Scalable Methods for Representing, Characterizing, and Generating Large Graphs

Ali Pinar

Sandia National Laboratories

Siam Annual Meeting

July 12, 2010

Pittsburgh, PA

Sandia National Laboratories

Characterizing and Generating Graphs

- **Goal:** design methods to characterize and identify a low dimensional representation of graphs
- **Impact:** enabling predictive simulation; monitoring dynamics on graphs; sampling and recovering network structure from limited observations
- **Areas to explore:**
  - **Enabling technologies:** develop novel algorithms and tailor existing ones for complex networks
  - **Modeling and generation:** Identify the right parameters for graph representation and develop algorithms to compute these parameters and generate graphs from these parameters
  - **Comparison:** Given two graphs how do we tell they are similar?
- Funded by DOE O. Science ASCR Applied Math program.
  - **Team:** Tamara Kolda, Jaideep Ray, Daniel Dunlavy, Cynthia Phillips, Bruce Hendrickson, Matthew Grace, David Gleich, Isabelle Stanton
- A related talk: “Compressively Sensed Complex Networks” by Jaideep Ray, MS80, Thursday 11am.

Sandia National Laboratories

What is a good metric/granularity for comparing graphs or evaluating models?

- Metrics:
  - Isomorphism:
    - looks for a permutation to make the graphs identical
    - too hard and too perfect.
  - Alignment:
    - similar to isomorphism, but tolerates imperfectness
    - good to identify correlations.
    - still hard
  - Feature-based comparison:
    - Measure features and compare
    - not rigorous enough…yet
- Granularity:
  - All edges: present and absent
  - Structures
  - Compact representations

Recursive Matrix Structure

- Generates a probability matrix, by starting with a Kronecker basis, and increasing the size using Kronecker products.

$$\Theta_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \Theta_2 = \begin{pmatrix} a\Theta_1 & b\Theta_1 \\ c\Theta_1 & d\Theta_1 \end{pmatrix} = \begin{pmatrix} a^2 & ab & ba & b^2 \\ ac & ad & bc & bd \\ ca & cb & da & db \\ c^2 & cd & dc & d^2 \end{pmatrix}$$

$$\Theta_3 = \begin{pmatrix} a\Theta_2 & b\Theta_2 \\ c\Theta_2 & d\Theta_2 \end{pmatrix}$$

- The *(i,j)* entry is the probability that an edge exists between vertex *i* and vertex *j*.
- An instance is generated from these probabilities.

Sandia National Laboratories

Fitting R-Mat parameters to a graph

- Leskovec et al. Proposed an MCMC algorithm
  - Seeks a permutation and tunes the parameters at the same time.
- Objective function: log-likelihood

$$l(\Theta) = \log P(G|\Theta) = \log \sum_{\sigma} P(G|\Theta, \sigma) P(\sigma)$$

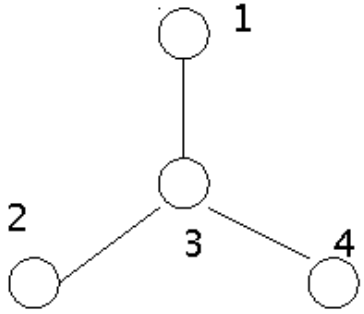
$$P(G|\mathcal{P}, \sigma) = \prod_{(u,v)\in E} \mathcal{P}[\sigma_u, \sigma_v] \prod_{(u,v)\notin E} (1 - \mathcal{P}[\sigma_u, \sigma_v])$$

- Bayesian information criteria is used to determine the size of the basis.

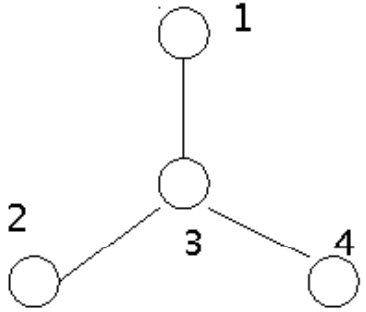
$$BIC = -2l(\Theta) + 2k \log(|V|)$$

where k is the number of variables in the model.

