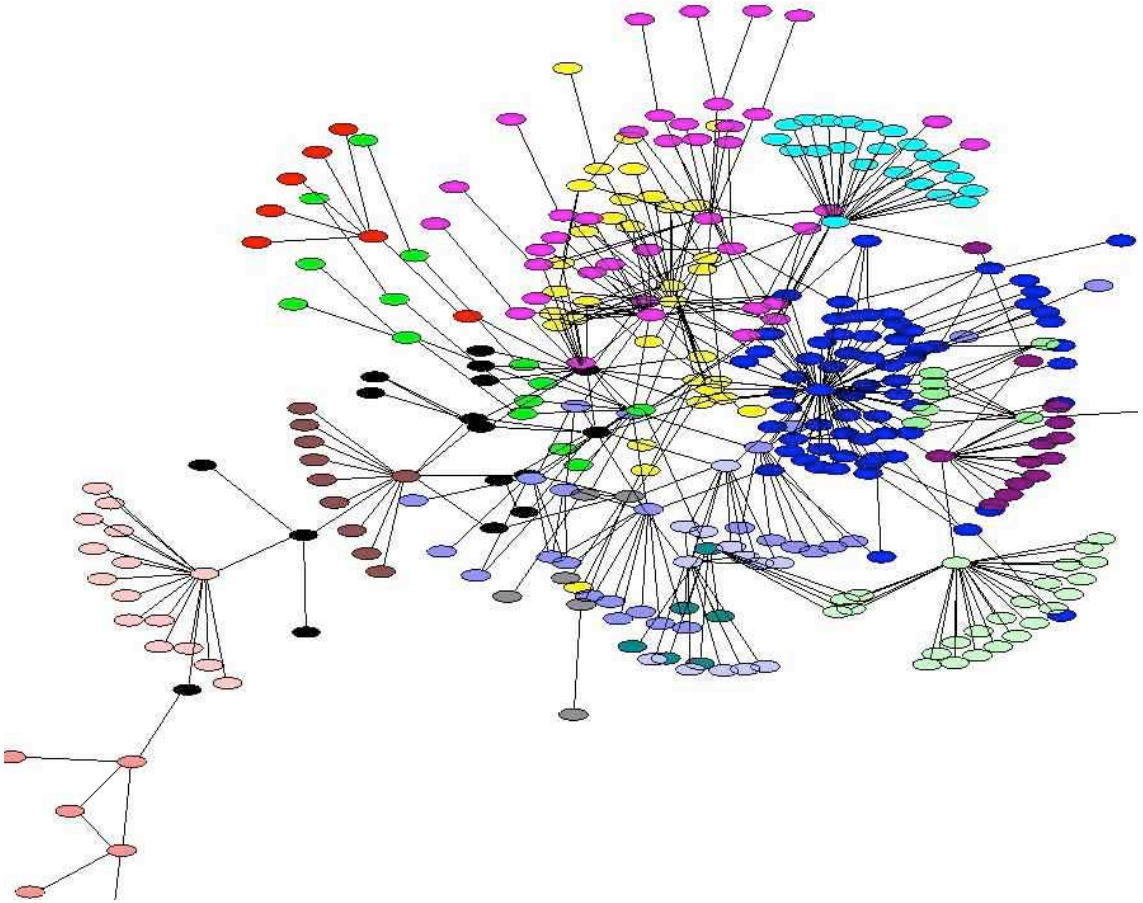
MAE 298

May 28, 2009



“Model selection and validaton”

Literature on validation of network models

- is rather limited

Four papers:

- M. Middendorf, E. Ziv, and C. H. Wiggins, “Inferring network mechanisms: The Drosophila melanogaster protein interaction network”, *PNAS* **102** (9), 2005. (About 59 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 38 citations.)

- V. Filkov, Z.M. Saul, S. Roy, R.M. DSouza, P.T. Devanbu, “Modeling and verifying a broad array of network properties”, *Europhys. Lett.* **86**, 2009.

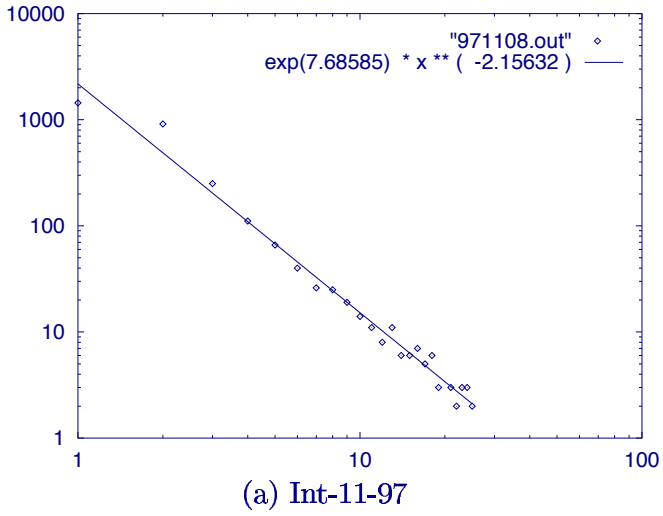
- J. Wang and G. Provan, “Generating Application-Specific Benchmark Models for Complex Systems”, *Proc. Twenty-Third AAAI Conf on Artificial Intelligence*, 2008.

Model validation: Overarching issues

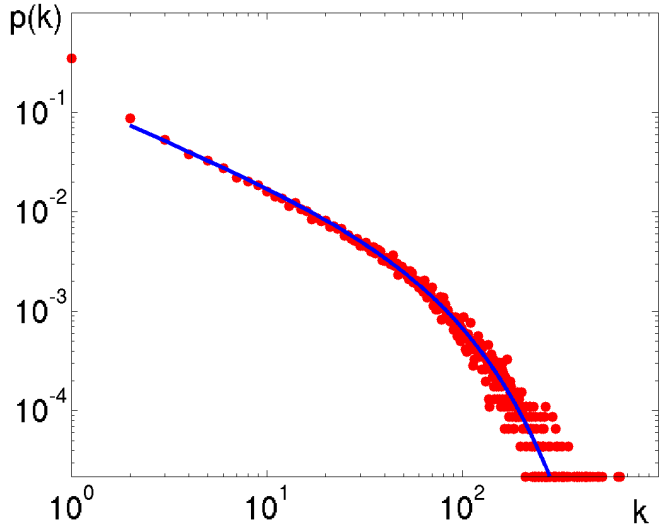
- Many models give rise to same large-scale statistics (e.g., degree distribution, diameter, clustering coefficient).
- Data sets have multiple attributes. Fitting one or two of them is not always sufficient.
- Data: Limited availability (expense or proprietary nature); small data sets

In the beginning – Power Laws

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



(a) Int-11-97



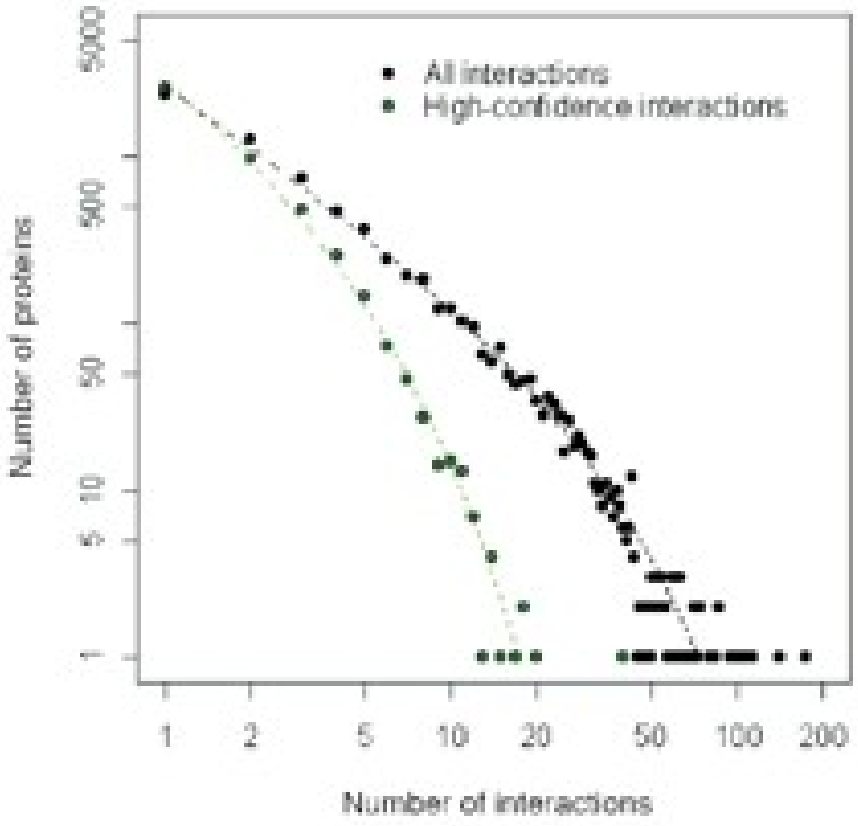
- M. Mitzenmacher, “The Future of Power Law Research” *Internet Mathematics*, **2** (4), 2006. (Editorial piece)
  - – A call to move beyond observation and model building to validation and control.
  - – Power laws ‘the signature of human activity’
- Clauset, Shalizi, Newman, “Power-law distributions in empirical data”, to appear *SIAM Review*. (http://arxiv.org/abs/0706.1062)
  - – Techniques to detect if actually have a power law, and if so, to extract exponents.

