Machine Learning and Modeling for Social Networks

Olivia Woolley Meza, Izabela Moise, Nino Antulov-Fatulin, Lloyd Sanders
Introduction to Networks
• Motivation
• Basic concepts and definitions
  • Adjacency matrix, paths, connected components
• Centrality
  • Degree, closeness, Page Rank, betweenness
• Structural features (of social networks)
  • Heterogeneity, assortativity, clustering, small world, communities
• Network models
  • Random graphs, generative models
Multiple interconnected social media platforms

In 1736 Euler posted the following problem: *Is it possible to have a walk in the city of Königsberg, that crosses each of the seven bridges only once?*
Networks: abstraction and representation of relations

Solution: No! Unless a node is a starting or endpoint, it must have an even number of edges if every edge is traversed only once.
Social networks

• Jacob L. Moreno introduced sociograms in his 1934 book “Who Shall Survive?”

• Understand the individual through its relation to the group
Table 3.1 Basic statistics for a number of published networks. The properties measured are as follows: total number of vertices $n$; total number of edges $m$; mean degree $\langle k \rangle$; mean vertex–vertex distance $\ell$; type of graph, directed or undirected; exponent $\alpha$ 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 (3.3); clustering coefficient $C^{(2)}$ from (3.6); degree correlation coefficient $r$, section 3.6. The last column gives the citation for the network in the bibliography. Blank entries indicate unavailable data.

<table>
<thead>
<tr>
<th>Network</th>
<th>Type</th>
<th>$n$</th>
<th>$m$</th>
<th>$\langle k \rangle$</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>25,516,482</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, 415]</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, 322]</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, 181]</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>[310, 312]</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>[310, 312]</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>directed</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>[8, 9]</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>-</td>
<td>[136]</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>[320]</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>0.005</td>
<td>0.001</td>
<td>-0.029</td>
<td>[45]</td>
<td></td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td>-</td>
<td>-</td>
<td>3.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>[264, 265]</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>783,339</td>
<td>671,619</td>
<td>8.57</td>
<td>3.0/-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>[350]</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>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>[243]</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>27</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>0.10</td>
<td>0.080</td>
<td>-0.003</td>
<td>[415]</td>
<td></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>[365]</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>[317]</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>[394]</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, 353]</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>[213]</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>[211]</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>[203]</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>0.40</td>
<td>0.48</td>
<td>-0.326</td>
<td>[271]</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>[415, 420]</td>
</tr>
</tbody>
</table>

Basic definitions

- A graph $G$ is defined as $G(N,L)$
  - Set of nodes (vertices) $N$
    - Nodes can have attributes
  - Set of links (edges) $L$
    - Directed (arcs) or undirected (edges)
    - Unweighted or weighted (distance, traveling time, etc.)
    - Links of different types can exist (multiplex networks)
Network Representations: Adjacency Matrix

\[ a_{ij} = \text{existence of interaction between } i \text{ and } j \]

\[ w_{ij} = \text{weights of interaction between } i \text{ and } j \]

(e.g. number of communications per unit time)

Adjacency Matrix \( A \) has entries \( a_{ij} \)

\[
\begin{pmatrix}
0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}
\]

\[ a_{ij} = \begin{cases} 
1, & \text{if } w_{ij} > 0 \\
0, & \text{else} 
\end{cases} \]
Network Representations: Edge and Adjacency Lists

**Edge list**

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

**Adjacency list**

<table>
<thead>
<tr>
<th>1 2 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 1 5</td>
</tr>
<tr>
<td>3 2 4</td>
</tr>
<tr>
<td>4 3 5 6</td>
</tr>
<tr>
<td>5 1 2 4</td>
</tr>
<tr>
<td>6 4</td>
</tr>
</tbody>
</table>

![Graph representation with nodes and edges](image)

The adjacency matrix for the graph is:

$$
\begin{pmatrix}
0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
$$
Paths

- **Path** of length n = ordered collection of n+1 nodes and n links.
  
  - Eg: (A,B,C,E), (A,D), (C,D), (D,E,C) in G = (N,L)

- **Circuit** = closed path (last node = first node)

Number of **walks** length k is given by powers or adjacency matrix
Geodesic paths

- The geodesic (shortest) path between $i$ and $j$ is minimum number of traversed edges

Distance $d(i,j) =$ shortest path between $i$ and $j$

Diameter $D$ of the graph $= \max(d(i,j))$
A graph $G=(N,L)$ is connected if and only if there exists a path connecting any two nodes in $G$. 

