ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

UNIVERSITY of MICHIGAN ■ COLLEGE of ENGINEERING

Subspace Clustering with Missing Data

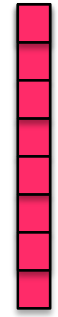
Laura Balzano

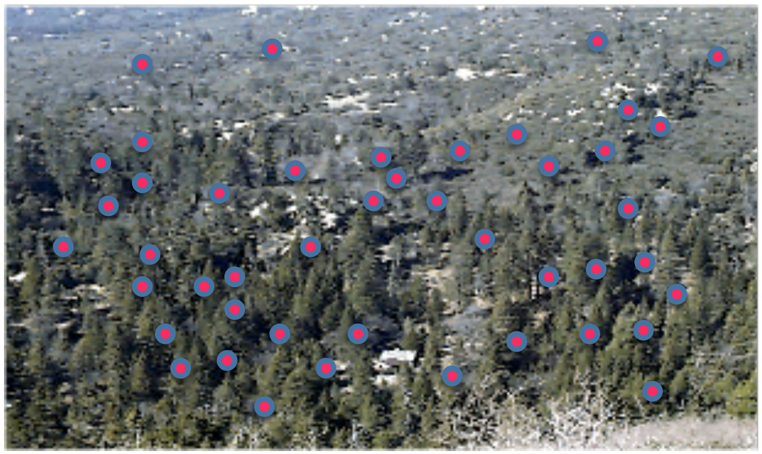
girasole@umich.edu

work with **Robert Nowak (UW), Brian Eriksson (Technicolor), Daniel Pimentel Alarcon (UW),** and **Arthur Szlam (Facebook NY).**

Subspace Representations

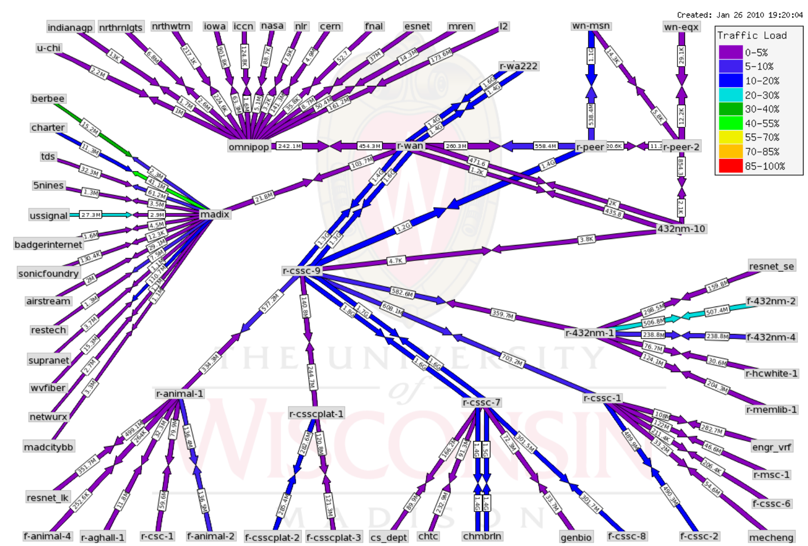
Monitor/sense with n nodes





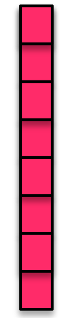
$v \in \mathbb{R}^n$ is a snapshot of the system state (e.g., temperature at each node)

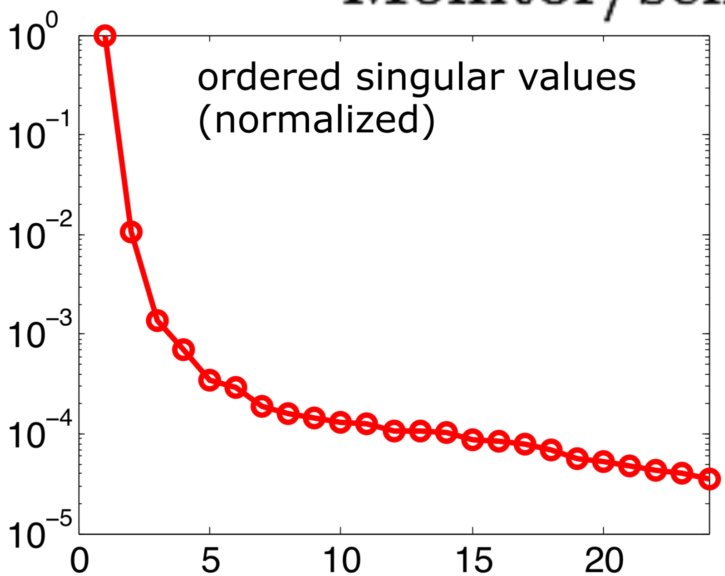
$v \in \mathbb{R}^n$ is a snapshot of the system state (e.g., traffic rates at each monitor)



Subspace Representations

Monitor/sense with n nodes

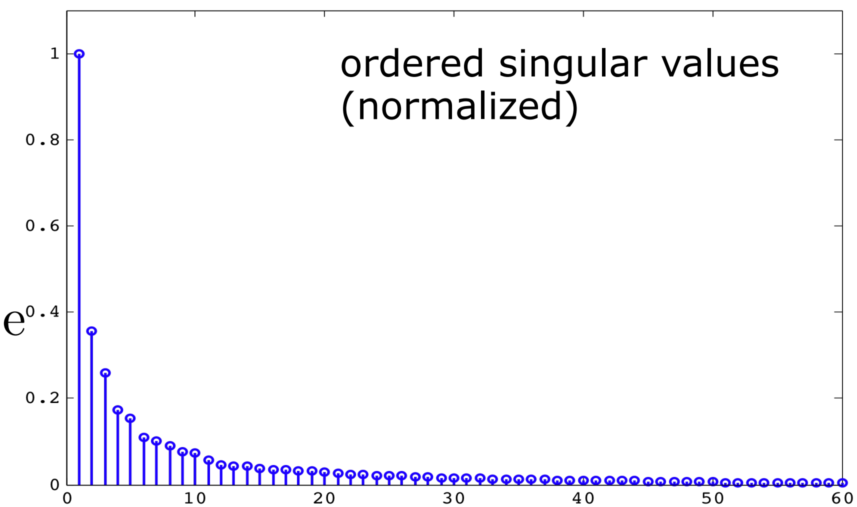




Temperature data from UCLA Sensornet

$v \in \mathbb{R}^n$ is a snapshot of the system state (e.g., temperature at each node)

$v \in \mathbb{R}^n$ is a snapshot of the system state (e.g., traffic rates at each monitor)



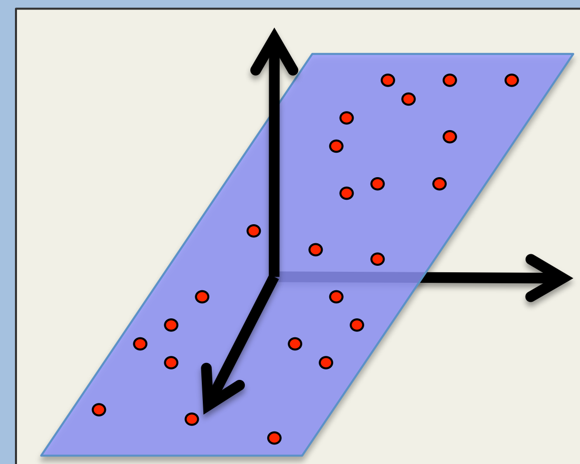
Byte Count data from UW network

Subspace Representations

Monitor/sense with n nodes

Each snapshot lies near a low-dimensional subspace

$$S \subset \mathbb{R}^n$$

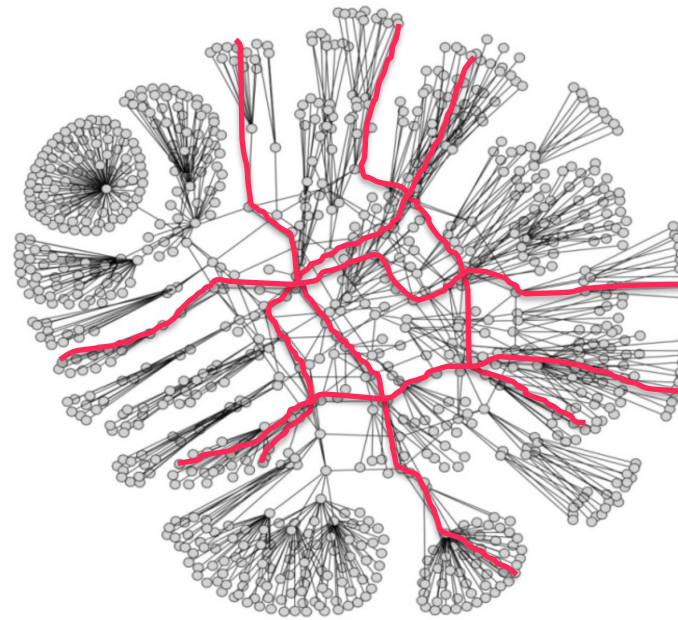


Using the **subspace as a model** for the data, we can leverage these dependencies for detection, estimation and prediction.

