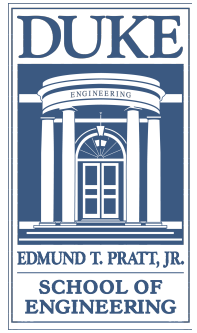
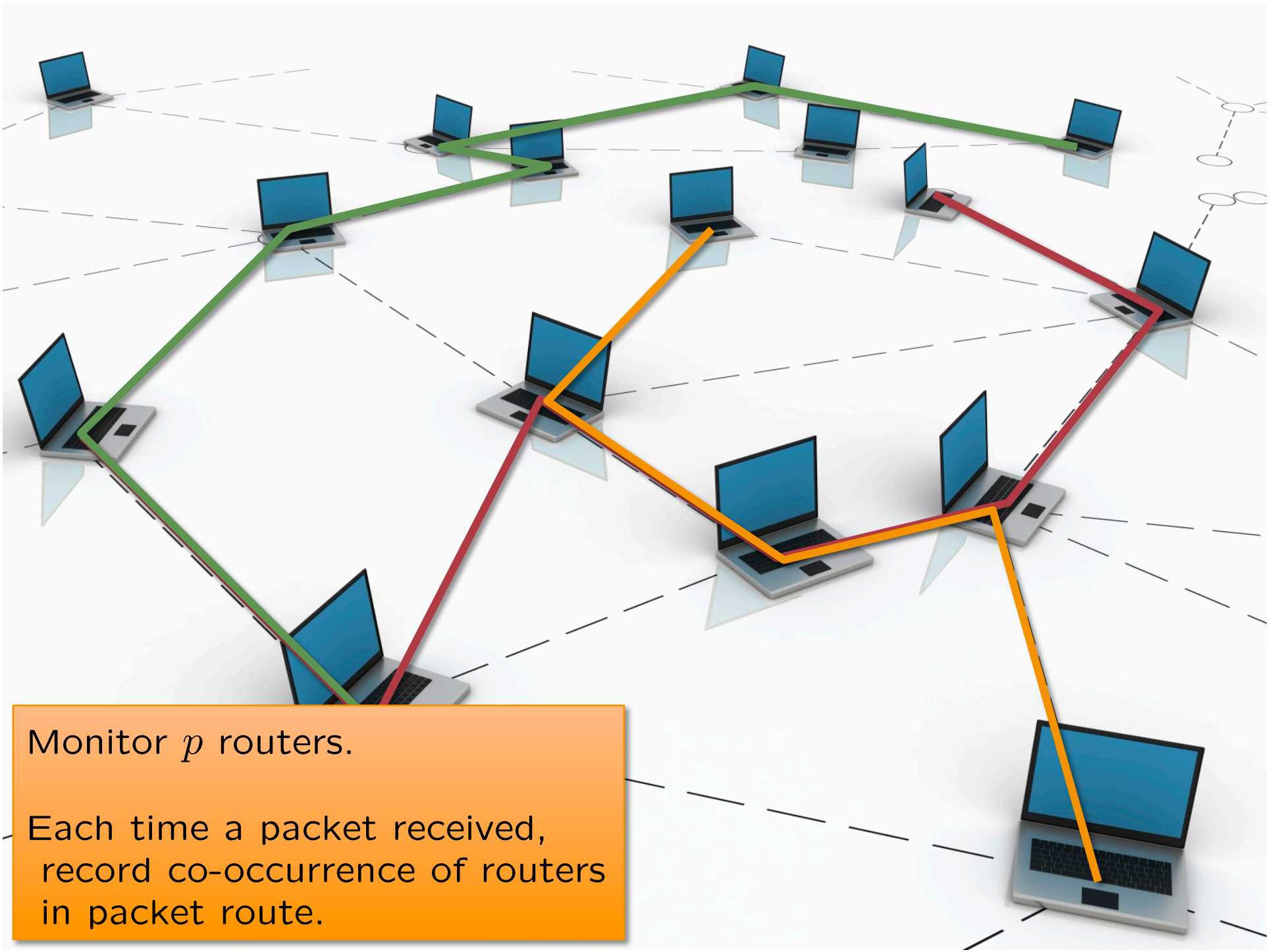
Hypergraph-based Inference on Network Routes

Rebecca Willett

With Jorge Silva, Maxim Raginsky, and Svetlana Lazebnik (UNC)

NIS
Network and Imaging Science Laboratory



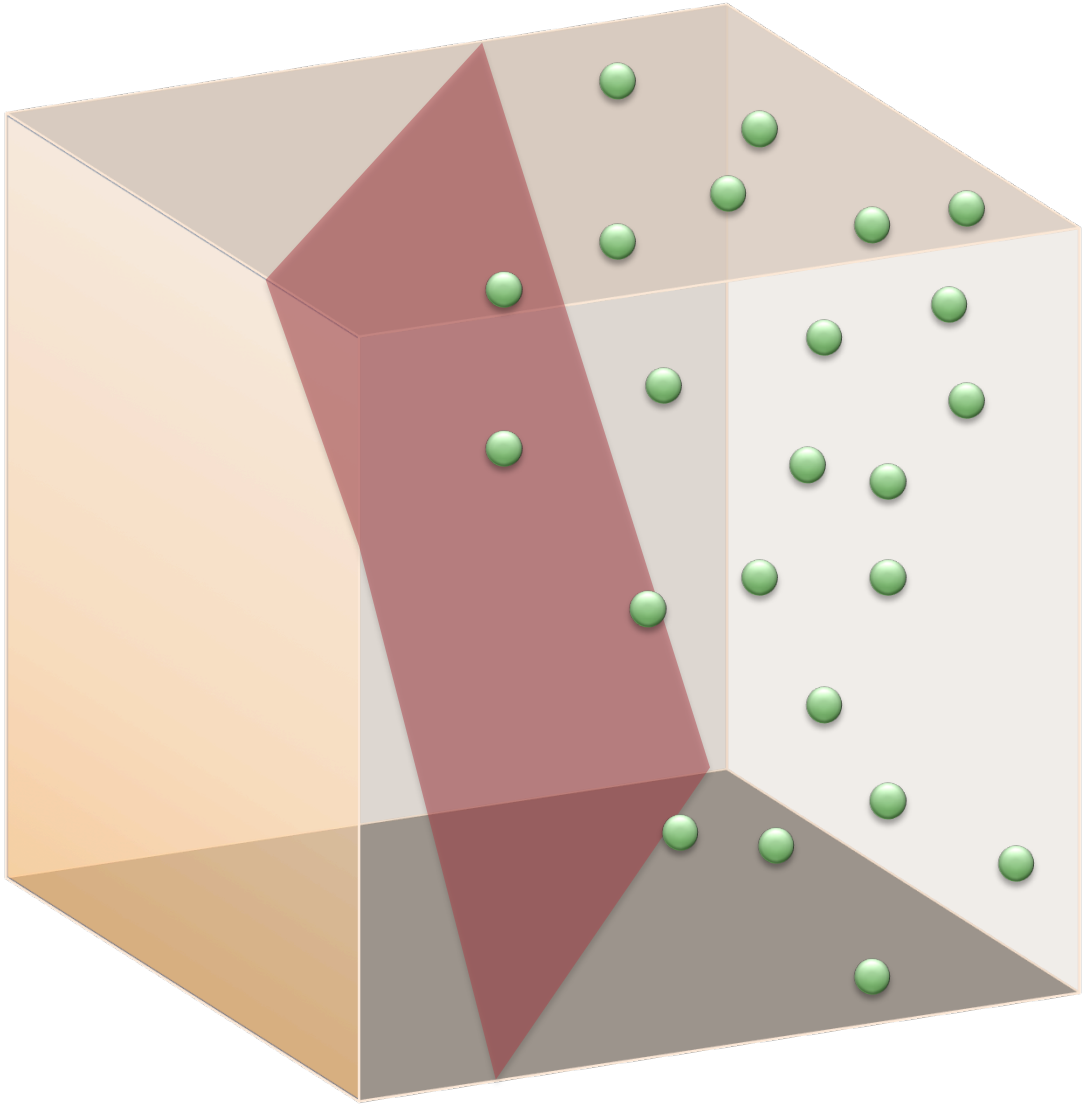

Monitor p routers.
Each time a packet received, record co-occurrence of routers in packet route.

