Propagating Uncertainty in Network Problems

Eric D. Kolaczyk

kolaczyk@math.bu.edu

Dept of Mathematics and Statistics, Boston University

Collaborators and Support

Collaborators:

- Qi (Hawk) Ding (Boston University, Math & Stat)
- Wes Viles (Boston University, Math & Stat)
- Xiaoyu Jiang (Boehringer Ingelheim)

- David L. Gold (SUNY-Buffalo, Biostatistics)
- Natallia Katenka (Boston University, Math & Stat)

- Mark Crovella (Boston University, Computer Science)
- Paul Barford (UWisconsin-Madison, Computer Science)

Work supported in part by ONR awards N00014-06-1-0096 and N000140910654 and NSF grant CNS-0905565.

Focus: Uncertainty

- Lots of networks and network-data work out there . . .
- . . . but often little to no assessment of uncertainty associated with the network and/or inferences made based on the network.
- Networks often based on ‘lower-level’ data . . . which is typically subject to uncertainty . . . so ‘higher-level’ network tasks will be similarly subject!

Focus of current work is *propagation of uncertainty* from low to high in various canonical settings.

This Talk

Illustrate with snap-shots of four projects in this space.

1. Propagation of uncertainty from network inference to network statistics.

2. Relative importance of multiple node characteristics in association networks.

3. Node label prediction under negatively-biased training labels.

4. Impact of length-bias sampling on community detection from communication data.

Project 1: Uncertainty in Network Statistics

Common *modus operandi* in network analysis:

- System of elements and their interactions is of interest.
- Collect elements and relations among elements.
- Represent the collected data via a network $G = (V, E)$.
- Characterize properties of the network
  (e.g., density, clustering, centrality, etc.)

Project 1: Uncertainty in Network Statistics

Common *modus operandi* in network analysis:

- System of elements and their interactions is of interest.
- Collect elements and relations among elements.
- Represent the collected data via a network $G = (V, E)$.
- Characterize properties of the network
  (e.g., density, clustering, centrality, etc.)

Sounds good . . . so what?

Case Study: Network Density

- Let $G = (V, E)$ be an association network, where

$$\{i, j\} \in E \quad \text{iff} \quad \rho_{ij} \neq 0$$

- Reject $H_0 : \rho_{ij} = 0$ in favor of $H_1 : \rho_{ij} \neq 0$ if $\hat{\rho}_{ij} \in \Gamma_\alpha$
- Estimate G by $\hat{G}$, where

$$\{i, j\} \in \hat{E} \quad \text{iff} \quad \hat{\rho}_{ij} \in \Gamma_\alpha$$

Question: What is the accuracy of the observed network density $\hat{\delta}$ in estimating the actual network density δ ?

Network Density: Preliminary Results

- Simple plug-in estimate is biased, i.e.,

$$\mathbb{E}[\hat{\delta}] = \delta(\pi - \alpha) + \alpha$$

- Can derive asymptotically unbiased estimators with normal limiting distribution (i.e., as $n, N_v \to \infty$).
- Derivation develops conditions under which power $\to 1$ and asymptotic independence holds among the $\hat{\rho}_{ij}$.
- Conditions too strong . . . current work easing asymptotic independence using Stein methods $+$ bootstrap for estimation of variance.

Project 2: Multiple Node Characteristics

- What makes two people friends/colleagues/accomplices ?

  $\Rightarrow$ The totality of their many characteristics.

- Often have only partial information on node characteristics.

- How do inferred networks differ with choice of subset of characteristics?

Goal: Comparative study of local and global network characteristics based on partial and full characteristic sets.

