A Controlled Sensing Approach to Graph Classification

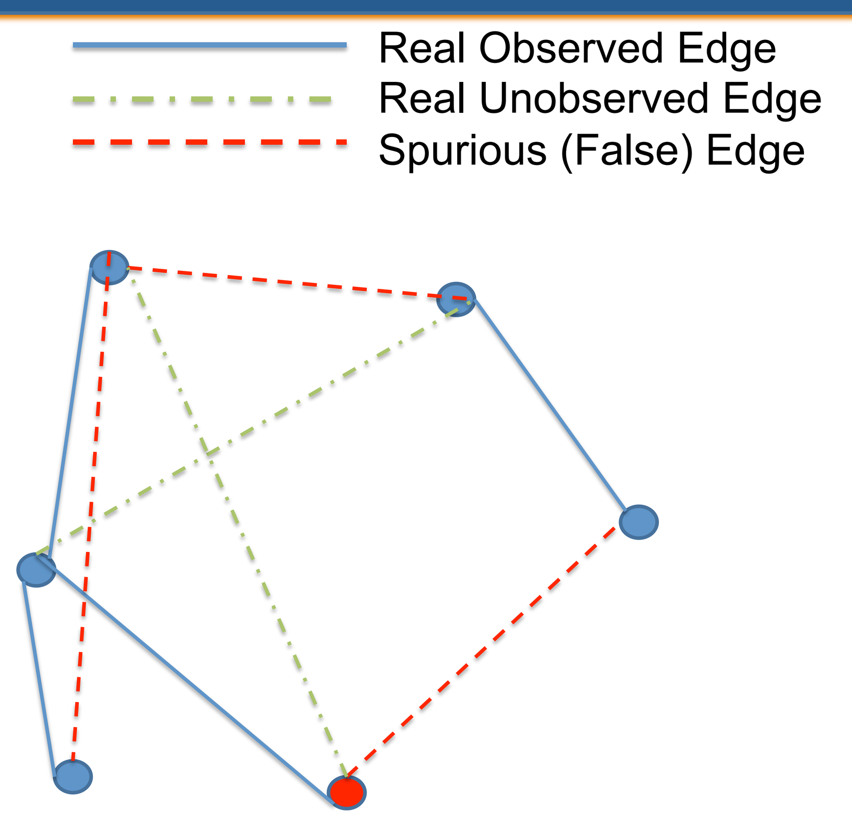
Jonathan G. Ligo[1], George K. Atia[2], Venugopal V. Veeravalli[1]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
[2]Department of Electrical Engineering and Computer Science, University of Central Florida

Graph Classification

- Classify a graph based on connectivity via probabilistic observations of edges
- Applications: Epidemic prediction/ detection, Social network analysis
- Objective: Balance cost of sampling with classification performance
- Framework: Sequential Hypothesis Testing with Control

Mathematical Model



- Fixed underlying graph $G=(V,E)$
- At each time select a node to observe (Red, Control)
- When node i is selected, observations y are subset of possible incident edges
  - Real edges are observed with probability p
  - Spurious edges are observed with probability $q<p$

$$P_G^i(y)=q^{|y\cap E^C|}(1-q)^{((N-1)-d_i)-|y^C\cap E^C|}p^{|y\cap E|}(1-p)^{d_i-|y^C\cap E|}$$