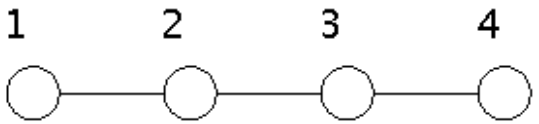
Averaging over all permutations is not accurate



Original graph



$$\begin{pmatrix} 1-\epsilon & \epsilon & 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon & 1-\epsilon & \epsilon \\ 1-\epsilon & 1-\epsilon & 1-\epsilon & 1-\epsilon \\ \epsilon & \epsilon & 1-\epsilon & 1-\epsilon \end{pmatrix}$$



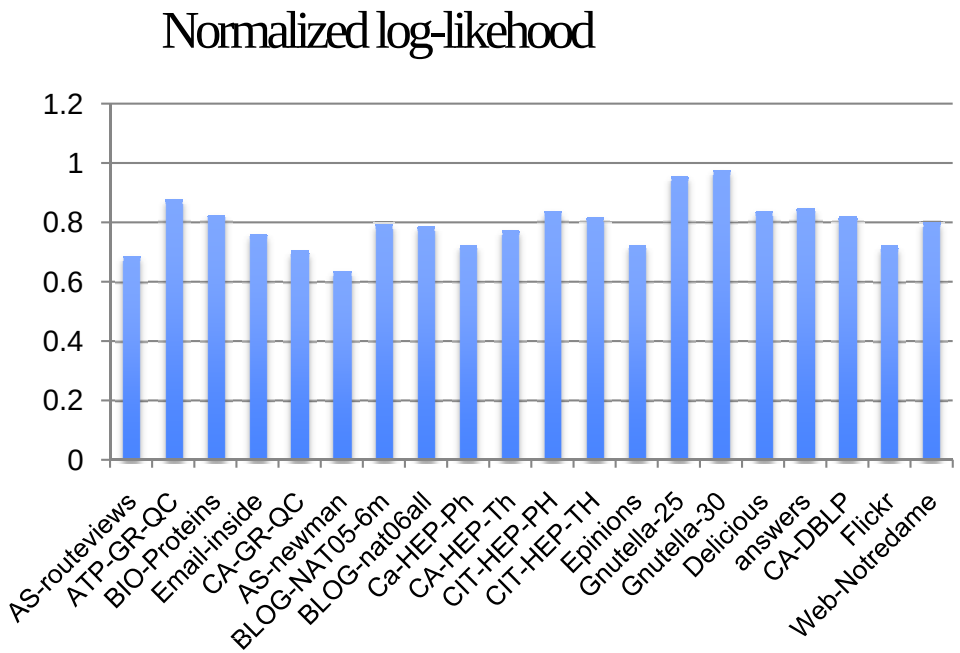
$$\begin{pmatrix} 1-\epsilon & 1-\epsilon & \epsilon & \epsilon \\ 1-\epsilon & 1-\epsilon & 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon & 1-\epsilon & 1-\epsilon \\ \epsilon & \epsilon & 1-\epsilon & 1-\epsilon \end{pmatrix}$$

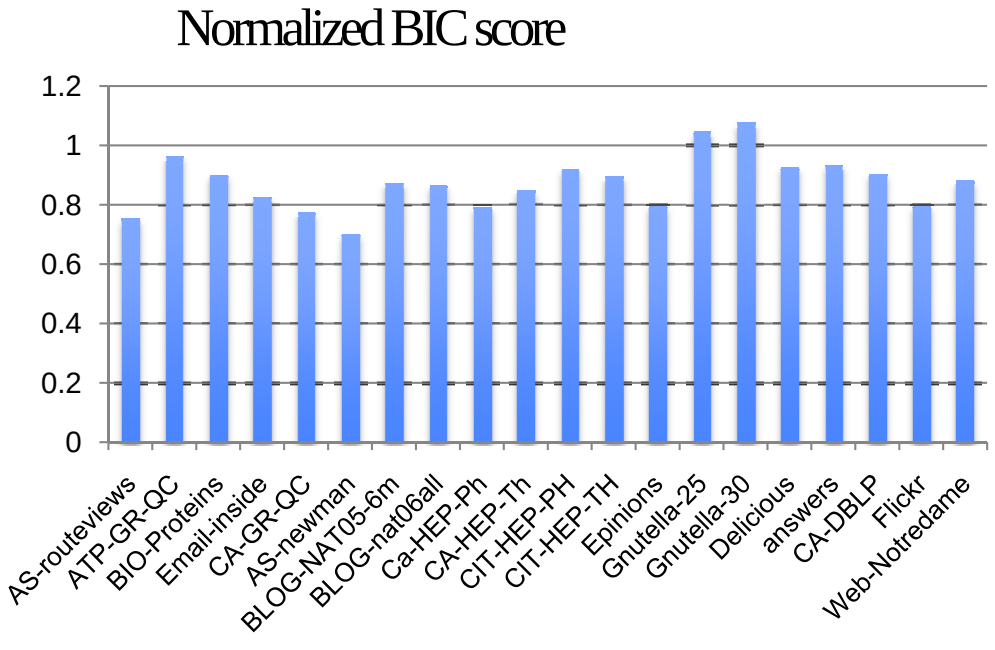
The average -based objective function cannot distinguish between itself and another graph. A better formulation is

