Three Lectures on Networks

Aaron Clauset
@aaronclauset
Associate Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute

lecture 1: what are networks and how do we talk about them?
what are networks?
what are networks?

• an approach
• a representation of complexity
• connect "micro" to "macro"
• structure above individuals / components
• structure below system / population
these lectures

- build intuition
- expose key concepts
- highlight some big questions
- teach a little math
- provide some examples
- give pointers to further study
- prep for other CSSS lectures
- not a substitute for technical coursework

About 5,810,000 results (0.04 sec)
Mark Newman
Professor of Physics
University of Michigan
External Faculty
Santa Fe Institute

http://www-personal.umich.edu/~mejn/
Network Analysis and Modeling

Instructor: Aaron Clauset or Daniel B. Larremore

This graduate-level course will examine modern techniques for analyzing and modeling the structure and dynamics of complex networks. The focus will be on statistical algorithms and methods, and both lectures and assignments will emphasize model interpretability and understanding the processes that generate real data. Applications will be drawn from computational biology and computational social science. No biological or social science training is required. (Note: this is not a scientific computing course, but there will be plenty of computing for science.)

Full lectures notes online (~150 pages in PDF)
https://aaronclauset.github.io/courses/5352/

Biological Networks

Instructor: Aaron Clauset

This undergraduate-level course examines the computational representation and analysis of biological phenomena through the structure and dynamics of networks, from molecules to species. Attention focuses on algorithms for clustering network structures, predicting missing information, modeling flows, regulation, and spreading-process dynamics, examining the evolution of network structure, and developing intuition for how network structure and dynamics relate to biological phenomena.

Full lectures notes online (~150 pages in PDF)
https://aaronclauset.github.io/courses/3352/
**Software**

- R
- Python
- Matlab
- NetworkX [python]
- igraph [python, R, c++]
- graph-tool [python, c++]
- GraphLab [python, c++]

**Standalone editors**

- UCI-Net
- NodeXL
- Gephi
- Pajek
- Network Workbench
- Cytoscape
- yEd graph editor
- Graphviz

**Network data sets**

- Colorado Index of Complex Networks

---

**The Colorado Index of Complex Networks (ICON)**

ICON is a comprehensive index of research-quality network data sets from all domains of network science, including social, web, information, biological, ecological, connectome, transportation, and technical networks.

Each network record in the index is annotated with and searchable or browsable by its graph properties, description, size, etc., and many records include links to multiple networks. The contents of ICON are curated by volunteer experts from Prof. Aaron Clauset’s research group at the University of Colorado Boulder.

Click on the **NETWORKS tab** above to get started.

---

**Entries found: 609  Networks found: 4419**
1. defining a network
2. describing a network
3. null models and statistical inference for networks
the two most fundamental questions in network science
what is a vertex?

$V$ distinct objects (vertices / nodes / actors)

when are two vertices connected?

$E \subseteq V \times V$

pairwise relations (edges / links / ties)
6 major classes of networks

- technological
- information
- transportation
- social
- biological
- economic
<table>
<thead>
<tr>
<th>network</th>
<th>vertex</th>
<th>edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet(1)</td>
<td>computer</td>
<td>IP network adjacency</td>
</tr>
<tr>
<td>Internet(2)</td>
<td>autonomous system (ISP)</td>
<td>BGP connection</td>
</tr>
</tbody>
</table>

