Lecture 9b: Exploration, testing , and prediction

Aaron Clauset
@aaronclauset
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

"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

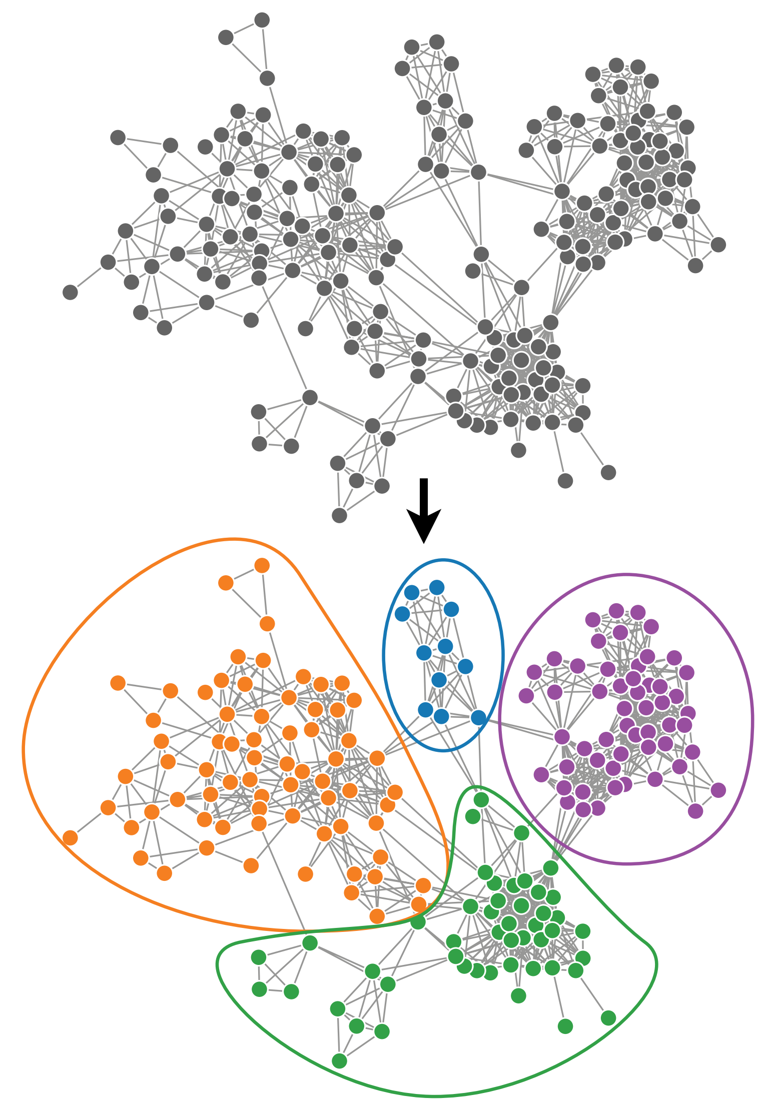
three roles of statistics

- data exploration
- model testing
- prediction

data exploration : community detection

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

- 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.)



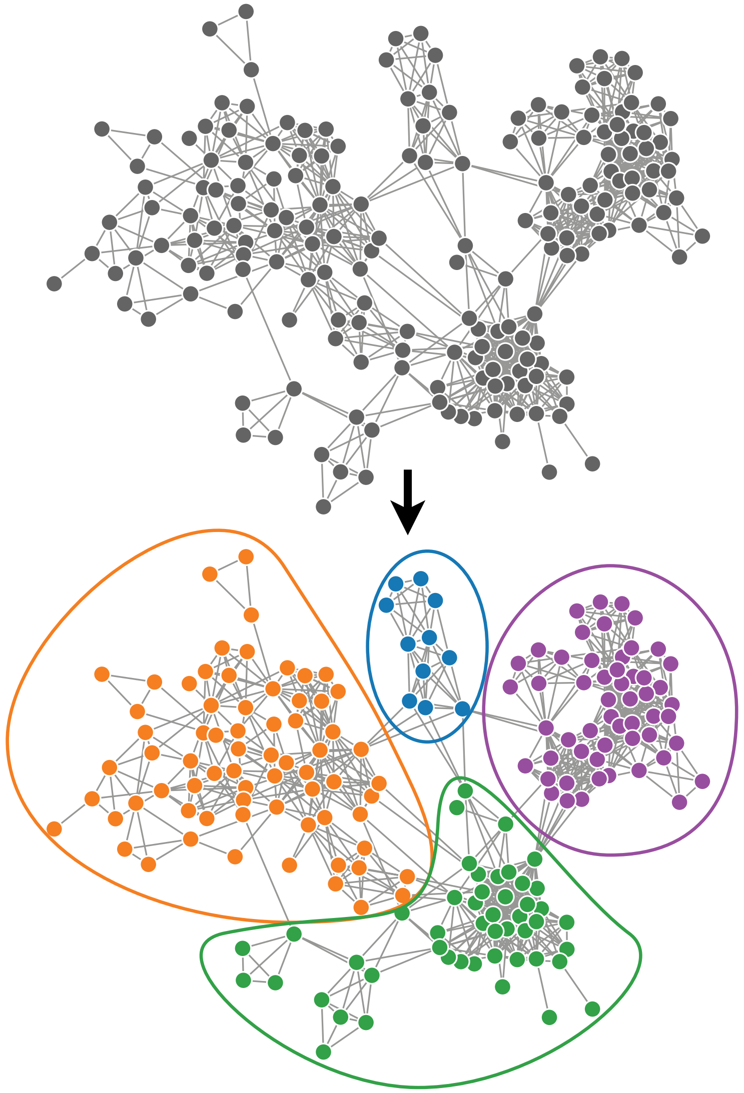
data exploration : community detection

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

- nearly all methods:

