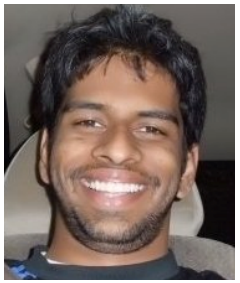
Efficient Spectral Clustering using Selective Similarities

Aarti Singh

Joint work with:



S. Balakrishnan

A. Krishnamurthy

M. Xu

ML MACHINE LEARNING DEPARTMENT

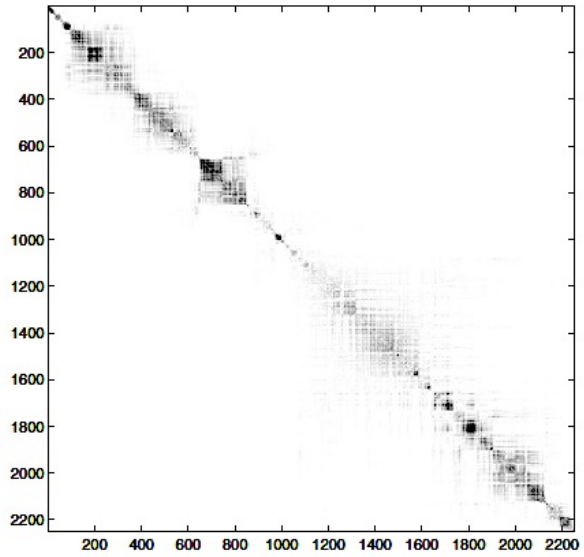
Carnegie Mellon.
School of Computer Science

Challenges of large-scale clustering

Identifying fiber-bundles in the brain

(Brun et al MICCAI'04)

Diffusion spectral imaging (DSI)
250,000 fiber micro-tracks



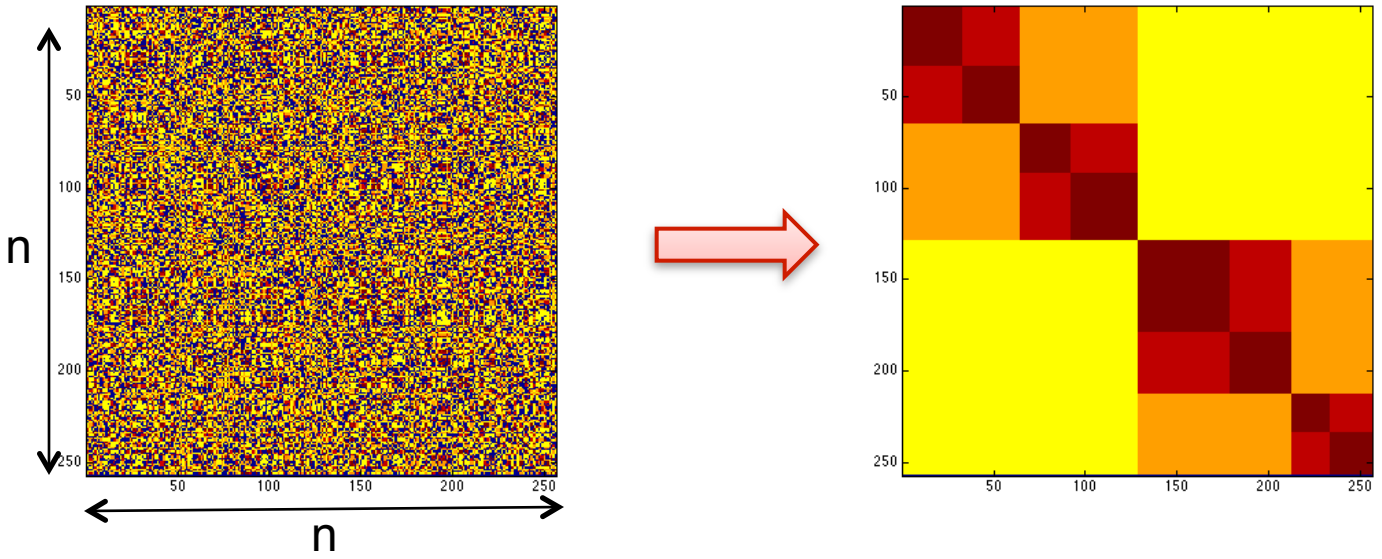
Pairwise similarities
between fiber microtracks