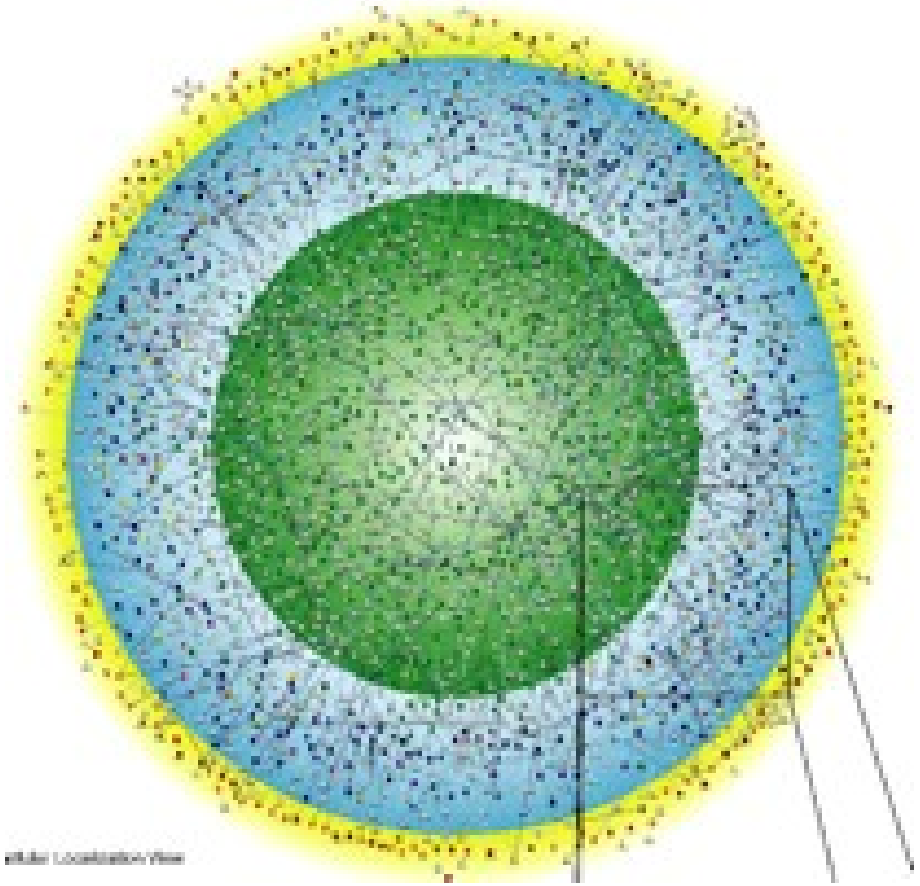
"Inferring network mechanisms: The Drosophila melanogaster protein interaction network"

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)





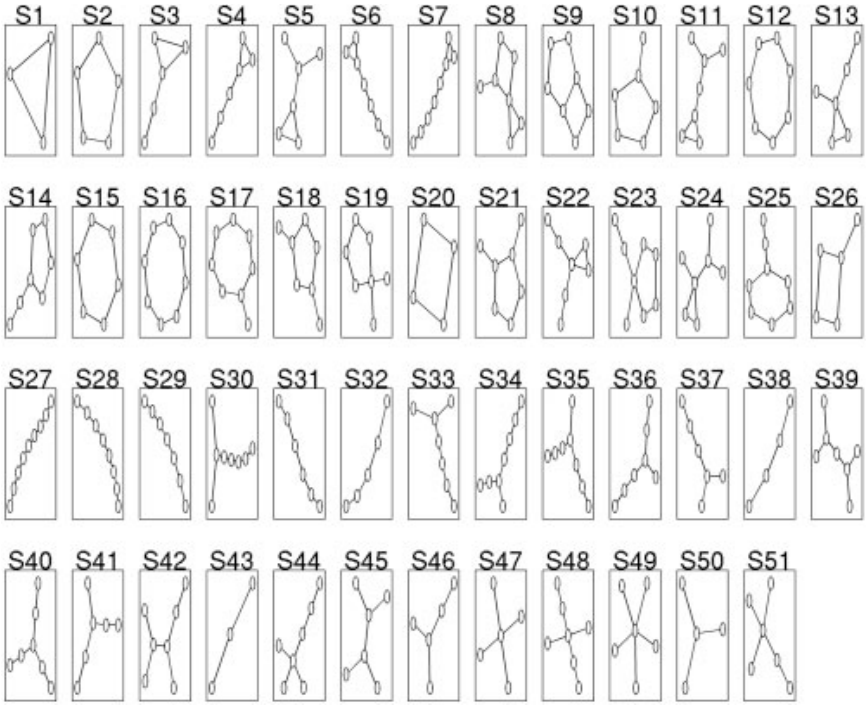
- Accept any edge with $p > 0.65$,
  3,359 vertices and 2,795 edges.

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.



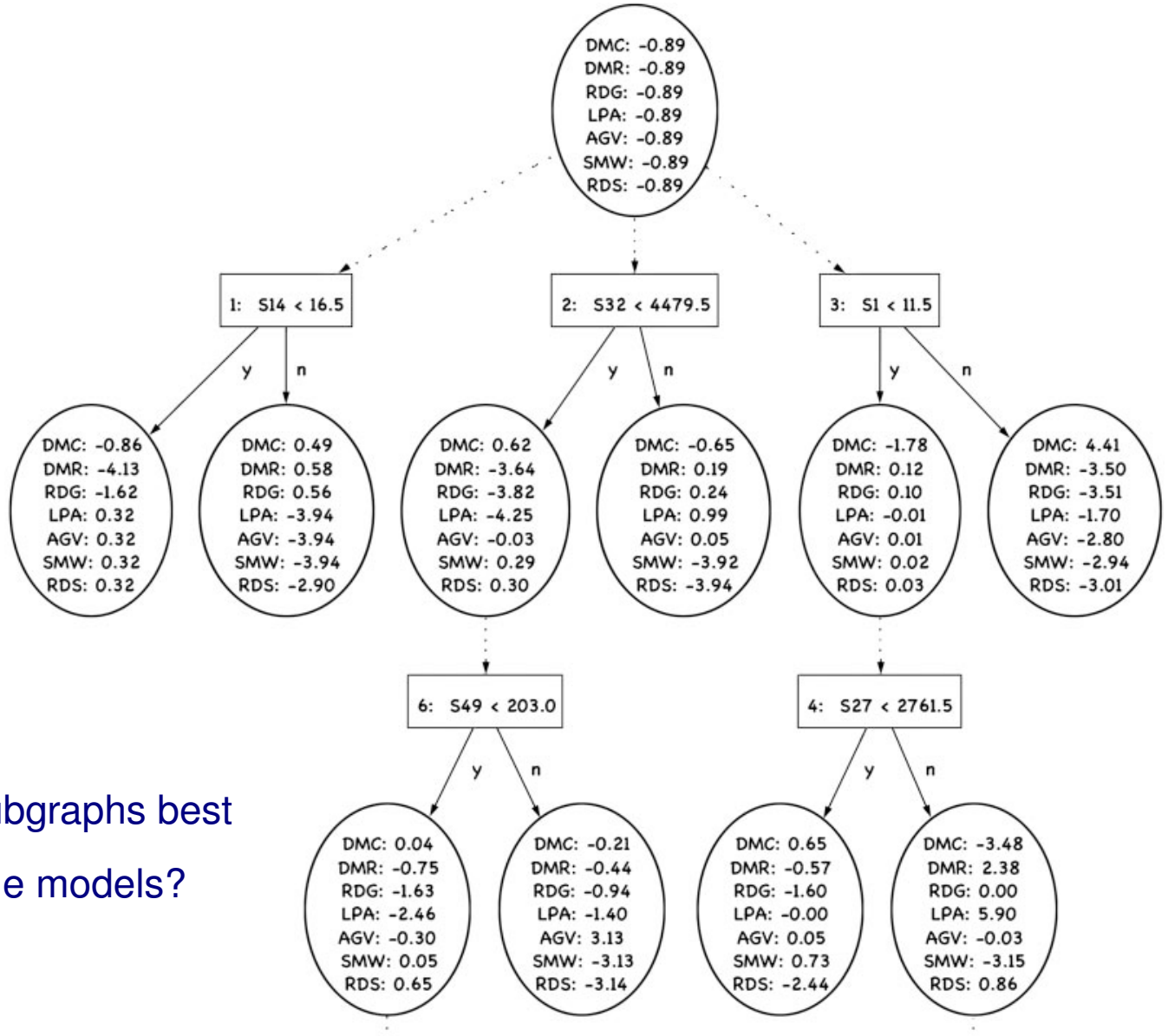
(Example subgraphs)

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).