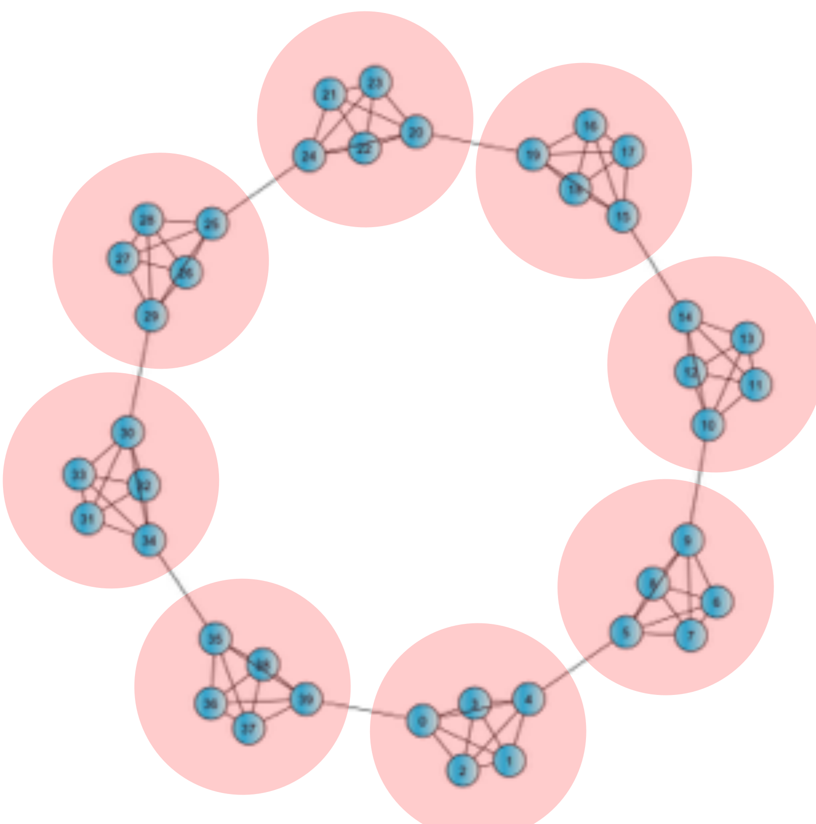
  estimate $\max_z f(z(G))$

  [WARNING: typically NP-hard]



the trouble with community detection

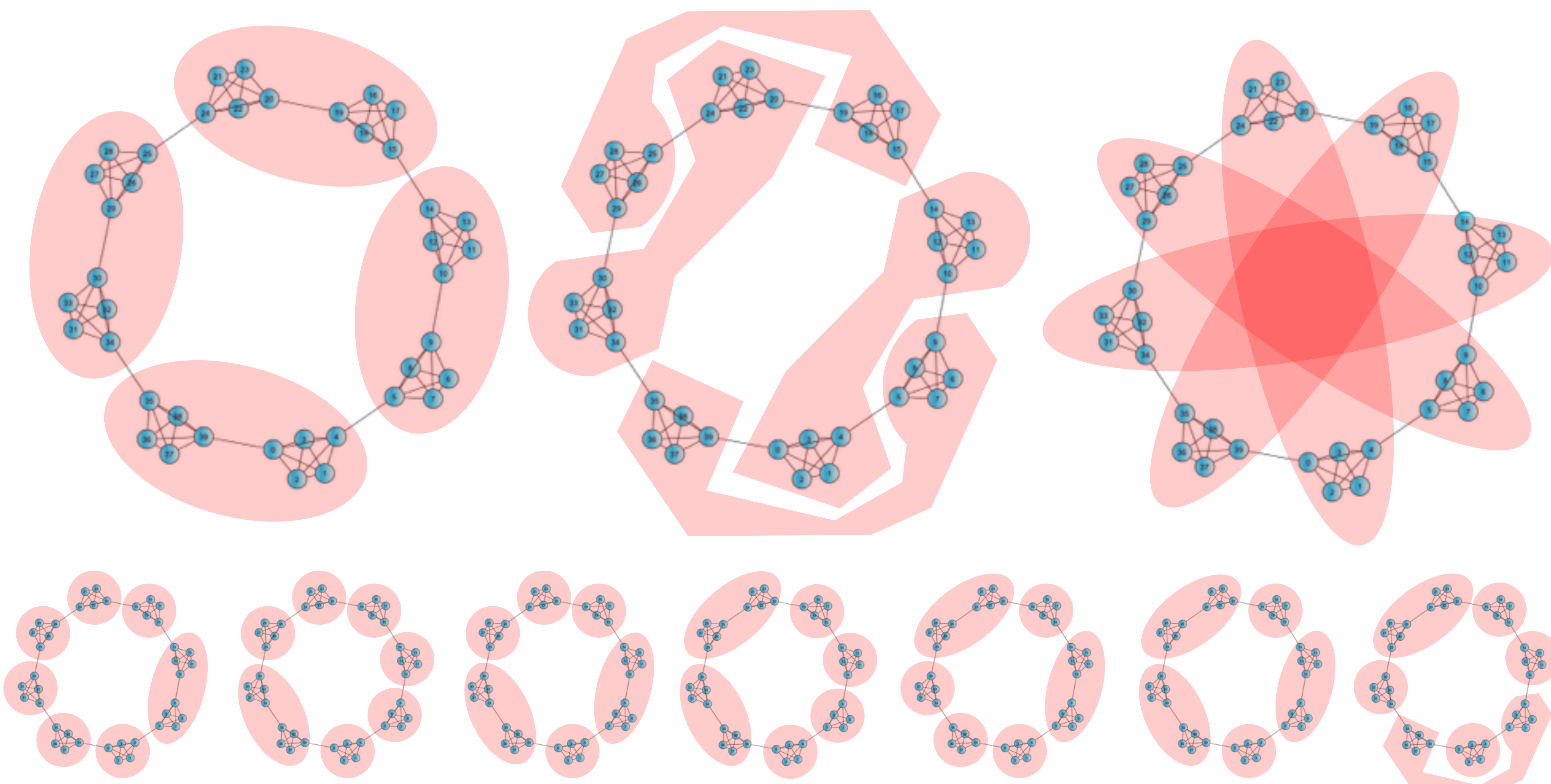
this is a pretty good division (under nearly any f)