- Lot of fibers
- Multiple Scales
- Noisy similarities
- Computation is expensive

Collaborator: Schneider Lab at Univ. Pittsburgh, Center for Neural Basis of Cognition at CMU

Robust and Efficient Clustering

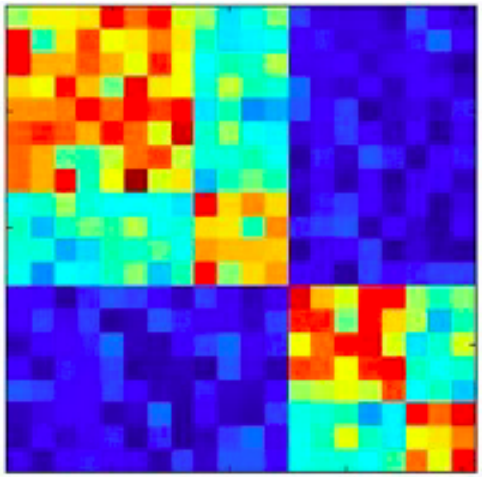
Goal: Given a noisy and incomplete pairwise similarity matrix, re-order rows/columns to infer groups with high within-cluster similarity and low between-cluster similarity.

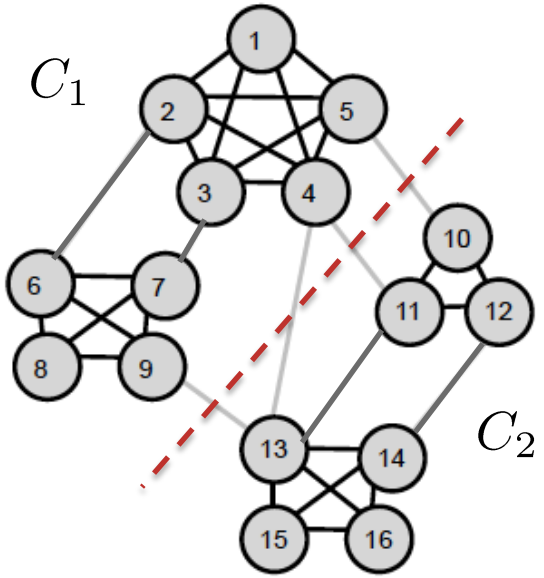


- **Robustness:** How much **noise** can a clustering algorithm tolerate while recovering all clusters up to a desired **resolution**?
- **Efficiency:** How many **similarity measurements** and/or **computation** are necessary for robust clustering?

Spectral clustering

W =





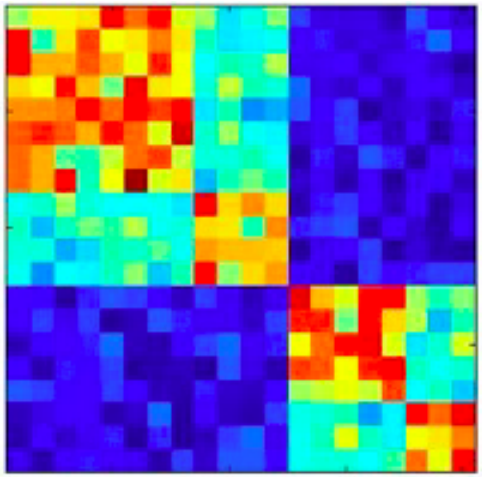
Balanced ratio-cut - Partition the graph into approximately equal size clusters such that weight of edges between them is minimized.