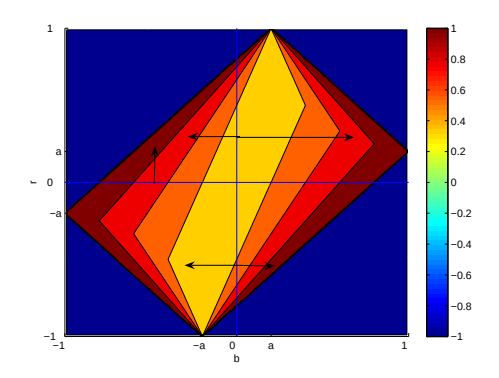
Case Study: Two Characteristics

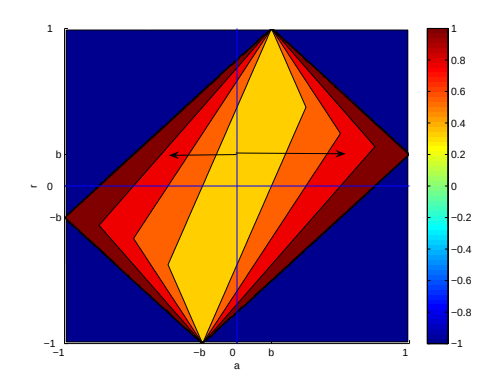
- Each node i associate with a pair of characteristics $\mathbf{X}_i = (X_i^{(1)}, X_i^{(2)})^T$.
- Denote covariance between elements of $\mathbf{X}_i$ and $\mathbf{X}_j$ as

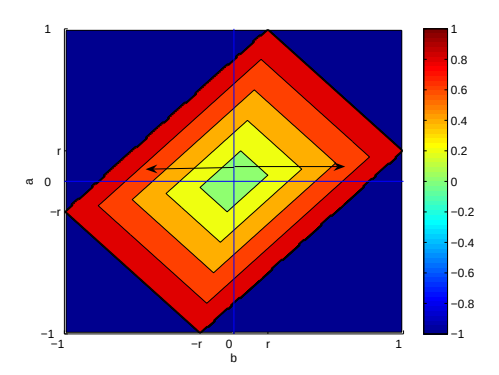
$$\Sigma = \begin{pmatrix} \Sigma_m & \Sigma_{c\cdot} \\ \Sigma_c & \Sigma_m \end{pmatrix} \tag{1}$$

- Define three networks
  1. $G^{(1)}$, based on corr$(X_i^{(1)}, X_j^{(1)})$
  2. $G^{(2)}$, based on corr$(X_i^{(2)}, X_j^{(2)})$
  3. G, based on cancorr$(\mathbf{X}_i, \mathbf{X}_j)$

Relationships Among Correlations Are Nontrivial

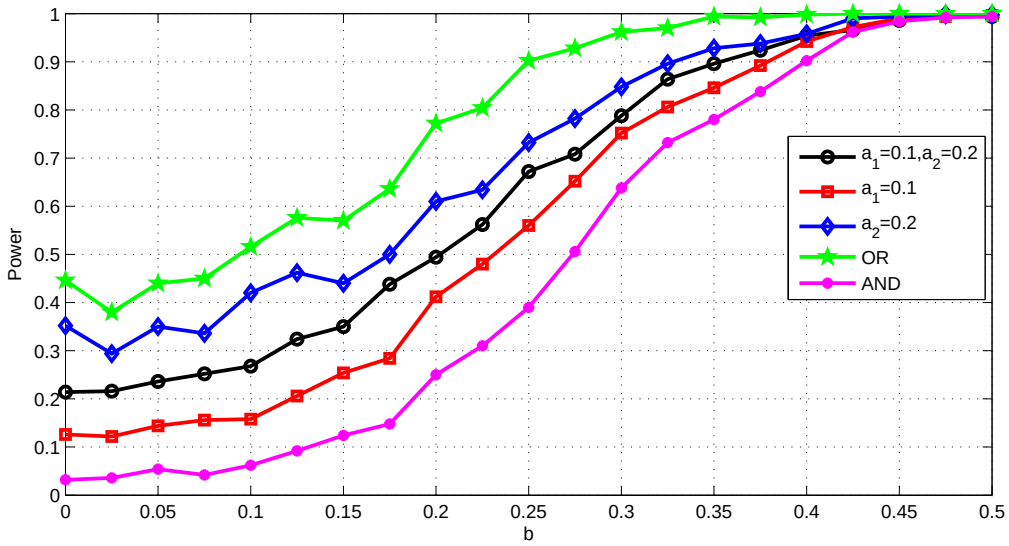






- Plots show heat-maps for cancorr as a function of the other correlation parameters.
- Constraints exist on range of the different correlations.