**IP-level Internet**

**ISP network**
<table>
<thead>
<tr>
<th>network</th>
<th>vertex</th>
<th>edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet(1)</td>
<td>computer</td>
<td>IP network adjacency</td>
</tr>
<tr>
<td>Internet(2)</td>
<td>autonomous system (ISP)</td>
<td>BGP connection</td>
</tr>
<tr>
<td>software</td>
<td>function</td>
<td>function call</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>web page</td>
<td>hyperlink</td>
</tr>
<tr>
<td>documents</td>
<td>article, patent, or legal case</td>
<td>citation</td>
</tr>
<tr>
<td>network</td>
<td>vertex</td>
<td>edge</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Internet(1)</td>
<td>computer</td>
<td>IP network adjacency</td>
</tr>
<tr>
<td>Internet(2)</td>
<td>autonomous system (ISP)</td>
<td>BGP connection</td>
</tr>
<tr>
<td>software</td>
<td>function</td>
<td>function call</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>web page</td>
<td>hyperlink</td>
</tr>
<tr>
<td>documents</td>
<td>article, patent, or legal case</td>
<td>citation</td>
</tr>
<tr>
<td>power grid transmission</td>
<td>generating or relay station</td>
<td>transmission line</td>
</tr>
<tr>
<td>rail system</td>
<td>rail station</td>
<td>railroad tracks</td>
</tr>
<tr>
<td>road network(1)</td>
<td>intersection</td>
<td>pavement</td>
</tr>
<tr>
<td>road network(2)</td>
<td>named road</td>
<td>intersection</td>
</tr>
<tr>
<td>airport network</td>
<td>airport</td>
<td>non-stop flight</td>
</tr>
</tbody>
</table>
network | vertex | edge
---|---|---
road network(1) | intersection | pavement
road network(2) | named road | intersection
<table>
<thead>
<tr>
<th>network</th>
<th>vertex</th>
<th>edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>road network(1)</td>
<td>intersection</td>
<td>pavement</td>
</tr>
<tr>
<td>road network(2)</td>
<td>named road</td>
<td>intersection</td>
</tr>
</tbody>
</table>

Interstate 40

- I-15 in Barstow, CA
- I-25 in Albuquerque, NM
- I-35 in Oklahoma City, OK
- I-30 in North Little Rock, AR
- I-55 in West Memphis, AR
- I-65 in Nashville, TN
- I-75 in Farragut, TN
- I-81 in Dandridge, TN
- I-85 in Greensboro, NC
- I-87 in Raleigh, NC
- I-95 in Benson, NC
<table>
<thead>
<tr>
<th>network</th>
<th>vertex</th>
<th>edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet(1)</td>
<td>computer</td>
<td>IP network adjacency</td>
</tr>
<tr>
<td>Internet(2)</td>
<td>autonomous system (ISP)</td>
<td>BGP connection</td>
</tr>
<tr>
<td>software</td>
<td>function</td>
<td>function call</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>web page</td>
<td>hyperlink</td>
</tr>
<tr>
<td>documents</td>
<td>article, patent, or legal case</td>
<td>citation</td>
</tr>
<tr>
<td>power grid transmission</td>
<td>generating or relay station</td>
<td>transmission line</td>
</tr>
<tr>
<td>rail system</td>
<td>rail station</td>
<td>railroad tracks</td>
</tr>
<tr>
<td>road network(1)</td>
<td>intersection</td>
<td>pavement</td>
</tr>
<tr>
<td>road network(2)</td>
<td>named road</td>
<td>intersection</td>
</tr>
<tr>
<td>airport network</td>
<td>airport</td>
<td>non-stop flight</td>
</tr>
<tr>
<td>friendship network</td>
<td>person</td>
<td>friendship</td>
</tr>
<tr>
<td>sexual network</td>
<td>person</td>
<td>intercourse</td>
</tr>
</tbody>
</table>
network
friendship network
sexual network