Estimating Subspaces with Missing Data

Rigid Structure from Motion object identification

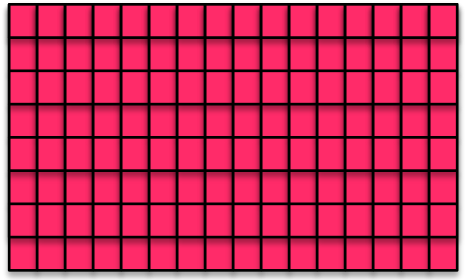
All images and markings from the "Hopkins 155" Dataset, R. Vidal lab, Johns Hopkins University.



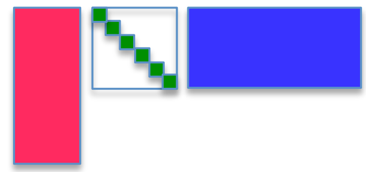
Network Topology Identification

Recommendation Systems

Estimating Subspaces with Missing Data

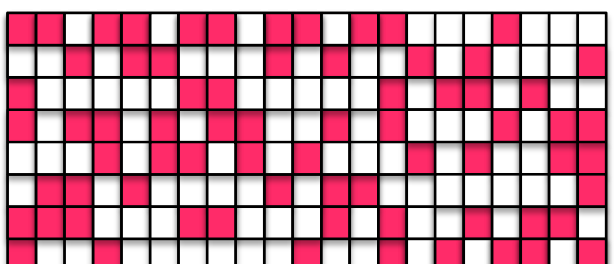


Consider an $n_1 \times n_2$ (where $n_1 < n_2$) matrix X of at most rank r. To identify the column space we may use the SVD.



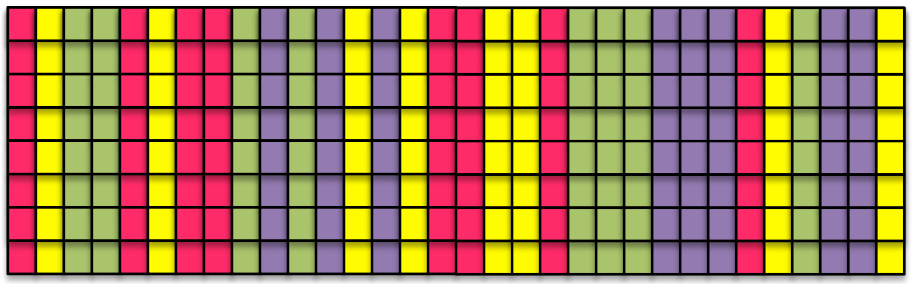
Now consider observing only a subset $\Omega \subset \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ of X, that has size $|\Omega| \geq O(rn_2 \log^2(n_2))$, and solving

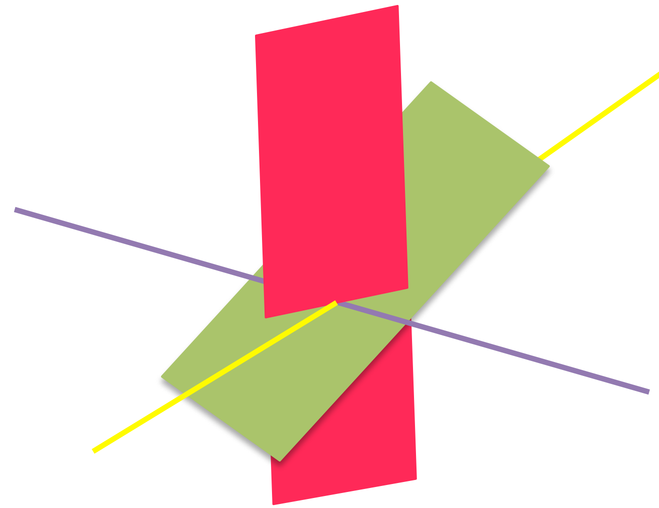
$$\text{minimize}_M \|(X - M)_\Omega\| + \lambda \|M\|_*$$



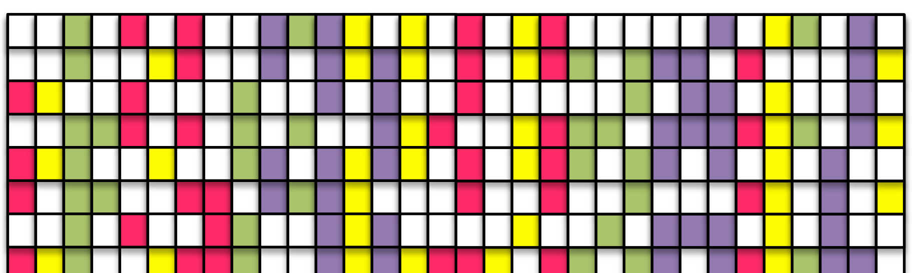
Theoretical results in "Low-Rank Matrix Completion" (LRMC) show that solving this optimization recovers X exactly.

Estimating Multiple Subspaces with Missing Data





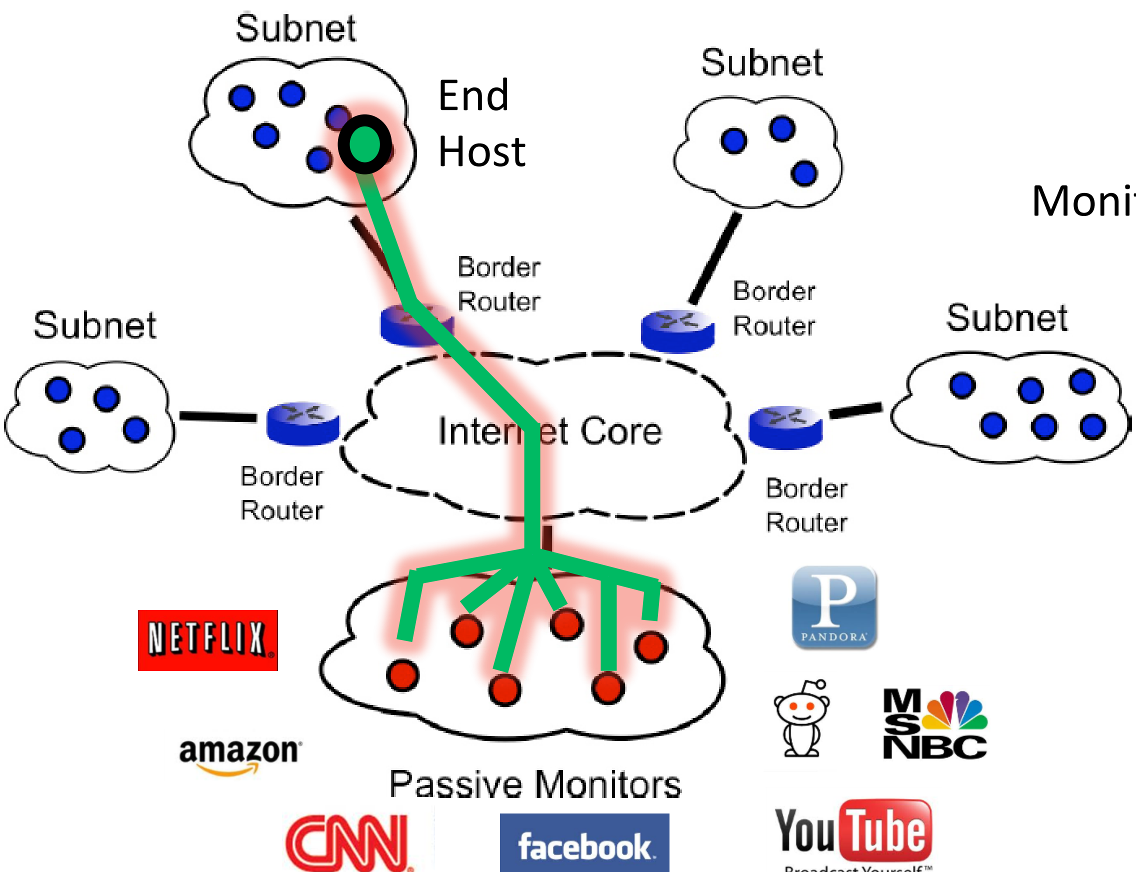
Now suppose our data come from at most k subspaces, also of at most rank r. Low rank matrix completion (LRMC) requires $O(krn_2 \log^2(n_2))$ measurements. If k is large this could be nearly full sampling. We wish to do better.