$$l(\Theta) = \log P(G|\Theta) = \log \max_{\sigma} P(G|\Theta, \sigma)$$

Sandia National Laboratories

Experiment 1: Is Erdos-Renyi a reasonable model for complex networks?

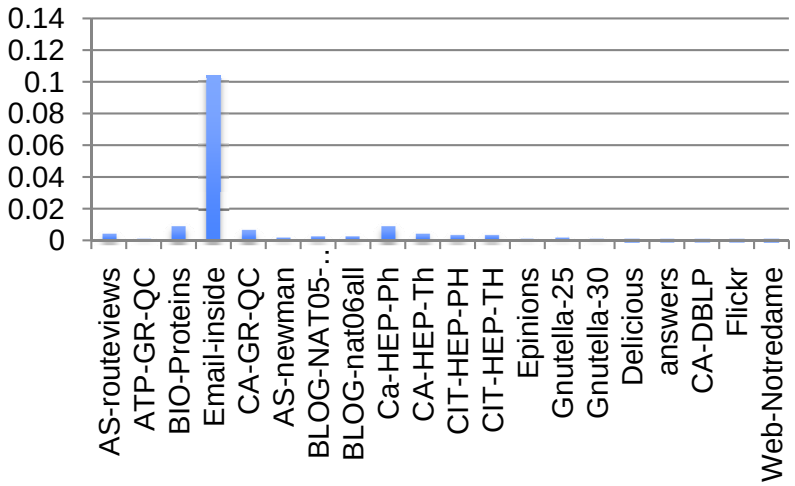




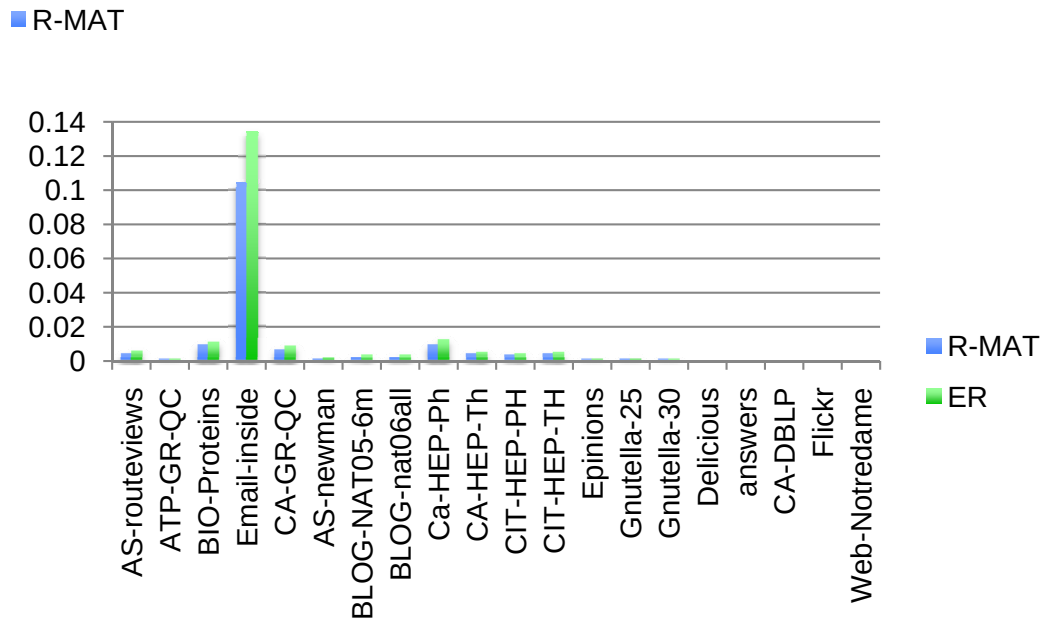
- The data says it is.
  - Not the best, but it is in the same ballpark with RMAT.
  - With better BIC scores for two of the graphs.
- First order logic
  - (false implies false) is true.
  - Where did we do wrong?

