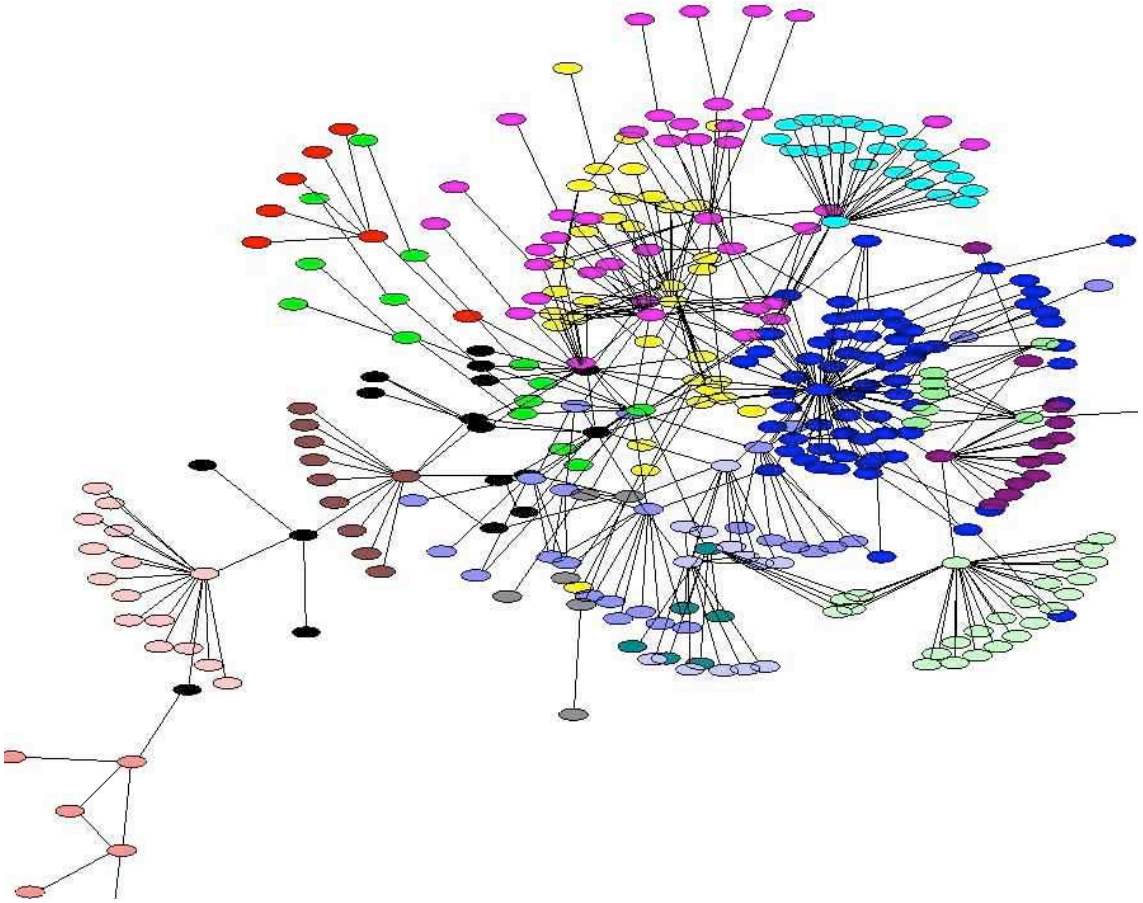
MAE 298, Lecture 16

March 10, 2008



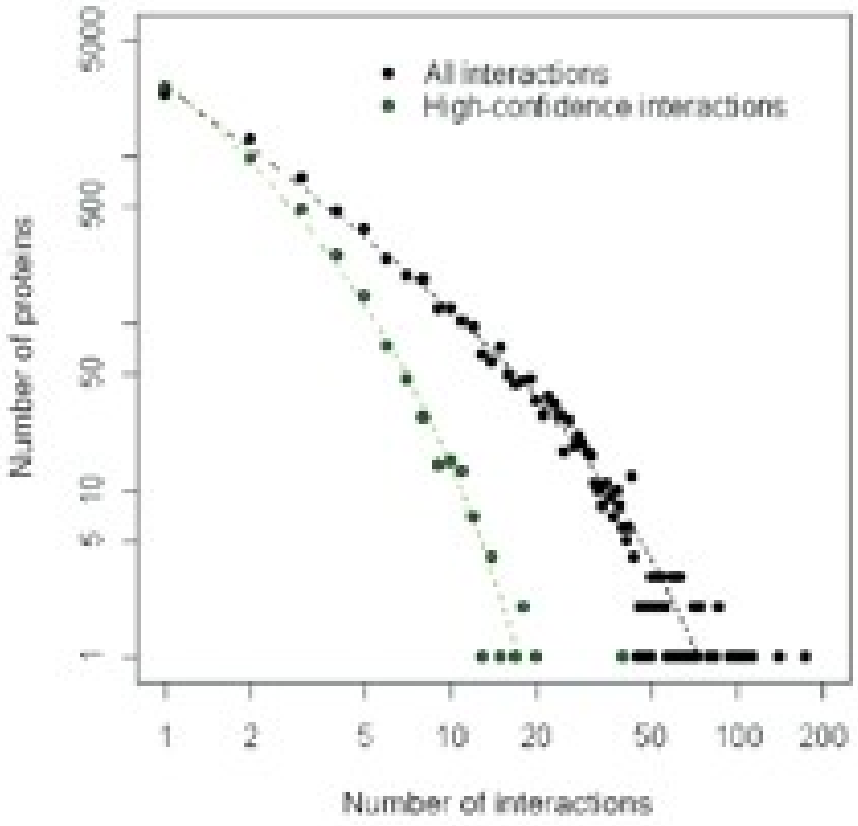
“Techniques for Model Selection”