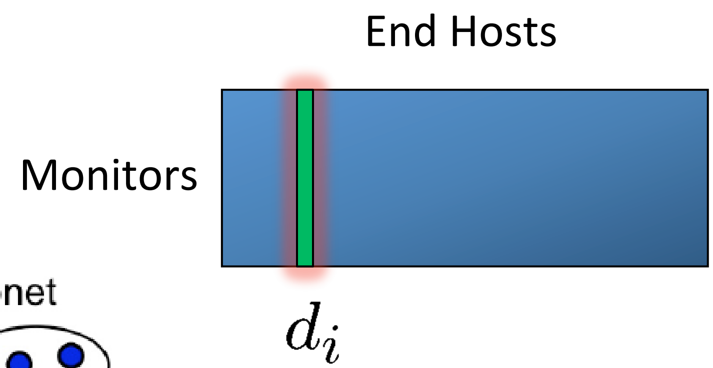


Outline

- Application of Network Topology Identification from incomplete hopcounts
- High Rank Matrix Completion (HRMC) algorithm and theory
- K-GROUSE and an EM algorithm
- Results

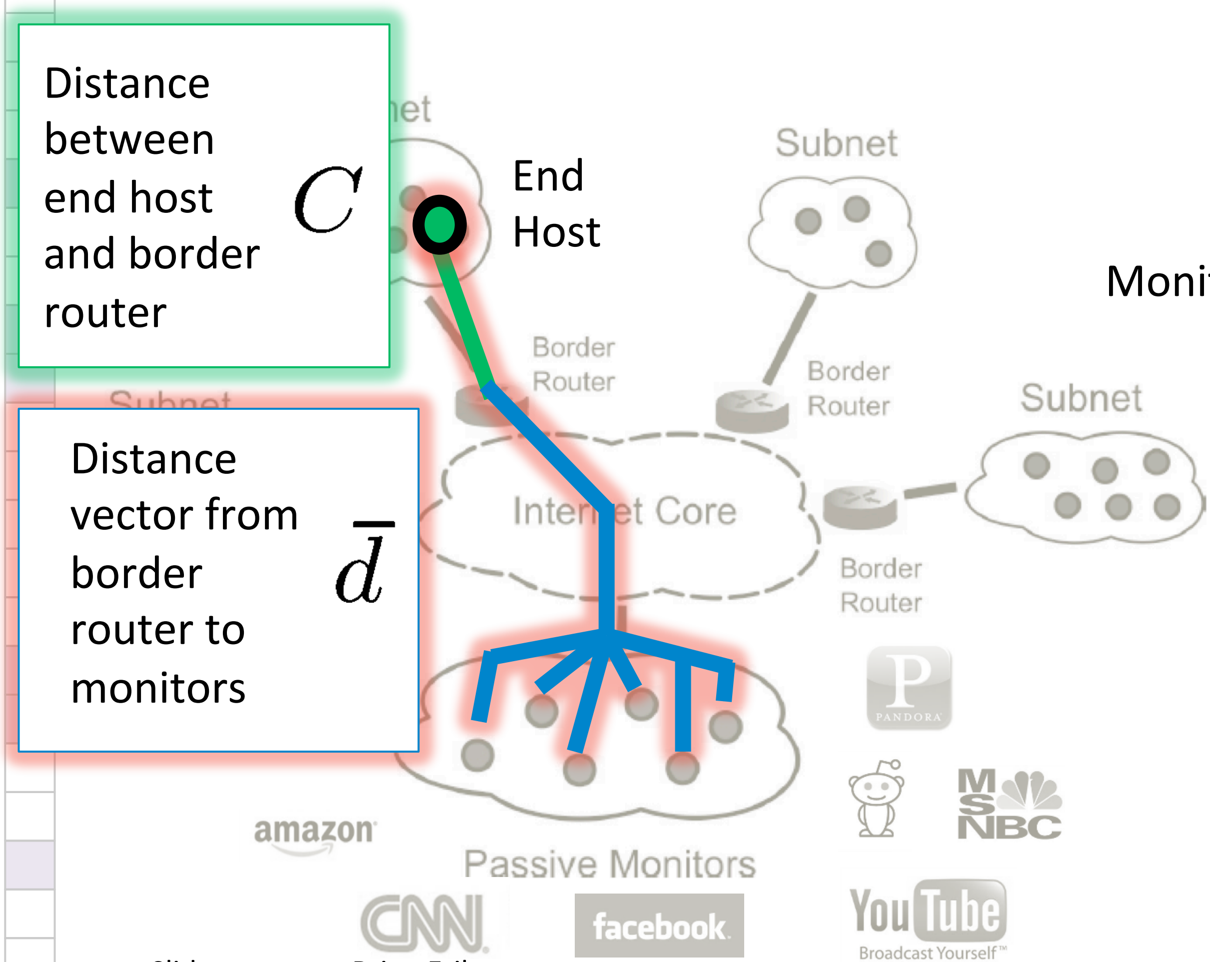
Network Topology Identification

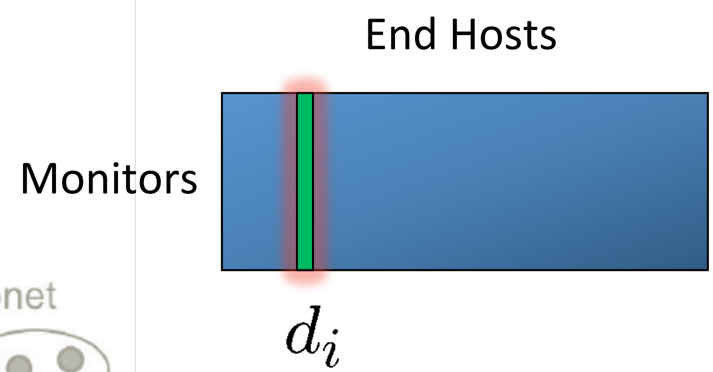




Slides courtesy Brian Eriksson

Network Topology Identification

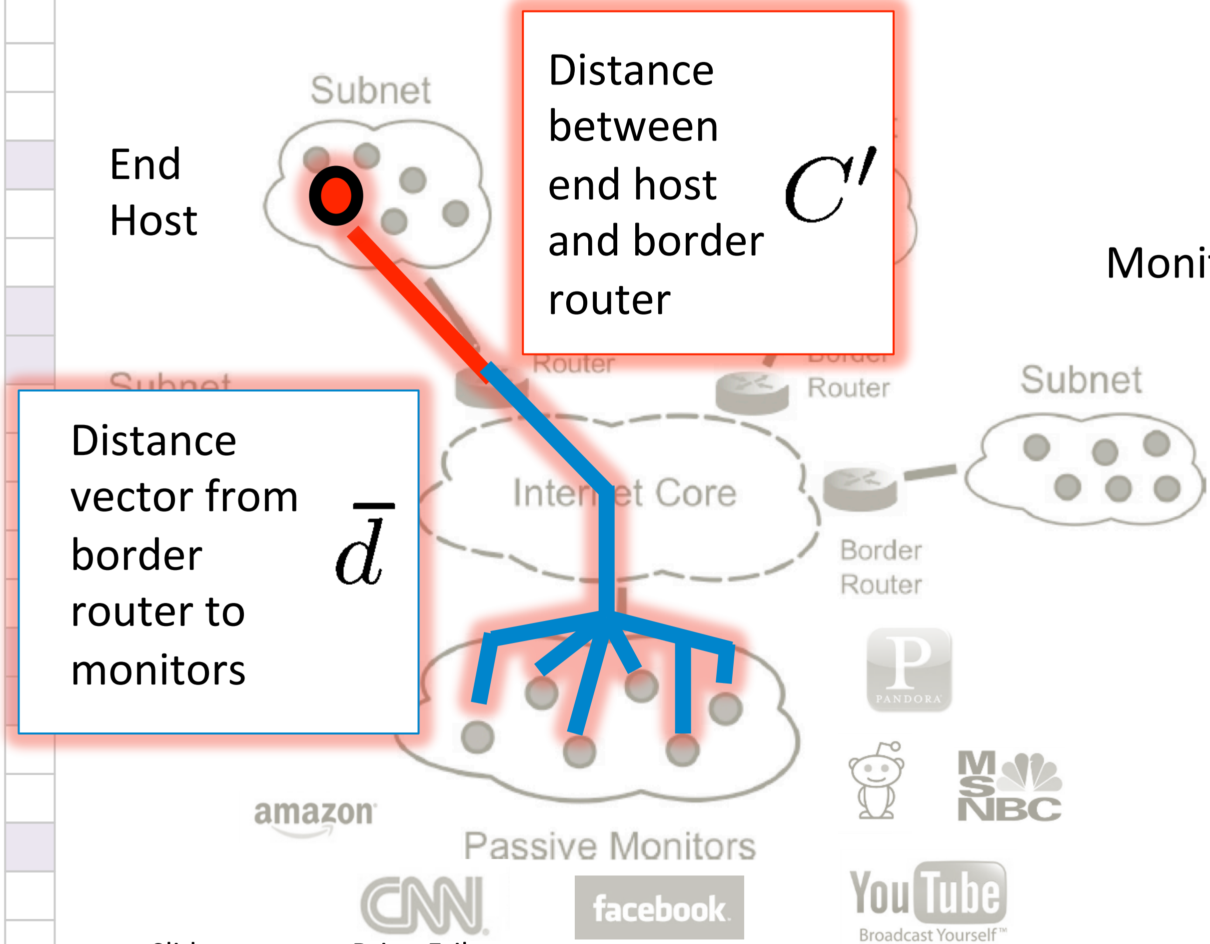


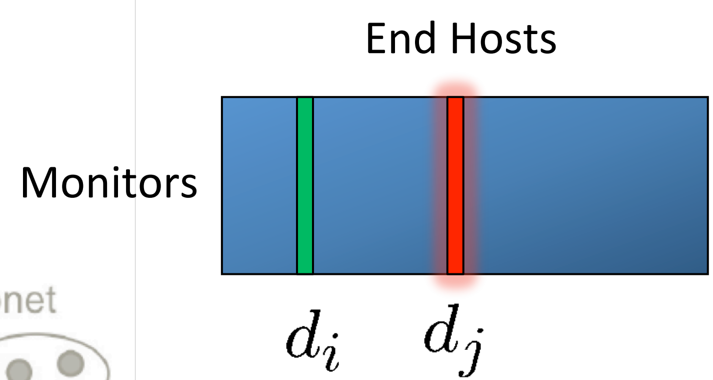


$$d_i = \bar{d} + C1$$

Slides courtesy Brian Eriksson

Network Topology Identification



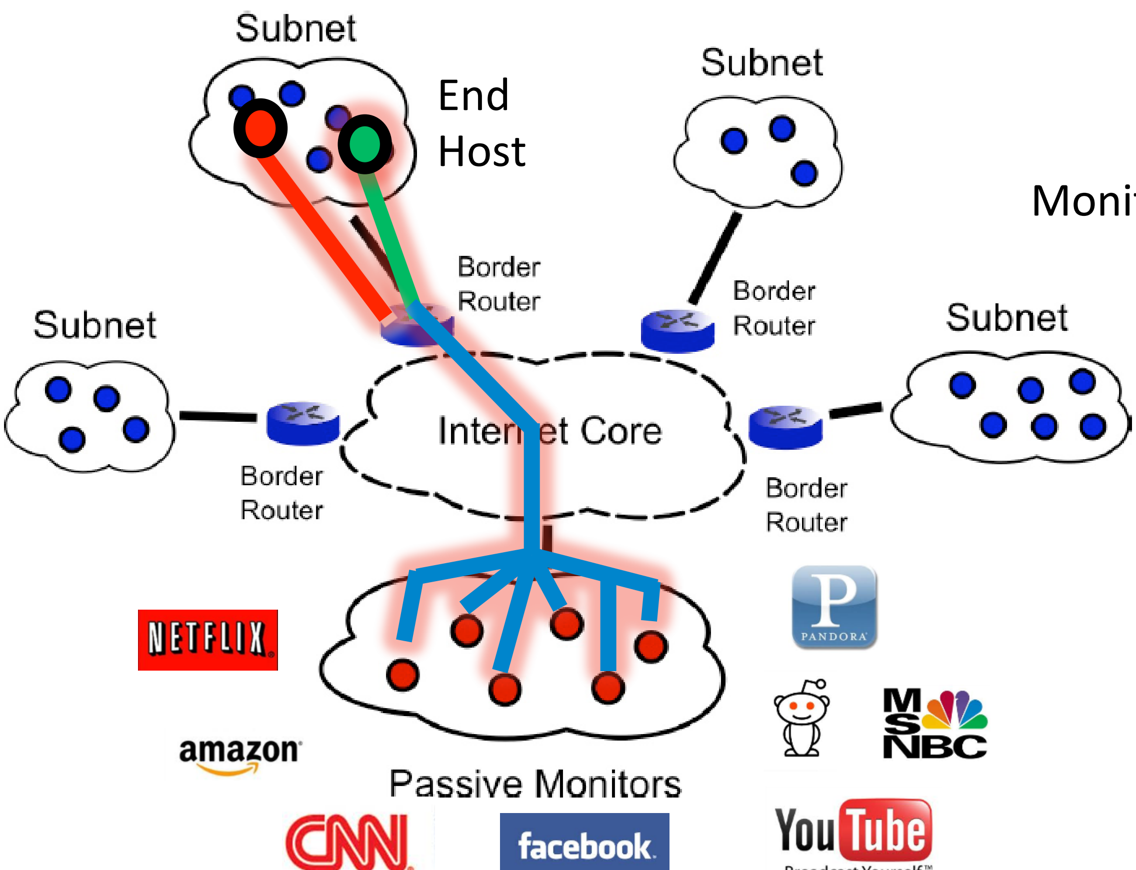


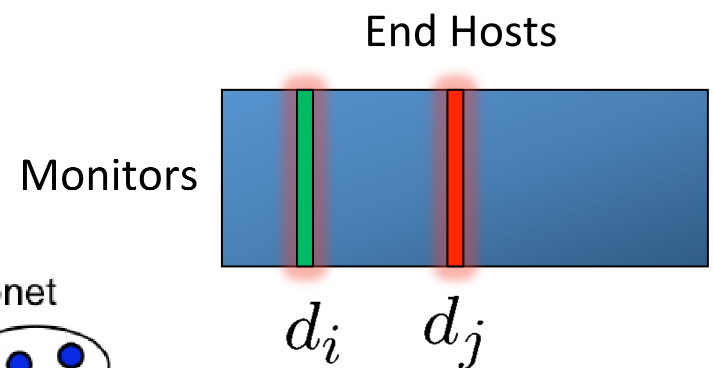
$$d_i = \bar{d} + C1$$

$$d_j = \bar{d} + C'1$$

Slides courtesy Brian Eriksson

Network Topology Identification





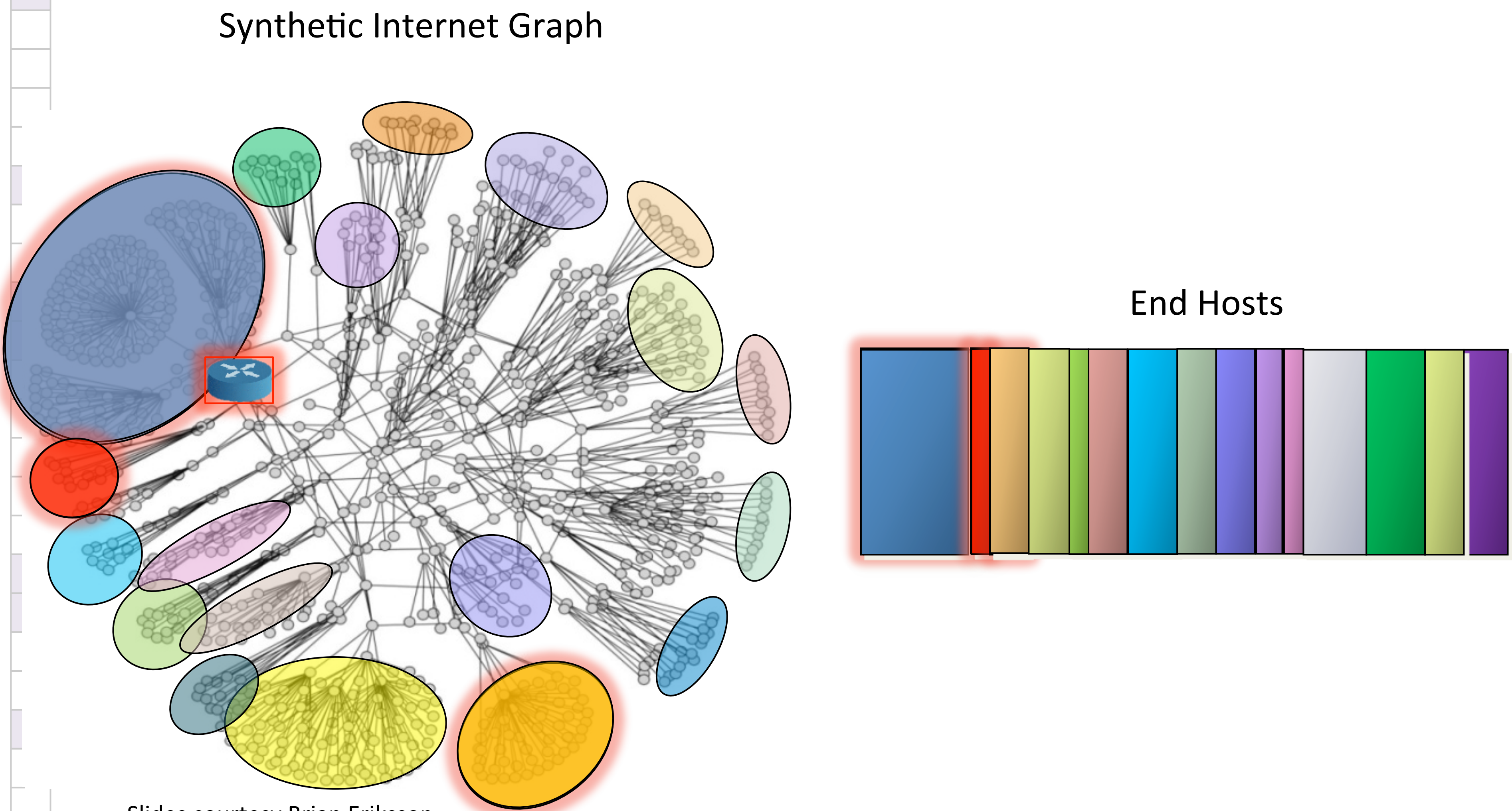
$$d_i = \bar{d} + C1$$

$$d_j = \bar{d} + C'1$$