vertex
person
person

edge
friendship
intercourse

high school friendships

sexual relationships
<table>
<thead>
<tr>
<th>network</th>
<th>vertex</th>
<th>edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet(1)</td>
<td>computer</td>
<td>IP network adjacency</td>
</tr>
<tr>
<td>Internet(2)</td>
<td>autonomous system (ISP)</td>
<td>BGP connection</td>
</tr>
<tr>
<td>software</td>
<td>function</td>
<td>function call</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>web page</td>
<td>hyperlink</td>
</tr>
<tr>
<td>documents</td>
<td>article, patent, or legal case</td>
<td>citation</td>
</tr>
<tr>
<td>power grid transmission</td>
<td>generating or relay station</td>
<td>transmission line</td>
</tr>
<tr>
<td>rail system</td>
<td>rail station</td>
<td>railroad tracks</td>
</tr>
<tr>
<td>road network(1)</td>
<td>intersection</td>
<td>pavement</td>
</tr>
<tr>
<td>road network(2)</td>
<td>named road</td>
<td>intersection</td>
</tr>
<tr>
<td>airport network</td>
<td>airport</td>
<td>non-stop flight</td>
</tr>
<tr>
<td>friendship network</td>
<td>person</td>
<td>friendship</td>
</tr>
<tr>
<td>sexual network</td>
<td>person</td>
<td>intercourse</td>
</tr>
<tr>
<td>metabolic network</td>
<td>metabolite</td>
<td>metabolic reaction</td>
</tr>
<tr>
<td>gene regulatory network</td>
<td>gene</td>
<td>regulatory effect</td>
</tr>
<tr>
<td>neuronal network</td>
<td>neuron</td>
<td>synapse</td>
</tr>
<tr>
<td>food web</td>
<td>species</td>
<td>predation or resource transfer</td>
</tr>
<tr>
<td>network</td>
<td>vertex</td>
<td>edge</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Internet(1)</td>
<td>computer</td>
<td>IP network adjacency</td>
</tr>
<tr>
<td>Internet(2)</td>
<td>autonomous system (ISP)</td>
<td>BGP connection</td>
</tr>
<tr>
<td>software</td>
<td>function</td>
<td>function call</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>web page</td>
<td>hyperlink</td>
</tr>
<tr>
<td>documents</td>
<td>article, patent, or legal case</td>
<td>citation</td>
</tr>
<tr>
<td>power grid transmission</td>
<td>generating or relay station</td>
<td>transmission line</td>
</tr>
<tr>
<td>rail system</td>
<td>rail station</td>
<td>railroad tracks</td>
</tr>
<tr>
<td>road network(1)</td>
<td>intersection</td>
<td>pavement</td>
</tr>
<tr>
<td>road network(2)</td>
<td>named road</td>
<td>intersection</td>
</tr>
<tr>
<td>airport network</td>
<td>airport</td>
<td>non-stop flight</td>
</tr>
<tr>
<td>friendship network</td>
<td>person</td>
<td>friendship</td>
</tr>
<tr>
<td>sexual network</td>
<td>person</td>
<td>intercourse</td>
</tr>
<tr>
<td>metabolic network</td>
<td>metabolite</td>
<td>metabolic reaction</td>
</tr>
<tr>
<td>gene regulatory network</td>
<td>gene</td>
<td>regulatory effect</td>
</tr>
<tr>
<td>neuronal network</td>
<td>neuron</td>
<td>synapse</td>
</tr>
<tr>
<td>food web</td>
<td>species</td>
<td>predation or resource transfer</td>
</tr>
<tr>
<td>faculty hiring</td>
<td>universities</td>
<td>faculty hiring</td>
</tr>
<tr>
<td>network</td>
<td>vertex</td>
<td>edge</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td>faculty hiring</td>
<td>universities</td>
<td>faculty hiring</td>
</tr>
</tbody>
</table>

**Computer Science faculty hiring**

- Orange: Northeast
- Green: Midwest
- Red: South
- Blue: West
- Purple: Canada
analyzing networks

what real networks look like…
analyzing networks

what real networks look like…

questions:

• how are the edges organized?
• how do vertices differ?
• does network location matter?
• are there underlying patterns?

what we want to know

• why do some edges exist, and not others?
• how does structure constrain dynamics?
• what does structure predict?
• how can we tell?
analyzing networks

what we want: understand its structure

\[ f : \text{object} \rightarrow \{\theta_1, \ldots, \theta_k\} \]

• what are the fundamental parts?
• how are these parts organized?
• where are the degrees of freedom \( \vec{\theta} \)?
• how can we define an abstract class?
• structure — dynamics — function?

what does \textit{local-level structure} look like?
what does \textit{large-scale structure} look like?
how does \textit{structure constrain} function?
analyzing networks

6 major approaches

1. **exploratory data analysis**: count & compare all the things (degree distributions, centrality scores, community detection, etc.)
analyzing networks

6 major approaches

1. **exploratory data analysis**: count & compare all the things (degree distributions, centrality scores, community detection, etc.)

