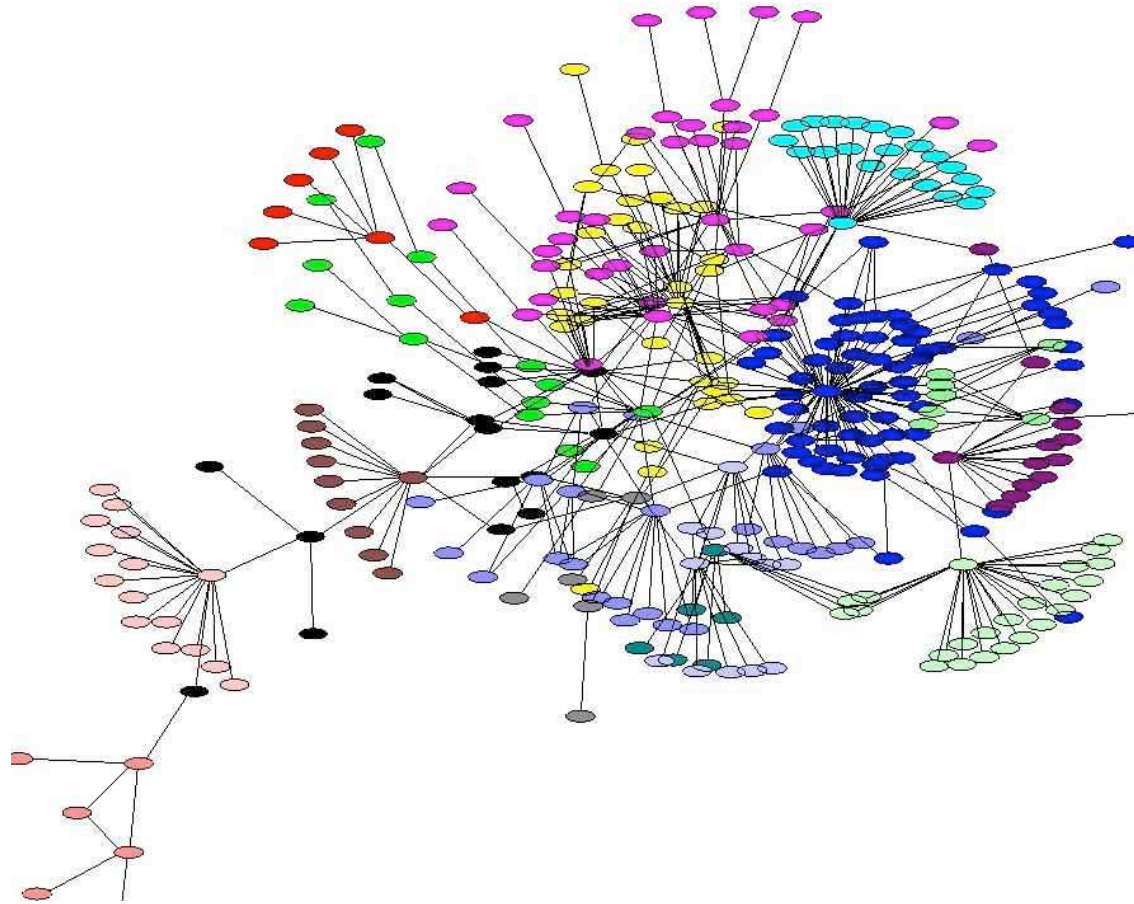


ECS 253 / MAE 253

May 24, 2023



“Intro to Biological Networks, Motifs, and Model selection/validation”

Biological Networks / Systems Biology

** In Biology, often edges can be activating or inhibiting! **

- Luca Cardelli, <http://lucacardelli.name/>
- Eivind Almaas, <http://www.ntnu.no/ansatte/eivind.almaas>
- Sergey Nuzhdin, <http://nlab.usc.edu/Site/Home.html>

At UCD:

- CS: Vladimir Filkov, Ilias Tagkopoulos, Patrice Koehl, Dan Gusfield, Dave Doty,
- Genome center / Biomed engineering: Savageau, Benham, Raychaudhuri, Saiz, Brady, Feihn
- Plant Biology: Dandekar, Maloof, ...

Intro to biological networks

See [almaasBioNets.pdf](#)

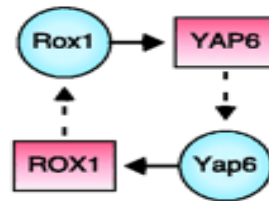
Network motifs

Network Analysis Detects Component Reuse (Network Motifs)

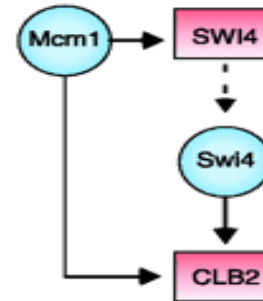
Autoregulation



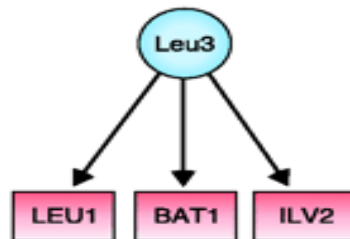
Multi-Component Loop



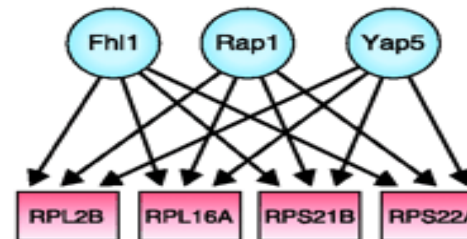
Feedforward Loop



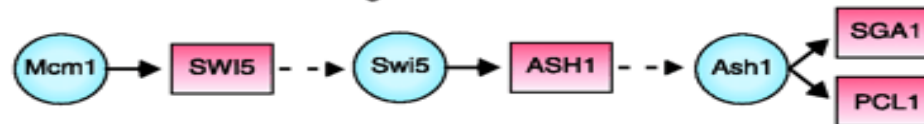
Single Input Motif



Multi-Input Motif



Regulator Chain



Examples of Biological Networks?