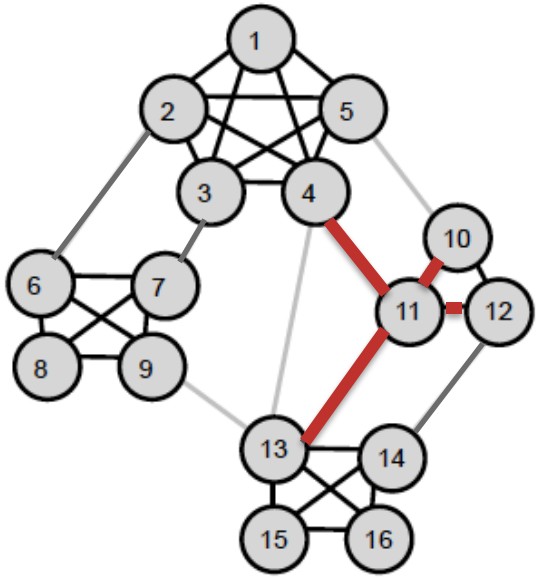
$$\min_{C_1,C_2} \sum_{i\in C_1, j\in C_2} W_{ij}\left(\frac{1}{|C_1|}+\frac{1}{|C_2|}\right)$$

NP Hard to solve!

Spectral Clustering – solves a relaxed version of the balanced graph cut.

Spectral clustering

W = 



Spectral Clustering – Second smallest eigenvector of the Graph Laplacian **L** approximates balanced cut

W : symmetric similarity matrix (n x n)

D : diagonal degree matrix $D_{ii} = \sum_{j=1}^{n} W_{ij}$

L = D – W : Graph Laplacian (unnormalized)

Note: **L1 = D1 - W1** = 0 smallest eigenvector **1** if graph is connected.

Spectral clustering Algorithm

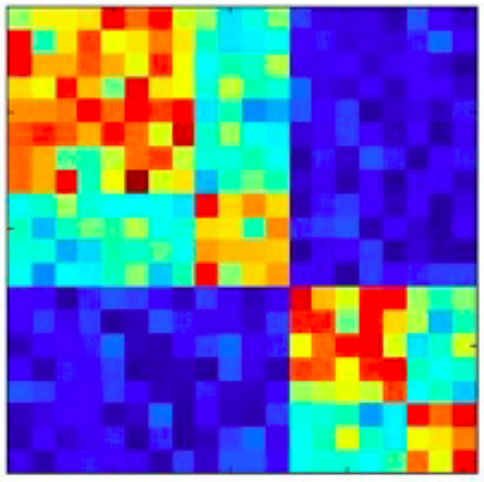
- Hierarchical Binary Spectral Clustering

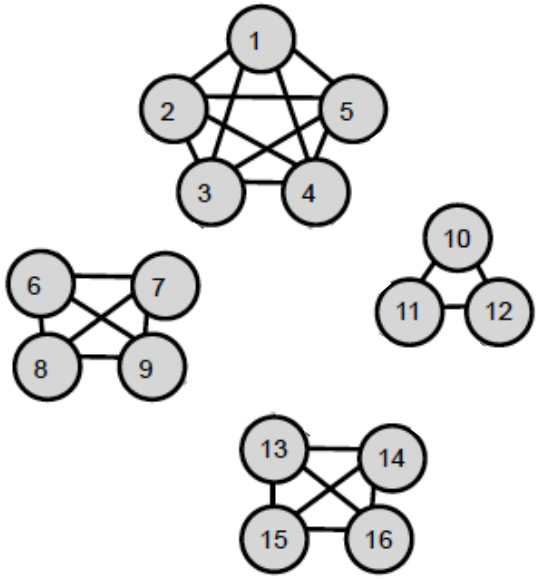
  Compute **L** = **D** – **W**

  $\mathbf{v}_2 \leftarrow$ second smallest eigenvector of **L**

  $C_1 = \{i : \mathbf{v}_2(i) \geq 0\}, C_2 = \{i : \mathbf{v}_2(i) < 0\}$

  Repeat on each cluster

W = 



What is the price of solving this relaxation of the balanced ratio-cut problem?