2. **simple regressions**: convert network structure into node-level features, and do traditional explanatory modeling
analyzing networks

6 major approaches

1. **exploratory data analysis**: count & compare all the things (degree distributions, centrality scores, community detection, etc.)

2. **simple regressions**: convert network structure into node-level features, and do traditional explanatory modeling

3. **null models**: use some kind of random graph to identify non-random patterns as deviations from the null
6 major approaches

1. **exploratory data analysis**: count & compare all the things (degree distributions, centrality scores, community detection, etc.)

2. **simple regressions**: convert network structure into node-level features, and do traditional explanatory modeling

3. **null models**: use some kind of random graph to identify non-random patterns as deviations from the null

4. **mechanisms / simulations**: explain structural or dynamical patterns as caused by specific process
6 major approaches

1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)

2. **simple regressions:** convert network structure into node-level features, and do traditional explanatory modeling

3. **null models:** use some kind of random graph to identify non-random patterns as deviations from the null

4. **mechanisms / simulations:** explain structural or dynamical patterns as caused by specific process

5. **predictive models:** fit parametric model of network structure & use it to predict missing or future data (edges, labels, etc.)
analyzing networks

6 major approaches

1. **exploratory data analysis**: count & compare all the things (degree distributions, centrality scores, community detection, etc.)

2. **simple regressions**: convert network structure into node-level features, and do traditional explanatory modeling

3. **null models**: use some kind of random graph to identify non-random patterns as deviations from the null

4. **mechanisms / simulations**: explain structural or dynamical patterns as caused by specific process

5. **predictive models**: fit parametric model of network structure & use it to predict missing or future data (edges, labels, etc.)

6. **network experiments**: manipulate structure and measure node-level or graph-level behavior as function of changes
### Representing Networks

#### 4 Representations

- **ridiculogram**
  - nice pictures, best for small networks
- **adjacency matrix**
  - mathematically convenient & useful mental model
- **adjacency list**
  - efficient computation
- **edge list**
  - efficient storage
a simple network

undirected
unweighted
no self-loops
a simple network

adjacency matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

adjacency list

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{2, 5}</td>
</tr>
<tr>
<td>2</td>
<td>{1, 3, 4}</td>
</tr>
<tr>
<td>3</td>
<td>{2, 4, 5, 6}</td>
</tr>
<tr>
<td>4</td>
<td>{2, 3}</td>
</tr>
<tr>
<td>5</td>
<td>{1, 3}</td>
</tr>
<tr>
<td>6</td>
<td>{3}</td>
</tr>
</tbody>
</table>

undirected
unweighted
no self-loops
beyond simple graphs

- directed edge
- weighted edge
- weighted node
- multi-edge
- self-loop

undirected
unweighted
no self-loops
### beyond simple graphs

**Adjacency Matrix**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>{1,1,2}</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>(\frac{1}{2})</td>
<td>{2,1}</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>{2,1}</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>{1,1,2}</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

**Adjacency List**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>({(5,1), (5,1), (5,2)})</td>
</tr>
<tr>
<td>2</td>
<td>({(1,1), (2, \frac{1}{2}), (3,2), (3,1), (4,1)})</td>
</tr>
<tr>
<td>3</td>
<td>({(2,2), (2,1), (4,2), (5,4), (6,4)})</td>
</tr>
<tr>
<td>4</td>
<td>({(2,1), (3,2)})</td>
</tr>
<tr>
<td>5</td>
<td>({(1,1), (1,1), (1,2), (3,4)})</td>
</tr>
<tr>
<td>6</td>
<td>({(3,4), (6,2)})</td>
</tr>
</tbody>
</table>
beyond simple graphs