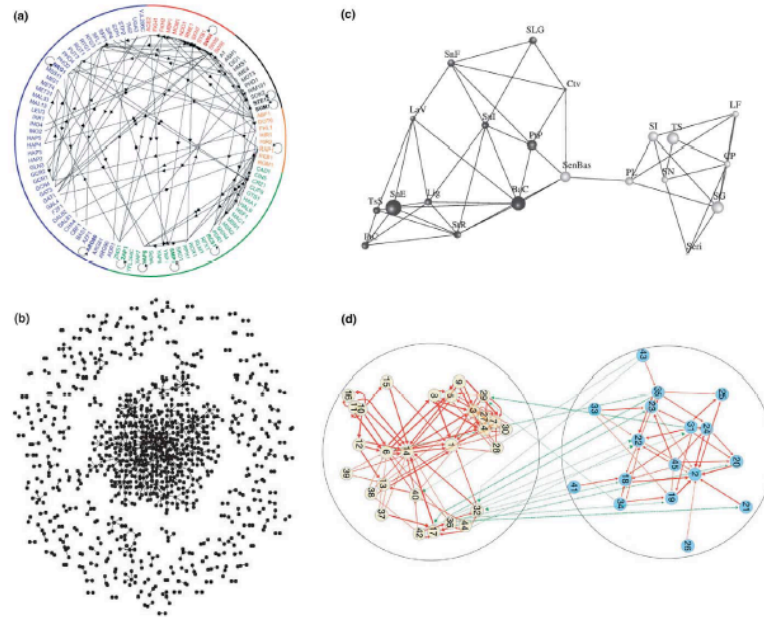
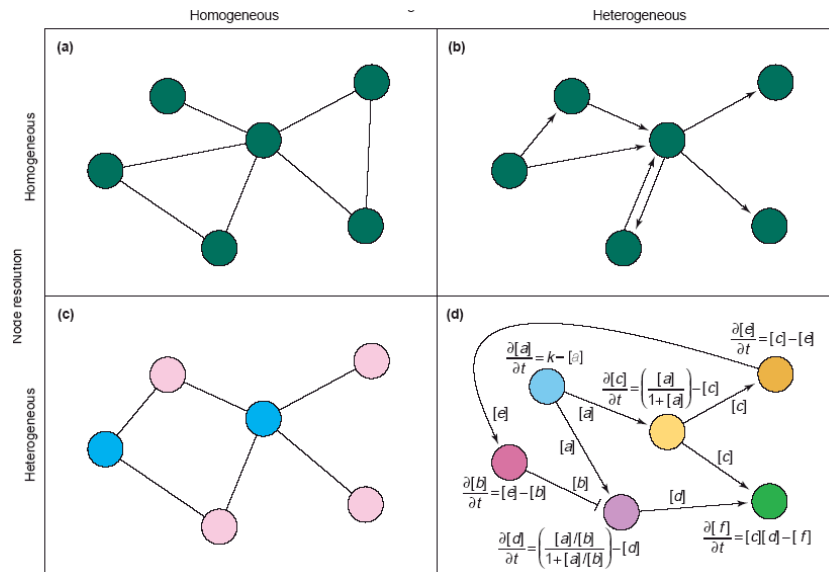
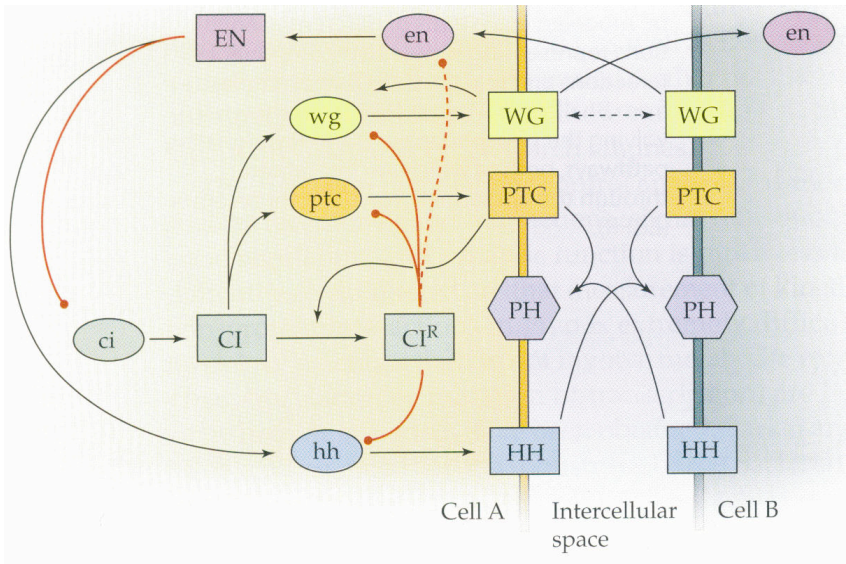
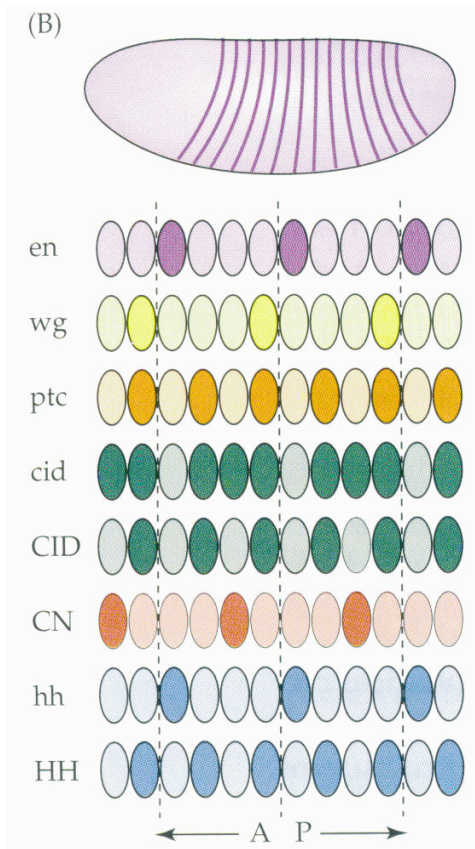