- **Connected (Tree)**
- **Not Connected (Forest)**
- **Connected with loops**
Centrality Measures
Degree, Strength, Closeness

\[ a_{ij} = \text{existence of interaction between } i \text{ and } j \]

\[ w_{ij} = \text{weight of interaction between } i \text{ and } j \text{ (e.g. number of communications per unit time)} \]

\[ d_{ij} = \text{distance between } i \text{ and } j \]

Node degree \[ k_i = \sum_j a_{ij} \]

Node flux / strength \[ F_i = \sum_j w_{ij} \]

Node closeness \[ D_i = \sum_j d_{ij} \]
Eigenvector centrality and PageRank


* Eigenvector centrality $x_i$ is higher the more high-scoring others a node is connected to:

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

Solution is dominated by the largest eigenvalue $\lambda_1$ as $t \to \infty$

$$Ax = \lambda_1 x \quad x_i = \lambda_1^{-1} \sum_{j=1}^{n} A_{ij} x_j$$

* PageRank $x_i$ downgrades common in-links and deals with directed links:

$$x_i = \alpha \sum_{j=1}^{n} A_{ij} \frac{x_j}{k_j^\text{out}} + \beta$$

$$x = \alpha AD^{-1}x + \beta 1$$

$$D = \max(k_{\text{out}}, 1)$$
Betweenness centrality

альнойность

- Idea: Controlling network flows
- The number of shortest paths passing through a node $v$:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$\sigma_{st} = \text{number of shortest paths from } s \text{ to } t$

$\sigma_{st}(v) = \text{number of shortest paths from } s \text{ to } t \text{ passing through } v$
Structural features (of social networks)
Giant Component

A giant component is a connected component which size scales with the size of the network.
Centrality heterogeneity

Figure 2: Log-log plot of outdegree (top) and indegree (bottom) distributions for (a) Flickr, (b) LiveJournal, (c) Orkut, and (d) YouTube.

source: Mislove et al. (2007)
Assortative mixing or homophily

- Birds of a feather flock together
- Can be any characteristic
- E.g. Degree assortativity:
  - Average nearest-neighbor degree for vertices with degree $k$

source: Mislove et al. (2007)
Transitivity and clustering

- My friends tend to be friends

- **Local clustering coefficient** $C(i)$: fraction of pairs of neighbors of a node that are also neighbors of each other. Equivalently, number of closed triples.

Source Costa (2008)

- Question: What is the local clustering coefficient for the node $i$?

- **Global clustering coefficient**: network average
Small world property

• Empirical puzzle: Social worlds that are highly clustered but at the same time global distances are short — e.g. there are at most 6 degrees of separation between any two randomly chosen individuals.

- A small-world network is a network where the typical distance $L$ between two randomly chosen nodes grows logarithmically with total number of nodes $N$.

Modularity and community structure

Community detection vs Graph partitioning

- **Graph partitioning** specifies the number of subgroups or number of nodes in each subgroup
  - Hierarchical clustering, $k$ means
- **Community detection** infers the subgroups from the network structure
  - Divisive algorithms (recursively removing highest betweenness edges)
  - Random walk algorithms (maximizing the time random walkers spend within a community)
  - Modularity
Modularity

\[
Q = \sum_{s=1}^{m} \left[ \frac{l_s}{L} - \left( \frac{d_s}{2L} \right)^2 \right]
\]

- Basic idea: High fraction of links within group compared to chance (a null model)
- Community detection: Find partition with maximal modularity \( Q \)
Network Models

1. Random graphs
2. Configuration models
3. Generative models
1. Random graphs (Erdos-Renyi)

- Start with a number of nodes $n$ (not connected)
- Define probability of connection $p$
- For all the possible couples of nodes a link is created with probability $p$
Degree and clustering are easily computable

- The degree distribution is the Binomial distribution
  \[ \Pr(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \]
  In the limit of large \( n \)
  \[ \Pr(k) \approx \frac{(n-1)^k}{k!} p^k e^{-c} = \frac{c^k}{k!} e^{-c} \]
- The average degree is: \( \langle k \rangle = p(n-1) \)
- Clustering coefficient \( C \) is simply \( p \) (the probability of any pair existing)
- No heterogeneous degree distributions
- No small-world scaling with clustering
Percolation transition

- The formation of the Giant Component is not a smooth process
- Emerges suddenly when $<k>=1$
- This phenomenon is called 1st order phase-transition

Source: A. Clauset Network lectures
http://tuvalu.santafe.edu/~aaronc/courses/5352/
2. Configuration model

- Fix the degree sequence or degree distribution
- Find a network that samples uniformly over all other properties
- E.g. assign degrees to nodes and add “stubs”
- Uniformly at random sample stubs and connect them
- Problem: Creates self and duplicate edges (works better as network size grows)
- This process can be generalized to any property (see also Exponential Random Graph models)
3. Generative models: Preferential attachment

* Algorithm:

* Start with a random connected graph

* At each time step create a new node and attach it to each node $i$ with probability $p_i$ proportional to the node degree $k_i$

$$p_i = \frac{k_i}{\sum_j k_j}$$

* Generates power-law tails (richer-get-richer)

$$P(K) \sim k^{-3}$$
Network packages


- Python: NetworkX [https://networkx.github.io/](https://networkx.github.io/)

- iGRAPH [http://igraph.org/redirect.html](http://igraph.org/redirect.html) (originally R, now also python and C/C++)
Representing and visualizing networks

- Gephi (http://gephi.org) -> Easy and common
- Pajek (http://pajek.imfm.si/doku.php) -> Easy to use
- NWB (http://nwb.cns.iu.edu) -> Good for Analysis
- Visone (http://visone.info)
- JUNG (http://jung.sourceforge.net) -> library
- Net Draw (http://www.analytictech.com/netdraw/netdraw.htm)
- Pegasus (http://www.cs.cmu.edu/~pegasus) -> for huge data
References


- Laszlo Barabasi web site: http://nd.edu/~alb/