B. H. Good, Y.-A. de Montjoye and A. Clauset, "The performance of modularity maximization in practical contexts." *Physical Review E* **81**, 046106 (2010).

data exploration : community detection

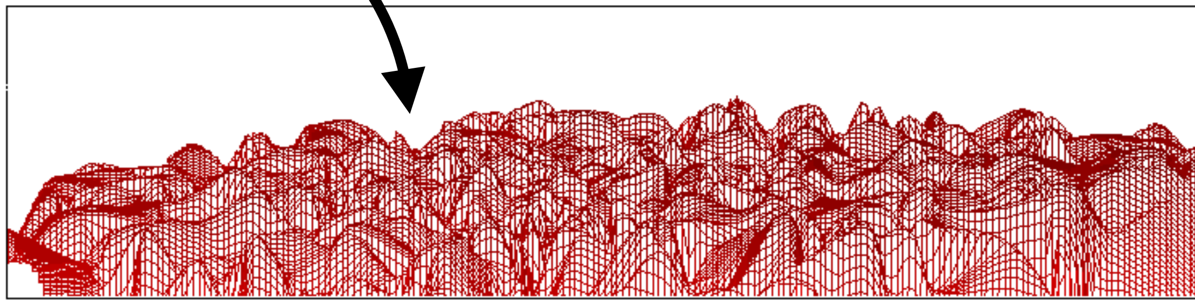
so are all of these (and many more)



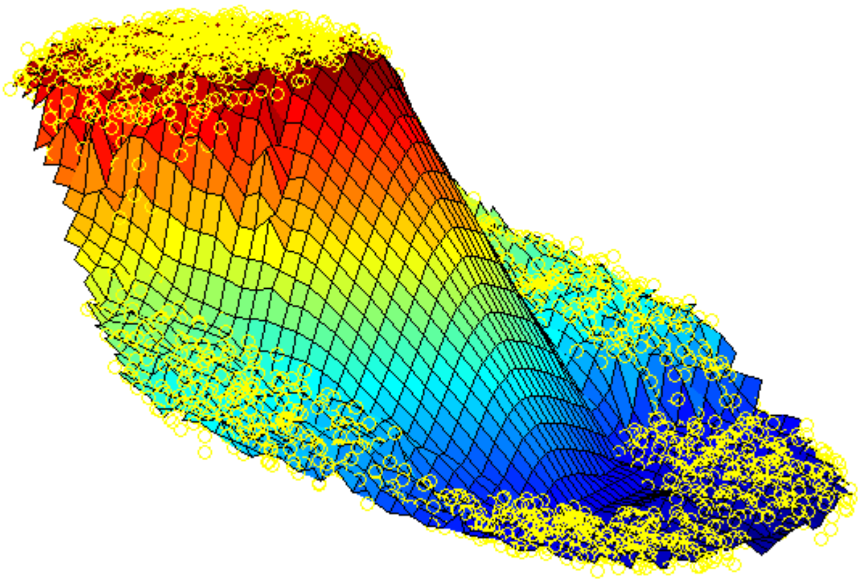
data exploration : community detection

- 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?