Figure 1. The use of network concepts to explore the structure and function of a variety of biological systems from genes (a) and proteins (b) to individuals within a population (c) and species within an ecosystem (d). (a) The network of regulatory interactions in the yeast *Saccharomyces cerevisiae*, where genes encoding transcription factors interact by binding the regulatory regions of other regulatory genes [16]. (b) The protein interaction network in which proteins that physically interact are connected by edges [17]. (c) The genetic relationship of populations of the cactus *Lophocereus schottii* [18]. In this graph, edge length represents the fraction of the total genetic variation explained by the connected populations. (d) Predator-prey interactions in the Chesapeake Bay food web [19]. Reproduced, with permission from [16] (a), [17] (b), [18] (c) and [19] (d).

c. What are appropriate network descriptions?

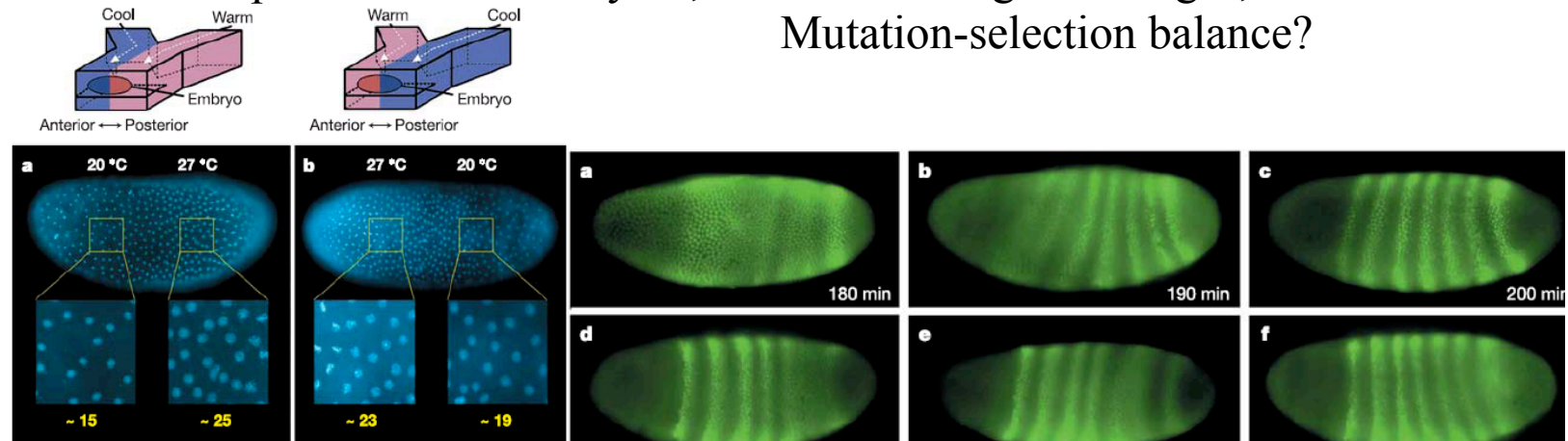




dynamic description
 Garry O'Dell, Nature 2001:
 19 equations, 54 parameters,

1/2000 random solutions → correct spatial structure.

Most parameters can vary 10,000 fold causing no changes,
 Mutation-selection balance?



Literature on validation of network models

- is rather limited
-

Four useful papers:

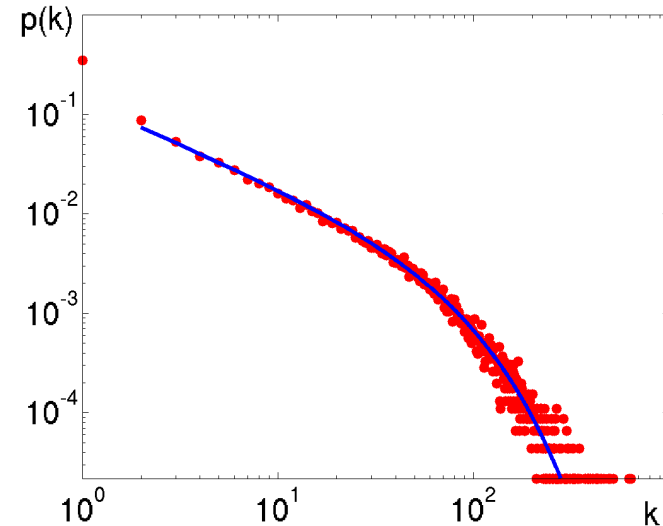
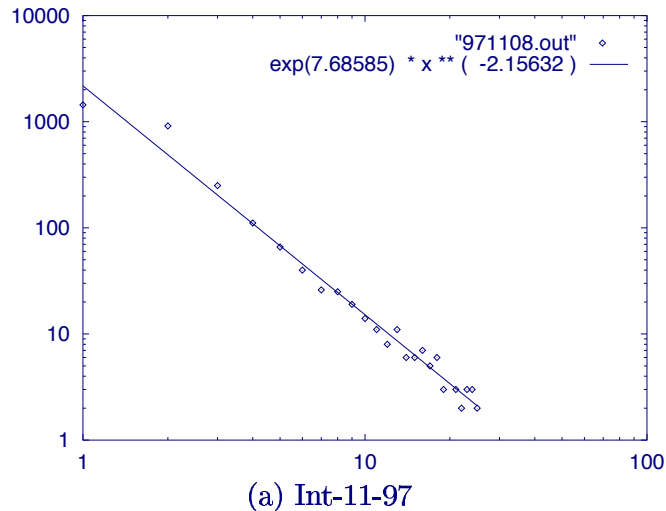
- M. Middendorf, E. Ziv, and C. H. Wiggins, “Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network”, *PNAS* **102** (9), 2005. (About 180 citations.)
- D. Alderson, L. Li, W. Willinger, and J. C. Doyle, “Understanding Internet Topology: Principles, Models, and Validation”, *IEEE/ACM Trans. on Networking*, **13** (6), 2005. (About 170 citations.)
- V. Filkov, Z.M. Saul, S. Roy, R.M. DÕSouza, P.T. Devanbu, “Modeling and verifying a broad array of network properties”, *Europhys. Lett.* **86**, 2009.
- J. Wang and G. Provan, “Generating Application-Specific Benchmark Models for Complex Systems”, *Proc. Twenty-Third AAAI Conf on Artificial Intelligence*, 2008.

Model validation: Overarching issues

- Many models give rise to same large-scale statistics (e.g., degree distribution, diameter, clustering coefficient).
- Data sets have multiple attributes. Fitting one or two of them is not always sufficient.
- Data: Limited availability (expense or proprietary nature); small data sets

In the beginning – Power Laws

- 1999 - 2005, explosion of observations of “power laws” in networks also of “small-worlds”.



- M. Mitzenmacher, “The Future of Power Law Research” *Internet Mathematics*, **2** (4), 2006. (Editorial piece)
 - A call to move beyond observation and model building to validation and control.
 - Power laws ‘the signature of human activity’
- Clauset, Shalizi, Newman, “Power-law distributions in empirical data”, *SIAM Review* 51, 661-703 (2009).
 - Techniques to detect if actually have a power law, and if so, to extract exponents.

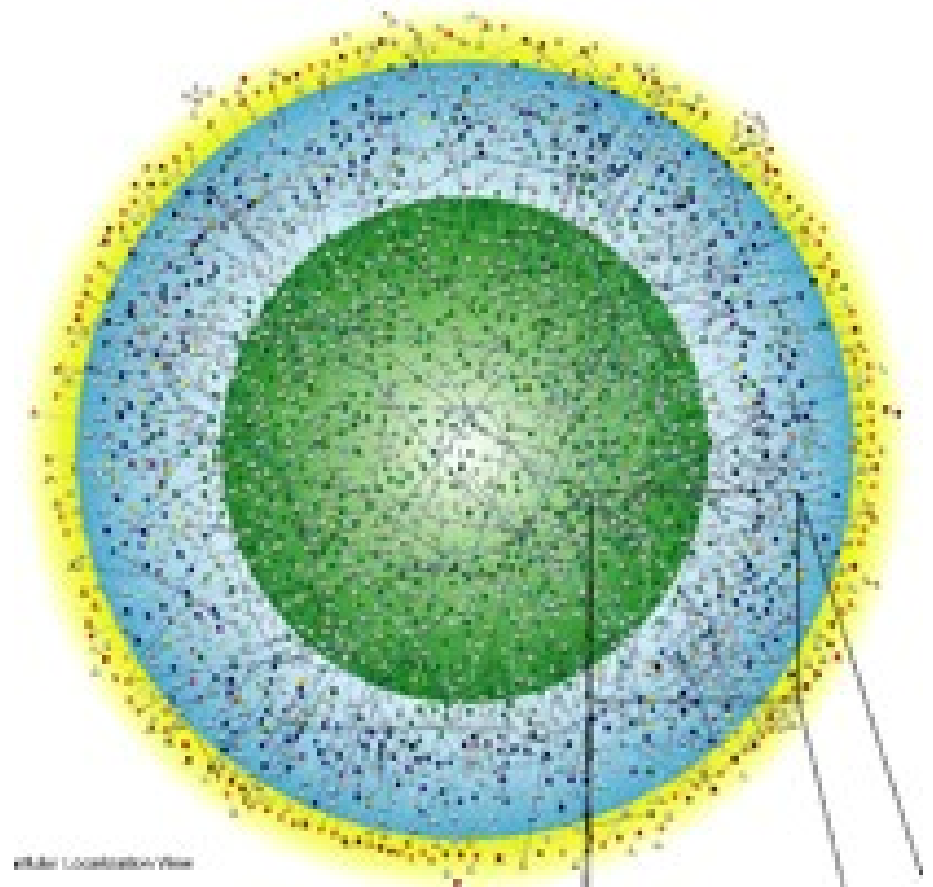
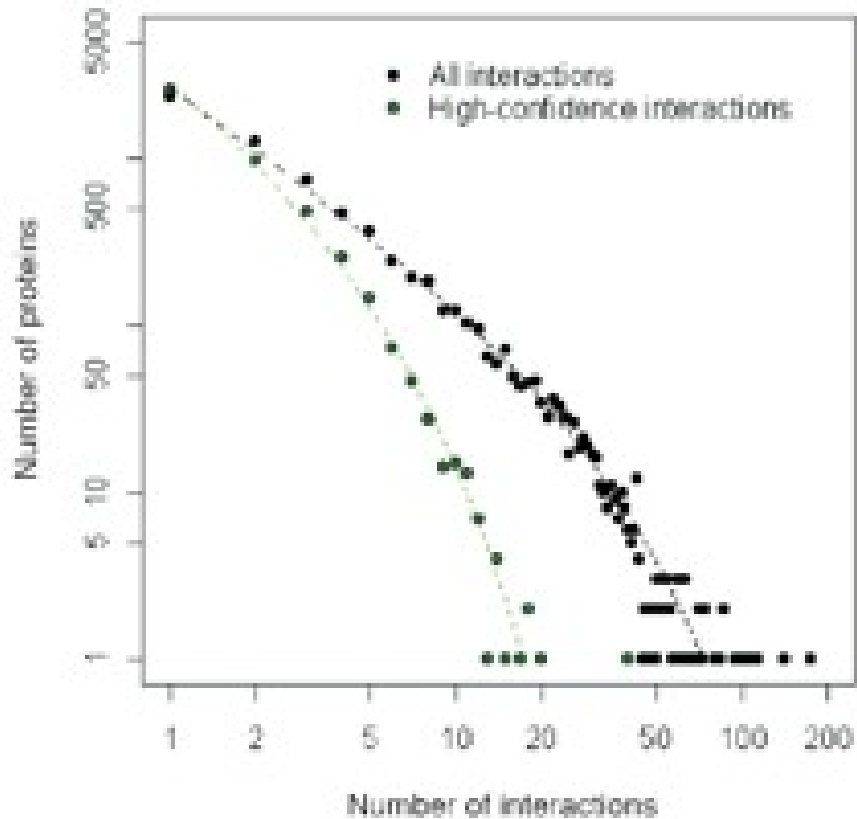
“Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network”