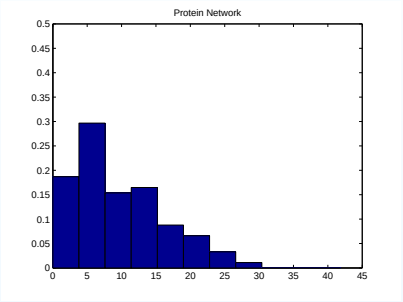
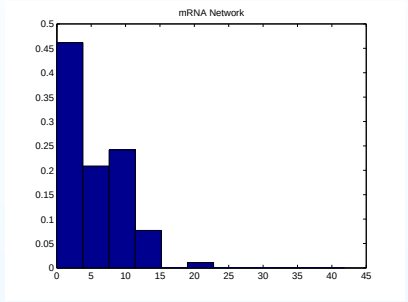
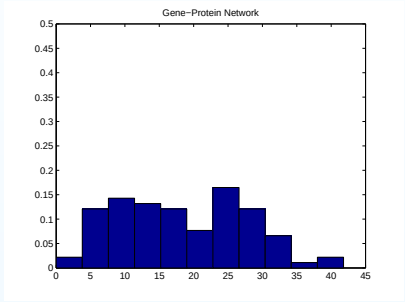
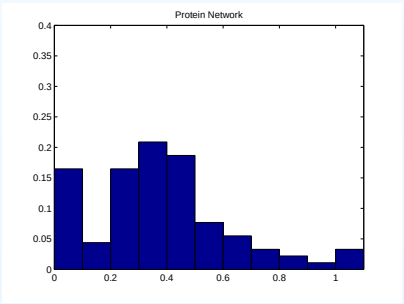
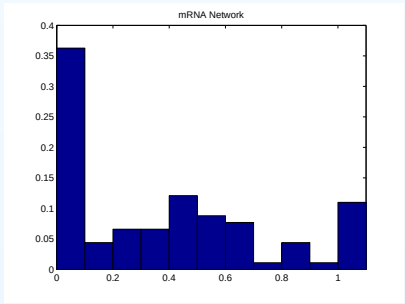
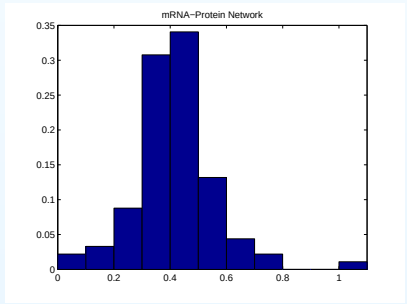
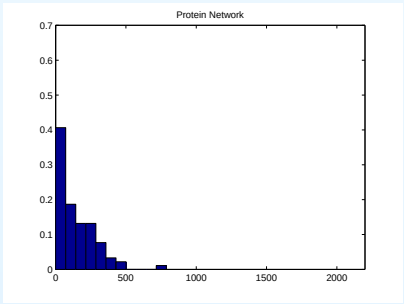
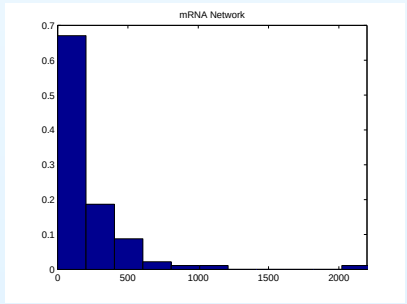
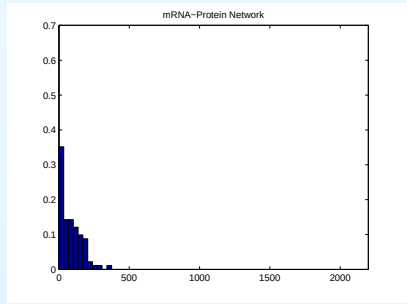
Power in Detecting an ‘Edge’