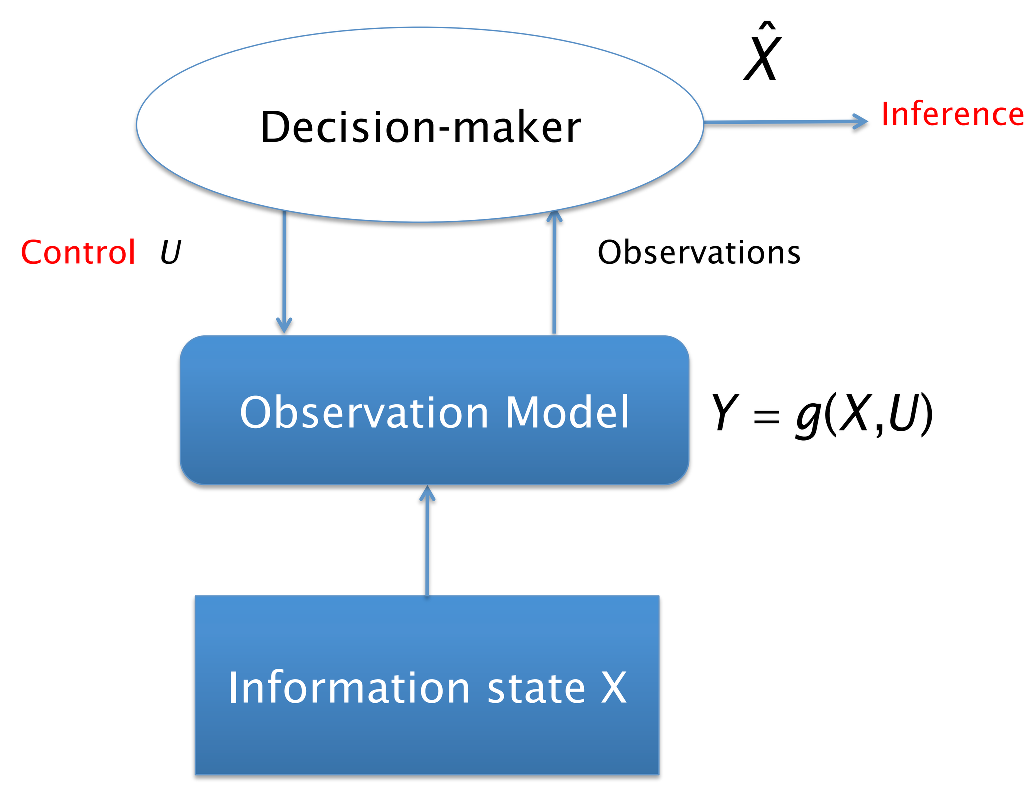
- Degree of a node: Number of edges incident to node
- Average node degree: $\bar{d}_G=\frac{2|E|}{|V|}$

Acknowledgements

This work was supported by the U.S. Defense Threat Reduction Agency through subcontract 147755 at the University of Illinois from prime award HDTRA1-10-1-0086.

Controlled Sensing

- Framework for hypothesis testing with control [1,2]
- Control affects quality of observations rather than evolution of information state



Problem Formulation

- Define two classes of graphs

$$\mathcal{G}_0=\{G:|V|=N,\bar{d}_G\le\eta\}$$
$$\mathcal{G}_1=\{G:|V|=N,\bar{d}_G>\eta\}$$

- Threshold for high connectivity η
- Binary Composite Hypothesis Test

$$H_0:G\in\mathcal{G}_0$$
$$H_1:G\in\mathcal{G}_1$$

- Proposed controlled sensing algorithm gives asymp. optimal error decay with sample size by [1,2]

Graph Estimation

- Simple Maximum-Likelihood Approach
- Let e_{ij} be possible edge between vertices i,j
- Let $\mathcal{T}_i(k)=\{\text{Times node } i \text{ is sampled up to time } k\}$

$$\mathcal{T}_{ij}(k)=\mathcal{T}_i(k)\cup\mathcal{T}_j(k)$$
$$l_{ij}=\#\text{of times edge } e_{ij} \text{ is observed}$$

- $e_{ij}\in\hat{G}(y^k,u^k)$ if

$$p^{l_{ij}(k)}(1-p)^{|\mathcal{T}_{ij}(k)|-l_{ij}(k)}>q^{l_{ij}(k)}(1-q)^{|\mathcal{T}_{ij}(k)|-l_{ij}(k)}$$

- Solvable in linear time in N

Proposed Algorithm [3]