Sandia National Laboratories

Experiment 2: How accurate are the fits



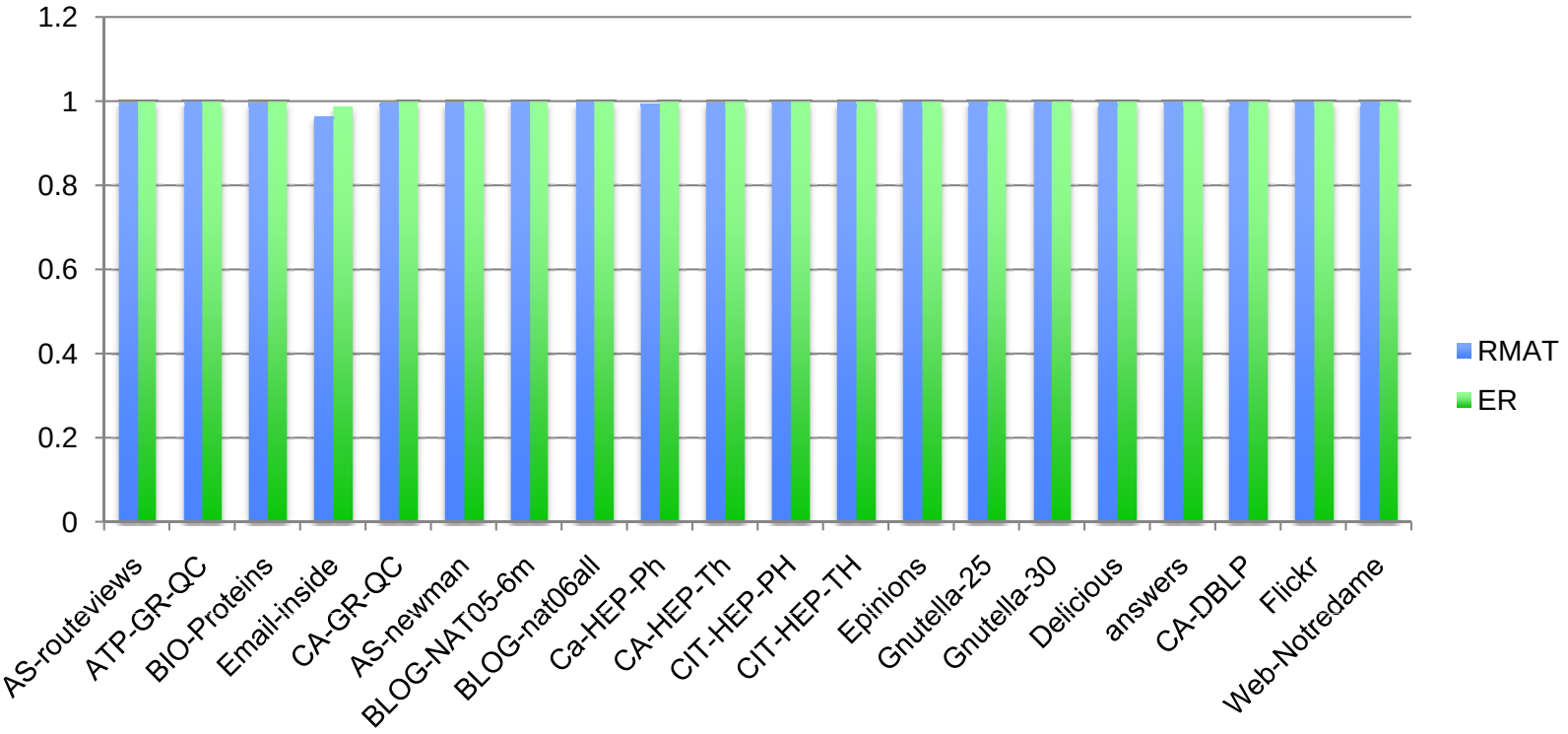
$$P(G|\mathcal{P},\sigma) = \prod_{(u,v)\in E} \mathcal{P}[\sigma_u,\sigma_v] \prod_{(u,v)\notin E} (1-\mathcal{P}[\sigma_u,\sigma_v])$$