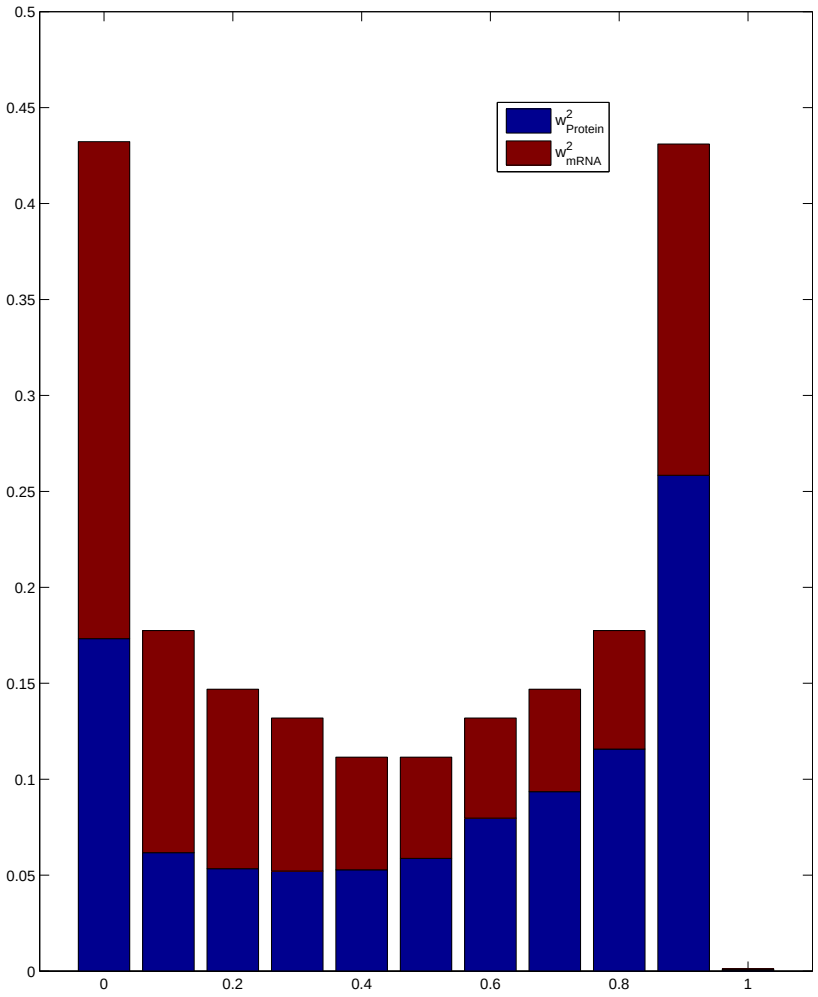
Power versus corr$(X_i^{(1)}, X_j^{(2)})$.

Message: Incomplete information on or improper handling of multiple characteristics leads to excessive/insufficient detection of edges.

Illustration: NCI-60 Gene/Protein Expression Data

	Protein Network	Gene Network	Protein-Gene Network
Deg			
Clu			
Btw			

How Much Protein/Gene in an Edge?



Project 3: Node Label Prediction

- Given a graph $G = (V, E)$ and labels y_i for some of the vertices $i \in V^{obs}$.
- Goal is to predict y_i for remaining vertices $i \in V^{miss}$.
- Various methods
  - Nearest neighbor
  - Markov random field
  - Kernel methods
  - Etc.

Uncertainty Due to Negative Label Bias

What if the node labels are not 100% reliable?

Example: Protein function prediction.