<table>
<thead>
<tr>
<th>attributes of</th>
<th>edges</th>
<th>nodes</th>
<th>network</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unweighted</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>weighted</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>signed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>undirected</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>directed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>multigraph</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>timestamps</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**beyond simple graphs**

<table>
<thead>
<tr>
<th>attributes of</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>edges</td>
<td>nodes</td>
</tr>
<tr>
<td>unweighted</td>
<td>metadata / attributes</td>
</tr>
<tr>
<td>weighted</td>
<td>locations / coordinates</td>
</tr>
<tr>
<td>signed</td>
<td>state variables</td>
</tr>
<tr>
<td>undirected</td>
<td></td>
</tr>
<tr>
<td>directed</td>
<td></td>
</tr>
<tr>
<td>multigraph</td>
<td></td>
</tr>
<tr>
<td>timestamps</td>
<td></td>
</tr>
</tbody>
</table>
beyond simple graphs

<table>
<thead>
<tr>
<th>attributes of</th>
<th>edges</th>
<th>nodes</th>
<th>network</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unweighted</td>
<td>metadata / attributes</td>
<td>sparse</td>
</tr>
<tr>
<td></td>
<td>weighted</td>
<td>locations / coordinates</td>
<td>dense</td>
</tr>
<tr>
<td></td>
<td>signed</td>
<td>state variables</td>
<td>bipartite</td>
</tr>
<tr>
<td></td>
<td>undirected</td>
<td></td>
<td>projection</td>
</tr>
<tr>
<td></td>
<td>directed</td>
<td></td>
<td>connected</td>
</tr>
<tr>
<td></td>
<td>multigraph</td>
<td></td>
<td>disconnected</td>
</tr>
<tr>
<td></td>
<td>timestamps</td>
<td></td>
<td>acyclic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>temporal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>multiplex</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>hypergraph</td>
</tr>
</tbody>
</table>
describing networks

aka, summarizing a network’s structure

\[ f : G \to \{x_1, \ldots, x_k\} \]

summary statistics
describing networks

**degree:**
the first order description of a network
describing networks

degree:
number of connections \( k \)

\[ k_i = \sum_j A_{ij} \]
describing networks

number of edges

\[ m = \frac{1}{2} \sum_{i=1}^{n} k_i = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} = \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ji} \]

degree:
number of connections \( k \)

\[ k_i = \sum_{j} A_{ij} \]
describing networks

**degree:**
number of connections $k$

\[ k_i = \sum_j A_{ij} \]

**number of edges**

\[ m = \frac{1}{2} \sum_{i=1}^{n} k_i = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} = \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ji} \]

**mean degree**

\[ \langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2m}{n} \]
degree:
number of connections $k$

$$k_i = \sum_j A_{ij}$$

degree sequence $\{1, 2, 2, 2, 3, 4\}$

degree distribution $Pr(k) = \left[\left(1, \frac{1}{6}\right), \left(2, \frac{3}{6}\right), \left(3, \frac{1}{6}\right), \left(4, \frac{1}{6}\right)\right]$
node degrees

"low" degree

"high" degree

node degrees: what impact does having fewer or more connections have?

more information? more exposure to disease? more robust connectivity? more influence? less bandwidth? etc.