p routers being monitored
n unlabeled training observations

1. How can we detect anomalous co -occurrences?

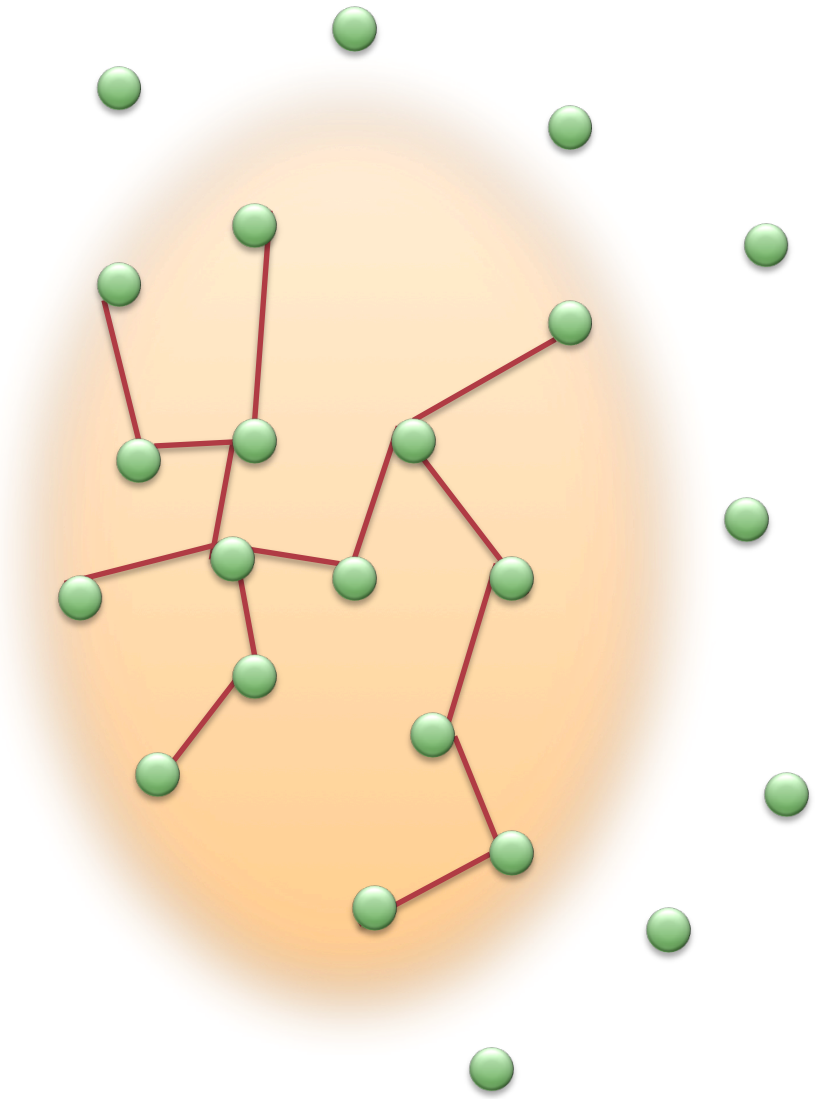
2. How can we determine which routers are most predictive?

One-class SVM

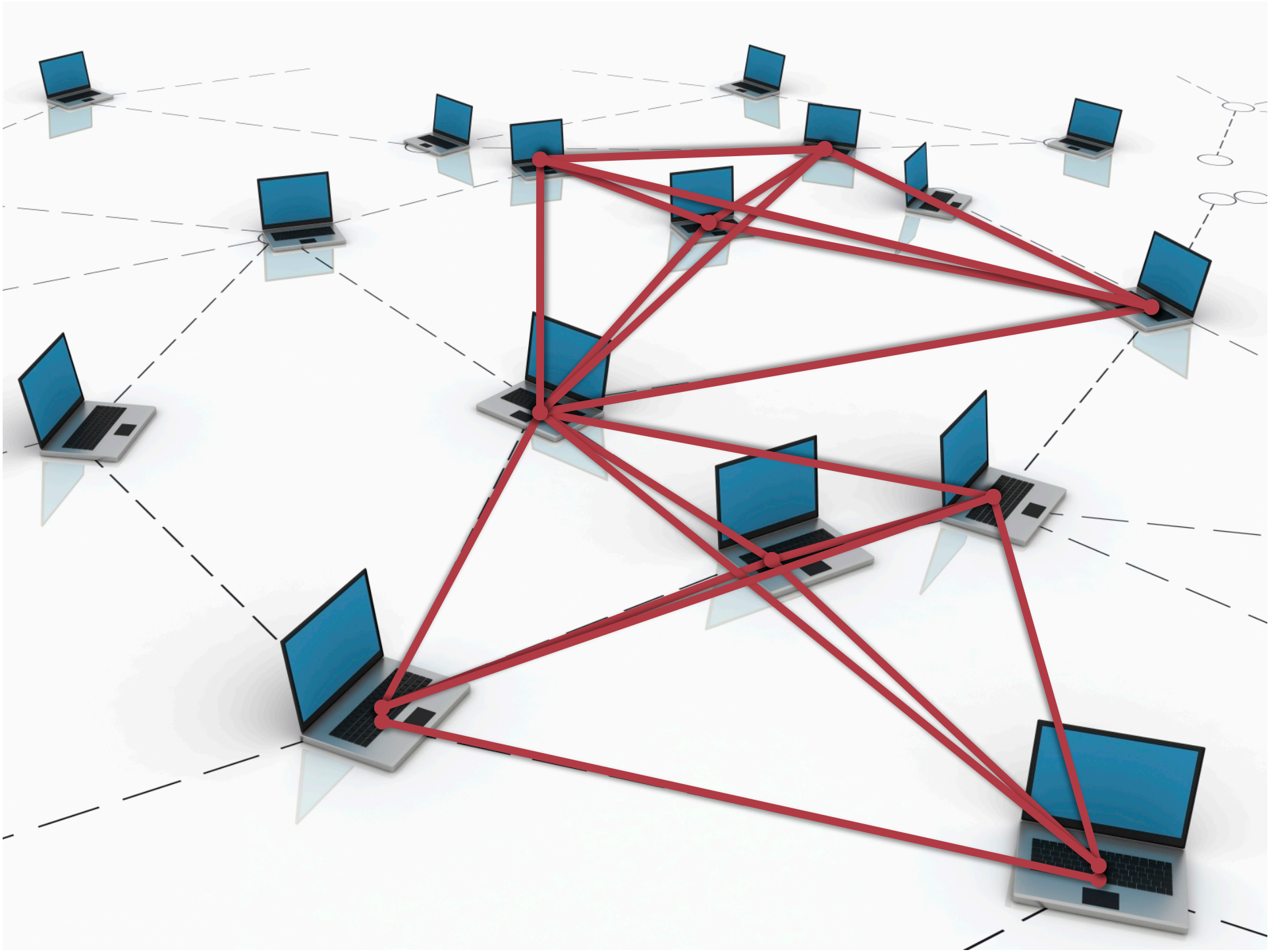


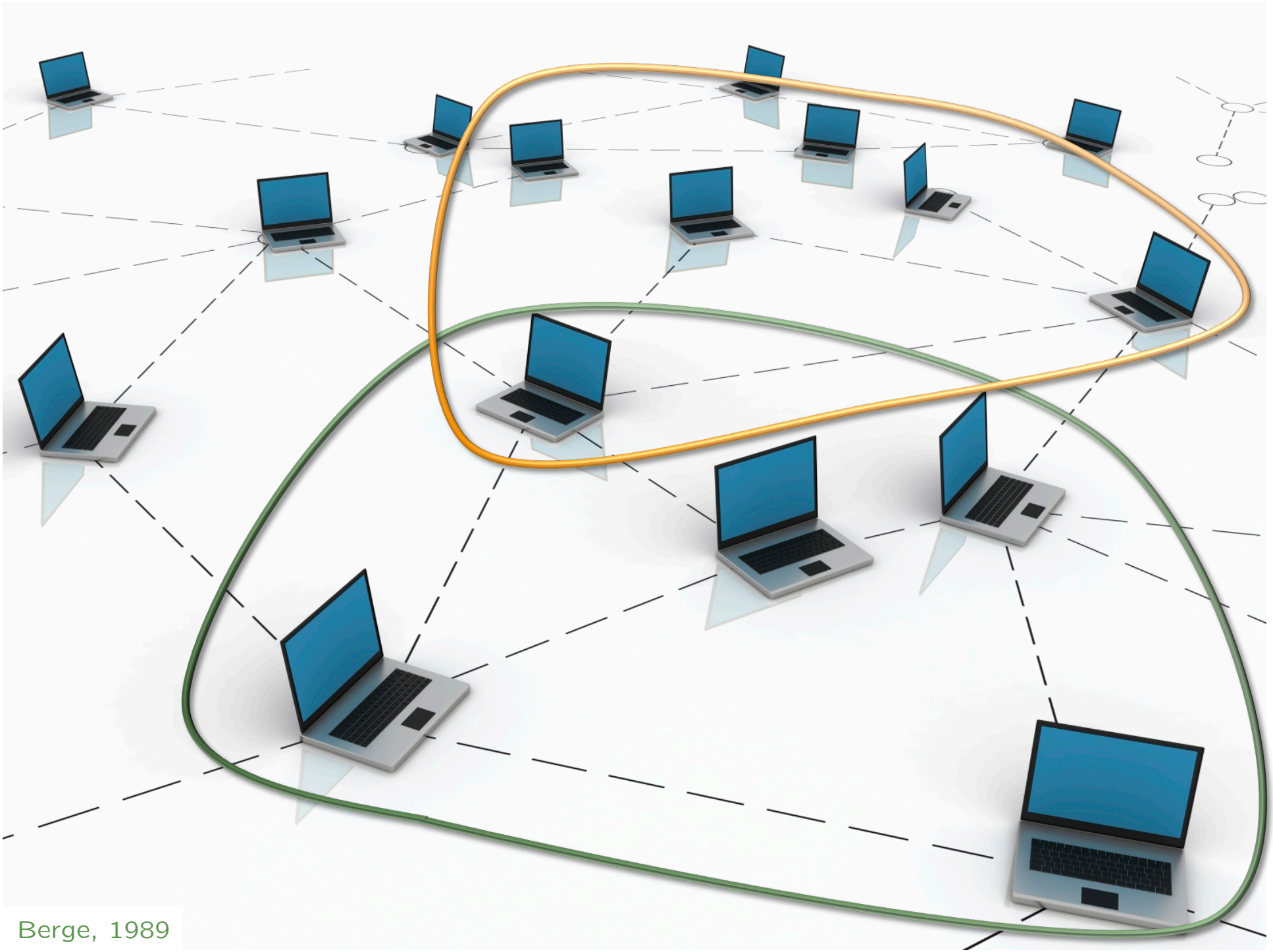
Scholkopf, Platt, and Shawne-Taylor, 2001