- Predict protein function using protein-protein interaction (PPI) networks.
- Network
  - Nodes represent proteins.
  - Edges indicate affinity for binding among proteins.
  - Labels show whether or not protein has a given biological function.
- Problem: Negative labels are highly biased, due to
  - clustering of scientific inquiry,and
  - positive nature of scientific publication.

Network-based Auto-probit Modeling

Need probabilistic modeling of error due to negative bias in order to propagate uncertainty in a conceptually rigorous fashion.

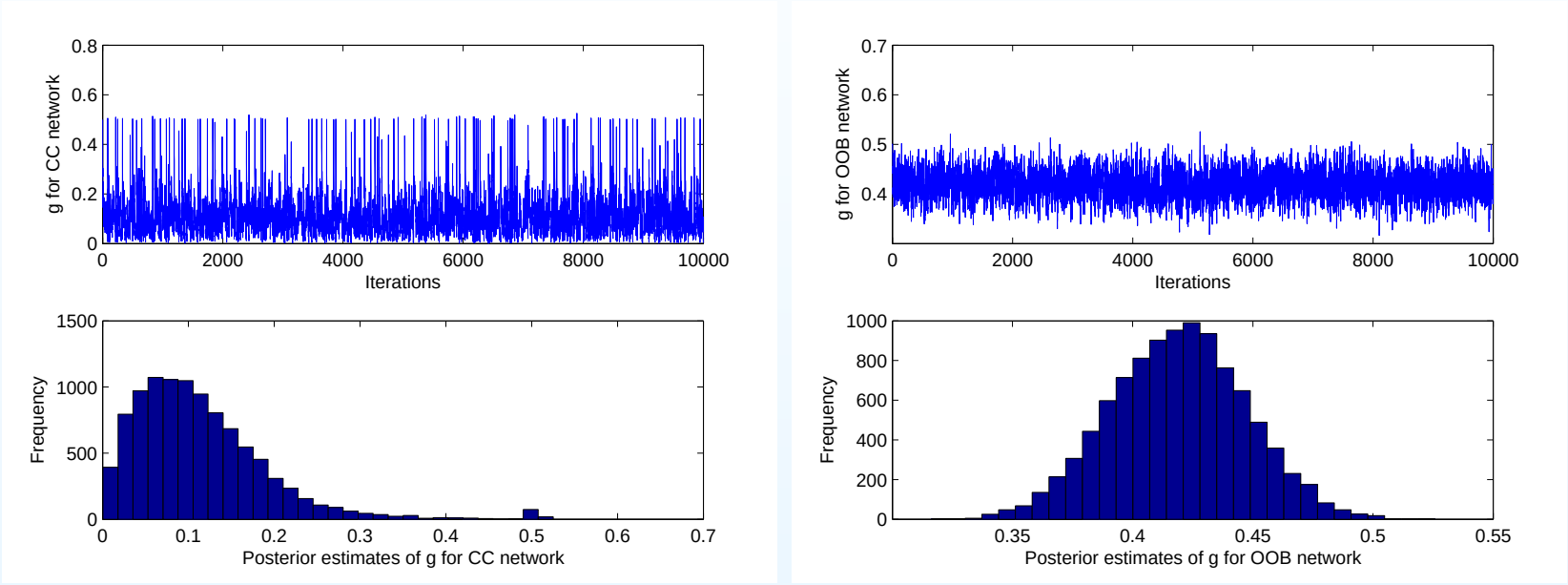
We develop

- a network-based auto-probit model,
- coupled with a biased coin-flip for label status

Task: Given training data at some proteins, predict others, accounting for unknown rates of uncertainty in labels.

Posterior Inference for g

Trace plots and histograms of the posterior estimates of g.



[Left]: CC network; [Right]: OOB network.