Middendorf, Ziv, and Wiggins *PNAS* **102**, 2005

- Study the *Drosophila* protein interaction network
- Use machine learning techniques (*discriminative classification*) to compare with seven proposed models to determine which model best describes data.
- *Classification* rather than *statistical tests* on specific attributes.

Data:

Giot et al, Science 302, 1727 (2003)



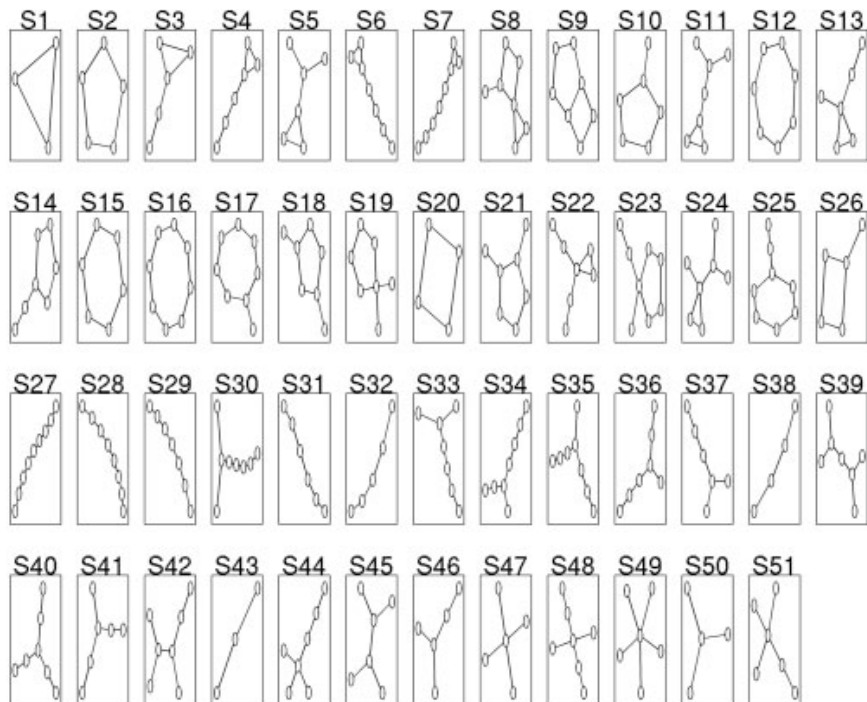
- Accept any edge with $p > 0.65$,
3,359 vertices and 2,795 edges.

7 candidate models

- DMC – duplication-complementation-mutation (Vasquez et al)
- DMR – duplication-mutation with random mutations
- RDS – random static (Erdos-Renyi)
- RDG – random growing graph (Callaway et al.)
- LPA – Linear pref attachment (Barabasi-Albert)
- AGV – Aging vertices
- SMW – Small world (Watts-Strogatz)

The procedure

- Generate 1000 random instances of a network with $N=3359$ and $E=2795$ for each of the seven models (7000 random instances in total). (Training data)
- “Subgraph census” – classify each network by exhaustive search for all possible subgraphs up to a given size. (“Motifs”)
- Classify each of the 7 mechanisms by raw subgraph counts.