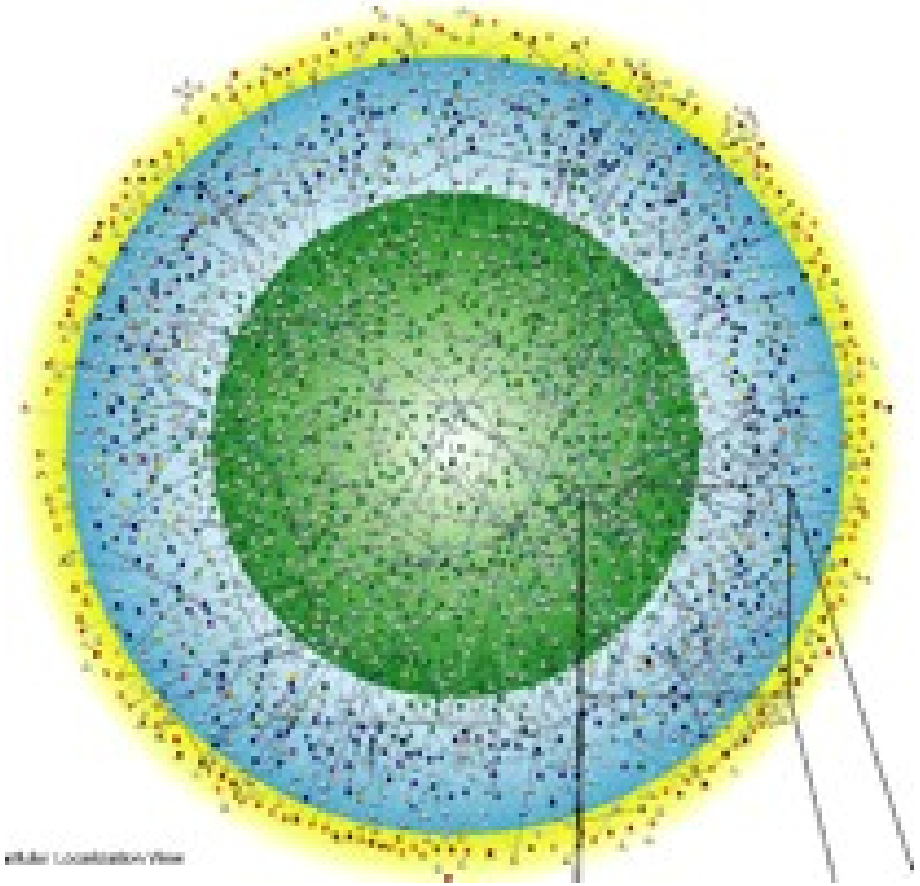
"Inferring network mechanisms: The Drosophila melanogaster protein interaction network"

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.

Data:
Giot et al, Science 302, 1727 (2003)





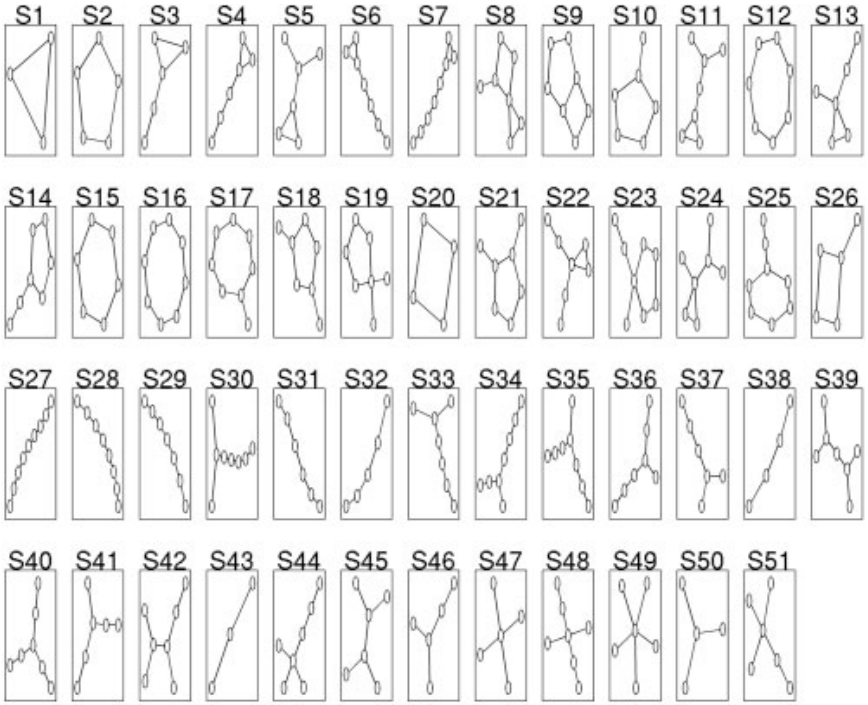
- 3,359 vertices and 2,795 edges.