Pseudo *in vitro* Evaluation

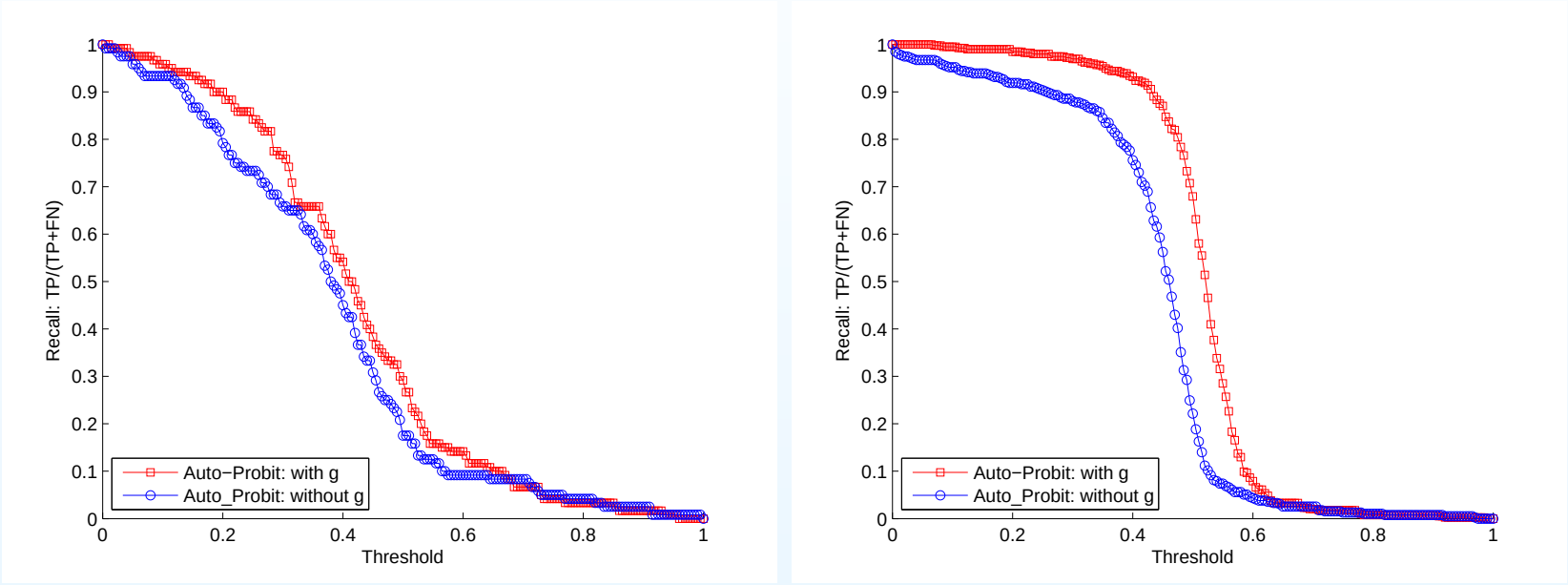
We demonstrate the model's capability of producing more accurate predictions by incorporating the annotation uncertainty

- Compare the auto-probit model with and without g
- Fit the models and estimate parameters based on "old" GO annotations (updated in June '06)
- Predict the target functions and evaluate against "new" annotations (updated in November '07)
- Evaluate model performance by