K-point minimum spanning trees



Hero, 2007





Berge, 1989

Observations

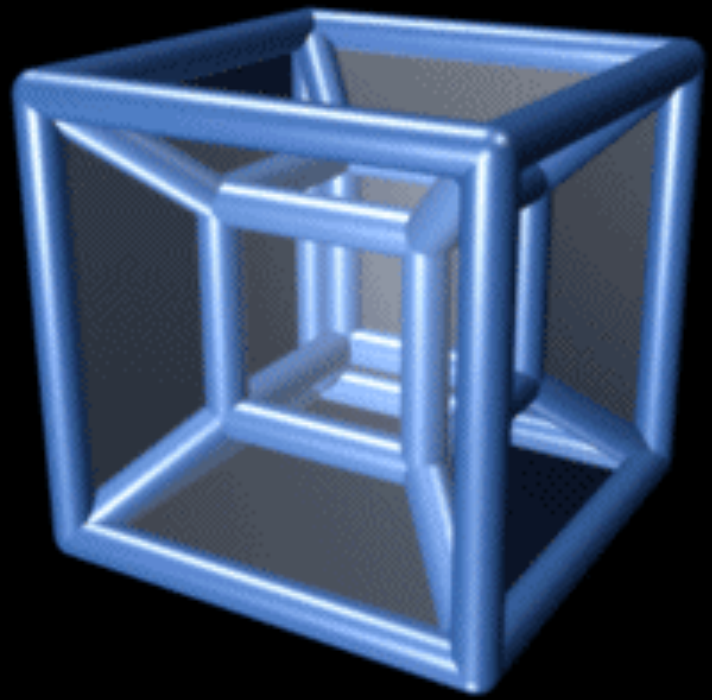
$$\boldsymbol{x} \in \{0, 1\}^p$$

$x_j = 1$ if j^{th} entity participates in co-occurrence / hyperedge

Challenge: 2^p possible hyperedges (and p is really big)

Orthogonal series

2^p possible hyperedges
= 2^p hypercube vertices
= 2^p basis functions
= intractable...



Mixture model

$$g(\boldsymbol{x}) = (1-\pi)f(\boldsymbol{x}) + \pi\mu(\boldsymbol{x})$$

Nominal distribution (unknown)

Mixing parameter (unknown)

Anomalous distribution (known)

Anomalous hyperedges *(what we want to estimate)*:

$$\mathcal{A}^* = \{\boldsymbol{x} : (1-\pi)f(\boldsymbol{x}) < \pi\mu(\boldsymbol{x})\}$$

Variational approximation

$$f(\boldsymbol{x}) = \prod_{j=1}^{p} f_j(x_j)$$

Nominal distribution

Marginals

Because $x_j \sim$ Bernoulli, $f_j(x_j) \equiv \theta_j^{x_j}(1-\theta_j)^{1-x_j}$

(Sufficiently rich for many high-dimensional problems)

Leads to $O(np)$
Variational Expectation-Maximization algorithm
and computation of annotations

Beylkin, Garcke, and Mohlenkamp, 2007
McLachlan and Krishnan, 1996
Peterson and Anderson, 1987

Unimodal: variational approximation

$$f(\boldsymbol{x}) = \prod_{j=1}^{p} f_j(x_j)$$

Nominal distribution

Marginals

Multimodal: mixture of variational approximations

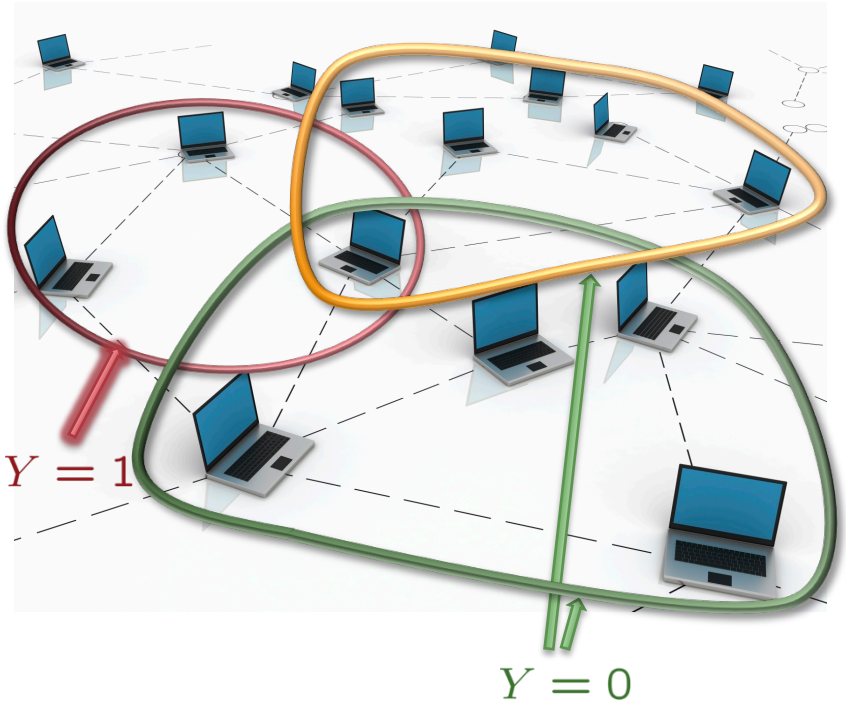
$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{p} f_{k,j}(x_j)$$

Weight on k^{th} component

Hidden data

Consider unobserved binary r.v.:

$$Y \equiv \begin{cases} 1, & \boldsymbol{X} \sim \mu \\ 0, & \boldsymbol{X} \sim f \end{cases}$$



Let $\eta(\boldsymbol{x}) \equiv \mathbb{P}(Y = 1 | \boldsymbol{X} = \boldsymbol{x}, f, \pi)$

Note $\mathcal{A}^* = \{\boldsymbol{x} : (1-\pi) f(\boldsymbol{x}) < \pi \mu(\boldsymbol{x})\}$

$= \{\boldsymbol{x} : \eta(\boldsymbol{x}) > 1/2\}$

Good estimate of $\eta \Rightarrow$ good estimate of $\mathcal{A}^*$

Variational EM Algorithm

E-step

$$\widehat{\eta}^{(t+1)}(\boldsymbol{x}_i) = \frac{\widehat{\pi}^{(t)}\mu(\boldsymbol{x}_i)}{(1-\widehat{\pi}^{(t)})\widehat{f}^{(t)}(\boldsymbol{x}_i) + \widehat{\pi}^{(t)}\mu(\boldsymbol{x}_i)}$$

M-step

$$\widehat{\pi}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\eta}_i^{(t+1)}$$

$$\widehat{\theta}_j^{(t+1)} = \frac{\sum_{i=1}^{n}(1-\widehat{\eta}_i^{(t+1)})x_{i,j}}{\sum_{i=1}^{n}(1-\widehat{\eta}_i^{(t+1)})}$$

$$\widehat{f}^{(t+1)}(\boldsymbol{x}_i) = \prod_{j=1}^{p}(\widehat{\theta}_j^{(t+1)})^{x_{i,j}}(1-\widehat{\theta}_j^{(t+1)})^{1-x_{i,j}}$$

- Densities only need to be evaluated at training hyperedges (not all 2^p possible hyperedges)
- Since marginals Bernoulli, complexity linear (not quadratic) in n

Wolfe, 1970

Annotations

$$\mathcal{A}_i = \{\boldsymbol{x} : f(\boldsymbol{x}) < f(\boldsymbol{x}_i)\}$$

hyperedges more anomalous than $\boldsymbol{x}_i$

smaller $\mathcal{A}_i \Leftrightarrow \boldsymbol{x}_i$ more anomalous

$$\text{pFDR}(\mathcal{A}_i) = \mathbb{P}(\boldsymbol{X} \sim f | \boldsymbol{X} \in \mathcal{A}_i)$$