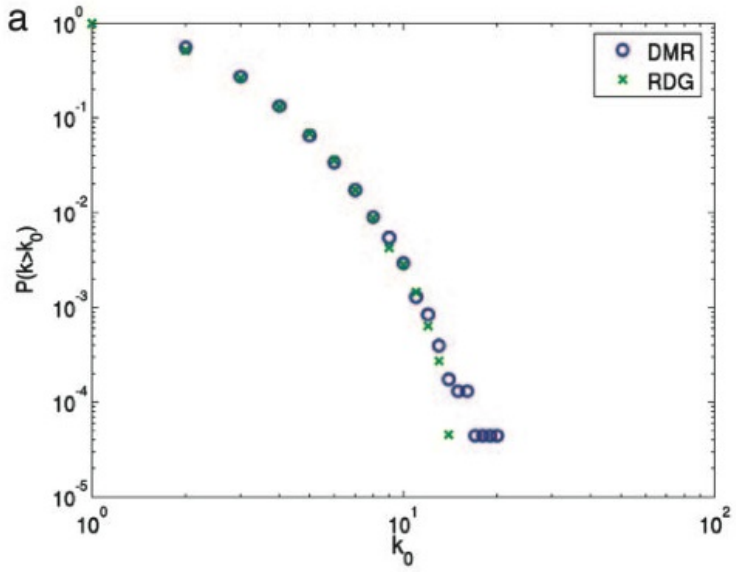
Which subgraphs best distinguish the models?

Validating classifier

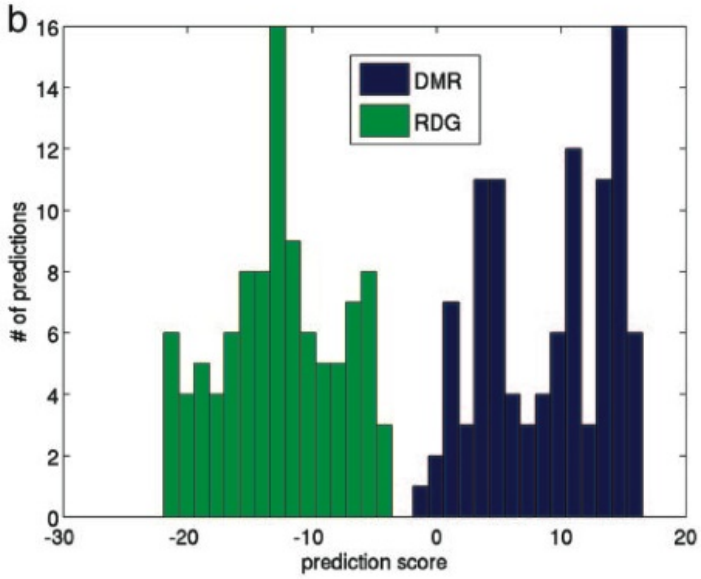
	Prediction						
Truth	DMR	DMC	AGV	LPA	SMW	RDS	RDG
DMR	99.3	0.0	0.0	0.0	0.0	0.1	0.6
DMC	0.0	99.7	0.0	0.0	0.3	0.0	0.0
AGV	0.0	0.1	84.7	13.5	1.2	0.5	0.0
LPA	0.0	0.0	10.3	89.6	0.0	0.0	0.1
SMW	0.0	0.0	0.6	0.0	99.0	0.4	0.0
RDS	0.0	0.0	0.2	0.0	0.8	99.0	0.0
RDG	0.9	0.0	0.0	0.1	0.0	0.0	99.0

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

Validating classifier



	DMR	RDG
$\langle C \rangle$	$2.6\ 10^{-4} \pm 1.3\ 10^{-4}$	$5.4\ 10^{-4} \pm 3.7\ 10^{-4}$
$\langle \ell \rangle$	10.4 ± 0.1	9.6 ± 0.04



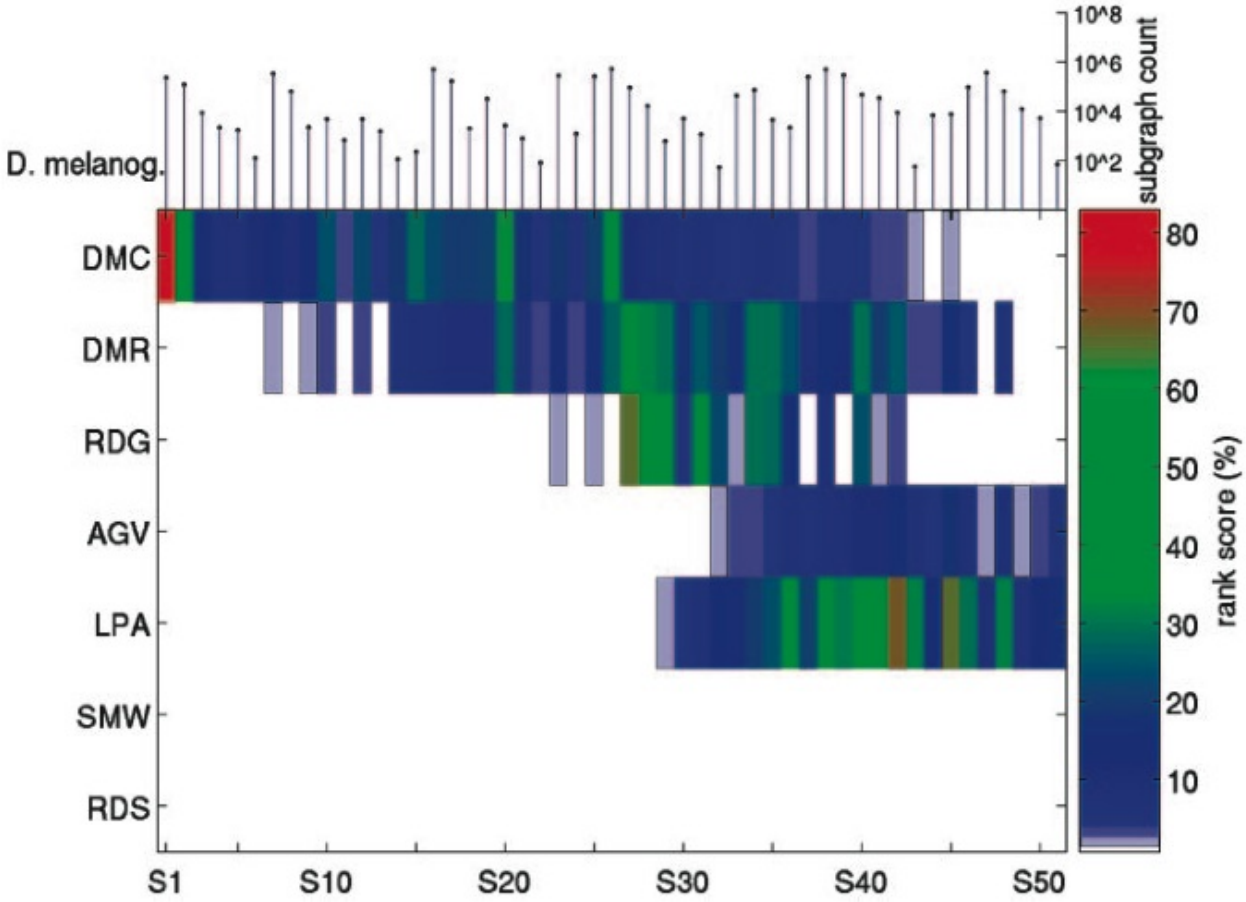
- (a) DMC and RDG produce similar statistical distributions.
- (b) Classifier can discriminate between the two models.