* scare quotes because 'low' and 'high' are relative terms
degree distributions

政治博客（2004年）

\[
\begin{align*}
n &= 1490 \\
m &= 19090 \\
\langle k \rangle &= 25.6
\end{align*}
\]

<table>
<thead>
<tr>
<th>( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Pr}(k) )</td>
<td>0.271</td>
<td>0.072</td>
<td>0.052</td>
<td>0.034</td>
<td>0.026</td>
<td>0.033</td>
<td>0.020</td>
<td>0.017</td>
<td>0.011</td>
</tr>
</tbody>
</table>

这张网络可以通过icon.colorado.edu获取。

degree distributions

Political blogs (2004)

Simple pdf:

\[ \Pr(K = k) \text{ vs. } k \]

This network available via icon.colorado.edu

degree distributions

simple pdf:

\[ \log_{10} \Pr(K = k) \] vs. \[ k \]

political blogs (2004)

\[ \langle k \rangle = 25.6 \]
\[ k_{\max} = 351 \]

\[ \log_{10} \Pr(k) \] vs. \[ k \]

\[ \log_{10} \Pr(k) \] vs. \[ \log_{10} k \]

this network available via icon.colorado.edu

degree distributions

political blogs (2004)

simple pdf:

\[ \log_{10} \Pr(K = k) \text{ vs. } \log_{10} k \]

\[ \langle k \rangle = 25.6 \]

\[ k_{\text{max}} = 351 \]

\[ \Pr(k) \]
degree distributions

political blogs (2004)

complementary cdf:

\[ \Pr(K \geq k) = \sum_{j=k}^{n} \Pr(K = j) \]

this network available via icon.colorado.edu

degree distributions

political blogs (2004)

\[ n = 1490 \]
\[ m = 19090 \]
\[ \langle k \rangle = 25.6 \]

- 90% (1349) have \( k \leq 67 \)
  connecting to 53% of all \( m \)

this network available via icon.colorado.edu

**degree distributions**

- Political blogs (2004)
  - \( n = 1490 \)
  - \( m = 19090 \)
  - \( \langle k \rangle = 25.6 \)

- 90% (1349) have \( k \leq 67 \) connecting to 53% of all \( m \)
- only 1% (14) have \( k > 169 \) connecting to 10% of all \( m \)

---

this network available via icon.colorado.edu

degree distributions

political blogs (2004)

\[ n = 1490 \]
\[ m = 19090 \]
\[ \langle k \rangle = 25.6 \]

- 90% (1349) have \( k \leq 67 \) connecting to 53% of all \( m \)
- only 1% (14) have \( k > 169 \) connecting to 10% of all \( m \)

this network available via icon.colorado.edu

degree distributions

political blogs (2004)

\( n = 1490 \)
\( m = 19090 \)
\( \langle k \rangle = 25.6 \)

Lorenz curve

- fraction of all edges \( F \) held by "richest" fraction of all nodes \( P \)
- Gini coefficient: \( G = 2A \)
  \[= 0.69\]

* the 80-20 rule: 80% of "wealth" held by 20% of "population"
exploring degree distributions

the complementary CDF:
\[ \Pr(K \geq k) = 1 - \Pr(K < k) \]
- fraction with degree \( K \geq k \)
- monotonic
- smoother than PDF
- better reveals curvature

"loglog" plots:
- good for high variance quantities

Lorenz curves & Gini:
- captures level of inequality
exploring degree distributions

- nearly all real networks exhibit a heavy-tailed degree distribution
- very few networks exhibit convincing power-law degree distributions
- some distributions exhibit power-law tails
- power laws are cool! but knowing one from garbage requires statistics*

- does the specific distributional form matter?
  think carefully about whether it does (it may not).

* data from 100 networks from 4 scientific domains
  Broido & Clauset, Nat. Comm. 10 (2019)
**exploring degree distributions**

**degree structure** is the first-order description of network structure

- drives interesting phenomena ("friendship paradox", spreading dynamics, etc.)
- explains many other network patterns (various centralities, disassortativity, etc.)

how can we tell if degrees explain a pattern?