- **Error per entry ($|V|^2$) is extremely small.**

Sandia National Laboratories

More on RMAT vs. Erdos-Renyi

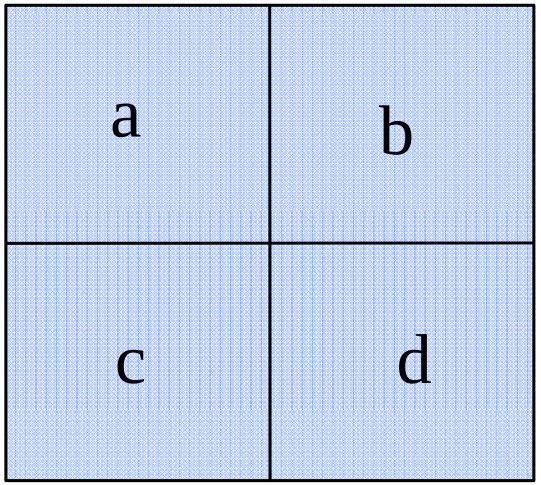


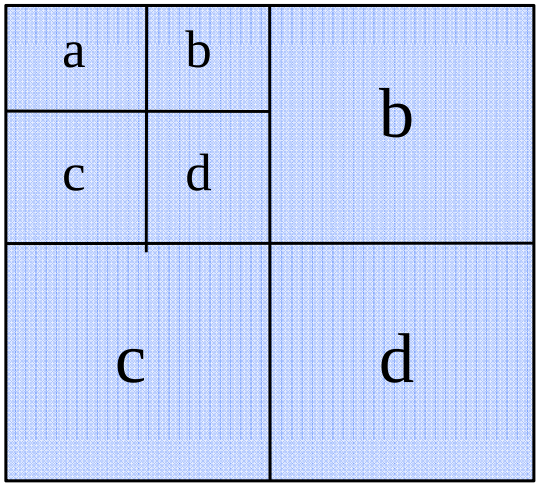
- **Error per edge is extremely large.**

$$P(G|\mathcal{P},\sigma) = \prod_{(u,v)\in E} \mathcal{P}[\sigma_u,\sigma_v] \prod_{(u,v)\notin E} (1-\mathcal{P}[\sigma_u,\sigma_v])$$

Sandia National Laboratories

Generating RMAT graphs in practice





- RMAT is generate a dense |V|x|V| matrix, which cannot, does not need to be constructed explicitly.
- Going over all entries is not feasible.
- In practice, |E| edges are inserted based on probabilities.
- Caveat: Some edges may be chosen multiple times.

Repeated Edges in RMAT generation: Theory

Theory

$$D(k+1) = \frac{1}{1-Q(k)}$$

$$Q(k) = Q(k-1) + \sum_{i=1}^{N} p_i n_i(k-1) \frac{p_i}{1-Q(i,k-1)}$$

$$Q(i,k) = Q(i,k-1) + \sum_{j=1 j \neq i}^{N} p_j n_j(k-1) \frac{p_j}{1-Q(j,k-1)}$$

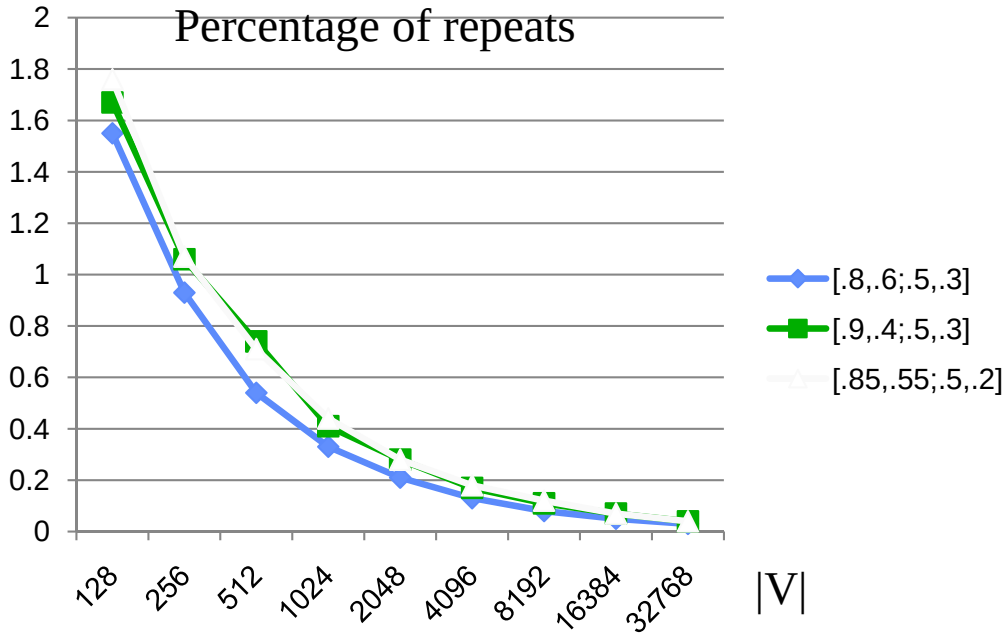
$$n_i(k) = \prod_{j=0}^{k-1} (1 - Q(i,j) - p_i)$$