All end hosts in the same subnet lie on the same 2-d subspace.

Slides courtesy Brian Eriksson

Network Topology Identification



Slides courtesy Brian Eriksson

Outline

- Application of Network Topology Identification from incomplete hopcounts
- High Rank Matrix Completion (HRMC) algorithm and theory
- K-GROUSE and an EM algorithm
- Results

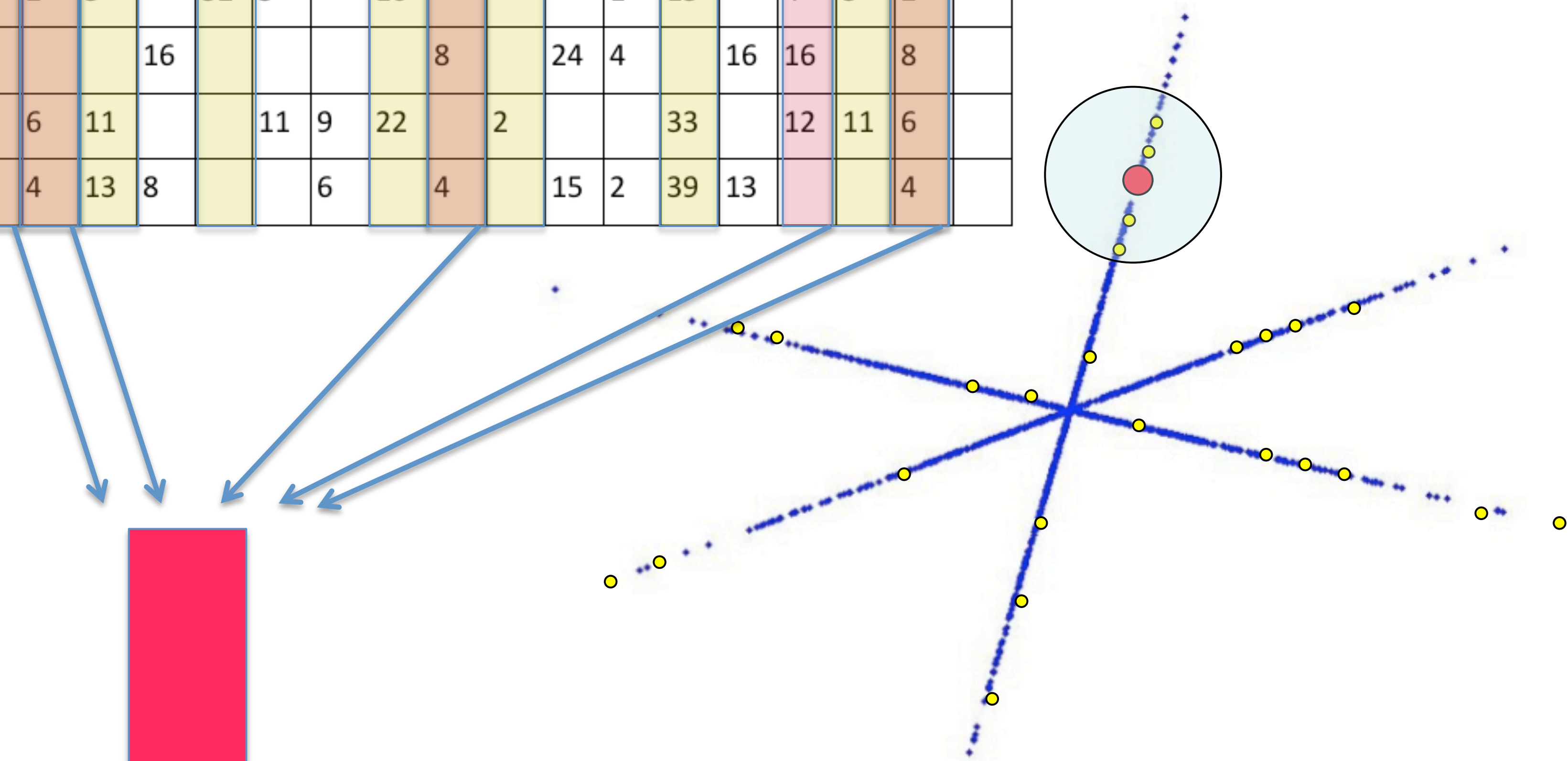
"High Rank Matrix Completion" Algorithm

with Brian Eriksson and Rob Nowak

IP addresses

monitors

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

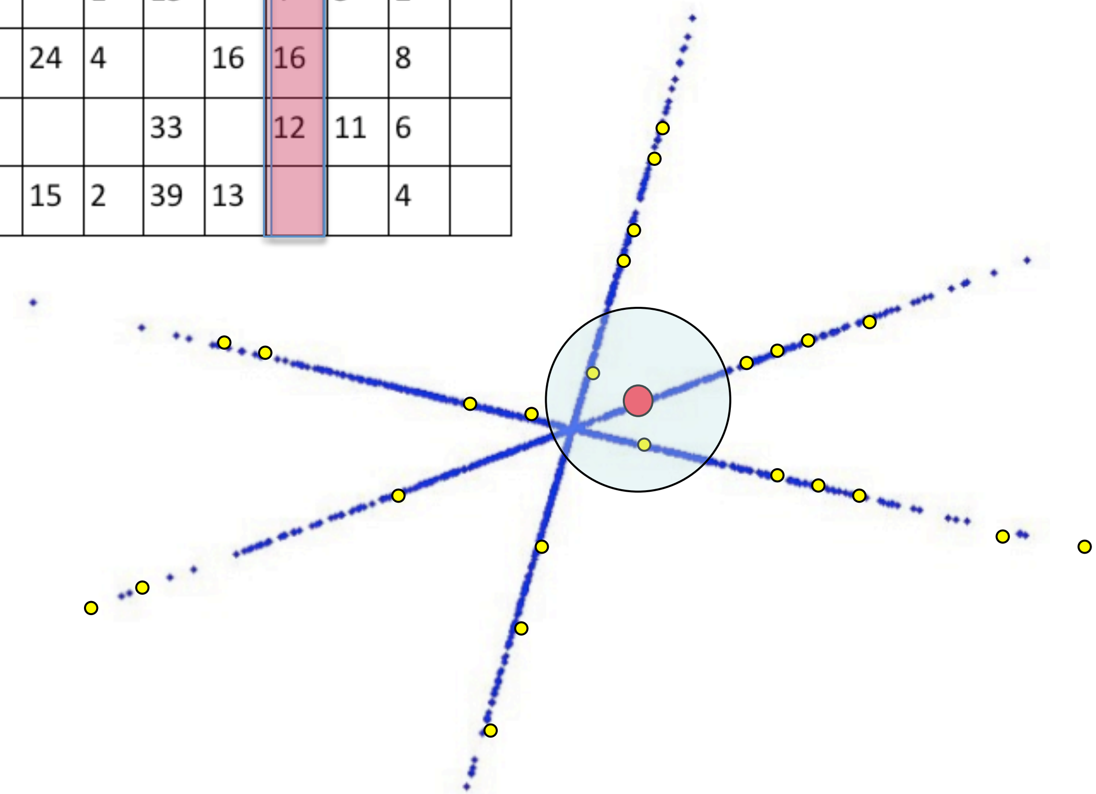


University of Michigan

"High Rank Matrix Completion" Algorithm

with Brian Eriksson and Rob Nowak

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	



"High Rank Matrix Completion" Algorithm

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

with Brian Eriksson and Rob Nowak

- Use enough seeds to guarantee every subspace has one seed with its neighborhood entirely in the subspace
- Find other columns that are in the seed's neighborhood (despite sampling)
- Guarantee matrix completion succeeds
- Show subspaces can be pruned to the correct set
- Guarantee remaining data points (not seeds or neighbors of seeds) can be assigned to the correct cluster

UNIVERSITY OF MICHIGAN

"High Rank Matrix Completion" Theory

with Brian Eriksson and Rob Nowak

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

Theorem: Let X be an $n_1 \times n_2$ matrix whose columns lie in the union of $k \ll n_2$ subspaces, of rank at most r, which are incoherent and not "too close" to one another. Let $n_2 = O(n_1^{\log n_1})$. Then with high probability, the matrix X can be perfectly reconstructed from $O(rn_2 \log^2(n_2))$ observations.