B. H. Good, Y.-A. de Montjoye and A. Clauset, "The performance of modularity maximization in practical contexts." *Physical Review E* **81**, 046106 (2010).

model testing : scale-free networks

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

M. Middendorf, E. Ziv and C. H. Wiggins, *Proc. Natl. Acad. Sci. USA* **102**(9), 319203197 (2005).

model testing : scale-free networks

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

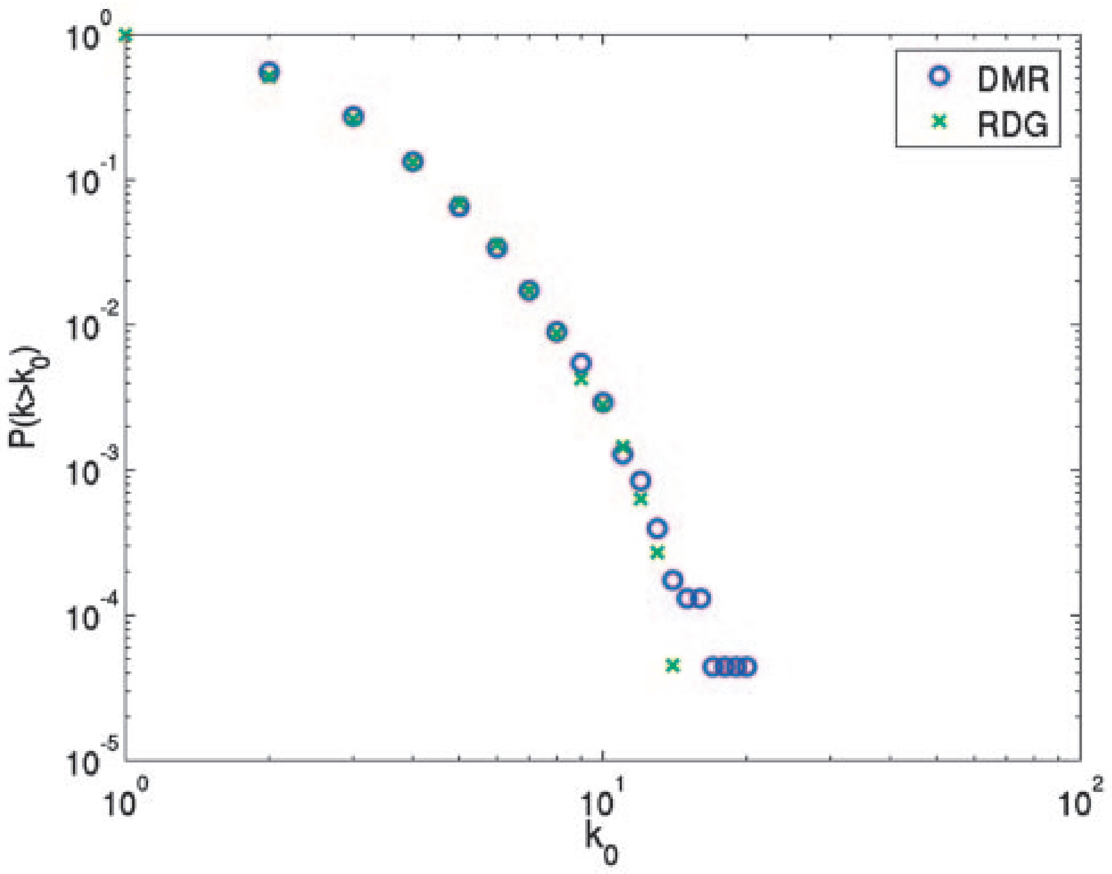
Manuel Middendorf[†], Etay Ziv[‡], and Chris H. Wiggins[§¶||]

- *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)