$$\text{Recall} = \frac{TP}{TP + FN} = \text{Sensitivity}.$$

Results: Using GO Annotation Uncertainty

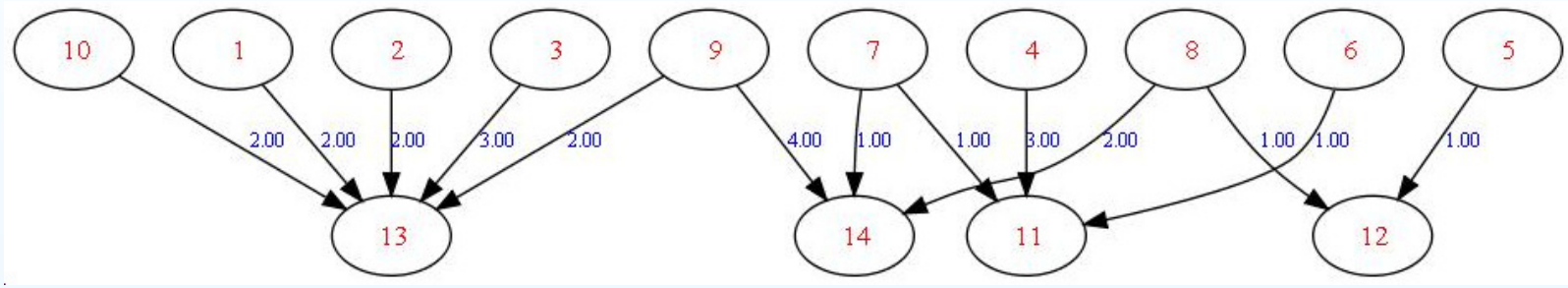
Plot of recall (sensitivity) versus prediction threshold



CC (left) and OOB (right) networks

Project 4: Sampling Bias and Community Detection

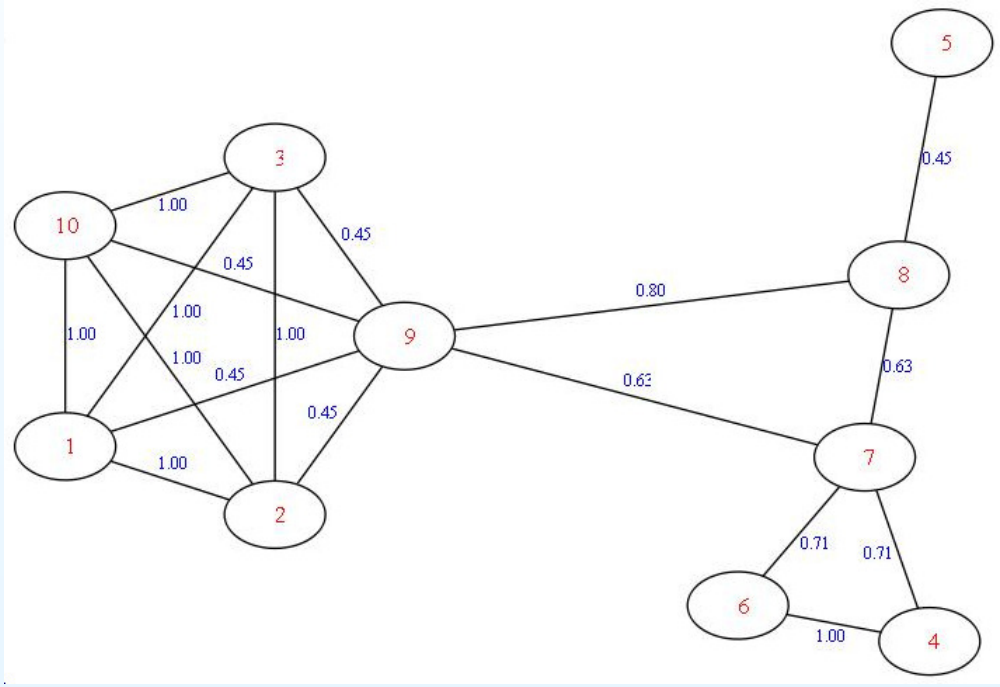
A bipartite representation of traffic flows in Géant.



- Top: Source IP addresses
- Bottom: Destination IP addresses
- Weights: Number of flows sent

Sampling Bias and Community Detection (cont)

A one-mode projection onto communication sources.



Question: What is the impact on the source graph topology, and the resulting community detection, of standard router-based flow sampling?

Discussion

- Common theme: Uncertainty in low-level data and its impact on high-level network tasks.
- Impact of accounting for low-level uncertainty can be quite dramatic.
- Need for substantially greater recognition of the issue, formulation of canonical problems, and work on solutions.