D(k): expected number of draws to choose the kth distinct object.
Q(k): expected sum of probabilities of the first k objects
Q(i,k): expected sum of probabilities of the k objects, given object i is not among them

n(i,k): probability that the ith object is not chosen after k selections.

Summary: In theory, not too many repetitions are expected.

Experiment 4: Repeated edges in RMAT generation: practice and implications



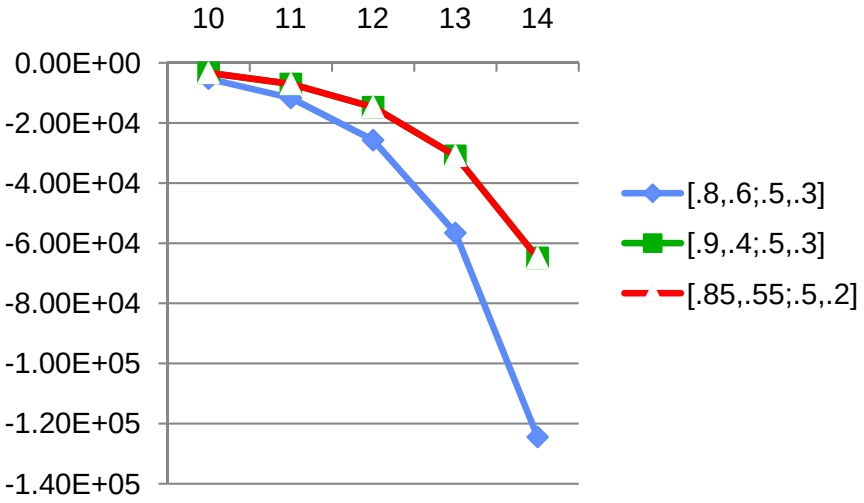
- There is not much repetition in practice either.

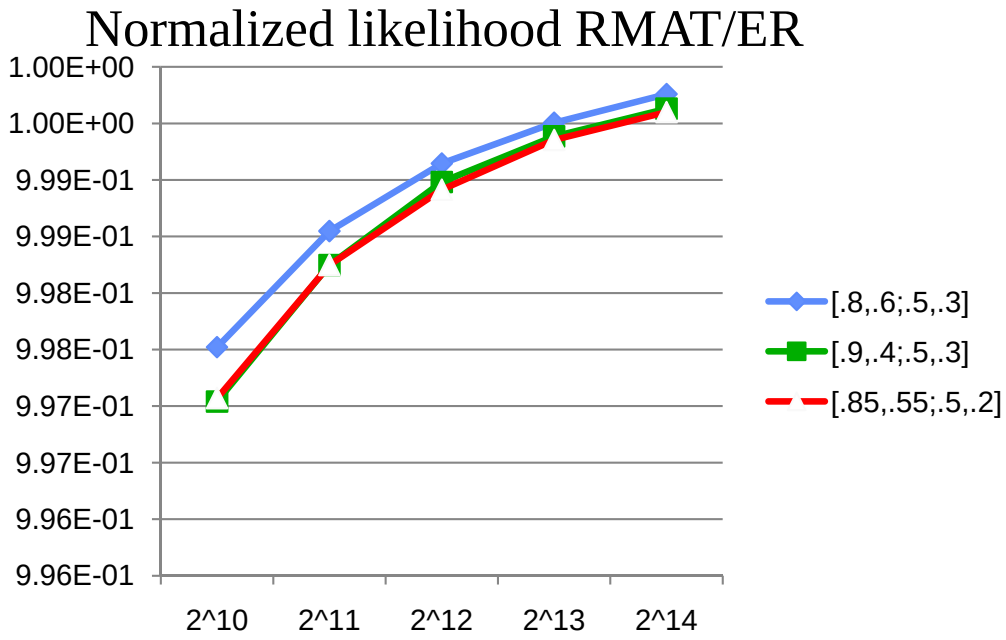
- Implications:
  - •Good efficiency in generating an RMAT graph.
  - •Two RMAT graphs generated form the same basis share very few edges, which implies
    - •Either two graphs generated form the same basis are not similar (experiments show they have similar features).
    - •Or we should not use individual edges as a unit of comparison.