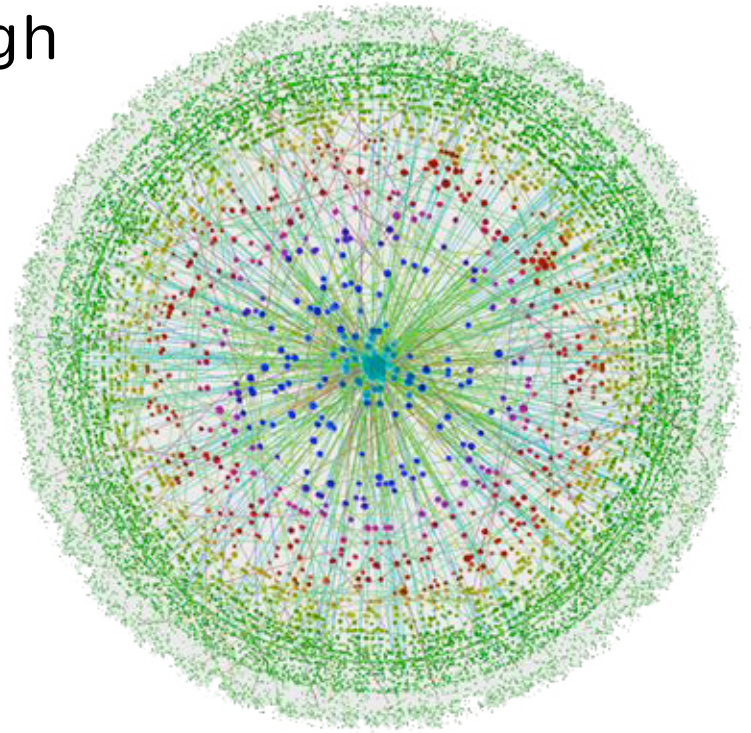
$$\begin{aligned} \gamma_i &\equiv 1 - \text{pFDR}(\mathcal{A}_i) \\ &= \frac{\pi \mathbb{U}(\mathcal{A}_i)}{(1-\pi)\mathbb{F}(\mathcal{A}_i) + \pi \mathbb{U}(\mathcal{A}_i)} \end{aligned}$$

Easy to estimate! (Monte Carlo efficient because of variational approx.)

Scott and Kolaczyk, 2007
Storey, 2003

NICO data

$n = 250$ router-level paths through
$p = 1105$ nodes
3 sources, 83 destinations
average path length of 15



Collected using `traceroute`,
but *order* information ignored

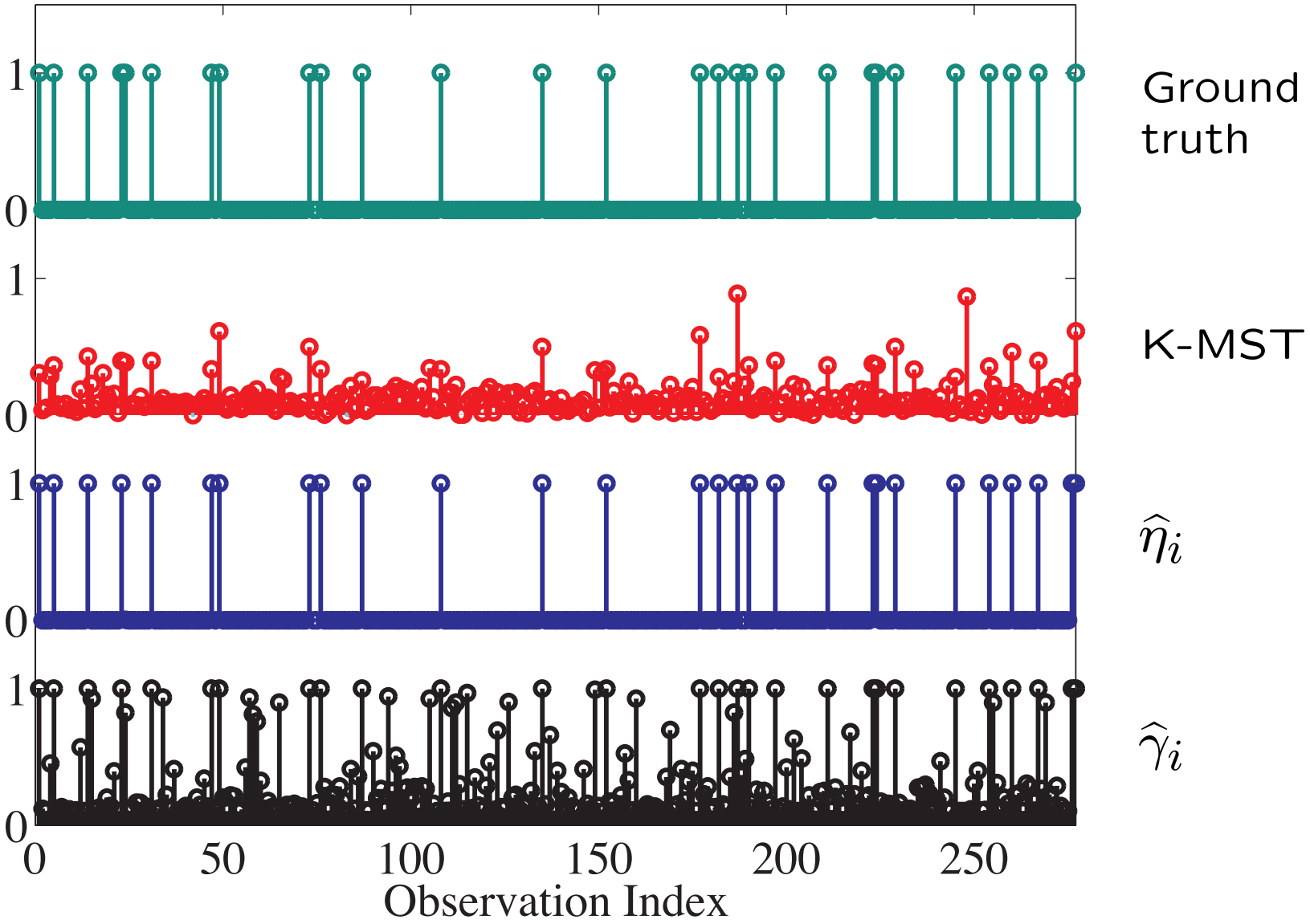
Injected 25 random anomalies
with average path length of 15

Anomalous distribution: $\mu(\boldsymbol{x}) = \prod_{j=1}^{p} \theta_j^{x_j}(1-\theta_j)^{1-x_j}$