- the **configuration model** : random graphs with specified (empirical) degree sequence (use *Python package from Dutta*)
- assess whether a node-level or network-level measure is big, typical, or small — it’s a *null model*
end of lecture 1

lecture 2: describing network structure

lecture 3: null models & inference for networks
directed networks

\[ A_{ij} \neq A_{ji} \]

citation networks
foodwebs*
edemiological
others?
directed acyclic graph
directed graph
WWW
friendship?
flows of goods, information
economic exchange
dominance
neuronal
transcription
time travelers
bipartite networks

no within-type edges

authors & papers
actors & movies/scenes
musicians & albums
people & online groups
people & corporate boards

people & locations (checkins)
metabolites & reactions
genes & substrings
words & documents
plants & pollinators
bipartite networks

Authors & papers
Actors & movies/scenes
Musicians & albums
People & online groups
People & corporate boards

People & locations (checkins)
Metabolites & reactions
Genes & substrings
Words & documents
Plants & pollinators
temporal networks

any network over time

discrete time (snapshots), edges \((i, j, t)\)
continuous time, edges \((i, j, t_s, \Delta t)\)
degree distributions

degree "wealth"

what fraction of total wealth $W$ is owned by richest fraction $P$
degree distributions

degree "wealth"

what fraction of total wealth $W$ is owned by richest fraction $P$

$\Pr(k) \propto e^{-\lambda k}$

exponential distribution

Lorenz curve
degree distributions

**degree "wealth"**

what fraction of total wealth $W$ is owned by richest fraction $P$

$$Pr(k) \propto \frac{1}{k} e^{-\left(\frac{\ln k - \mu}{\sigma \sqrt{2}}\right)^2}$$

log-normal distribution

Lorenz curve
degree "wealth"

what fraction of total wealth $W$ is owned by richest fraction $P$

$$\Pr(k) \propto k^{-\alpha}$$

power-law distribution

80/20 rule

Lorenz curve
degree distributions

is this a power law?

political blogs*

power-law distributions

\[ \Pr(k) = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\min} \]

- let's do some math
- (a nice warm up for other things, later)
power-law distributions

\[ \Pr(k) = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\min} \]

- normalization (probability density function)

\[ 1 = \int_{k_{\min}}^{\infty} \Pr(k) \, dk \quad \text{pdf} \]

- complementary cumulative distribution function

\[ P(k) = \int_{k}^{\infty} \Pr(y) \, dy \quad \text{ccdf} \]

power-law distributions

\[ \text{Pr}(k) = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

- normalization (probability density function)*
  \[
  1 = \int_{k_{\text{min}}}^{\infty} \text{Pr}(k) \, dk \quad \rightarrow \quad \text{Pr}(k) = \frac{\alpha - 1}{k_{\text{min}}} \left( \frac{k}{k_{\text{min}}} \right)^{-\alpha} \]

- complementary cumulative distribution function
  \[
  P(k) = \int_{k}^{\infty} \text{Pr}(y) \, dy \quad \rightarrow \quad P(k) = \left( \frac{k}{k_{\text{min}}} \right)^{-\alpha+1} \]

- power laws have unusual properties, imply unusual underlying mechanisms

* the math here is easier for the continuous variables, but qualitatively similar results hold for discrete variables. Also, yes, vertex degree is discrete not continuous.

power-law distributions

\[ \Pr(k) = C \, k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

- high-variance

\[ \langle k^m \rangle = \int_{k_{\text{min}}}^{\infty} k^m \Pr(k) \, dk \]
power-law distributions

\[ \Pr(k) = C \, k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

- high-variance

\[ \langle k^m \rangle = \int_{k_{\text{min}}}^{\infty} k^m \Pr(k) \, dk \]

\[ = k_{\text{min}}^m \left( \frac{\alpha - 1}{\alpha - 1 - m} \right) \]

- infinite mean \( 1 < \alpha < 2 \)

- infinite variance \( 2 < \alpha < 3 \)

- much, much heavier tails than exponential, normal, etc.