M. Middendorf, E. Ziv and C. H. Wiggins, *Proc. Natl. Acad. Sci. USA* **102**(9), 319203197 (2005).

model testing : scale-free networks

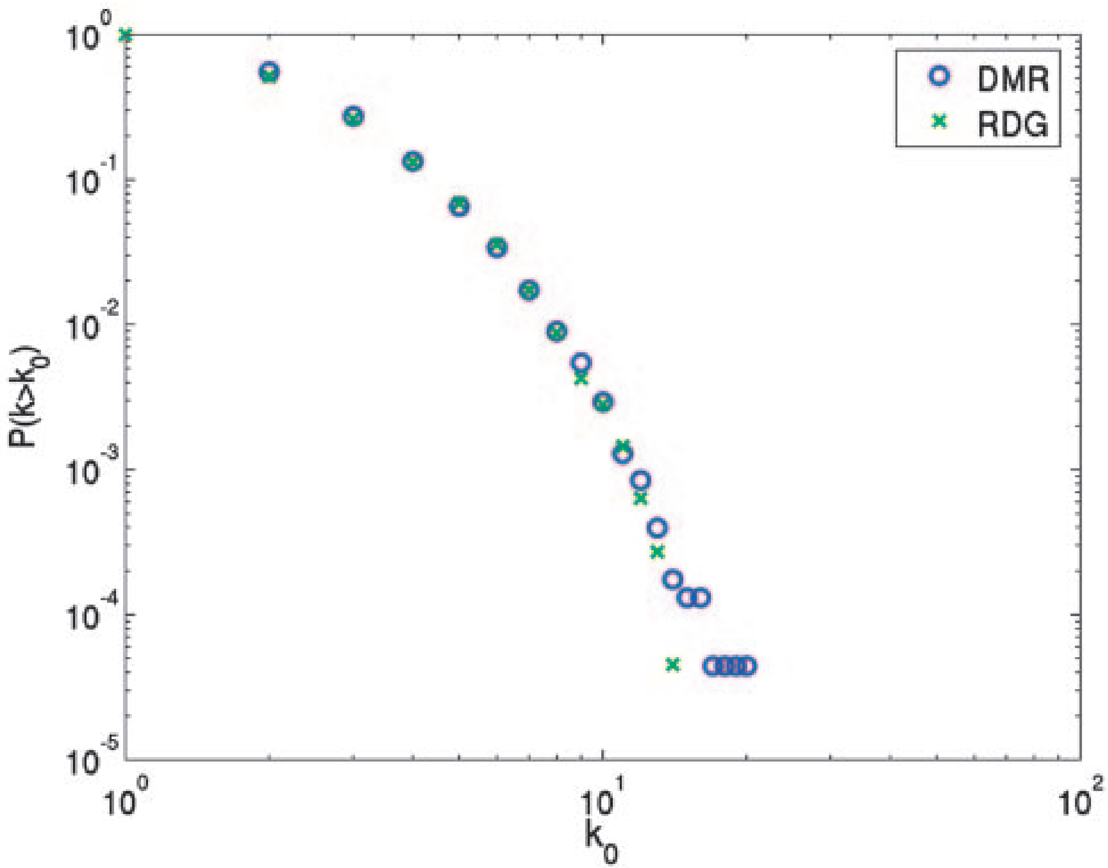
- *the problem:* all models fit the observed degree distribution

