- $C_1 = C_2 = 1/1$
- $C_3 = 1/6$
- $C_4 = C_5 = 0$
- $C = \frac{1}{5}(1 + 1 + 1/6) = 13/30 = 0.433$
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
  \[ \hat{p} = \frac{e}{\frac{1}{2} n (n - 1)} \]

- 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!
The Watts-Strogatz model, I

From [Watts and Strogatz, 1998]

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 [Barabási and Albert, 1999]

- “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 $m_0$ 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).

random

preferential attachment
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.


References II