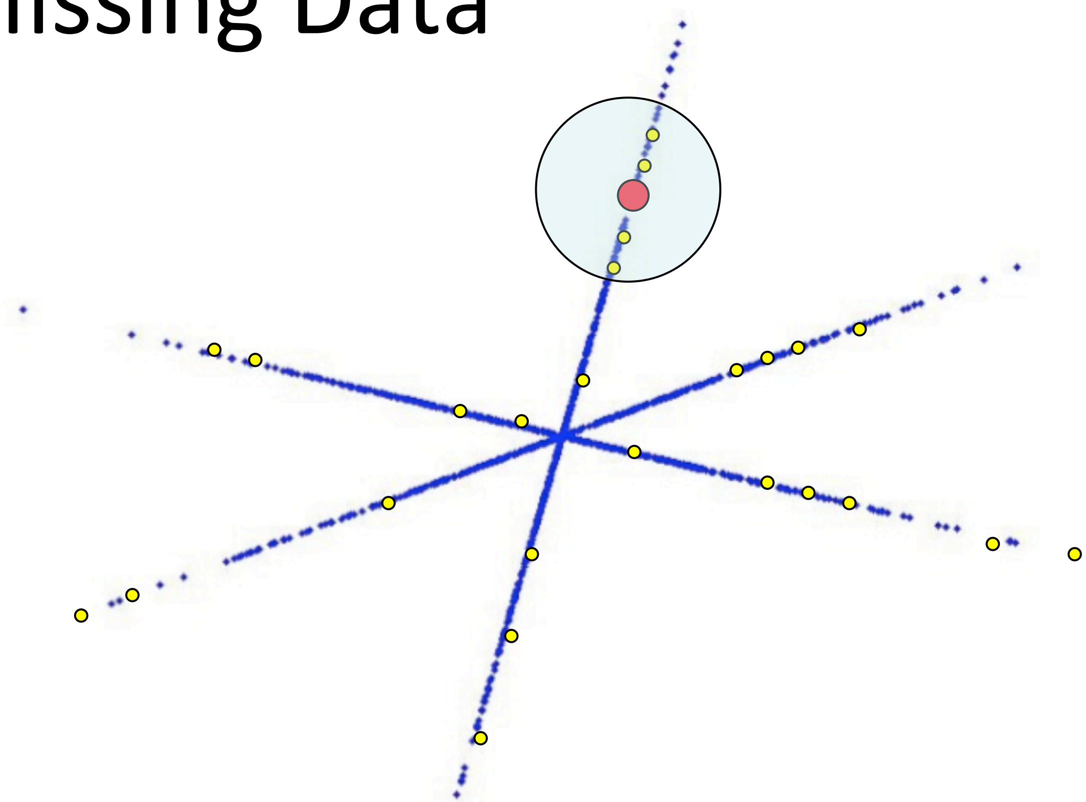
Why so many data points?

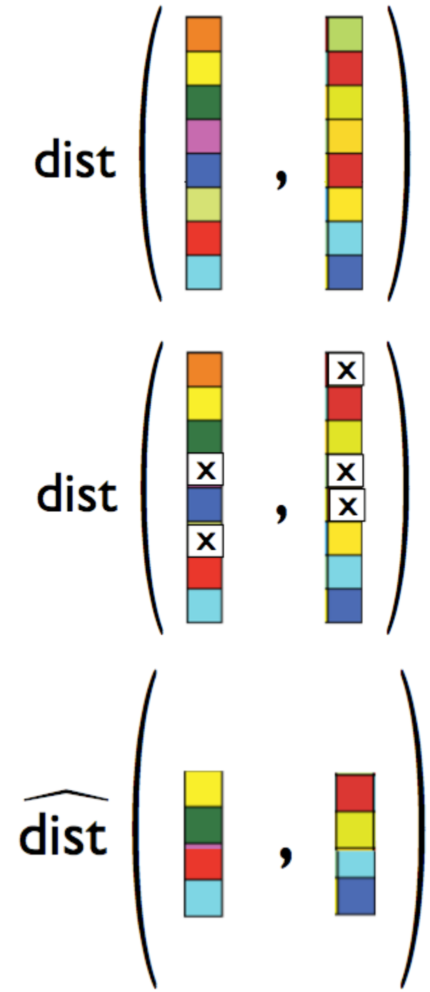
	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

with Brian Eriksson and Rob Nowak

- Use enough seeds to guarantee every subspace has one seed with its neighborhood entirely in the subspace
- **Find other columns that are in the seed's neighborhood (despite sampling)**
- Guarantee matrix completion succeeds
- Show subspaces can be pruned to the correct set
- Guarantee remaining data points (not seeds or neighbors of seeds) can be assigned to the correct cluster

Finding neighborhoods with Missing Data





n_1 ambient dimension, q # overlapping entries, μ_0 incoherence parameter.

$$\frac{1}{2}\text{dist}^2 \le \frac{n_1}{q}\widehat{\text{dist}}^2 \le \frac{3}{2}\text{dist}^2 \quad \text{w.p} \ge 1 - 2\exp\left(\frac{-q}{2\mu_0^2}\right)$$

Why so many data points?

Theorem 2. *Let q represent the random variable of number of entries observed in common for two arbitrary vectors in $\mathbb{R}^{n_1}$. For some $t \geq 1$, if the number of observations per vector are such that*

$$m \geq n_1^{1/2} \max\left\{2t, 8\log\left(\frac{1}{\delta}\right)\right\}^{1/2}$$

then

$$\mathbb{P}(q \geq t) \geq 1 - \delta\ .$$

On the other hand, if the observation probability is such that $m = g(n_1) = O(\sqrt{n_1})$, then for n_1 such that $n_1 \geq g(n_1)^2$, we have that

$$\mathbb{P}(q \geq t) \leq \exp(-t/2 + 1)\ .$$

Outline

- Application of Network Topology Identification from incomplete hopcounts
- High Rank Matrix Completion (HRMC) algorithm and theory
- K-GROUSE and an EM algorithm
- Results

Faster algorithms

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

If the subspaces were known, we could estimate the column assignments.

If the column assignments were known, we could estimate the subspace using low-rank matrix completion.

A faster algorithm (with Arthur Szlam)

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

k-GROUSE

- initialization of k subspaces, either randomly or using zero-filled distances.
- Assign partially observed vectors to subspaces, and consider this assignment the new clustering.
- Use the new clustering to estimate subspaces using low-rank matrix completion.