7 candidate models

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

The procedure

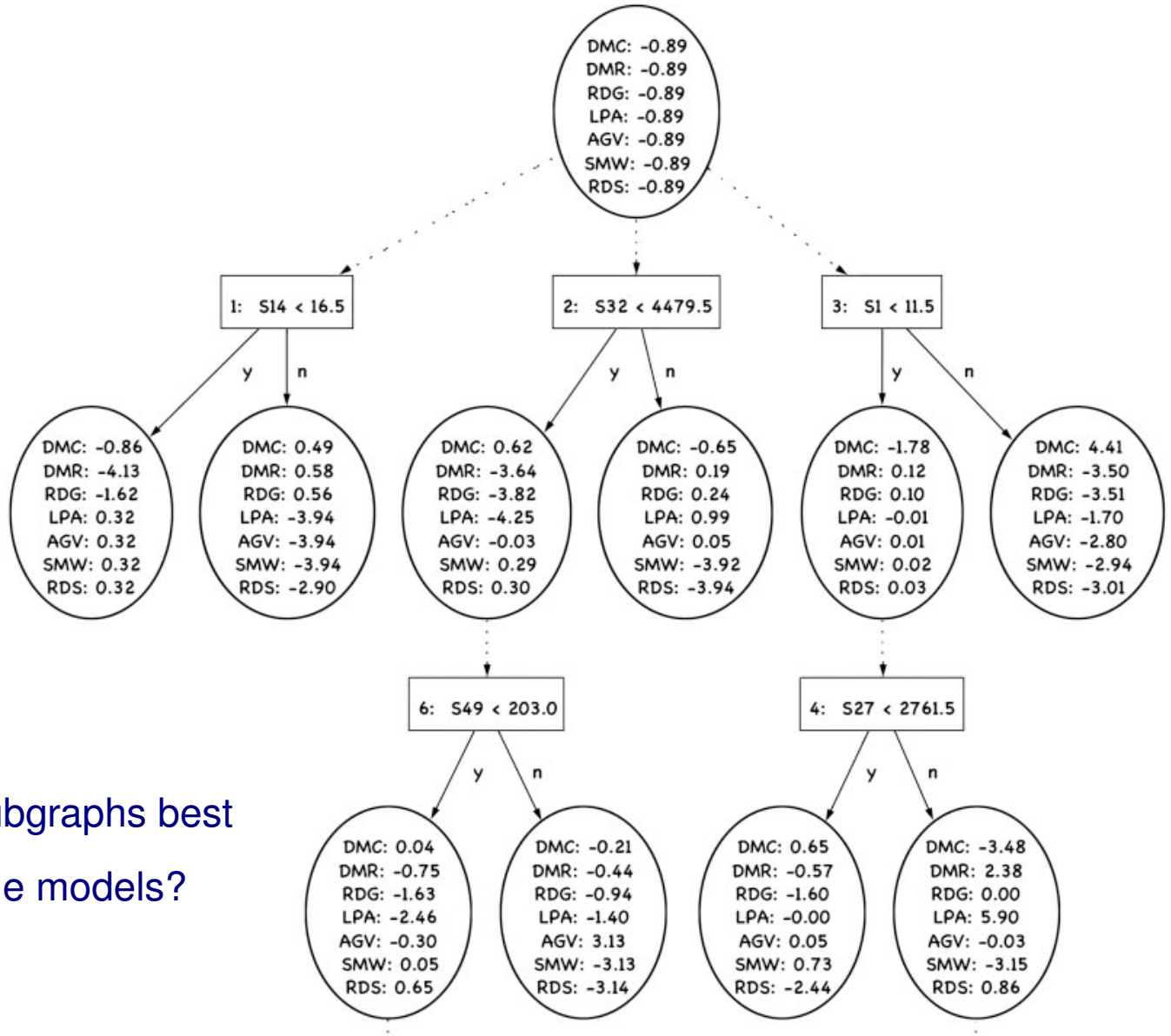
- Generate 1000 random instances of a network with N=3359 and E=2795 for each of the seven models (7000 random instances in total). (Training data)
- "Subgraph census" – classify each network by exhaustive search for all possible subgraphs up to a given size.
- Classify each of the 7 mechanisms by raw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



(Example subgraphs)

Build classifier from the training data (Learning Algorithm)

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



Which subgraphs best distinguish the 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!

Comments

- A **relative** not absolute 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.

- Model **selection** not validation.