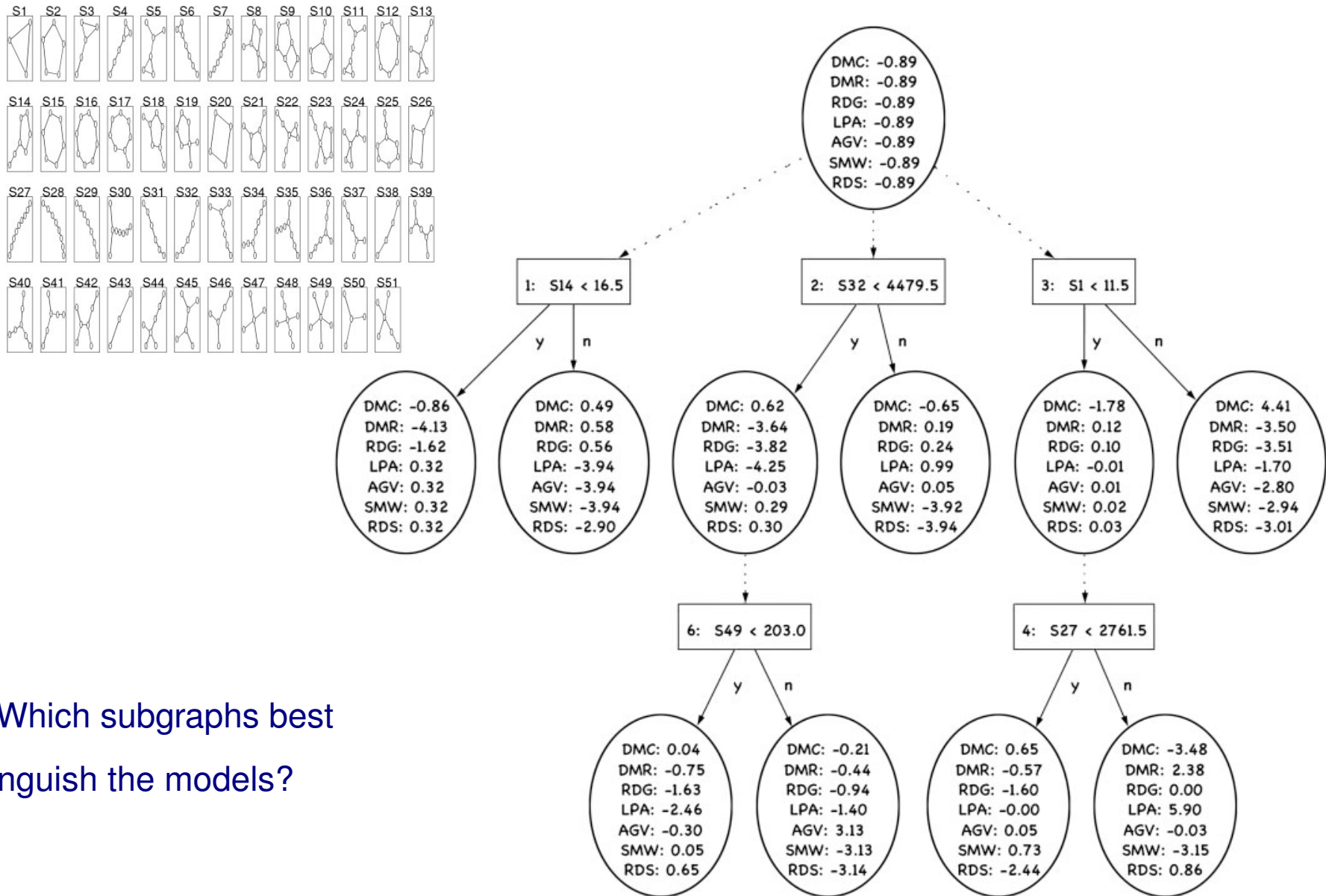
(Example subgraphs)

Notes on procedure

- Similar to techniques in social sciences (p^* , exponential random graph models).
- Network “motifs”, Milo et al *Science*, 2002. But motifs only up to $n = 3$ or $n = 4$ nodes.
- Note the term “clustering” here refers to machine learning technique to categorize data, not “clustering coefficient” (transitivity).

Build classifier from the training data (Learning Algorithm)

- Alternating Decision Tree (ADT), (Freund and Schapire, 1997).