After classifier built, use it to characterize individual network realizations

(Walk the Drosophila data through the ADT)

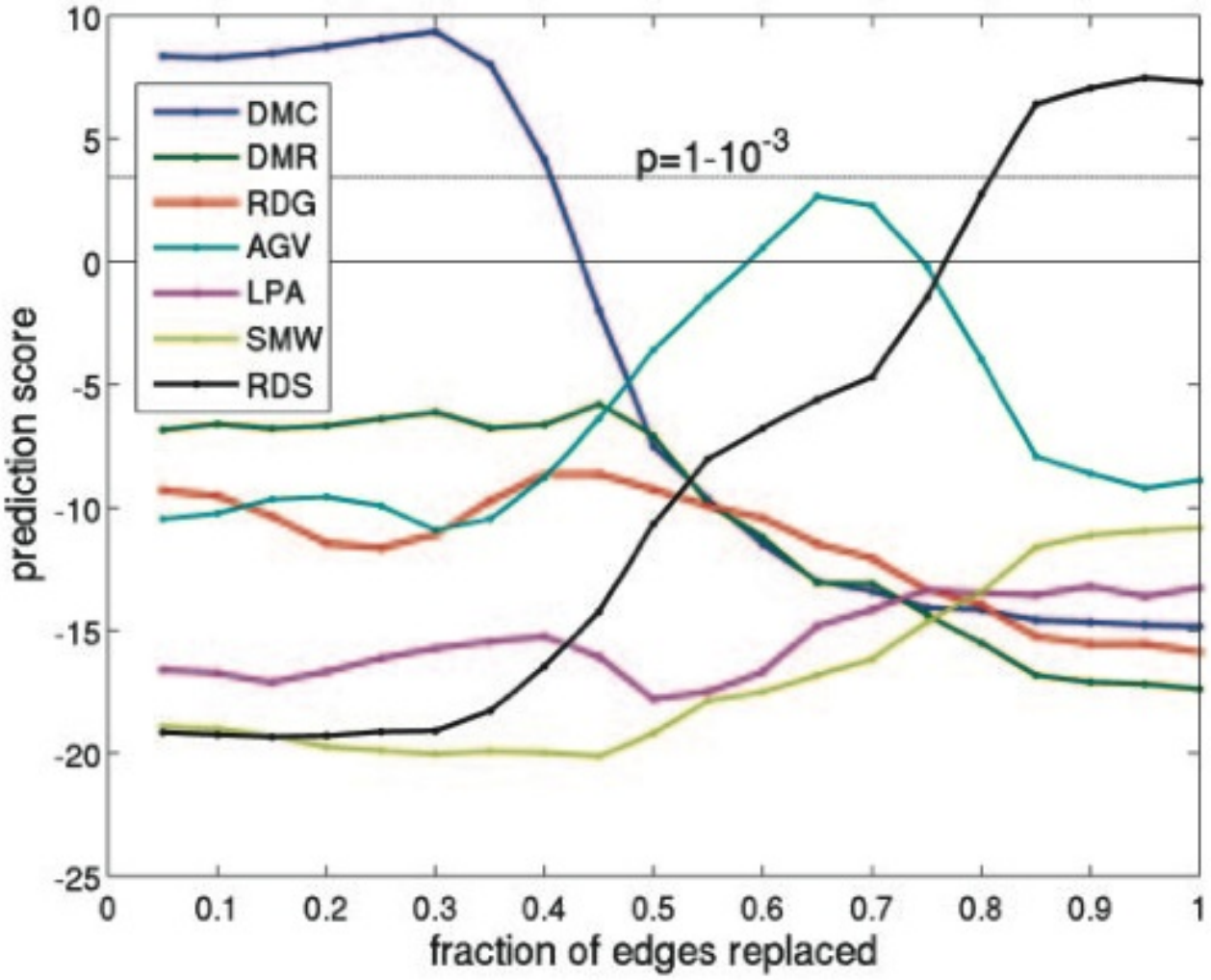
- A given network's subgraph counts determine paths in the ADT (decision nodes are rectangles)
- The ADT outputs a real-valued prediction score, which is the sum of all weights over all paths.
- The final weight for a model is related to probability that particular network realization was generated by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



- Classifier robust.
- Also robust to $p = 0.5$ and different subgraph counts, $n = 7, 8$.

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.)

"Understanding Internet Topology: Principles, Models, and Validation"

D. Alderson, L. Li, W. Willinger, and J. C. Doyle, *IEEE/ACM Trans. on Networking*, **13** (6), 2005.

- I chose this paper since they explicitly claim to deal with **validation**.
- Do you think they did?

Overview

- “First principles” approach to router-level Internet modeling.
- Consider capability of real routers (annotated graph).
- Consider **core** versus **edge** requirements.
- From this design “optimal” networks.
- Compare with sampled topologies of actual internet, and show that constraint-capacity performance curve of real routers fit with their hypothesis.

Motivation – Need accurate models of the internet

- Testing and evaluating protocols
- Protecting against and detecting attacks
- Improved designs and resource provisioning

$\longrightarrow$ need annotated graphs, with bandwidth capacity explicit (also router buffer capacity).

- Given topology generated from a model, which statistical properties to test?
- Ascribing meaning to model details. (Why would a random construction relate to an engineered network?)

Router level connectivity

- Layer 2 (data-link layer) connectivity

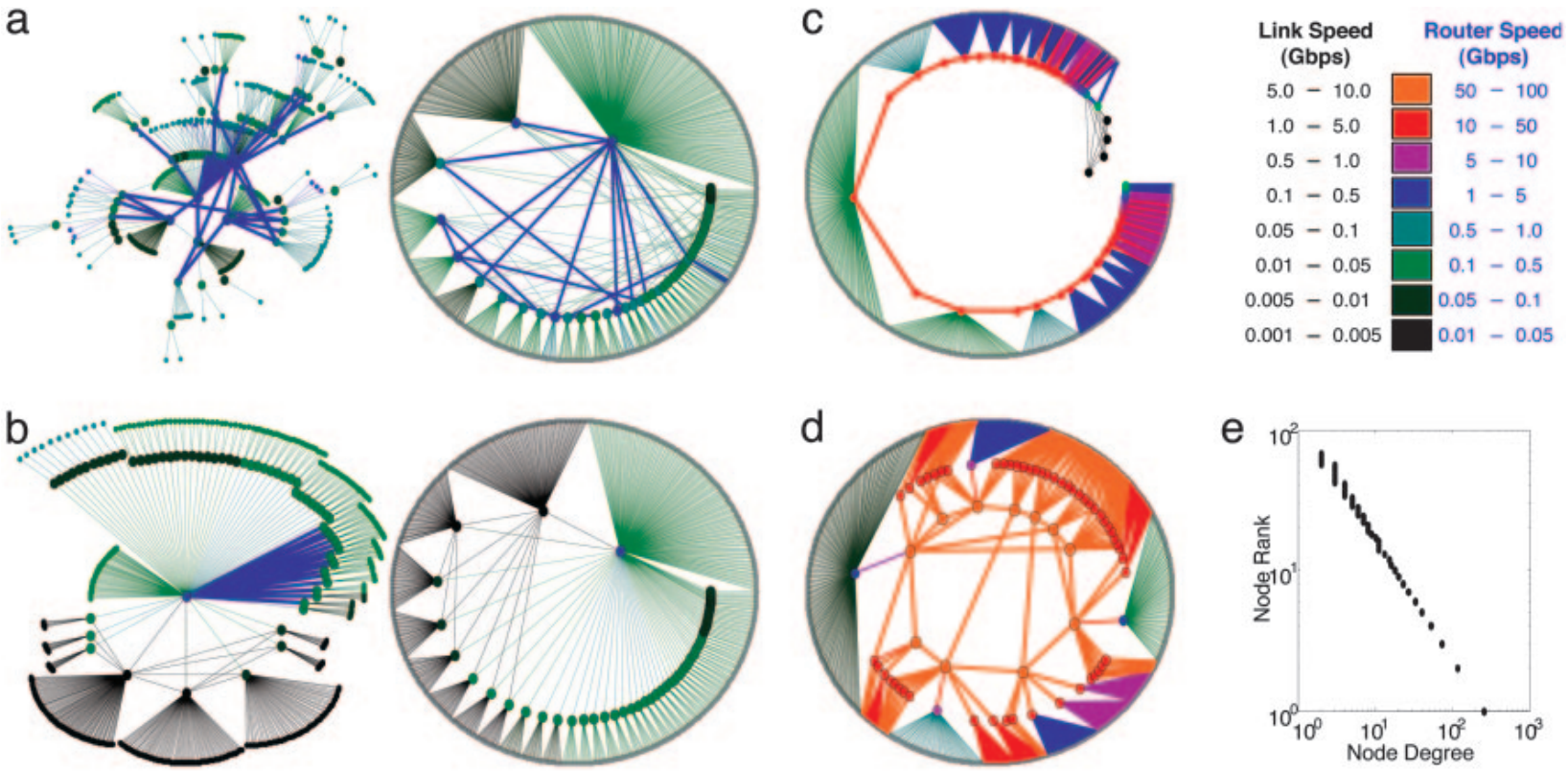
Open Systems Interconnection (OSI) Reference Model