At each time k,

1. Find maximum-likelihood estimate of graph $\hat{G}=\hat{G}(y^k,u^k)\in\mathcal{G}_i$
2. Estimate hypothesis $\hat{i}(k)$ from $\hat{G}$
3. Stop if (stopping rule)

$$\min_{\tilde{G}\in\mathcal{G}_j}\log\frac{P_{\hat{G}}(y^k,u^k)}{P_{\tilde{G}}(y^k,u^k)}>\log\beta \text{ where } j\neq i$$

where $P_G(y^k,u^k)$ is the induced joint distribution of the observations and controls.
Else, select next node u_{k+1} to observe according to distribution q^* solving

$$\max_{q(u)u\in V}\min_{\tilde{G}:\tilde{G}\in\mathcal{G}_j}\sum_{u=1}^{N}q(u)D(P_{\hat{G}}^u,P_{\tilde{G}}^u)$$

where β is a design parameter (control policy)

Stopping Rule

- Minimizer found by moving edges from $\hat{G}^C$ to $\hat{G}$ ($\hat{G}\in\mathcal{G}_0$) or vice versa($\hat{G}\in\mathcal{G}_1$)
- Consider $\hat{G}\in\mathcal{G}_0$
- Define cost of moving edge $e_{ij}\in\hat{G}^C$

$$\delta_{ij}=l_{ij}(k)\log\frac{q}{p}+\left(|\mathcal{T}_{ij}(k)|-l_{ij}(k)\right)\log\frac{1-q}{1-p}$$

- Minimum value is sum of $\left\lceil(\eta-\bar{d}_{\hat{G}})\frac{N}{2}\right\rceil$ weights
- Analogous for $\hat{G}\in\mathcal{G}_1$ (swap $\hat{G},\hat{G}^C$), negate δ_{ij}
- Solvable in $O(N\log N)$ time

Control Policy

- Two player zero-sum game
  - Player 1: Choose control to maximize avg. KL-distance between estimate and $\mathcal{G}_j$
  - Player 2: Choose graph under $\mathcal{G}_j$ to minimize avg. KL-distance
- Pose in terms of incidence matrix ($\hat{G}\in\mathcal{G}_0$)

$$\max_{q\in P_N}\min_{x\subseteq IS}qM_{\hat{G}^C}x$$

$P_N=$ Probability distributions on N nodes

$IS=$ Edges to insert into $\hat{G}$ to be in $\mathcal{G}_1$

$M_{\hat{G}^C}=$ Incidence Matrix of $\hat{G}^C$

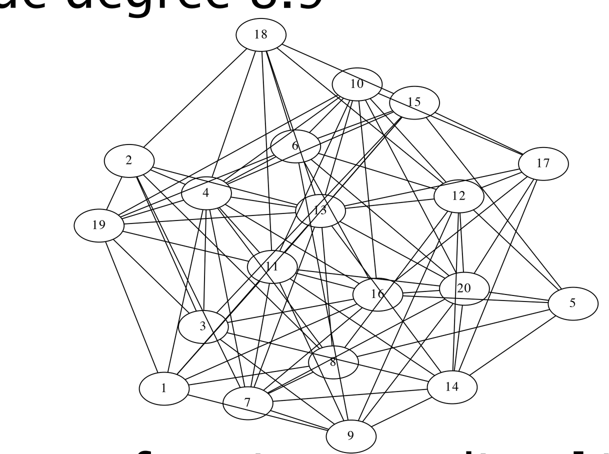
- LP Relaxation

$$\max_{q\in P_N}\min_{\{x\in\mathbb{R}^{|E_{\hat{G}^C}|}:x\ge0,\|x\|_1=\lceil|\eta-\bar{d}_{\hat{G}}|N/2\rceil\}}qM_{\hat{G}^C}x$$