Prior Justification

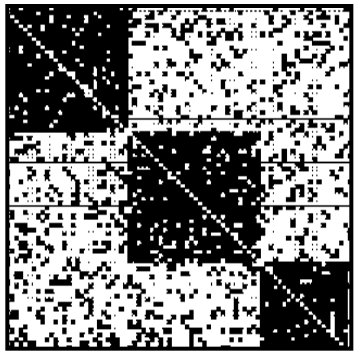
High-level justifications: Connection to graph cut, random walks on graph, electric network theory, Laplace-Beltrami operator on manifold

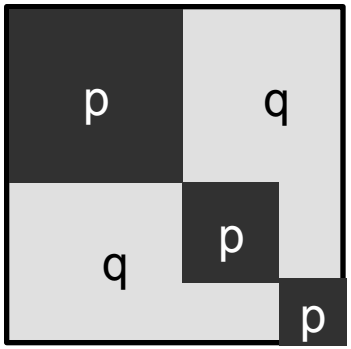
- don't translate to cluster recovery guarantees

Perturbation analysis: Eigenvectors are stable in ℓ_2 norm under small similarity perturbations

- Fraction of misclusterings $\longrightarrow$ 0 Ng et al (2001), Huang et al (2009)

Stochastic Block Model/ Planted Partition Model

W = 

E[**W**] = 

McSherry (2001), Rohe-Chatterjee-Yu (2010), Sussman et al (2011)

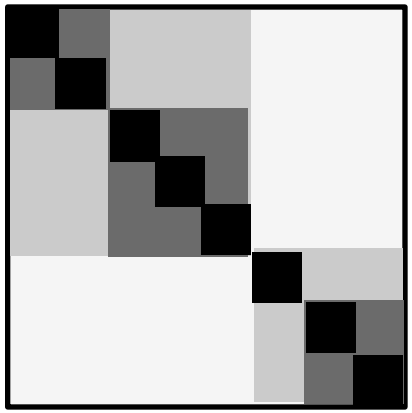
probability of within-cluster edge, p > probability of between-cluster edge, q

Laplacian eigenvectors of Ultrametrics

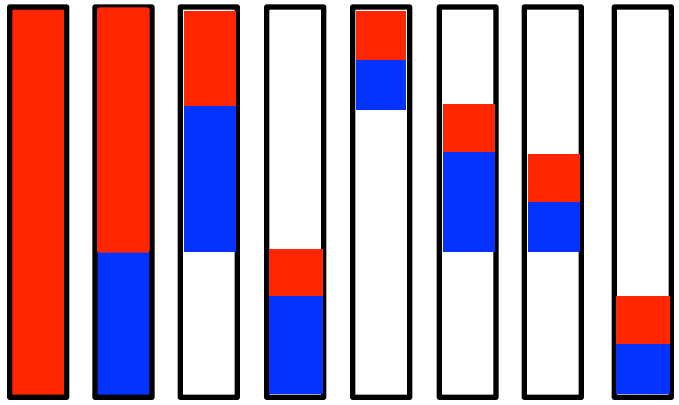
Why should Laplacian eigenvectors of hierarchically block matrices reveal cluster structure?

Ultrametric: Noiseless constant block similarities.