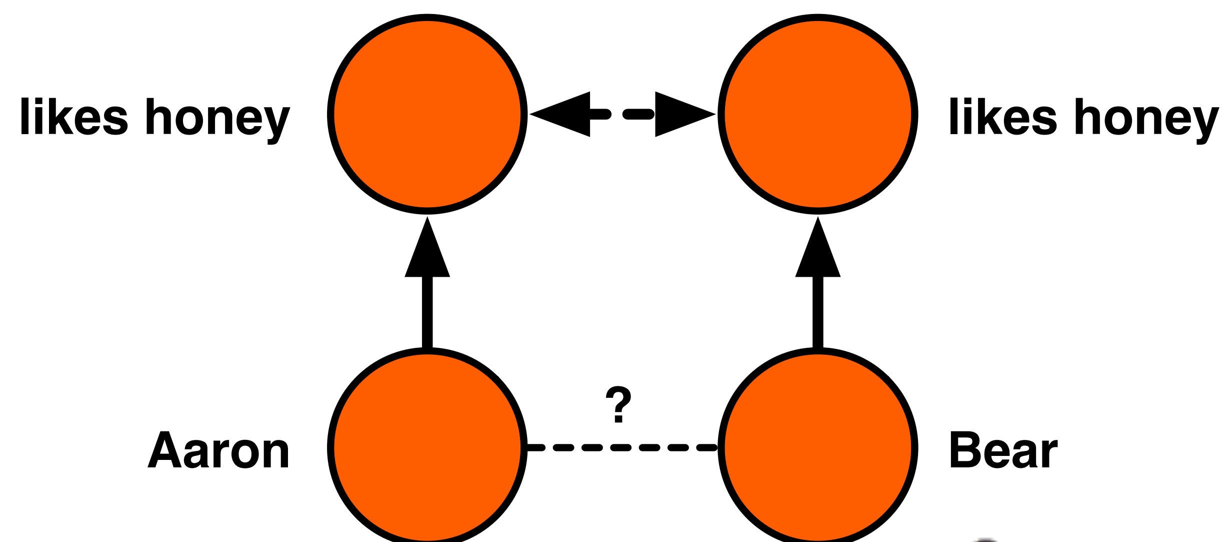


M. Middendorf, E. Ziv and C. H. Wiggins, *Proc. Natl. Acad. Sci. USA* **102**(9), 319203197 (2005).

model testing : scale-free networks

- *the problem:* all models fit the observed degree distribution

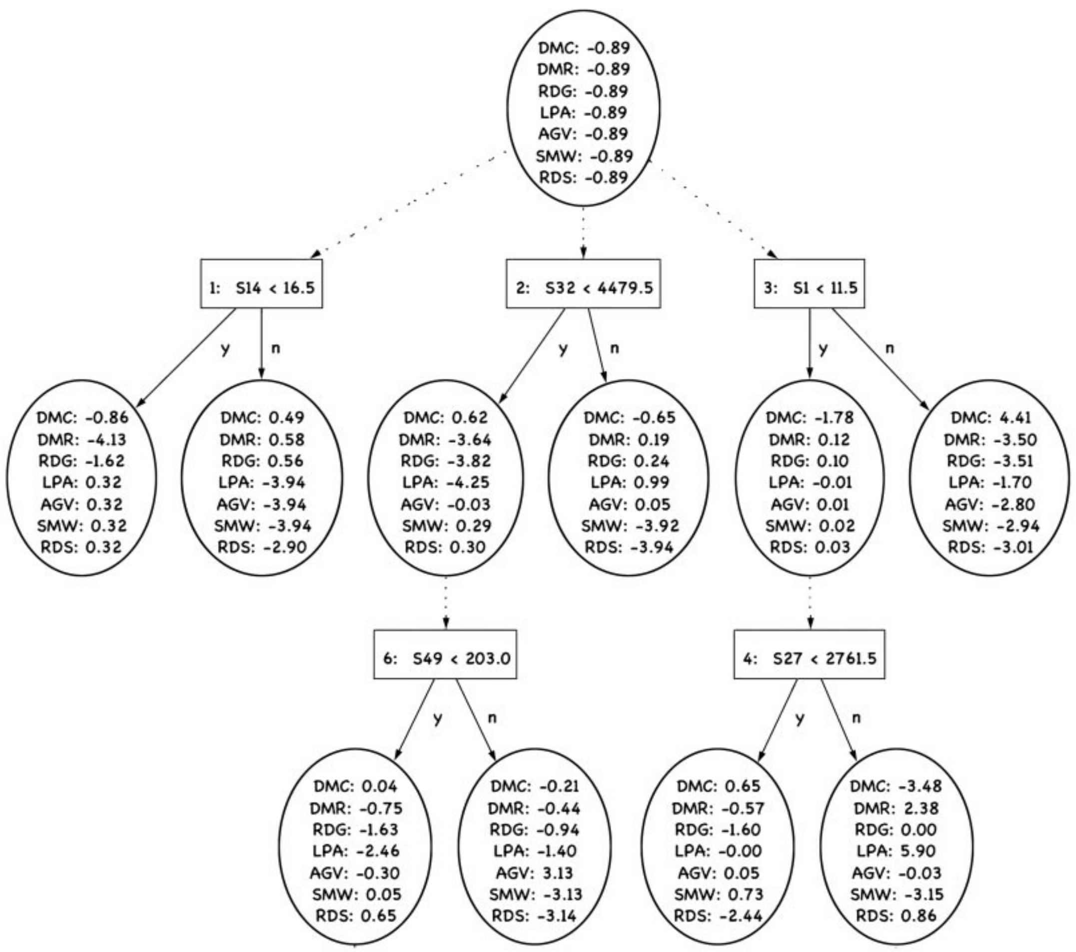




M. Middendorf, E. Ziv and C. H. Wiggins, *Proc. Natl. Acad. Sci. USA* **102**(9), 319203197 (2005).

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:



	Prediction						
Truth	DMR	DMC	AGV	LPA	SMW	RDS	RDG
DMR	99.3	0.0	0.0	0.0	0.0	0.1	0.6
DMC	0.0	99.7	0.0	0.0	0.3	0.0	0.0
AGV	0.0	0.1	84.7	13.5	1.2	0.5	0.0
LPA	0.0	0.0	10.3	89.6	0.0	0.0	0.1
SMW	0.0	0.0	0.6	0.0	99.0	0.4	0.0
RDS	0.0	0.0	0.2	0.0	0.8	99.0	0.0
RDG	0.9	0.0	0.0	0.1	0.0	0.0	99.0

M. Middendorf, E. Ziv and C. H. Wiggins, *Proc. Natl. Acad. Sci. USA* **102**(9), 319203197 (2005).

model testing : scale-free networks

- *then pass the classifier the real PPIN*

	Eight-step subgraphs ($p^* = 0.65$)		Subgraphs with up to seven edges ($p^* = 0.65$)	
Rank	Class	Score	Class	Score
1	DMC	8.2 ± 1.0	DMC	8.6 ± 1.1
2	DMR	-6.8 ± 0.9	DMR	-6.1 ± 1.7
3	RDG	-9.5 ± 2.3	RDG	-9.3 ± 1.6
4	AGV	-10.6 ± 4.2	AGV	-11.5 ± 4.1
5	LPA	-16.5 ± 3.4	LPA	-14.3 ± 3.2
6	SMW	-18.9 ± 0.7	SMW	-18.3 ± 1.9
7	RDS	-19.1 ± 2.3	RDS	-19.9 ± 1.5