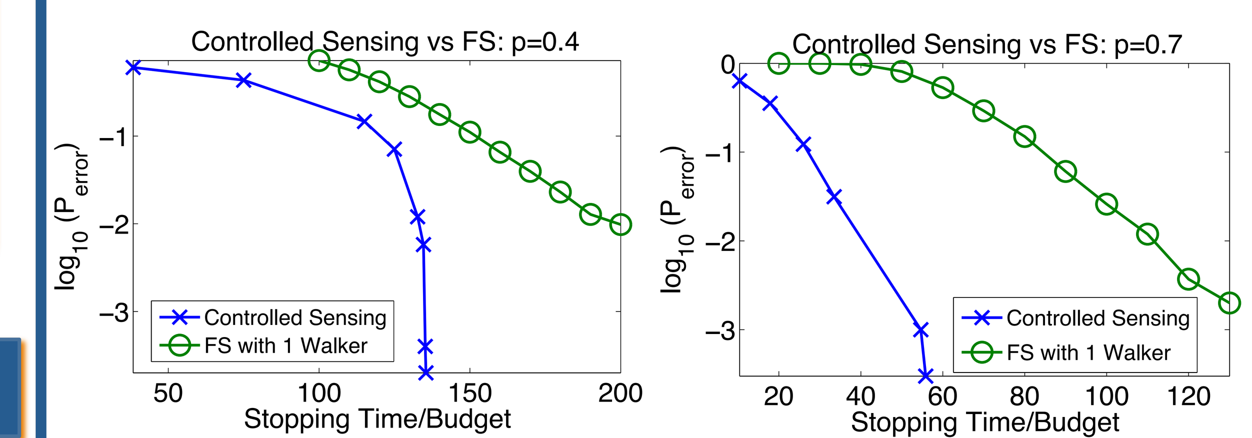
- Analogous when $\hat{G}\in\mathcal{G}_1$ (swap $\hat{G},\hat{G}^C$)

Numerical Results [3]

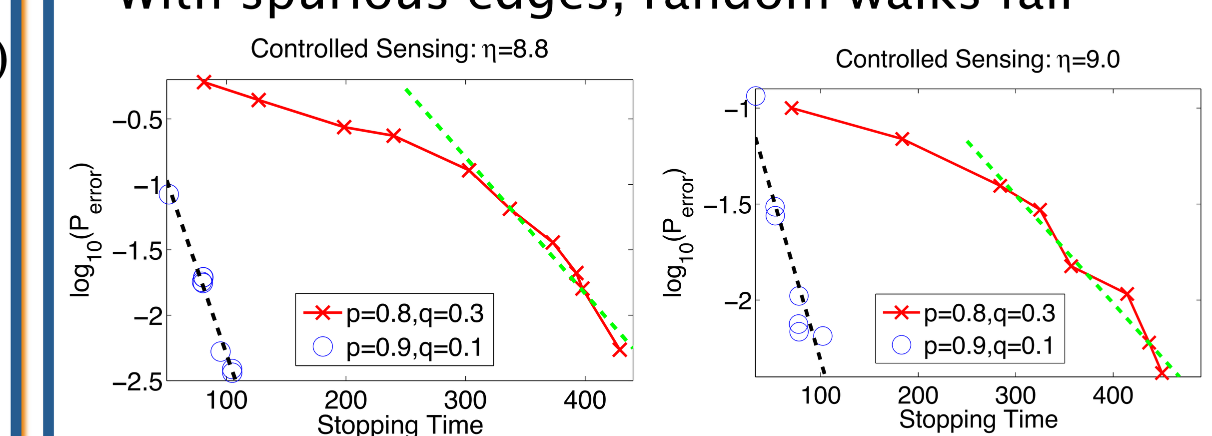
- Consider the following 20 node graph with average node degree 8.9



- Comparison to frontier sampling [4], a weighted random walk technique with no spurious edges, $\eta=8.8$



- With spurious edges, random walks fail



Conclusions and Future Work

- Proposed an asymptotically optimal sequential hypothesis test with control to classify graphs by connectivity
- Future work involves validation on large data sets, computationally simple approximations of the controlled sensing scheme

References

[1] H. Chernoff, "Sequential design of experiments," *Ann. Math. Statist.*, vol. 30, pp. 755–770, 1959.
[2] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for multihypothesis testing," *To appear in IEEE Trans. Autom. Contr.*, October 2013.
[3] J. G. Ligo, G. K. Atia, and V. V. Veeravalli, "A Controlled Sensing Approach to Graph Classification," in *Proc. of the 38-th Int. conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, Vancouver, Canada, May 2013, IEEE.
[4] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, New York, NY, 2010, IMC '10, pp. 390–403, ACM.