$$\theta_j = 15/p$$

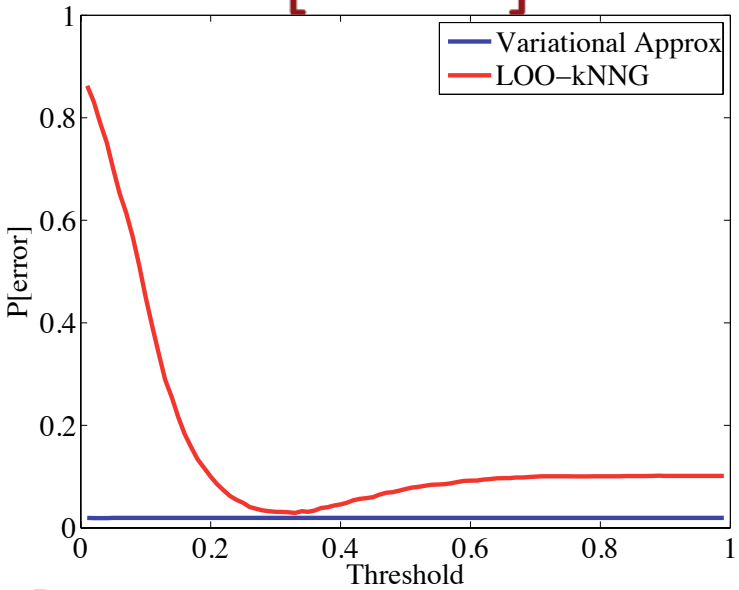
Rabbat, Figueiredo and Nowak, 2007

NICO results

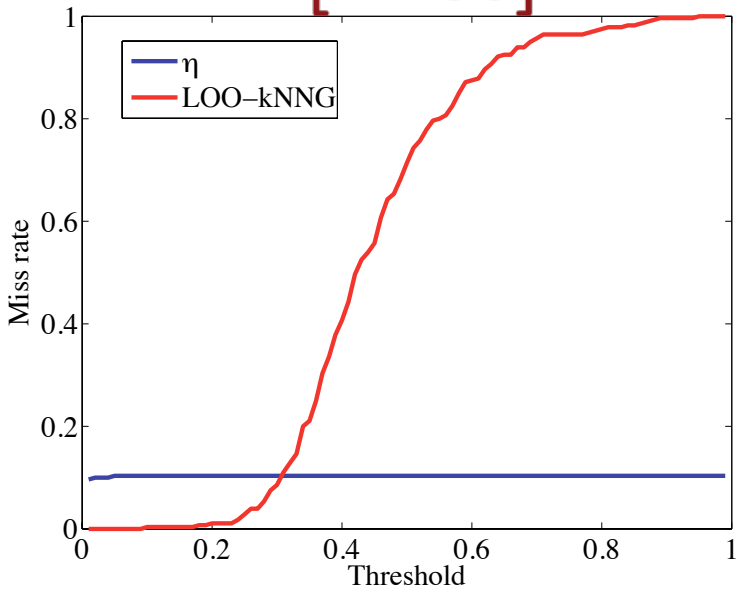


NICO results

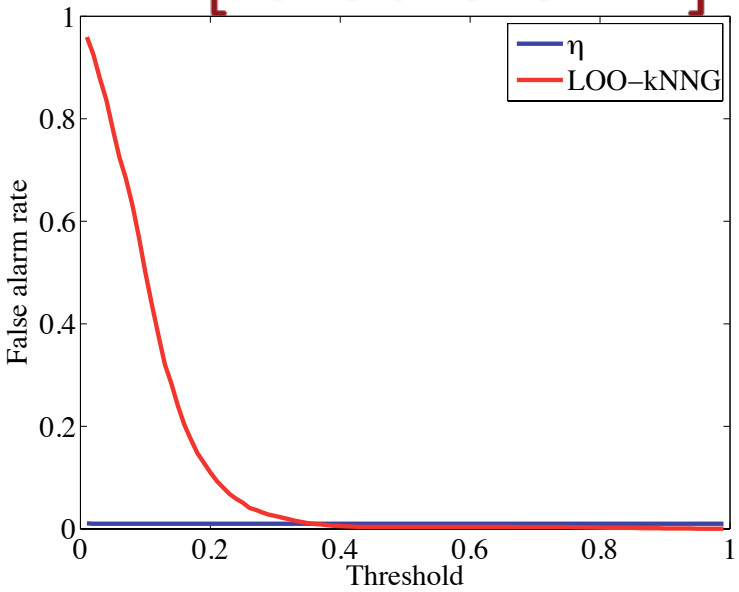
$P[\text{error}]$



$P[\text{miss}]$

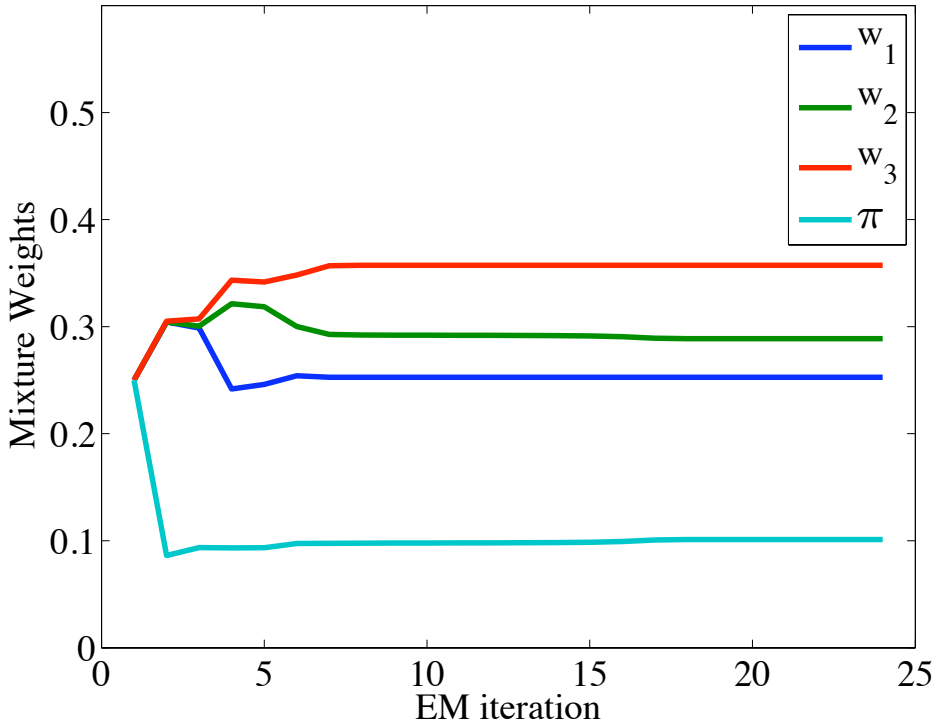


$P[\text{false alarm}]$

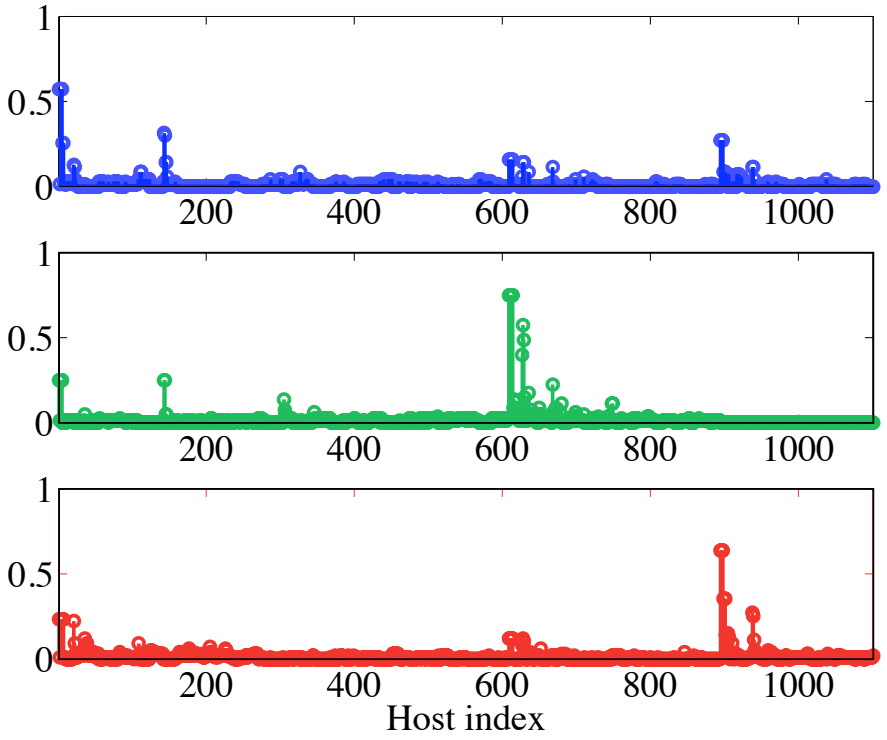


NICO results

Mixture weights

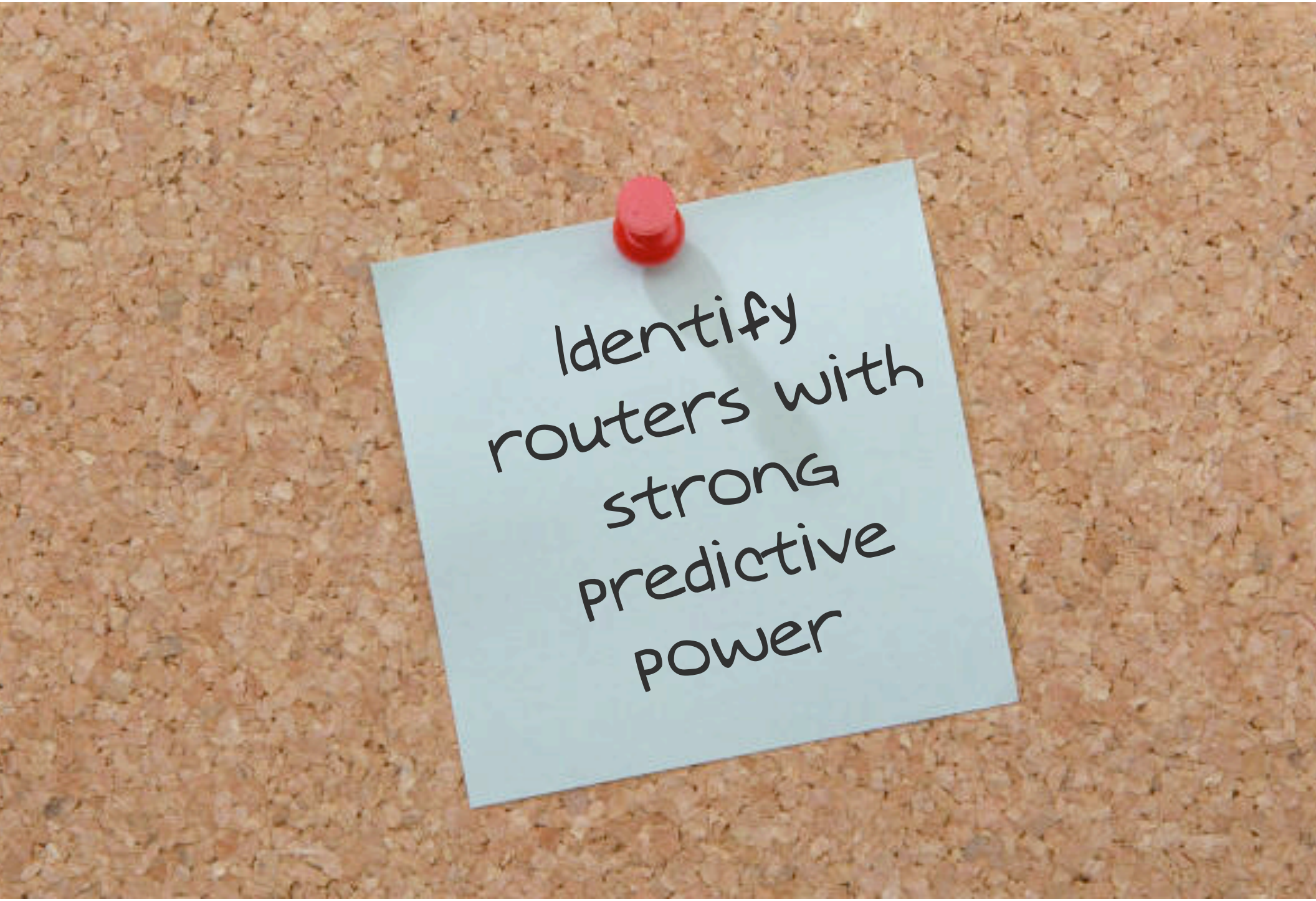


Variational approx. coeffs.