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

M. Middendorf, E. Ziv and C. H. Wiggins, *Proc. Natl. Acad. Sci. USA* **102**(9), 319203197 (2005).

prediction : link prediction

- *how can we evaluate how good a model is?*
- **cross-validation**

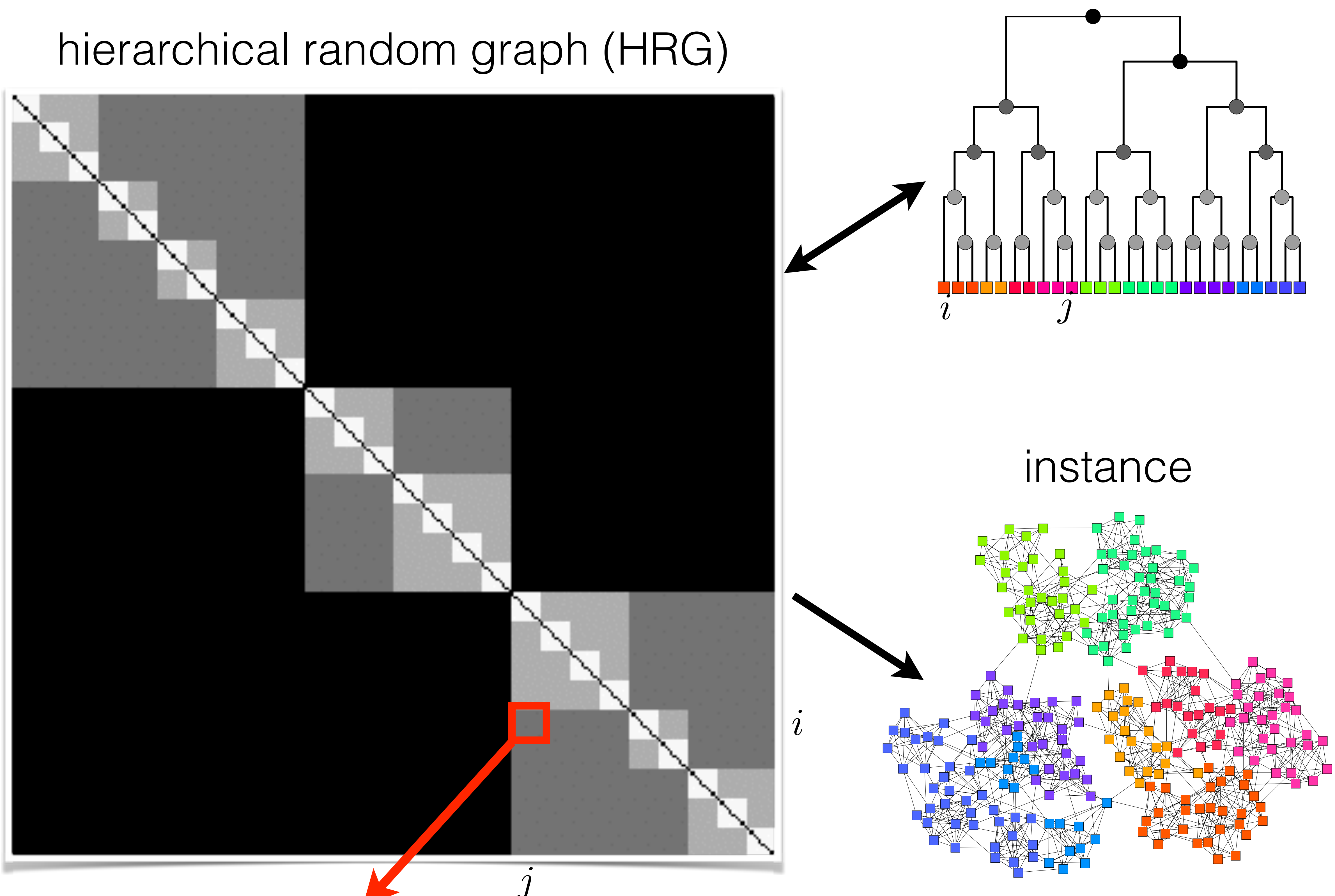
  hold out some data

  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**

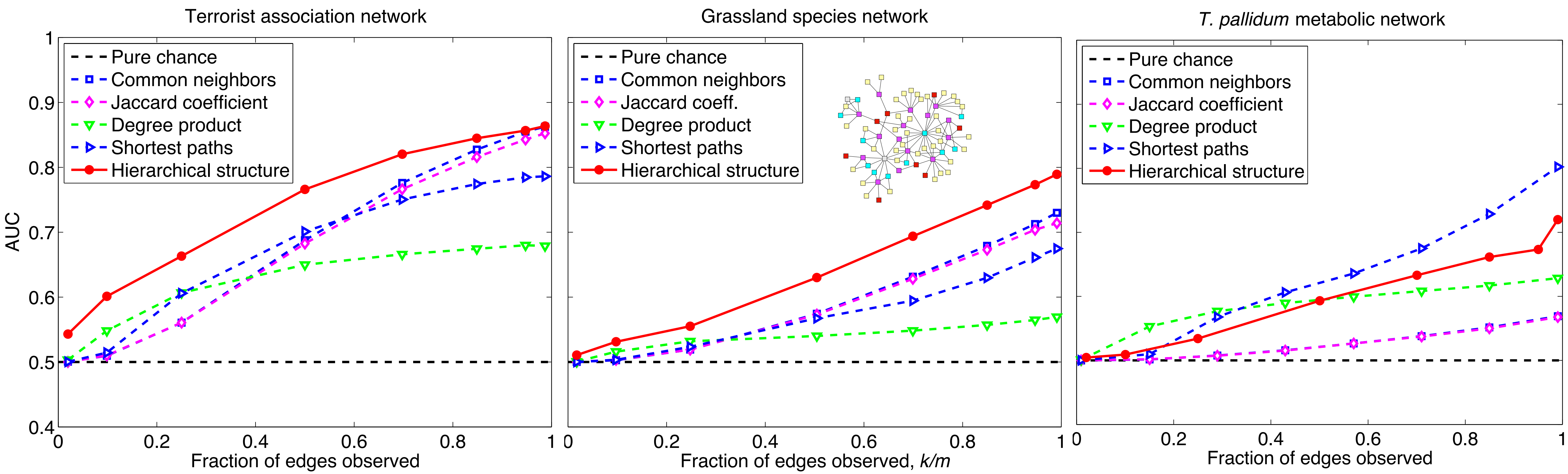
hierarchical random graph (HRG)