W =



Eigenvectors of Laplacian **L**
= Unbalanced Haar wavelets



Sharpnack, Singh (NIPS 2010)

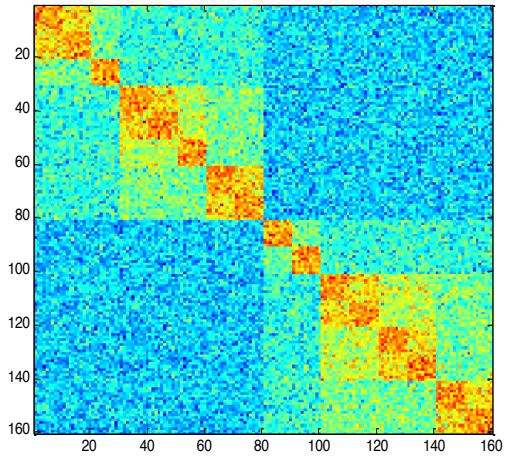
Not low-rank or compressible!

Signal+Noise model for Hierarchical Clustering

Observed Hierarchical Similarity matrix:

W = A + R

R ~ i.i.d. zero mean, subgaussian(σ^2) perturbation
(includes Bernoulli)



Signal $\mu = \min_{(i,j)\in C} A_{ij} - \max_{i\in C, k\notin C} A_{ik} - \eta \left(\max_{i\in C, k\notin C} A_{ik} - \min_{i\in C, k\notin C} A_{ik} \right)$

Within vs between cluster similarities

Cluster balance factor

Range of between cluster similarities

Signal-to-Noise Ratio, SNR = $\frac{\mu}{\sigma}$

Noise Thresholds for Spectral Clustering, NIPS 2011.

Robustness of Spectral Clustering

Spectral clustering limit: If

$$\text{SNR} \quad \left(\frac{\mu}{\sigma}\right)^2 \gtrsim \sqrt{\frac{\log n}{|C|}}$$

then, with probability > 1-1/n, **hierarchical binary spectral** clustering will exactly recover all clusters of size at least |C| in a binary hierarchy.

- Similar result for k-way partitional clustering
- SNR depends on the size of smallest cluster we want to resolve.
- Popular greedy merging strategies for hierarchical clustering such as single linkage, complete linkage or average linkage fail under this level of noise.

Minimax SNR for Clustering

$$\min_{\text{all clusterings, } \widehat{C}(\mathbf{W})} \quad \max_{\mathbf{W} \text{with SNR } \mu/\sigma} Pr(\widehat{C}(\mathbf{W}) \neq C)$$

Information Theoretic limit: If $\text{SNR} \quad \frac{\mu}{\sigma} \preccurlyeq \sqrt{\frac{\log n}{|C|}}$

then, for **any** clustering procedure, the probability of failing to recover clusters of size |C| remains bounded away from zero by a constant.

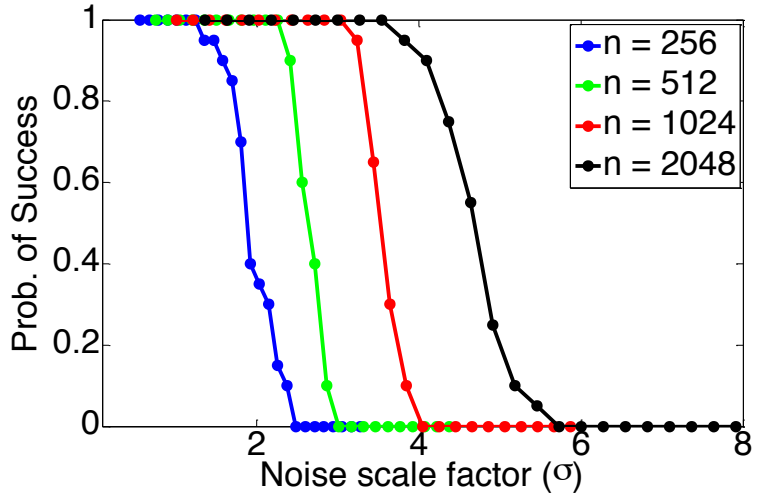
Spectral Clustering
Price for relaxation ?