Running times

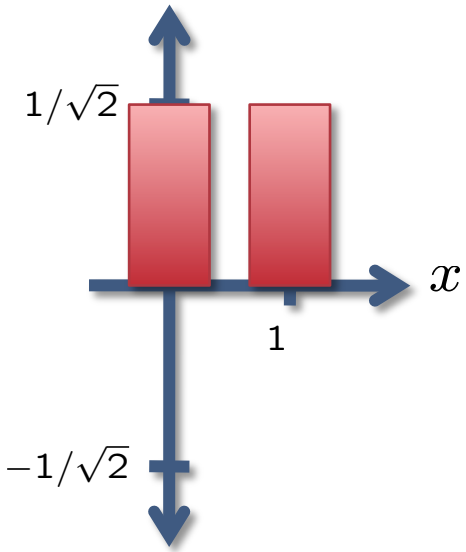
	Variational EM	OCSVM	K-MST
Simulation, n=100, p=2000	31.7	6.3	3.7
Simulation, n=300, p=2000	32.7	57.2	49.5
NICO data, n=277, p=1101	5	-	30
Enron data, n=800, p=75511	504.0	1182.5	$\gg$ 3600


Identify
routers with
strong
predictive
power

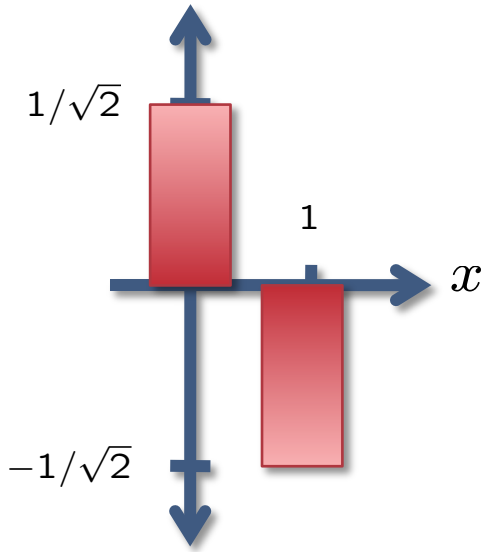
Our mission:
Estimate 2^p Walsh basis coefficients
with *fast* multiscale algorithm

Walsh functions in 1-d: $x \in \{0, 1\}$

$$\varphi_0(x) \triangleq \frac{1}{2^d}$$



$$\varphi_1(x) \triangleq \frac{1}{2^d}(-1)^x$$



Walsh, 1923
Talagrand, 1994
Dinur, Friedgut, Kindler and O'Donnell, 2007

Walsh basis in p dimensions

$s \in \{0,1\}^p$ ← Index of basis functions

$$\varphi_s(x) \stackrel{\triangle}{=} \prod_{j=1}^{p} \varphi_{s(j)}(x(j))$$

Similar to variational approx. from earlier. Each marginal = 1-d Walsh function.

Results in orthonormal basis of $\{0,1\}^p$

$$f = \sum_{s \in \{0,1\}^p} \theta_s \varphi_s$$

Liang and Krishnaiah, 1985
Ott and Kronmal, 1976

Orthogonal series estimation

1. Compute empirical basis coefficients:

$$\widetilde{\theta}_s = (1/n) \sum_{i=1}^{n} \varphi_s(X_i)$$

2. Threshold:

$$\widehat{\theta}_s = I_{\{T(\widetilde{\theta}_s) \geq \lambda\}}$$

3. Sum weighted basis functions:

$$\widehat{f} = \sum_{s \in \{0,1\}^p} \widehat{\theta}_s \varphi_s$$

Problem: 2^p basis coefficients!

Key insight

Walsh functions factorize

$$\varphi_s = \varphi_u \otimes \varphi_v$$

whenever concatenation $uv = s$

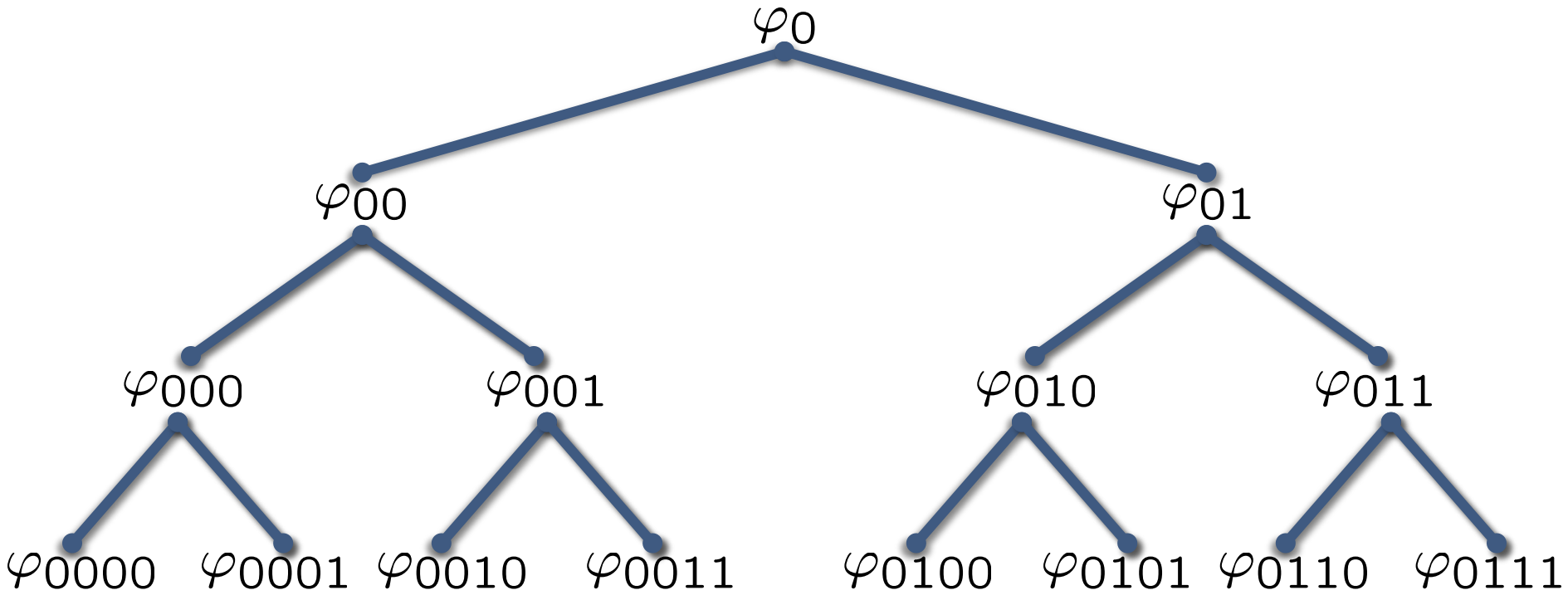
Key insight

Walsh functions factorize

$$\varphi_s(x) \triangleq \prod_{j=1}^{p} \varphi_{s(j)}(x(j))$$

$$= \prod_{j=1}^{k} \varphi_{u(j)}(x(j)) \prod_{j=k+1}^{p} \varphi_{v(j)}(x(j))$$

whenever concatenation $uv = s$



Determining if a prefix is "significant"

Main idea: Fix some $u \in \{0,1\}^k$. If coefficient θ_s is "insignificant" (i.e. going to be thresholded) whenever s has prefix u, then don't compute any θ_s explicitly.