University of Michigan

A faster algorithm (with Arthur Szlam)

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

"High Rank MC"

- calculates masked distances for each of O(k log k) seed points
- runs matrix completion on an nxn matrix O(k log k) times for at least $O((k \log k)(n_1^2 r))$ time.
- to prune subspaces, must consider every (k log k choose k) subset to find the best set, gives $O(k^k)$ operations.

k-GROUSE

- rough initialization of k subspaces using zero-filled distances
- iteratively chooses a random vector and updates the closest subspace in $O(kmr^2+n_1r)$ time per update.
- Empirically we need $O(rn_2)$ updates, so total time is $O(n_2kmr^3 + n_1n_2r^2)$

+EM algorithm (with Daniel Pimentel)

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

EM Algorithm

- Initialization (usually random) of k subspaces
- Computes the probability of each data point belonging to each of the subspaces
- Computes the maximum likelihood estimate of the means and covariances
- $O(n_1 n_2 kr)$ per iteration.

k-GROUSE

- rough initialization of k subspaces using zero-filled distances
- iteratively chooses a random vector and updates the closest subspace in $O(kmr^2+n_1 r)$ time per update.
- Empirically we need $O(rn_2)$ updates, so total time is $O(n_2 kmr\text{^}3 + n_1 n_2 r\text{^}2)$

UNIVERSITY OF MICHIGAN

Outline

- Application of Network Topology Identification from incomplete hopcounts
- High Rank Matrix Completion (HRMC) algorithm and theory
- K-GROUSE and an EM algorithm
- Results

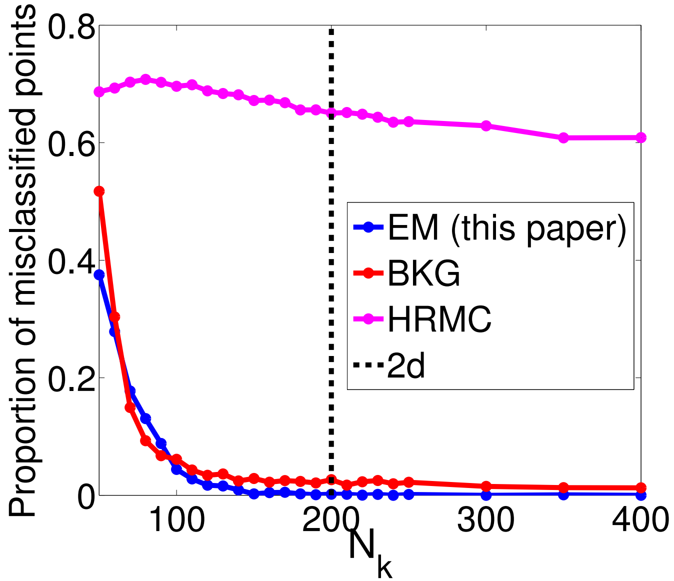
Speed results on simulated data

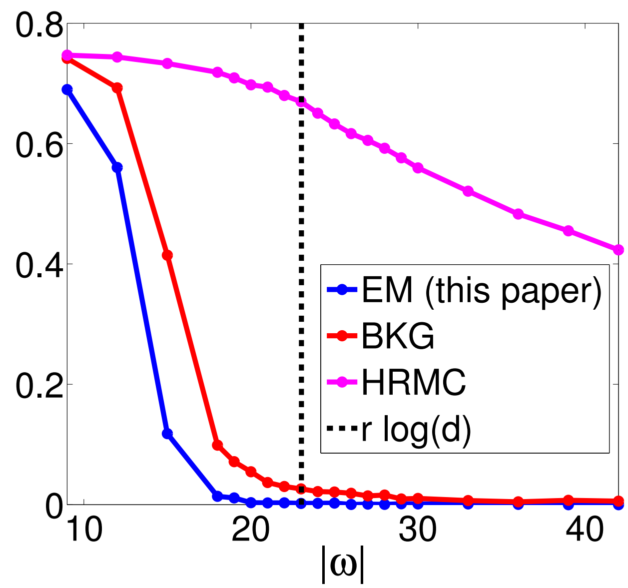
Algorithm	Run time average	Std Dev	% successful
High Rank MC 3k logk seeds	10395.0 (2.8 hours)	655.8	56 (of 100 trials)
High Rank MC 10k logk seeds	34162.3 (9.5 hours)	2086.5	100 (of 11 trials)
k-GROUSE	127.6 (2 minutes)	0.24	93 (of 100 trials)

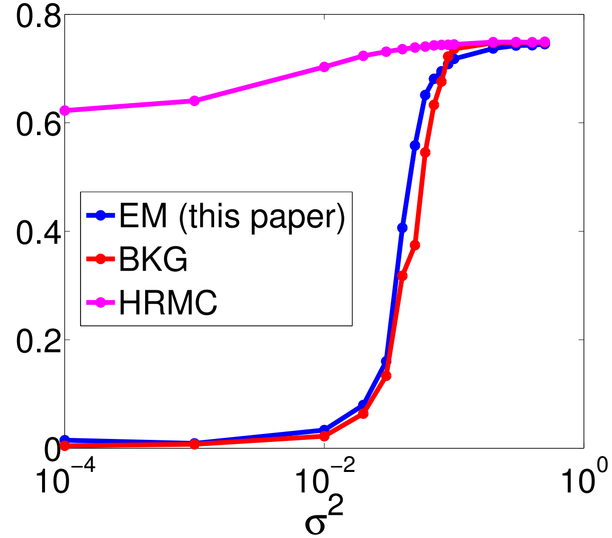
n_1=50, r=4, k=10, n_2=40000, sampling = 60%

The success probability of high-rank MC can be improved a great deal by increasing the # of seeds, which drastically increases the running time.

Results: Synthetic Experiment







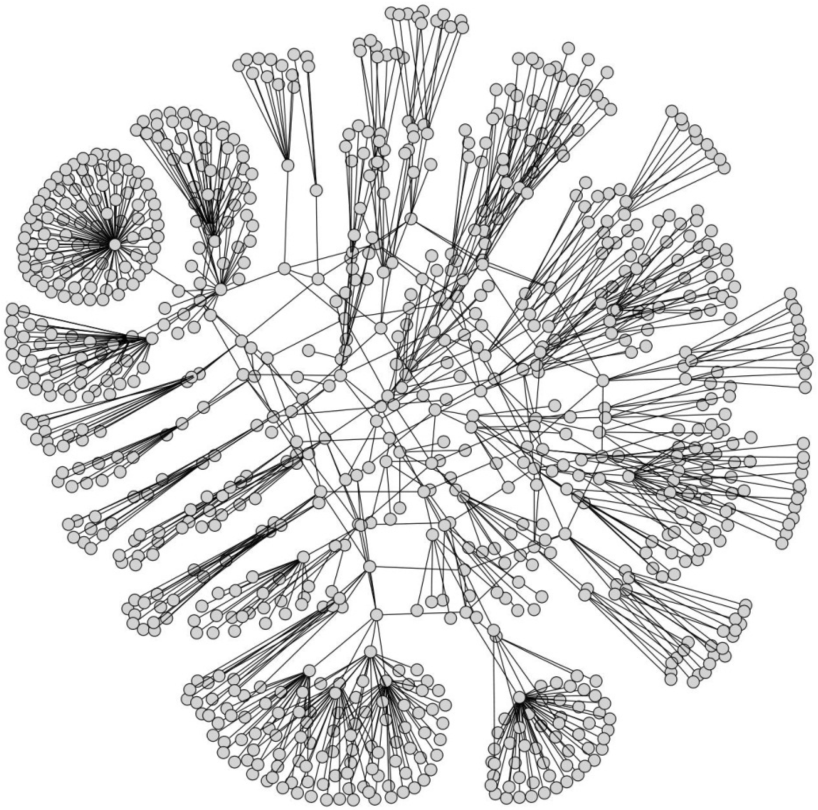
Ambient dimension n_1=100, k=4, r=5

Fig 1: 24 samples per vector

Fig 2: N_k = 210 samples per subspace

Fig 3: N_k = 300, 24 samples per vector

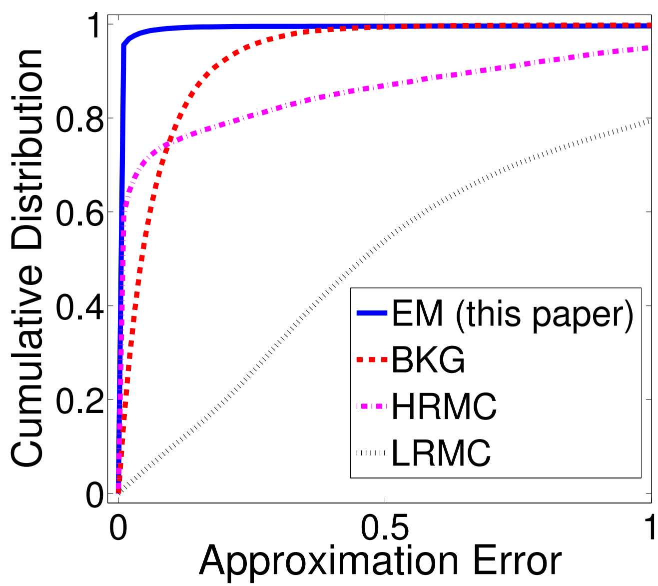
Results: Synthetic Network Distance Experiment



Synthetic Heuristically Optimized Topology (Li, et. al, Sigcomm 2004)

75 monitors and 2,700 end hosts in 12 subnets

Only 40% of the distances were observed



Union of Subspaces Open Questions

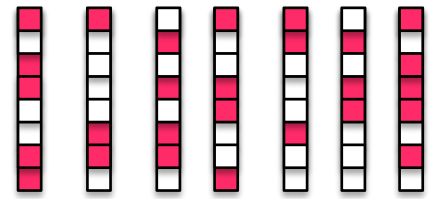
- Heuristic K-Subspaces algorithm (developed w/Arthur Szlam) and EM algorithm (w/ Daniel Pimentel) work very well in practice. Can we prove it?
- Gap between low rank matrix completion sampling (r log n) and requirements for overlap in calculating distances (root n)
  - We need a way to check the mask of missing entries to see whether those data would lie in a unique low-dimensional subspace.