- heavier than log-normal (asymptotically)

power-law distributions

\[ \Pr(k) = C k^{-\alpha} \text{ for } k \geq k_{\text{min}} \]

• "scale invariance" (aka "scale free")
  \[ \Pr(c \, k) = (\alpha - 1) k_{\text{min}}^{\alpha-1} (c \, k)^{-\alpha} \]

power-law distributions

\[ \Pr(k) = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

- "scale invariance" (aka "scale free")
  \[ \Pr(c \, k) = (\alpha - 1) k_{\text{min}}^{\alpha - 1} (c \, k)^{-\alpha} = c^{-\alpha} \left[ (\alpha - 1) k_{\text{min}}^{\alpha - 1} k^{-\alpha} \right] \propto \Pr(k) \]
- power law is only distribution with this property
- implies no natural "scale" of distribution
- implies signature form: straight line on log-log plot
  \[ \ln \Pr(k) = \ln C - \alpha \ln k \]

power-law distributions

\[ \Pr(k) = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

• exotic mechanisms
  • preferential attachment [Yule 1925, Simon 1955, Price 1976, etc.]
  • combinations of exponentials [Miller 1957, Reed & Hughes 2002]
  • phase transitions [many]
  • self-organized criticality (SOC) [Bak et al. 1988]
  • highly optimized tolerance (HOT) [Carlson and Doyle, 1999]
  • fragmentation [many]
  • multiplicative random walks (with lower limit) [many]
  • many, many others

**power-law distributions**

\[ \Pr(k) = Ck^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

- how do you know? statistics.
- estimating \( \alpha \) from data \( \{k_i\} \) via maximum likelihood

\[
\ln \mathcal{L}(\{k_i\} \mid \theta) = \ln \prod_{i=1}^{n} \Pr(k_i \mid \theta)
\]

power-law distributions

\[ \Pr(k) = C k^{-\alpha} \text{ for } k \geq k_{\text{min}} \]

- how do you know? statistics.
- estimating \( \alpha \) from data \( \{k_i\} \) via maximum likelihood

\[
\ln \mathcal{L}(\{k_i\} | \theta) = \ln \prod_{i=1}^{n} \Pr(k_i | \theta)
\]

- for the power-law distribution (log-likelihood)

\[
\ln \mathcal{L}(\{k_i\} | \alpha, k_{\text{min}}) = n \ln \left( \frac{\alpha - 1}{k_{\text{min}}} \right) - \alpha \sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\text{min}}} \right)
\]

- solving \( \partial \mathcal{L} / \partial \alpha = 0 \), yields MLE with standard error

\[
\hat{\alpha} = 1 + n \sqrt{\sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\text{min}}} \right)}
\]

\[
\hat{\sigma} = \frac{\hat{\alpha} - 1}{\sqrt{n}} + O(1/n)
\]

power-law distributions

\[ \Pr(k) = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

- how do you know? statistics.
- estimating \( \alpha \) from data \( \{k_i\} \) via maximum likelihood
  
  \[ \ln \mathcal{L}(\{k_i\} | \theta) = \ln \prod_{i=1}^{n} \Pr(k_i | \theta) \]

- for the power-law distribution (log-likelihood)
  
  \[ \ln \mathcal{L}(\{k_i\} | \alpha, k_{\text{min}}) = n \ln \left( \frac{\alpha - 1}{k_{\text{min}}} \right) - \alpha \sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\text{min}}} \right) \]

- solving \( \partial \mathcal{L} / \partial \alpha = 0 \), yields MLE

\[ \hat{\alpha} = 1 + n \left/ \sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\text{min}}} \right) \right. \]

with standard error

\[ \hat{\sigma} = \frac{\hat{\alpha} - 1}{\sqrt{n}} + O(1/n) \]

umm... we don't know this value
power-law distributions

\[ \Pr(k) = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}} \]

- we can choose \( k_{\text{min}} \) smartly [see SIAM Review 51; code is here *]
- but how do we know if the model is good? fitting is easy

**moral: always check your model’s goodness-of-fit**

- ways to do this:
  1. compute a \( p \)-value relative to a *reasonable* null model
  2. compare your model against *reasonable* alternatives
  3. compare synthetic data drawn from your model with your empirical data
  4. use your model to predict something *reasonable*

---


* http://santafe.edu/~aaronc/powerlaws/