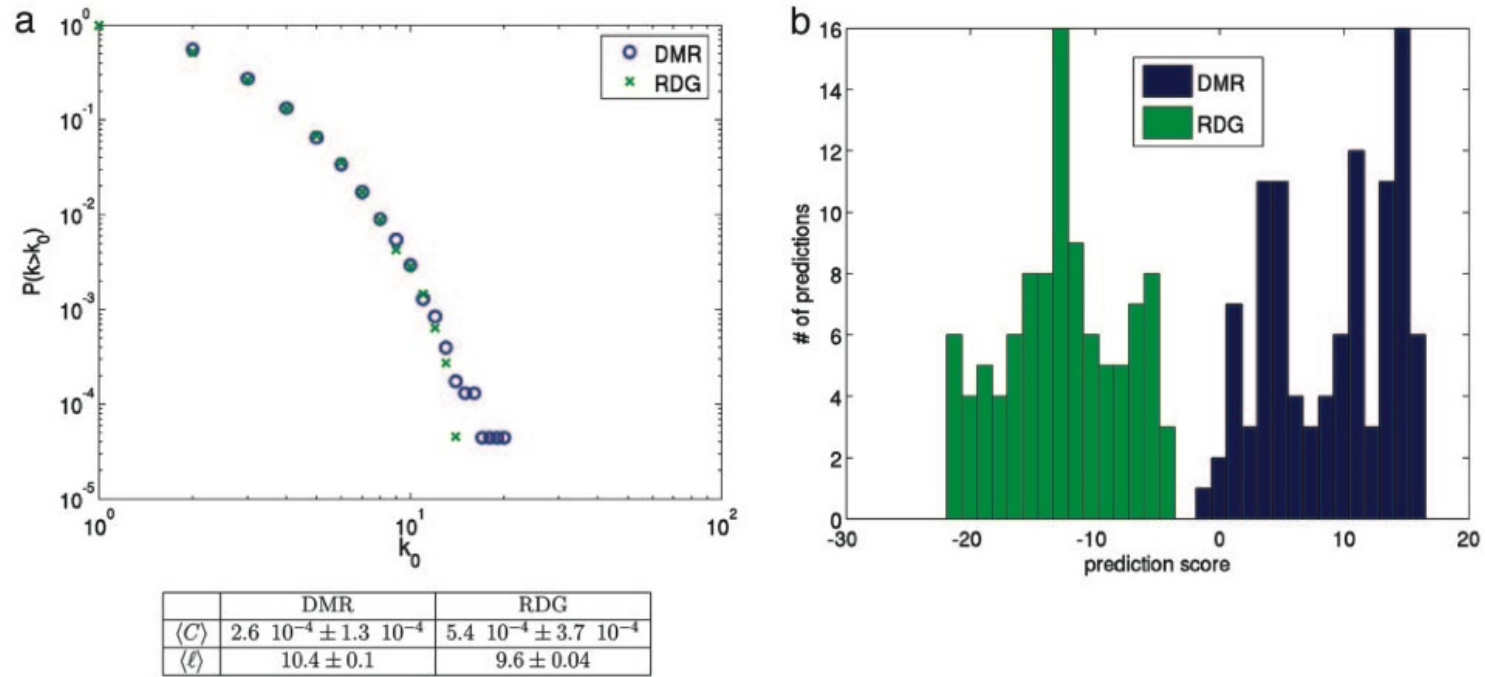
Which subgraphs best distinguish the models?

Validating classifier

Truth	Prediction						
	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

- Slight overlap in models which are variations on one-another.

Validating classifier



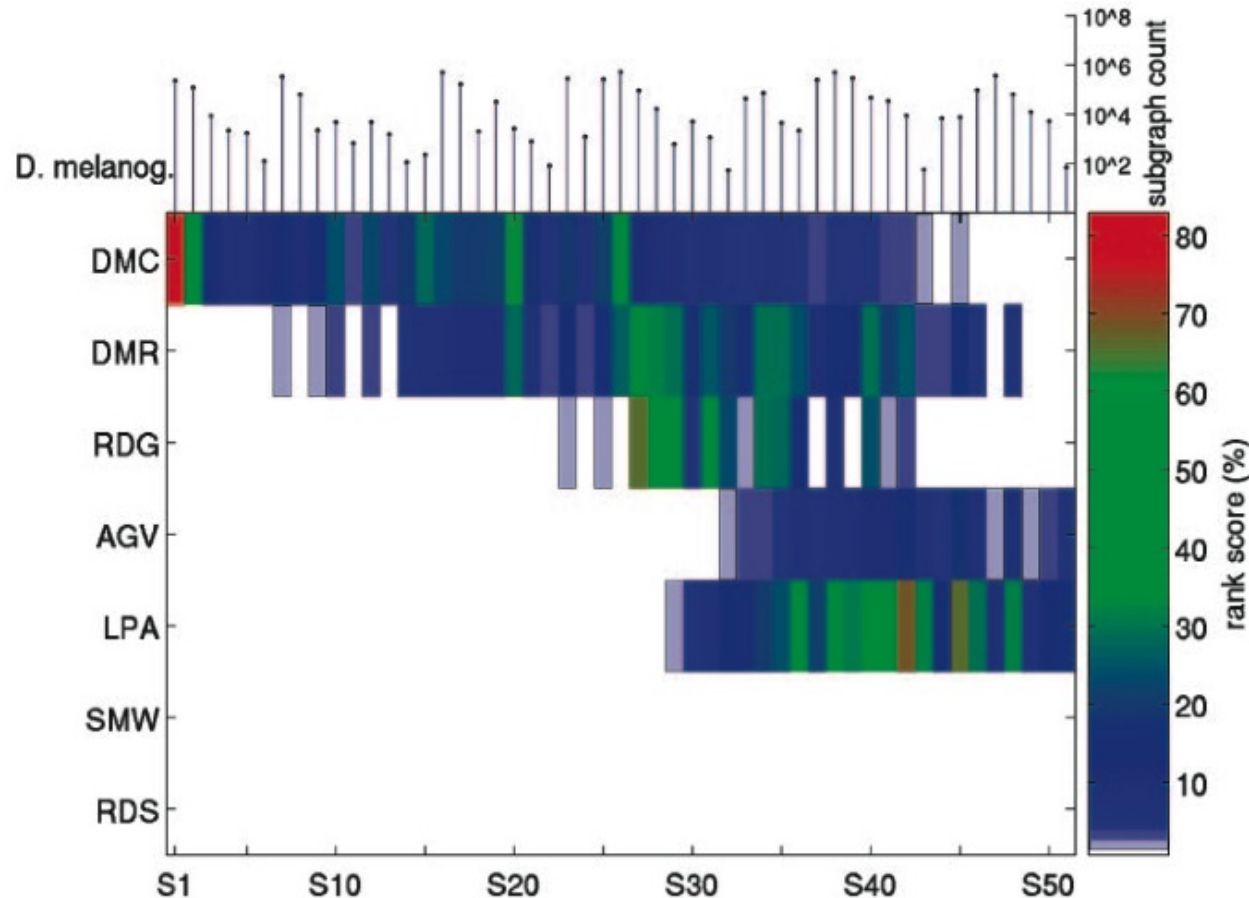
- (a) DMC and RDG produce similar statistical distributions.
- (b) Classifier can discriminate between the two models.