7 Application Layer
6 Presentation Layer
5 Session Layer
4 Transport Layer
3 Network Layer
2 Data Link Layer
1 Physical Layer

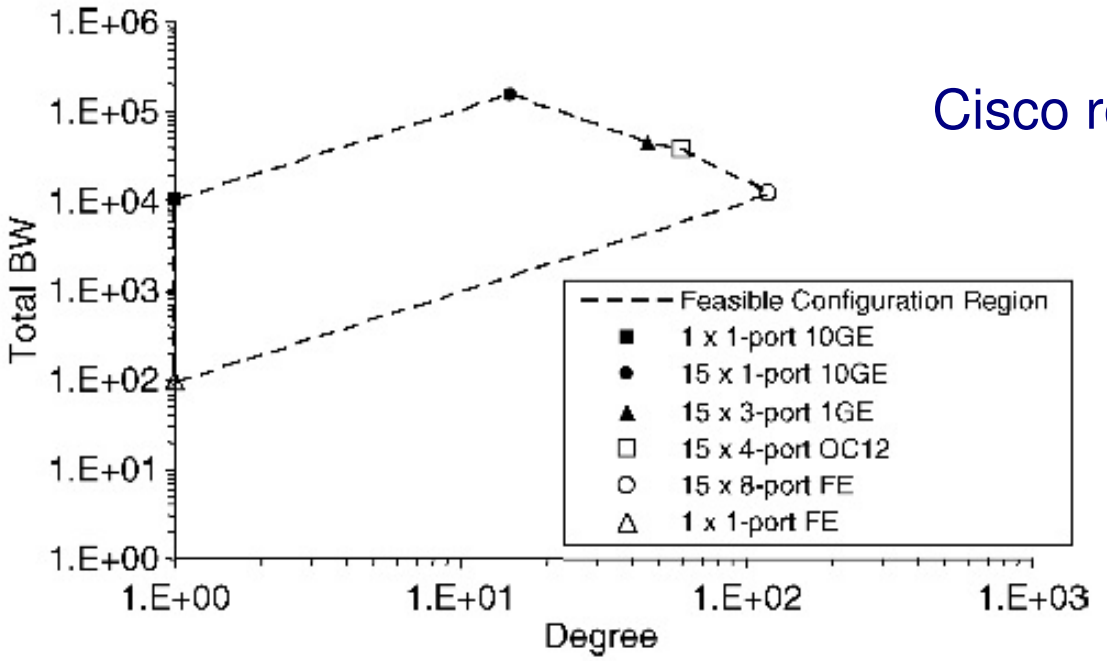
Past work – structural topology generators

- Random connectivity (e.g., Waxman model)
- Transit-stub models of Zegura (Georgia Tech Internetwork Topology Models)
- But this miss the broad-scale (power-law-like) distributions in connectivity presumed to be in real Internet.
- So people jumped on the “preferential attachment” bandwagon.

Degree distribution – not the whole story



“First principles” – start with constraints on routers



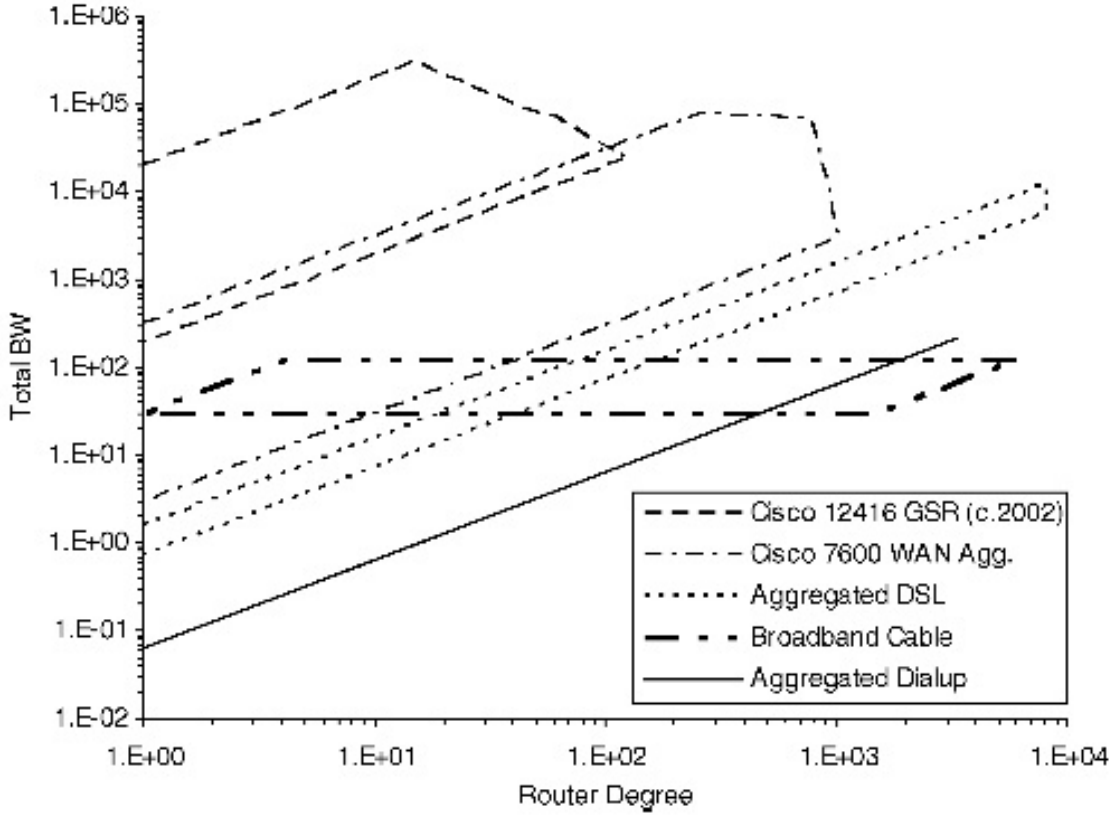
Cisco router w 15 slots

- Set number of interface cards. Initially with each new card added, increase overall bandwidth.
- Once number of connections exceeds interface cards, connections have to share limited bandwidth. (Increase connectivity and decrease maximum bandwidth available to each link.)
- Overhead in switching causes decrease in total bandwidth.

“Core” versus “edge”

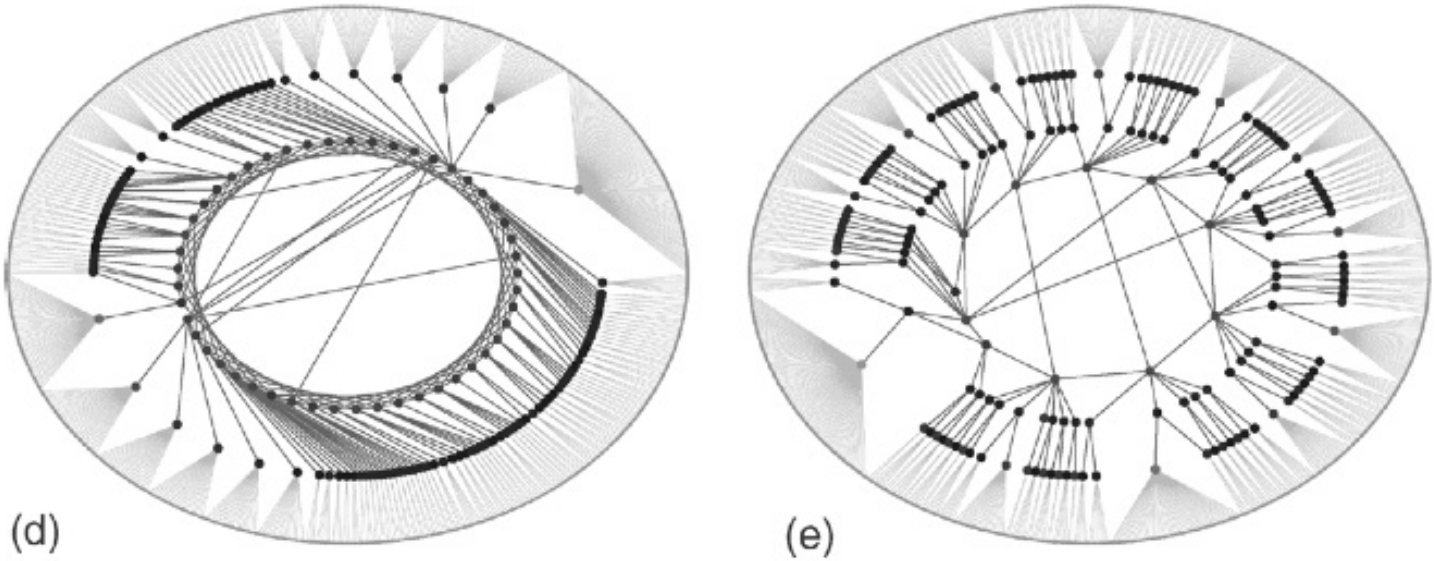
- Core routers – support highest link speed with limited connectivity (long-haul connectivity)

- Edge routers (access routers) – support wide range of low-speed connections (this traffic then aggregated and sent on to core) also wide range of technologies (dial-up, DSL, cable, etc) with wide range of pricing strategies.



