Lecture 9b: Exploration, testing, and prediction

Aaron Clauset

Assistant Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute
"Those who ignore Statistics are condemned to reinvent it." — Bradley Efron
"The first principle is that you must not fool yourself, and you are the easiest person to fool."
— Richard Feynman
"Those who ignore Statistics are condemned to reinvent it."
— Bradley Efron

"The first principle is that you must not fool yourself, and you are the easiest person to fool."
— Richard Feynman

"There are three kinds of lies: lies, damned lies, and statistics."
— unknown
"Those who ignore Statistics are condemned to reinvent it."
— Bradley Efron

"The first principle is that you must not fool yourself, and you are the easiest person to fool."
— Richard Feynman

"There are three kinds of lies: lies, damned lies, and statistics."
— unknown

"It’s easy to lie with statistics, but it’s easier to lie without them."
— Fred Mosteller
"Those who ignore Statistics are condemned to reinvent it."
— Bradley Efron

"The first principle is that you must not fool yourself, and you are the easiest person to fool."
— Richard Feynman

"It’s easy to lie with statistics, but it’s easier to lie without them."
— Fred Mosteller

"There are three kinds of lies: lies, damned lies, and statistics."
— unknown

"If your experiment needs statistics, you ought to have done a better experiment."
— E. Rutherford
"Those who ignore Statistics are condemned to reinvent it." — Bradley Efron

"The first principle is that you must not fool yourself, and you are the easiest person to fool." — Richard Feynman

"It's easy to lie with statistics, but it's easier to lie without them." — Fred Mosteller

"There are three kinds of lies: lies, damned lies, and statistics." — unknown

"If your experiment needs statistics, you ought to have done a better experiment." — E. Rutherford

"Far better an approximate answer to the right question... than an exact answer to the wrong question." — John W. Tukey
"Those who ignore Statistics are condemned to reinvent it."
— Bradley Efron

"The first principle is that you must not fool yourself, and you are the easiest person to fool."
— Richard Feynman

"It's easy to lie with statistics, but it's easier to lie without them."
— Fred Mosteller

"There are three kinds of lies: lies, damned lies, and statistics."
— unknown

"If your experiment needs statistics, you ought to have done a better experiment."
— E. Rutherford

"Far better an approximate answer to the right question... than an exact answer to the wrong question."
— John W. Tukey

"In God we trust. All others must bring data."
— W. Edwards Deming

"God must bring data, too."
— unknown
three roles of statistics

- data exploration
- model testing
- prediction
data exploration: community detection

- given a graph $G$
- divide its vertices into coherent groups $\mathcal{z}(G)$

- consummate data exploration!
- a common task in network analysis
- helped yield insight into real social, biological, technological systems
- scores of methods, many extremely powerful, some with guarantees (stochastic block model, Belief Propagation, etc.)
data exploration : community detection

• given a graph $G$
• divide its vertices into coherent groups $z(G)$

• nearly all methods:
  estimate $\max_z f(z(G))$

[WARNING: typically NP-hard]
the trouble with community detection

this is a pretty good division (under nearly any $f$)

data exploration: community detection

so are all of these (and many more)
data exploration : community detection

• there are an exponential number of good-looking local maxima

  each algorithm chooses one

• this is okay for data exploration!

• anything else requires caution

• risks: 'wrong' optima

• opportunities: community structure is genuinely interesting!

• difficulties: how do we select among all these good divisions?

Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network

Manuel Middendorf†, Etay Ziv‡, and Chris H. Wiggins§

- *observation*: many protein interaction networks have heavy-tailed (power-law?) degree distributions
Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network

Manuel Middendorf*, Etay Ziv‡, and Chris H. Wiggins§

- **observation**: many protein interaction networks have heavy-tailed (power-law?) degree distributions
- **claims**: as of 2005, FIVE different models proposed as generative mechanisms
- duplication mutation complementation (DMC), duplication mutation-random (DMR), linear preferential attachment (LPA), random growing networks (RDG), aging vertex networks (AGV)

model testing : scale-free networks

• the problem: all models fit the observed degree distribution
model testing: scale-free networks

- the problem: all models fit the observed degree distribution
model testing : scale-free networks