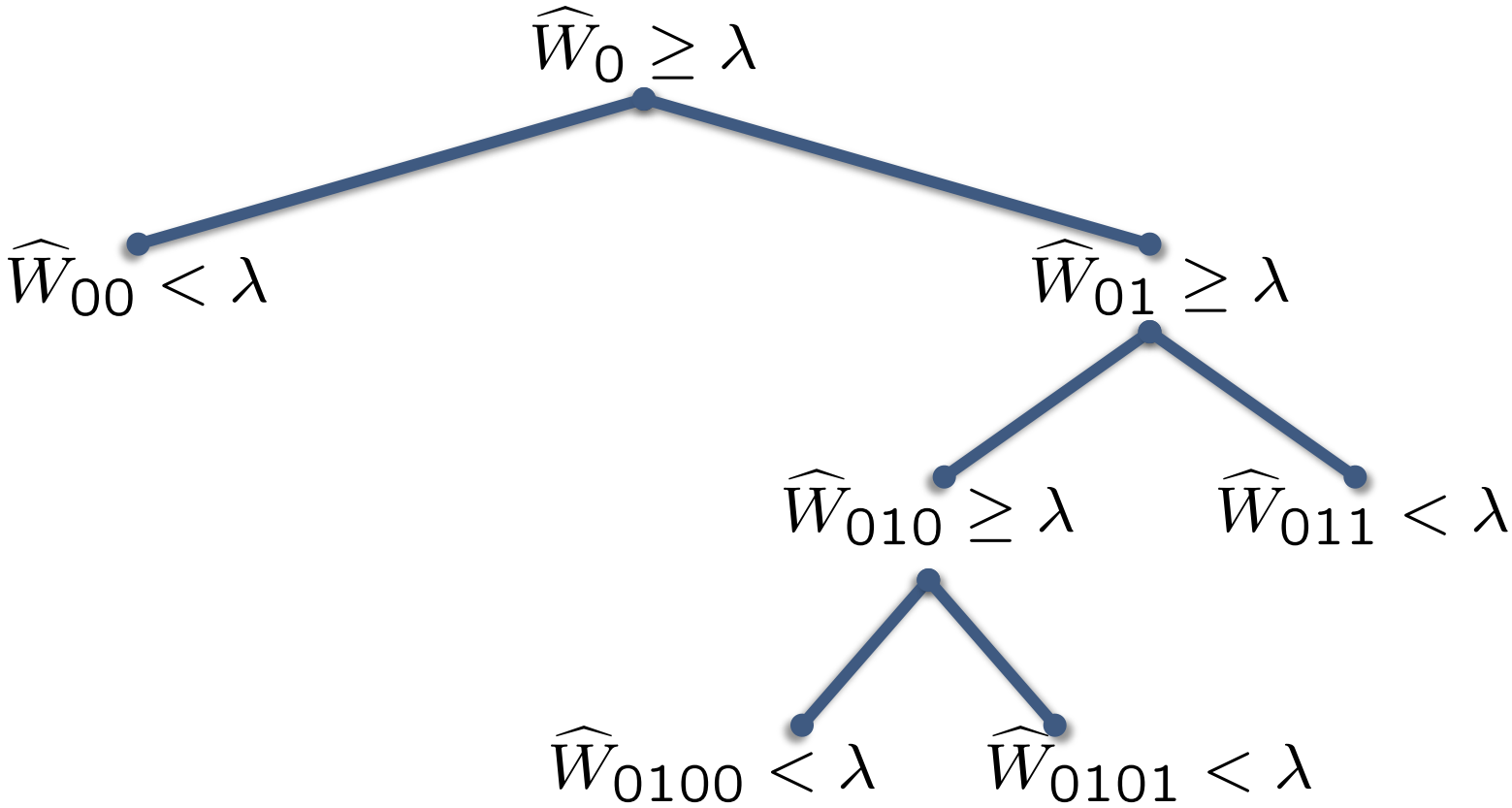
Consider energy of coefficients with index prefix u

$$W_u \stackrel{\triangle}{=} \sum_{v \in \{0,1\}^{p-k}} \theta_{uv}^2$$

If $W_u < \lambda$ then all $\theta_{uv}^2 < \lambda$, so u is not significant.

W_u can be estimated from data.

Hierarchical estimation



Raginsky, Lazebnik, Willett and Silva, 2008
Goldreich and Levin, 1989
Kushilevitz and Mansour, 1993

Performance

Theorem: If f is in a weak-ℓ_p ball in the Walsh basis, so that

$$|\theta_{(m)}| \leq R \cdot m^{-r-1/2},$$

and thresholds are sufficiently large, then

$$\sup_{f \in \mathcal{F}_d(p)} \mathbb{E}_f \|f - \hat{f}\|_2^2 \leq C 2^{-d} \left(\frac{\log n}{n} \right)^{2r/(2r+1)}$$

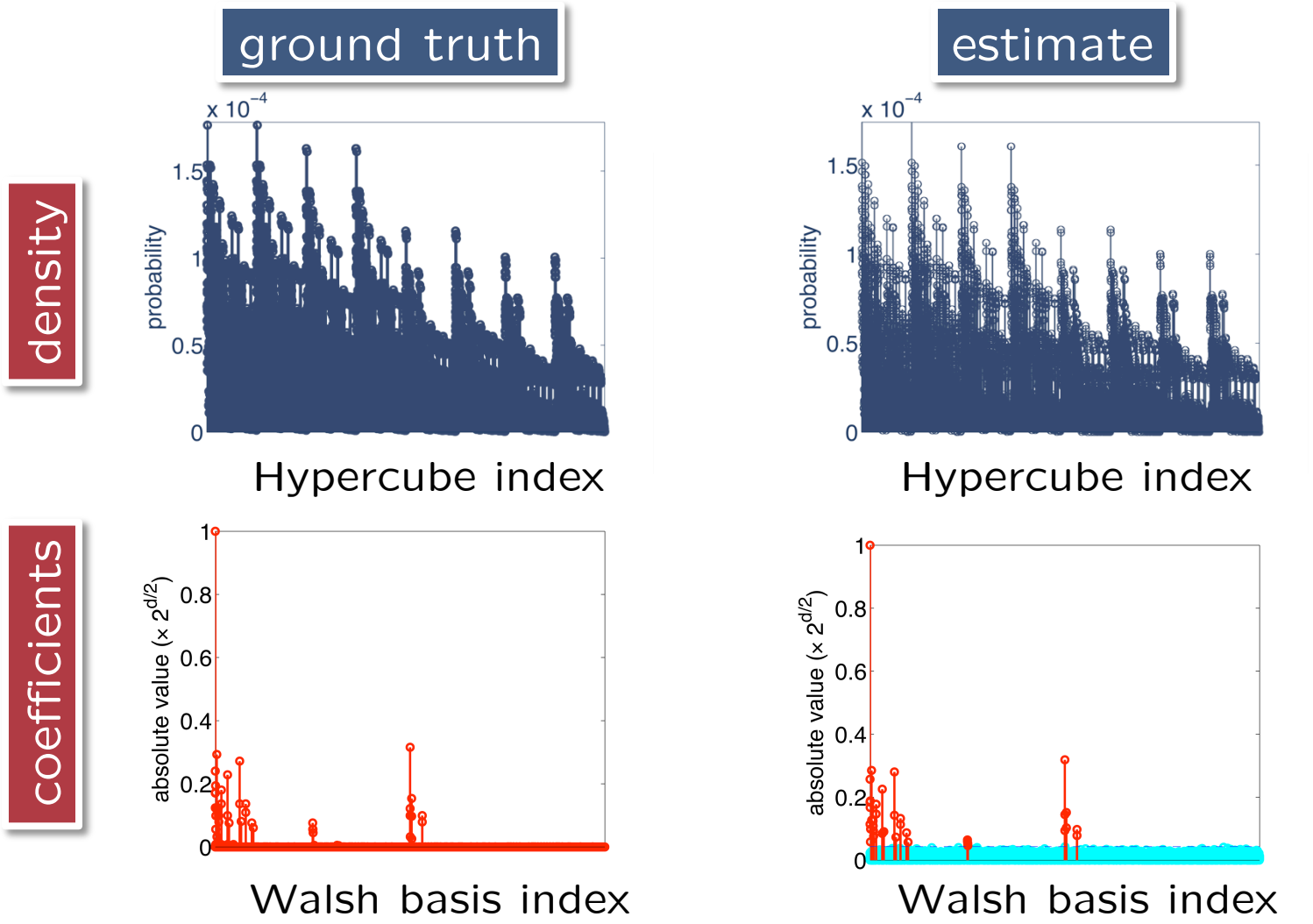
Minimax rate:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_d(p)} \mathbb{E}_f \|f - \hat{f}\|_2^2 \geq C 2^{-d/(2r+1)} n^{-2r/(2r+1)}$$

Raginsky, Lazebnik, Willett and Silva, 2008
Johnstone, 1994

Experimental validation

$p = 15$, $n = 10000$,
true density = product of marginals



Complexity

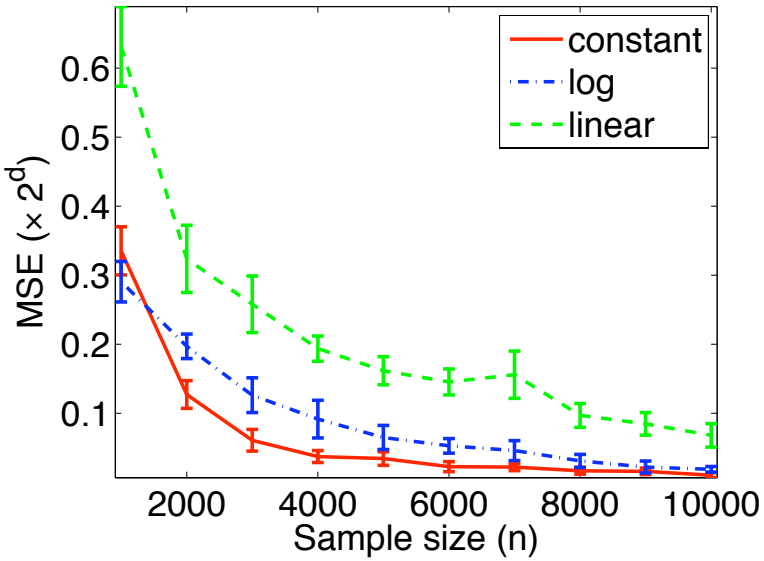
Theorem: For any $\delta \in (0, 1)$, if thresholds depend on δ and are sufficiently large, then with probability at least $1-\delta$ the computational complexity of the algorithm is