University of Michigan

Thank you!

Complete each column

using the incomplete data projection:



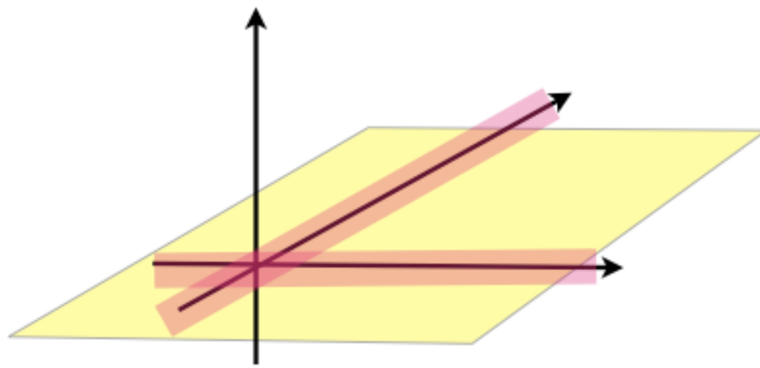
$$\|v_\Omega - P_{S_\Omega} v_\Omega\|_2^2$$

Theorem: If $|\Omega| = O(\mu(S) d \log d)$ and Ω is chosen uniformly with replacement, then with high probability and ignoring constant factors,

$$\frac{|\Omega| - d\mu(S)}{n} \|v - P_S v\|_2^2 \leq \|v_\Omega - P_{S_\Omega} v_\Omega\|_2^2 \leq \frac{|\Omega|}{n} \|v - P_S v\|_2^2$$

Low Rank Models: Union of subspaces

with Brian Eriksson and Rob Nowak

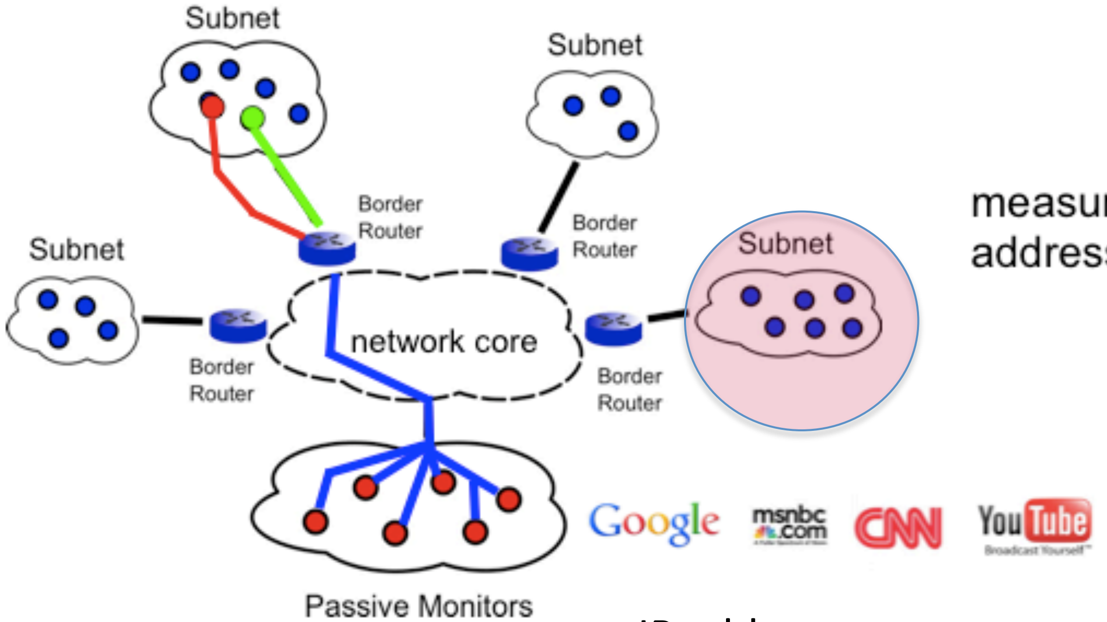


k=2

	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

Low Rank Models: Union of subspaces

with Brian Eriksson and Rob Nowak



measure distance from each IP address ● to each monitor ●

IP addresses

monitors

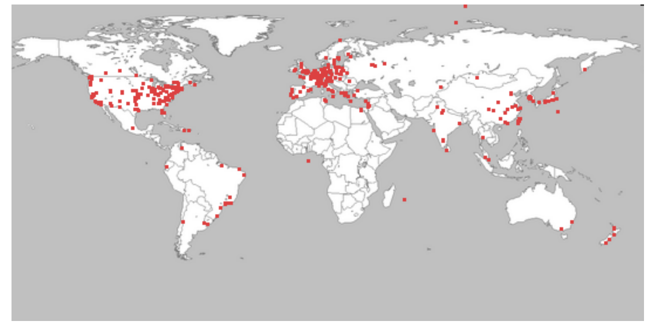
	10	5	10	8		20		15		10	10					20	8		5
1	2	7					12		24	14	2	3			12				1
			2	5		52	5		10				1	15		4	5	2	
8	16	4			16					8		24	4		16	16		8	
		3	6	11			11	9	22		2			33		12	11	6	
5	10		4	13	8			6		4		15	2	39	13			4	

Results: Real-World Delay Measurement Experiment

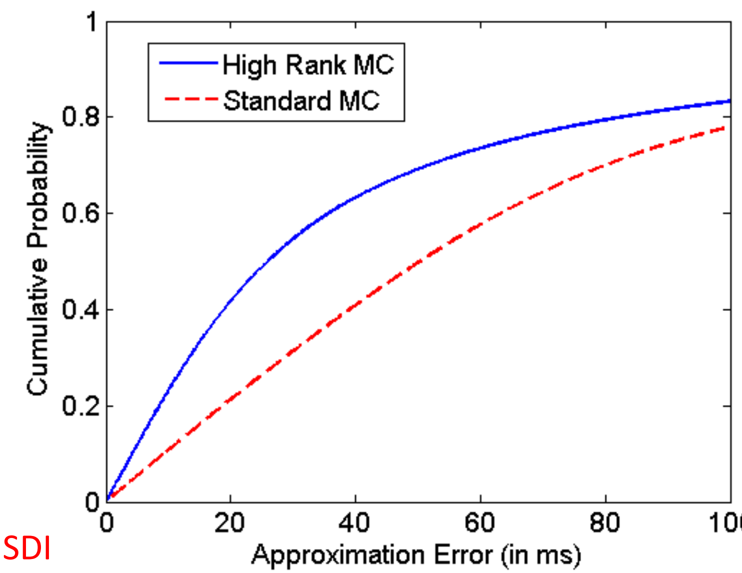
Delay measurements from 100 Planetlab monitors to over 12,000 P2P end hosts in an *unknown number of subnets.*

Only 80% of the delays were observed due to system limitations.

Using 20% of the delays for estimation purposes.



Planetlab Installation Sites



Slides courtesy Brian Eriksson

Jonathan Ledlie, Paul Gardner, and Margo Seltzer, Network Coordinates in the Wild, In Proceedings of NSDI 2007, Cambridge, MA, April 2007

University of Michigan

Assigning columns to subspaces

Let the number of samples per column be $|\Omega| = m$, and consider projecting a vector v_Ω onto subspaces S^i of rank r_i using P_{S^i}, $i = \{0, 1\}$. Let the parameters $\alpha_1, \alpha_0, \beta_1, \gamma_1 > 0$ and $\mu(S^i)$ be the incoherence of S^i.

$$C(m) = \frac{m(1-\alpha_1) - r_1 \mu(S^1) \frac{(1+\beta_1)^2}{1-\gamma_1}}{m(1+\alpha_0)} \qquad \theta_0 = \sin^{-1}\left(\frac{\|v - P_{S^0} v\|_2}{\|v\|_2}\right)$$

Theorem 1. *Let* $\delta > 0$ *and* $m > \frac{8}{3} r_1 \mu(S^1) \log\left(\frac{2r_1}{\delta}\right)$. *Assume that*

$$\sin^2(\theta_0) < C(m) \sin^2(\theta_1) \ .$$

Then with probability at least $1 - 4\delta$,

$$\|v_\Omega - P_{S^0_\Omega} v_\Omega\|_2^2 < \|v_\Omega - P_{S^1_\Omega} v_\Omega\|_2^2 \ .$$

In particular, if $v \in S^0$ *and thus* $\theta_0 = 0$, *and if* $\theta_1 > 0$, *then the result holds as long as* $C(m) > 0$.