Sandia National Laboratories

Experiment 5: Self-confidence

- Confidence: Can you recognize the graph you generated?





- The log likelihood metric does not distinguish a self-graph from an Erdos Renyi graph.

Sandia National Laboratories

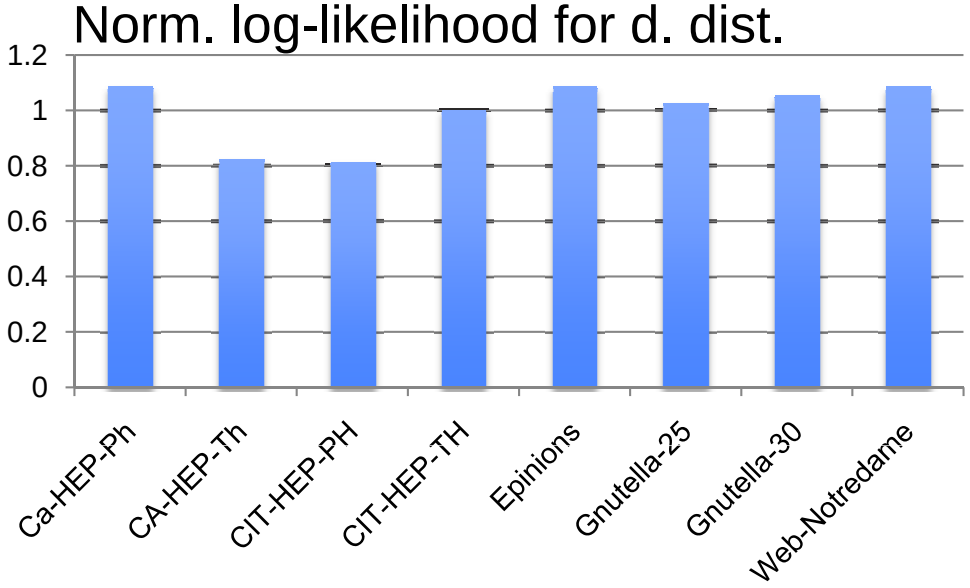
Unit of Comparison

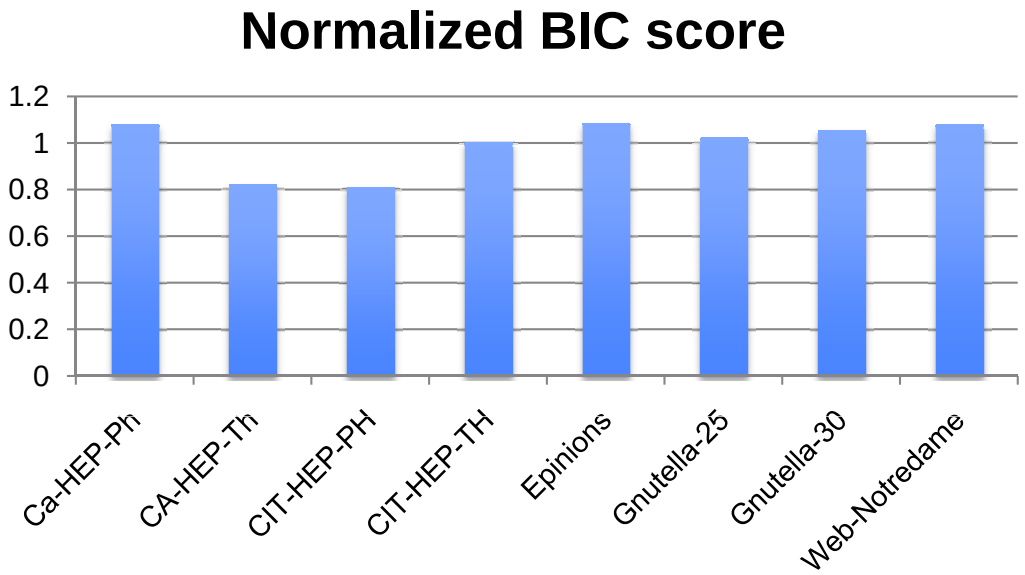
- The graphs we are studying are extremely sparse.
- A metric that is based an edge-by edge prediction will suffer from the skewed distribution of present and absent edges.

$$P(G|\mathcal{P},\sigma) = \prod_{(u,v)\in E} \mathcal{P}[\sigma_u,\sigma_v] \prod_{(u,v)\notin E} (1-\mathcal{P}[\sigma_u,\sigma_v])$$

- The dominant signal is the sparsity, edges only add a noise on top of the signal.
- Proposed alternative: comparison based on *carefully chosen* set of features.