"Heuristically Optimal Topologies"

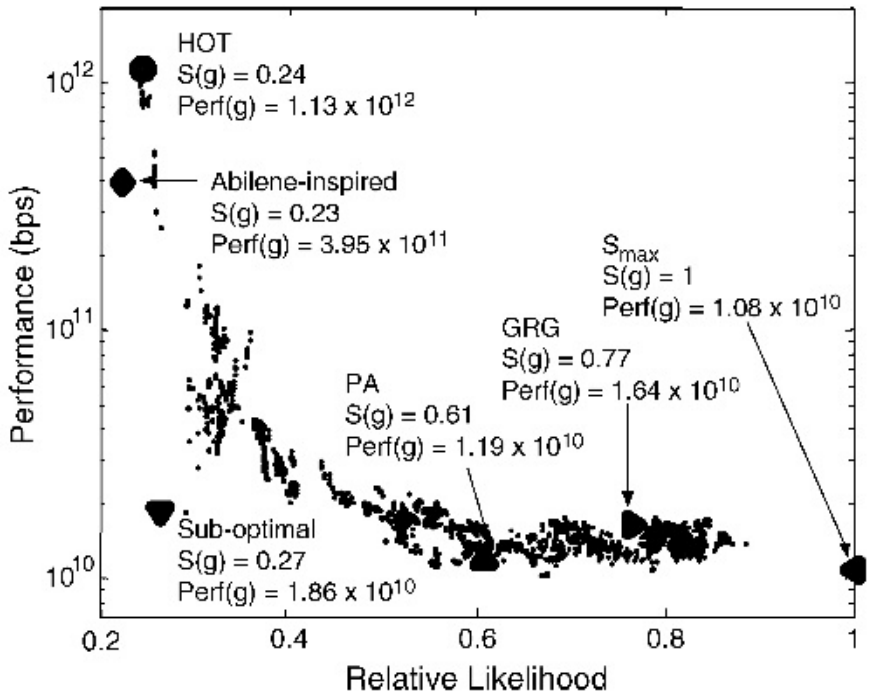
- Core routers – high bandwidth, low connectivity
- Edge routers – Many low bandwidth connections, aggregating traffic as close to edge as possible.



(d) HOT net (selectively rewired PA net), (e) HOT net started from core observed in real-world provider (Abilene), replacing non-Abilene nodes.

Metrics (no clear connection to validation)

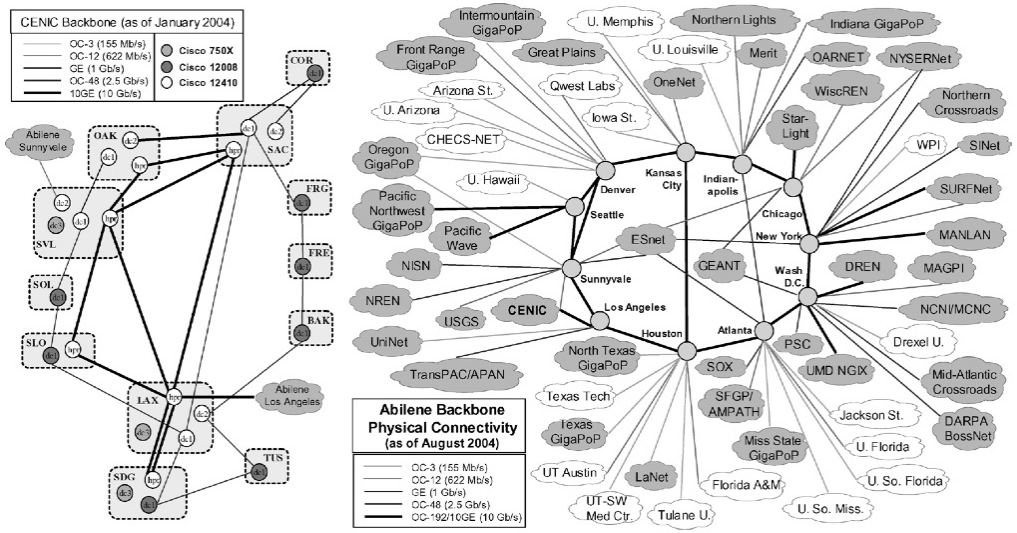
- Performance: maximum throughput of graph g, under gravity model of traffic.
- Likelihood: $S(g) = \frac{s(g)}{s_{\max}}$, where $s(g) = \sum_{\text{edges}} w_i w_j$.
  – degree-degree correlations
  – s(g) similar to assortativity/dissortativity in ways. High s(g) means hubs connected together and will form core.



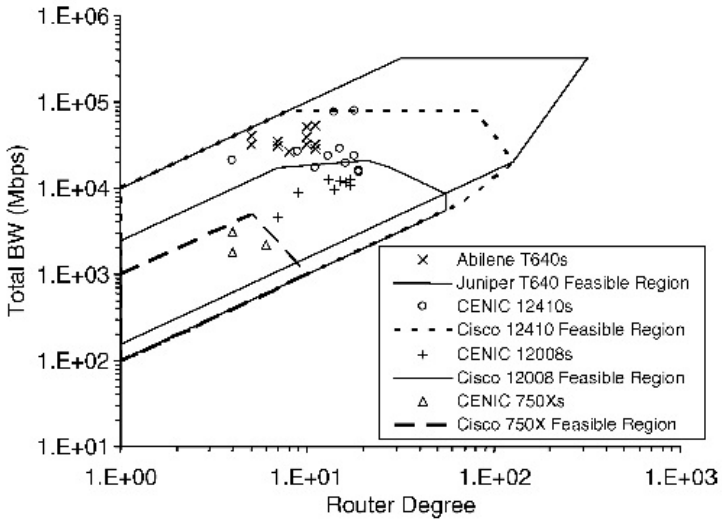
Empirical data, I

- Abilene network (Internet backbone for higher education use, carries 1% of traffic in North America).
- CENIC (California education network)

“Validation”



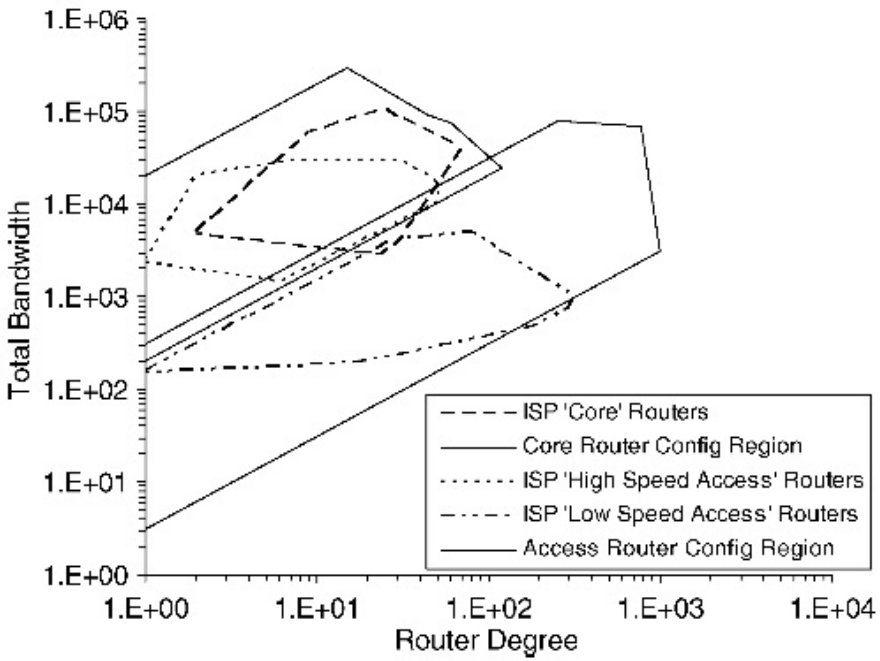
(a)