model



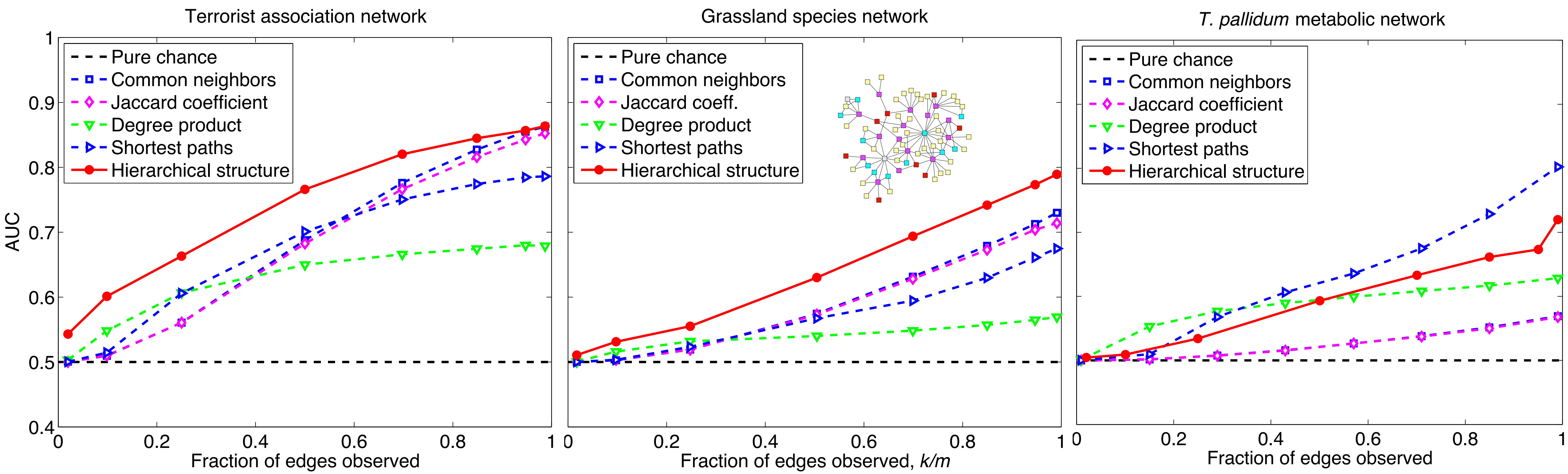
$$\Pr(i, j \text{ connected}) = p_r$$
$$= p_{(\text{lowest common ancestor of } i,j)}$$

prediction : link prediction

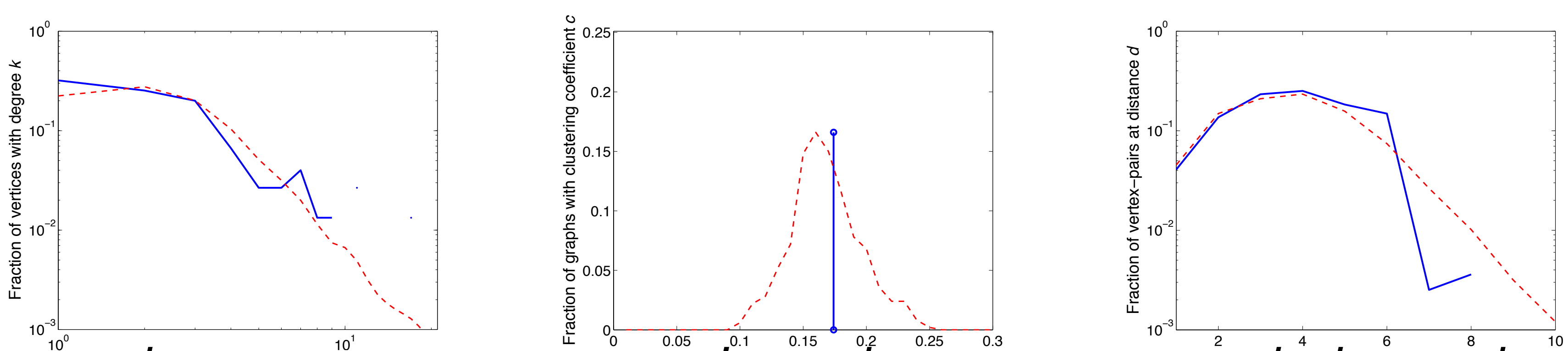


A. Clauset, C. Moore and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks." *Nature* **453**, 98 - 101 (2008).

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

fin