Fitting features translates to edge-level accuracy





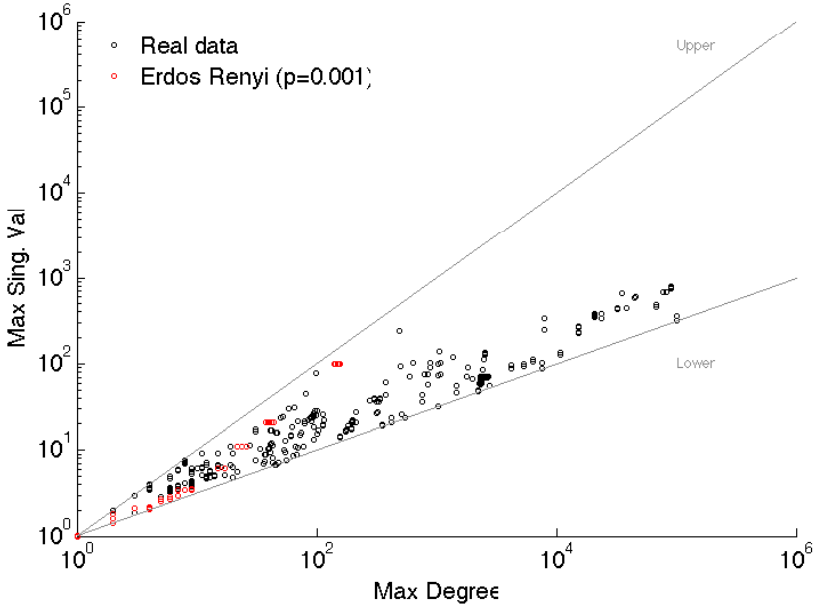
- Given the degree distribution we can predict edges.
- Expected number of edges between vertex *i* and vertex *j* would be $d_i d_j / |E|$
- In the experiments we used the exact degree distribution.

Sandia National Laboratories

Fitting R-Mat Parameters based on features

- RMAT has 4 independent parameters (3 for undirected graphs).
- Fitting with 4 features should help us compute these 4 parameters. We tried
  - Number of edges
  - In-degree distribution
  - Out-degree distribution
  - Largest singular value
- The first 3 metrics can be predicted by RMAT parameters. We used sampling for the last.
- Enables faster computations, better fit on features, and close fits on the log-likelihood function.

Selecting the features



- Parameter fitting or comparison of graphs based on features is sensitive to selection of the features.
- Features should be chosen to be independent, and span the space.
- Interesting result by Mihail and Papadimitriou
  - Largest eigenvalues of a graph with power law degree distribution can be predicted by the largest degrees.
  - Our experiments respectfully disagree.
  - We are trying to identify the source of the difference in predictions.

Sampling of Graphs

- Identifying dependencies among graph features requires statistical analysis.
- Real data sets, while essential, cannot help with controlled experiments.
- Sampling of graphs with a specified property will be essential for identifying dependencies between graph features.
- There are solid theoretical results for sampling from a given vertex degree distribution.
- We are developing techniques for joint degree distribution. We have
  - necessary and sufficient conditions for existence of a graph with a given distribution
  - an algorithm to construct an instance
  - a local perturbation technique to construct other instances.
  - proof that the state space is connected under this perturbation.
  - experiments that show promise.

Conclusions

- A bad metric can make anything look good.
- A metric that is based an edge-by edge prediction will suffer from the skewed distribution of present and absent edges.
- The dominant signal is the sparsity, edges only add a noise on top of the signal.
  - The real signal, structure of the graph is often lost behind the dominant signal.
- Proposed alternative: comparison based on *carefully chosen* set of features.
  - It is more efficient.
  - Sensitive to selection of features.
  - Finding independent set of features is an important area, and keep an eye on us for some important results.

Questions?

Sandia National Laboratories