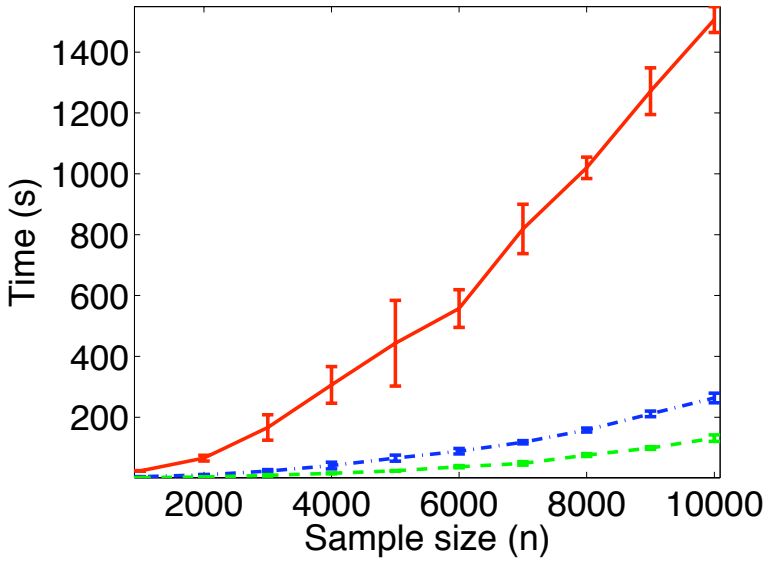
$$O(n^2 p(n/2^p \log n)^{1/(2r+1)} K)),$$

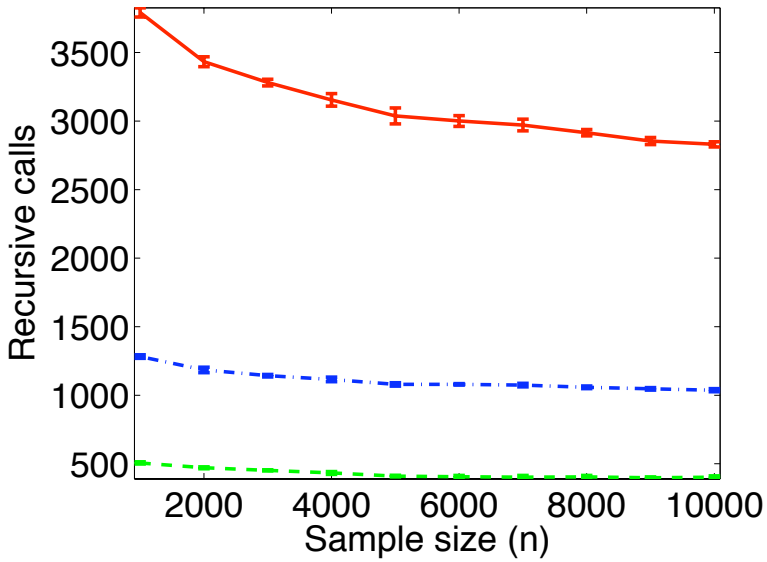
where $K = K(d)$ is a bound on the thresholds.

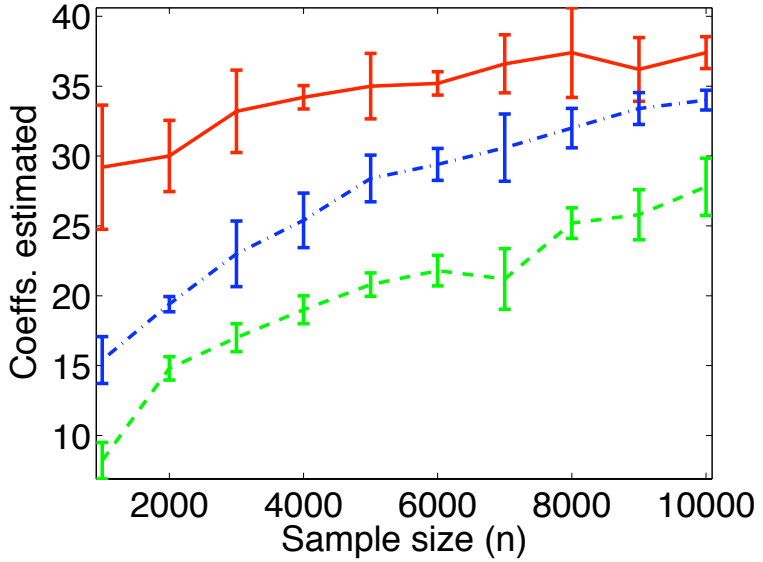
Raginsky, Lazebnik, Willett and Silva, 2008

MSE vs. Complexity





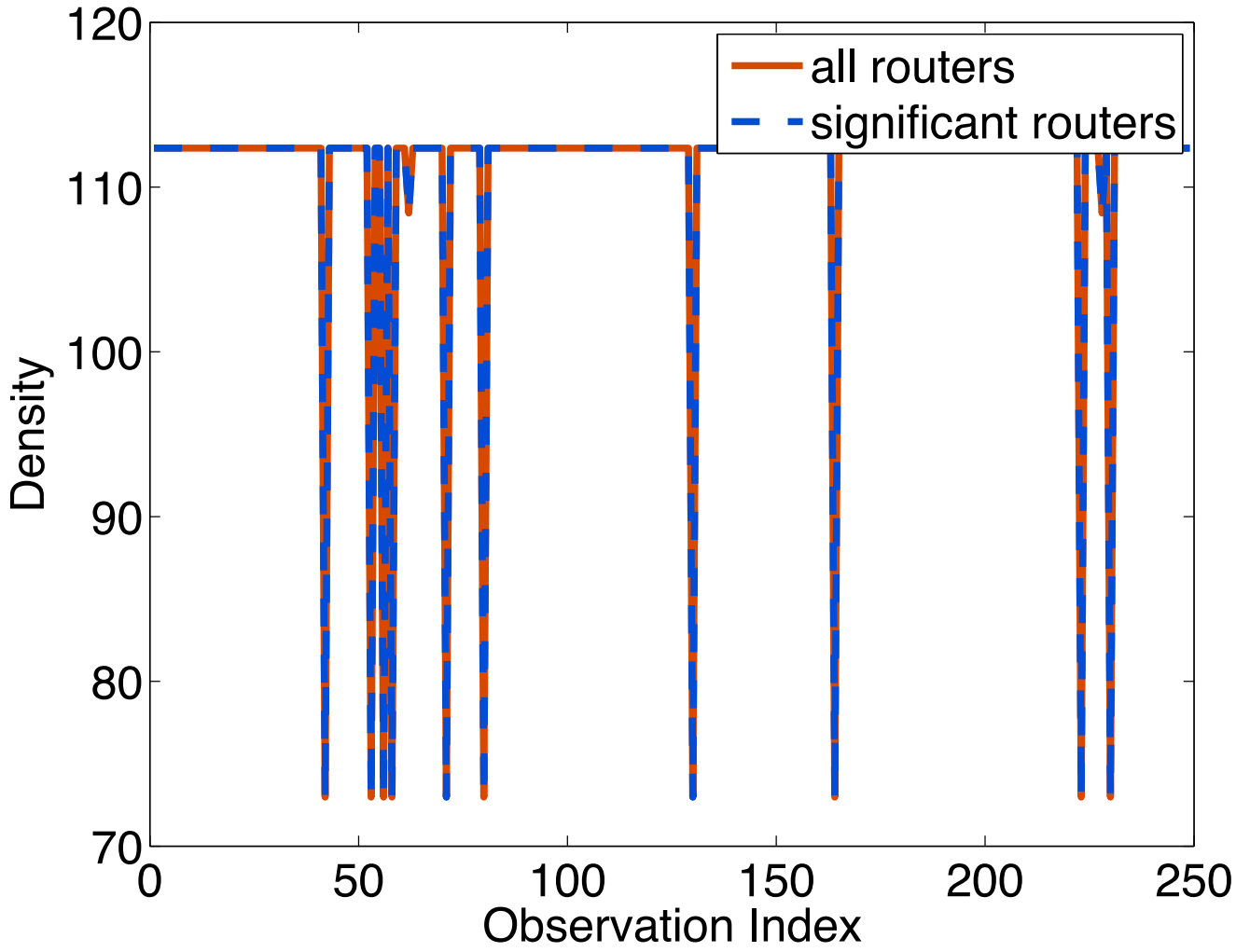




NICO Results

$p = 20$ routers,

$n = 250$ routes



VARIATIONAL APPROXIMATION FOR HYPERGRAPH ANOMALY DETECTION

- $O(np)$ computational complexity
- Implementable and effective even when p=10,000+ without any specialized data structures or hardware
- No parameter tuning or bandwidth selection
- False discovery rate can be quickly and accurately calculated

WALSH BASIS FOR DENSITY ESTIMATION

- Multiscale algorithm fast and accurate
- Allows predictive constellations of routers to be indentified

http://www.ee.duke.edu/nislab/

Questions?