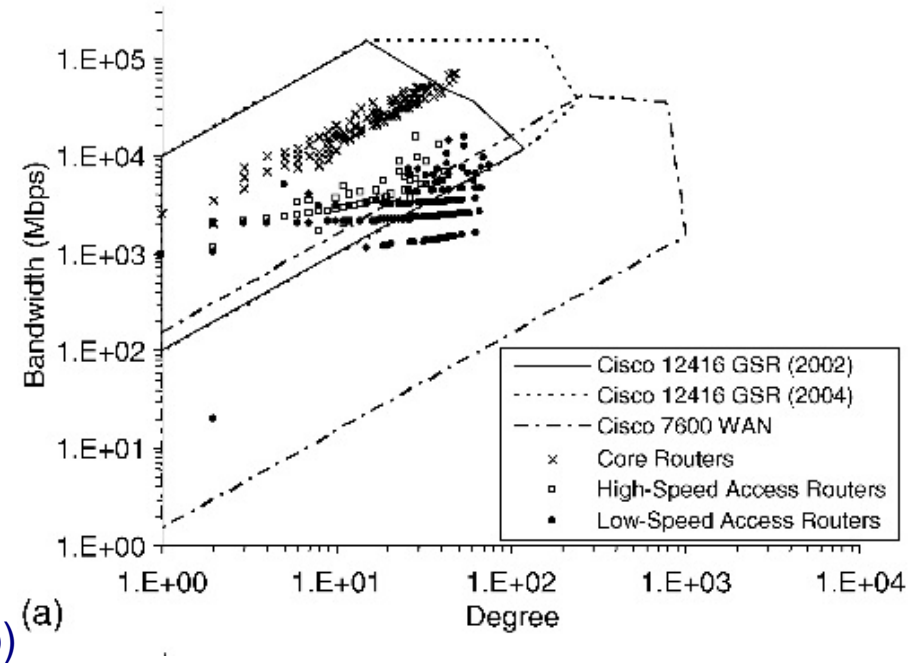
(b)

- (a) Qualitative agreement with HOT derived network
- (b) Observed routers sort of seem to segregate into core/edge of their first principles.

Empirical data, II



(a)



(b)

- (a) Anonymized ISP, (b) Rocketfuel sampled.
- "the point here is that even a heuristic process informed by a detailed understanding of router role and technology constraints can go a long way toward generating realistic annotated router-level maps. While not conclusive, what is remarkable about these results is that the simple assumptions for (fixed) link bandwidths in Table I result in bandwidth-degree combinations that are not inconsistent with our understanding of heuristically optimal network design."

Further points

- High variability could be due exclusively to edge. Variability in link technologies (DSL, dial-up, cable) and user demands (pricing).

- High variability may be due to errors in sampling approaches (e.g., traceroute)

- "Achilles' Heel"
  – hubs in core make networks vulnerable to targeted attack.
  – hubs in periphery have little impact on connectivity.
  – HOT networks more robust, even "damaged" HOT net better than intact PA network.

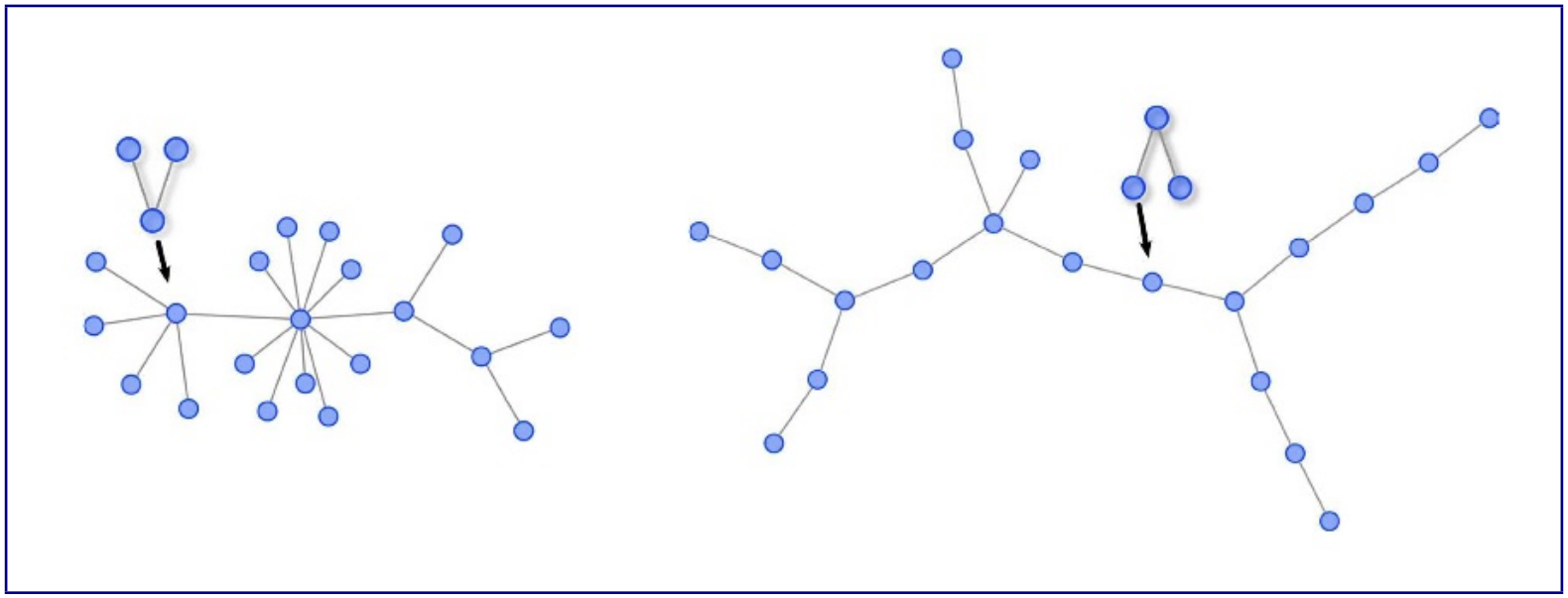
Comments

- Lacks objectivity (heavy self-citation)
- Matching the constraint-capacity performance curve does not **validate** that their first principles approach produces a true topology.
- Specific to internet topology (which is their point) / first principles.
- See also: W. Willinger, D. Alderson, and J.C. Doyle. "Mathematics and the Internet: A source of enormous confusion and great potential", *Notices of the American Mathematical Society*, 56(5):286-299, May 2009.

"Modeling and verifying a broad array of network properties"