- **the solution**: build a **classifier** that can **distinguish** networks generated by the 5 models + 2 controls **based on** their motif frequencies
- use decision trees + Adaboost (very powerful) to **learn which motifs** distinguish the models
- **validated** on synthetic graphs with known structure:

<table>
<thead>
<tr>
<th>Truth</th>
<th>DMR</th>
<th>DMC</th>
<th>AGV</th>
<th>LPA</th>
<th>SMW</th>
<th>RDS</th>
<th>RDG</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMR</td>
<td>99.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>DMC</td>
<td>0.0</td>
<td>99.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>AGV</td>
<td>0.0</td>
<td>0.1</td>
<td>84.7</td>
<td>13.5</td>
<td>1.2</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>LPA</td>
<td>0.0</td>
<td>0.0</td>
<td>10.3</td>
<td>89.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>SMW</td>
<td>0.0</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>99.0</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>RDS</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>99.0</td>
<td>0.0</td>
</tr>
<tr>
<td>RDG</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>99.0</td>
</tr>
</tbody>
</table>

**Prediction**

model testing : scale-free networks

- then pass the classifier the real PPIN

<table>
<thead>
<tr>
<th>Rank</th>
<th>Class</th>
<th>Score</th>
<th>Class</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DMC</td>
<td>8.2 ± 1.0</td>
<td>DMC</td>
<td>8.6 ± 1.1</td>
</tr>
<tr>
<td>2</td>
<td>DMR</td>
<td>−6.8 ± 0.9</td>
<td>DMR</td>
<td>−6.1 ± 1.7</td>
</tr>
<tr>
<td>3</td>
<td>RDG</td>
<td>−9.5 ± 2.3</td>
<td>RDG</td>
<td>−9.3 ± 1.6</td>
</tr>
<tr>
<td>4</td>
<td>AGV</td>
<td>−10.6 ± 4.2</td>
<td>AGV</td>
<td>−11.5 ± 4.1</td>
</tr>
<tr>
<td>5</td>
<td>LPA</td>
<td>−16.5 ± 3.4</td>
<td>LPA</td>
<td>−14.3 ± 3.2</td>
</tr>
<tr>
<td>6</td>
<td>SMW</td>
<td>−18.9 ± 0.7</td>
<td>SMW</td>
<td>−18.3 ± 1.9</td>
</tr>
<tr>
<td>7</td>
<td>RDS</td>
<td>−19.1 ± 2.3</td>
<td>RDS</td>
<td>−19.9 ± 1.5</td>
</tr>
</tbody>
</table>

- risks: we sometimes fall in love with our models
- opportunities: statistics offers powerful tools for model testing
- difficulties: requires learning new tools, and bravery

prediction : link prediction

• *how can we evaluate how good a model is?*

• **cross-validation**
  
  hold out some data
  
  fit the model to what remains
  
  quantify model’s ability to predict held-out data

• for networks, this usually means *link prediction*

• to do this well, we use **probabilistic generative models**
Pr($i, j$ connected) = $p_r$

= $P$(lowest common ancestor of $i, j$)
prediction : link prediction

prediction: link prediction

and reproduces motifs and other patterns

degrees

triangles

path lengths
prediction: link prediction

- Link prediction is a **hard** form of validation
- Simple and clear evaluation measure

**Risks**: overfitting
- Cross-validation *not* well-defined for networks we care about more than missing links

**Opportunities**: data driven with up-front assumptions
- Generative models quantify uncertainty, predict missing data

**Difficulties**: usually non-mechanistic (predictive but not explanatory)
- How do we test more complicated predictions?
"It’s easy to lie with statistics, but it’s easier to lie without them." — Fred Mosteller

• statistics are the foundation of a data-driven Network Science.

• exploration — what patterns need to be explained?
• model testing — how well can I capture those patterns?
• prediction — how well can I predict missing / future patterns?

• the BIG risk: we’ll reinvent statistics, slowly, haltingly
• the BIG opportunity: we’ll use modern Statistics to be better scientists, to find truth more quickly, accurately
• the BIG difficulty: Statistics is hard
"The first principle is that you must not fool yourself, and you are the easiest person to fool."
— Richard Feynman