After classifier built, use it to characterize individual network realizations

(Walk the Drosophila data through the ADT)

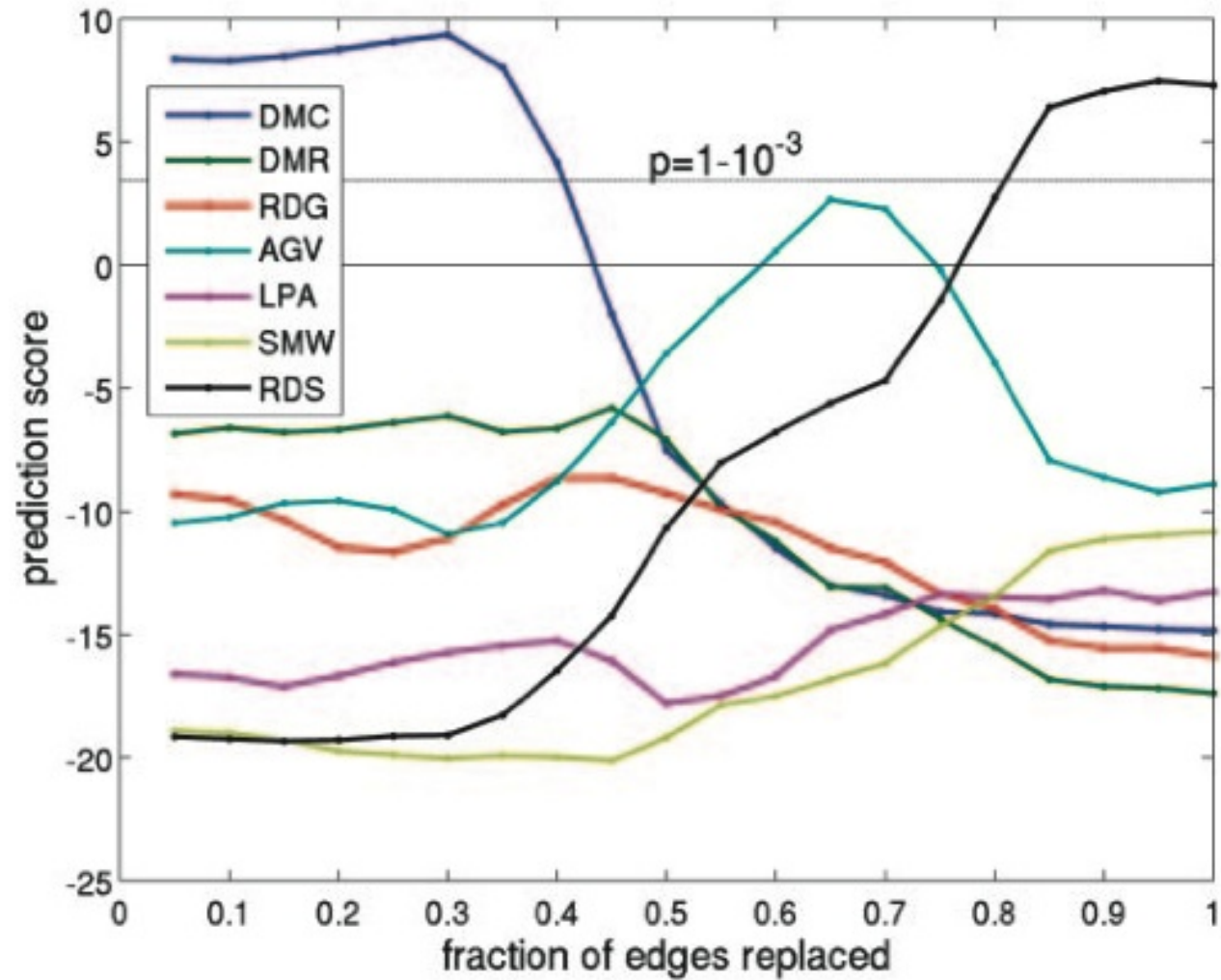
- A given network's subgraph counts determine paths in the ADT (decision nodes are rectangles)
- The ADT outputs a real-valued prediction score, which is the sum of all weights over all paths.
- The final weight for a model is related to probability that particular network realization was generated by that model.
- Model with the highest weight wins (best describes that particular network realization).
- **DMC wins for Giot Drosophila data!**

Comparison by subgraph counts



- Green is best (same median occurrence as in real *Drosophila* data).
- 0 means the subgraph is in data, but not in model.

Introducing noise – degree preserving edge rewiring



- Classifier robust.

Comments

- **Model selection not validation.** (Relative judgement) (i.e., which of these 7 models fits the data best?)
- Many of these 7 models considered produce similar macroscopic features (degree distribution, clustering, diameter, etc).
- Delve into microscopic details and let the data distinguish between the 7 models.
- **Must start with models that are accurate statistical fits to data!** (different type of model validation). (Acompanying commentary, Rice et al PNAS 2005, DMC does not reproduce giant component.)

Model Validation Lit Review: Conclusions

- New techniques being introduced (classifiers, PCA).
- Calls for necessity of validation (e.g., Mitzenmacher)
- Specifics may matter, constraint curves “first principles”.
- **Selection easier than validation!**