V. Filkov, Z.M. Saul, S. Roy, R.M. DSouza, P.T. Devanbu, *EPL* **86**, 2009.

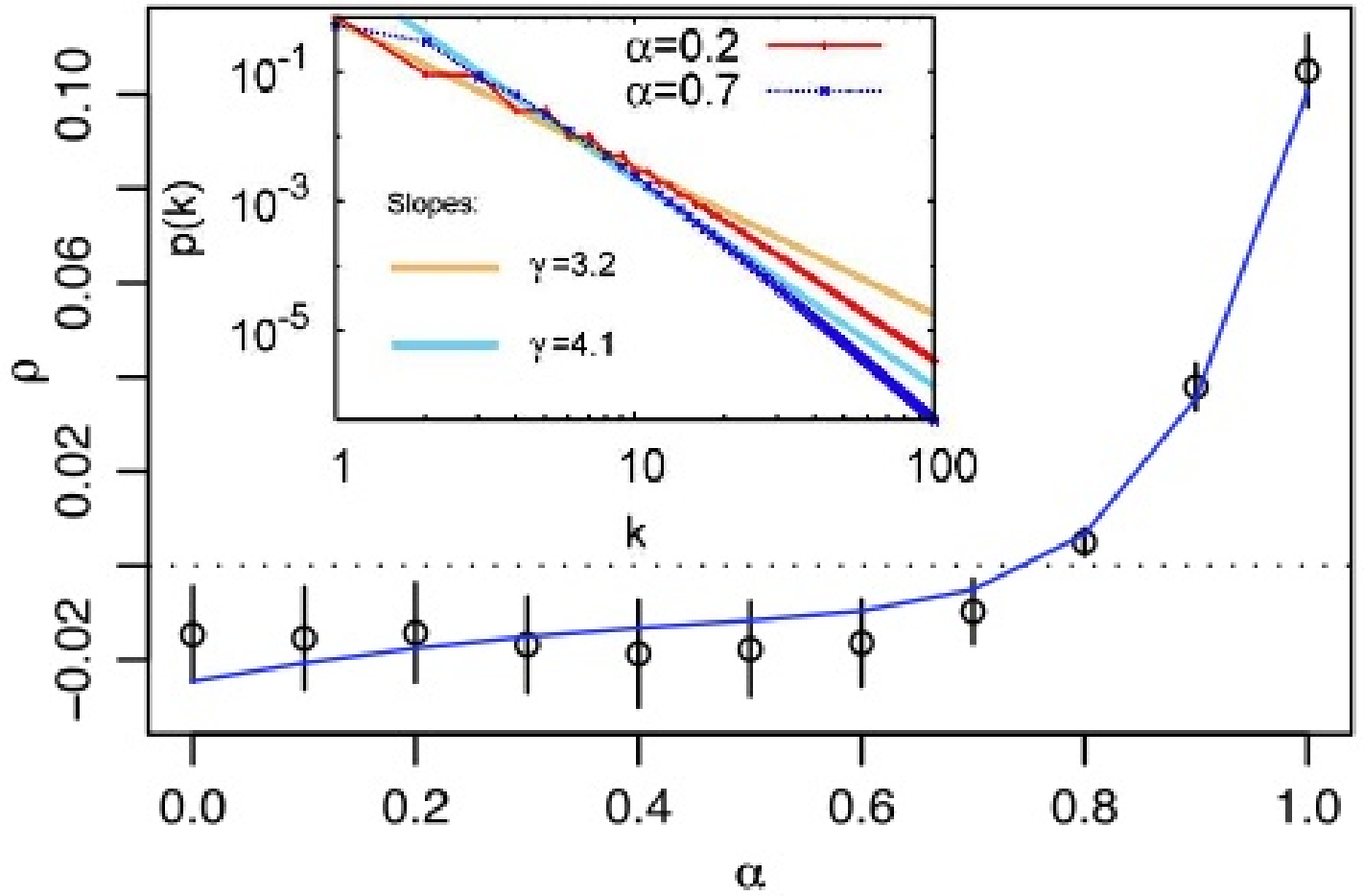
- Graphlet arrival model



Select parent node proportional to degree (PA).

With probability $(1-\alpha)$ connect at midpoint, with probability α at end.

Tunable degree distribution and assortativity

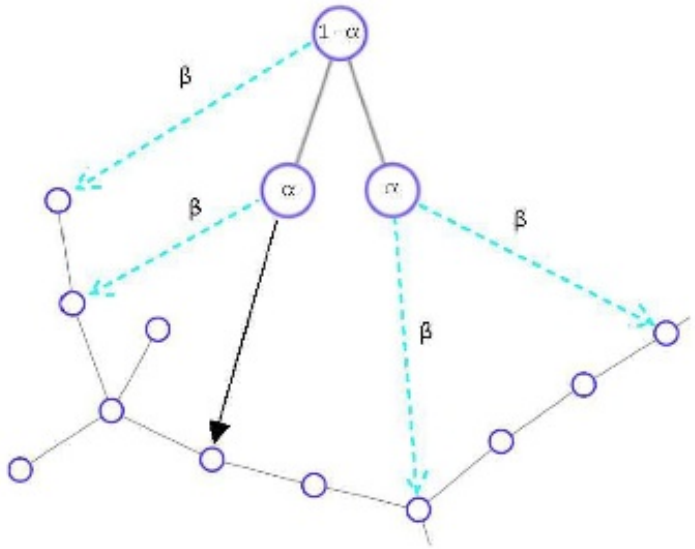


- Degree distribution, power law $p(d) \sim d^{-\gamma}$, with $\gamma = (6-\alpha)/(2-\alpha)$.

- Assortativity (Pearson's correlation coefficient for degree-degree of edges), calculated numerically from recurrence relation.

Allow each incoming “V” to make multiple connections

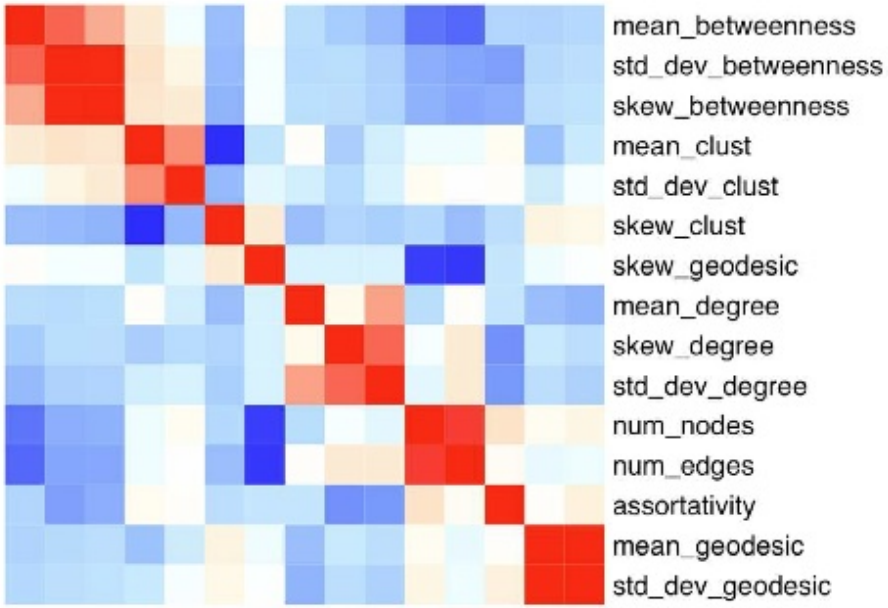
(Original model only generates trees)



With probability β create l edges (multiple attachment points), chosen at random.

Comparing across multiple attributes simultaneously

- Real data – 113 networks from software call graphs, social networks, protein-protein networks, gene networks.
- Select a set of attributes (15 initially)
- Compare pair-wise correlations (heatmap):



Red is correlated, blue is anti-correlated, white neutral.

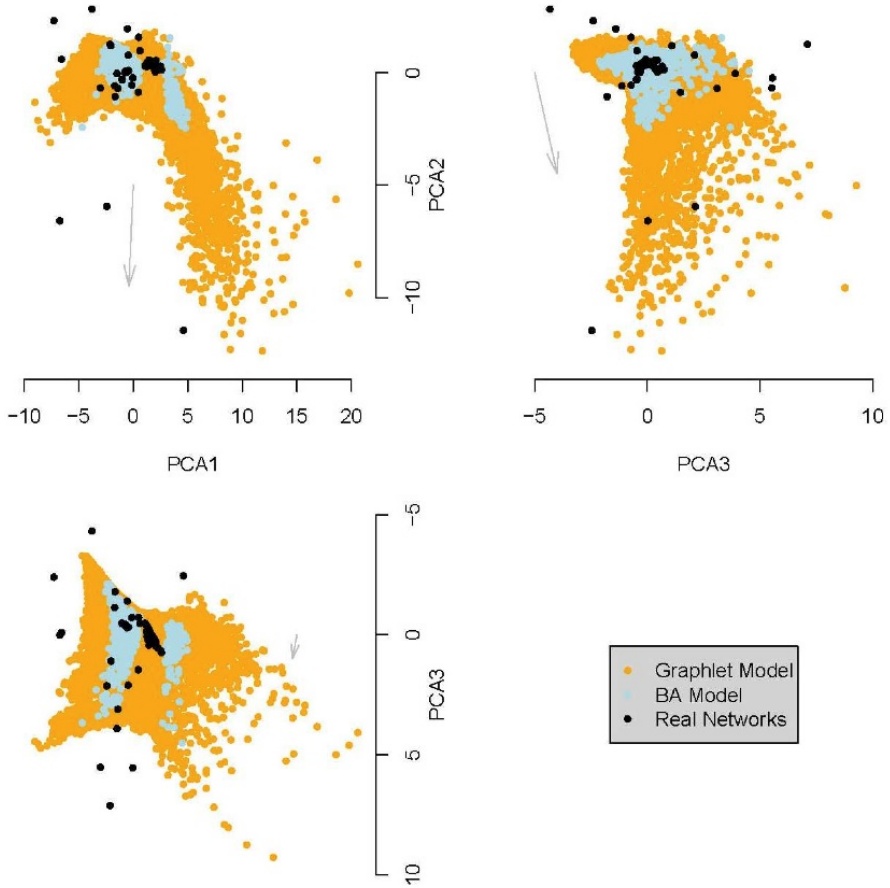
How effectively do they span a space of independent attributes?

(We retain 11 of the 15)

Comparing models to real-data

- Generate 60,500 sample graphlet model networks for a range of α, β, l and n.
- Generate 500 sample PA networks for a range of l.
- Visualize attribute space, including real networks and model networks, using dimension-reduction technique of Principal Component Analysis (PCA).
- PCA finds projection of n-dimensional data set onto a space of the same dimension, but where the new axis (principle components) are orthogonal and linear combinations of the original attributes.
- The PC1 axes is the one that demonstrates maximal variance of the original data set, PC2 the second largest variance, etc.

Results of PCA



- PCA1$\sim$ number of edges, mean and skew of geodesic, mean and std dev of clustering and mean and std dev of degree.
- PCA2 is std dev and mean of geodesic, and assortativity.
- For reference grey arrow is assortativity.

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!**