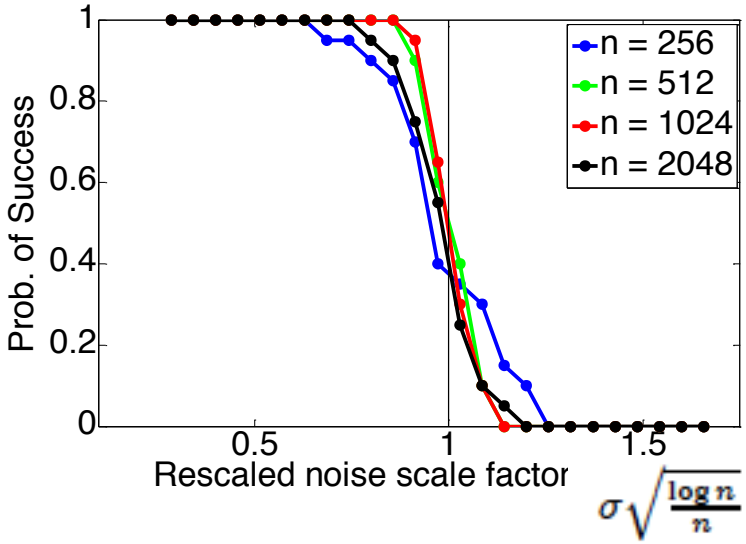
Balanced ratio-cut limit: If $\text{SNR}^{p} \quad \frac{\mu}{\sigma} \succcurlyeq \sqrt{\frac{\log n}{|C|}}$

then, with probability 1−1/n, the **combinatorial minimum balanced-cut** procedure exactly recover all clusters of size |C|.

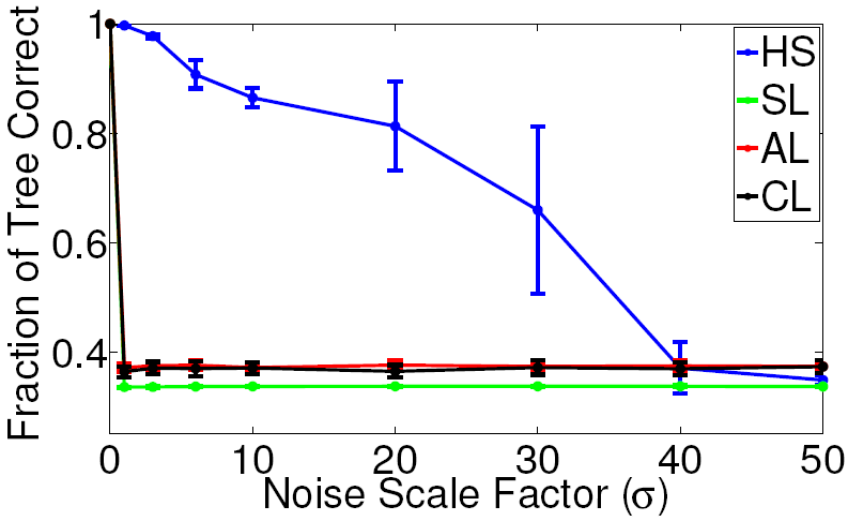
Simulation Results

Noise threshold – two balanced clusters





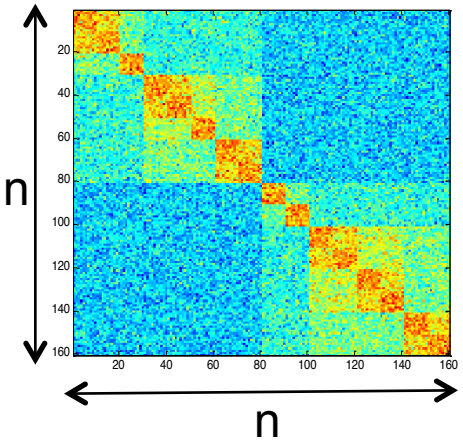
Robustness: comparison with other hierarchical clustering algorithms



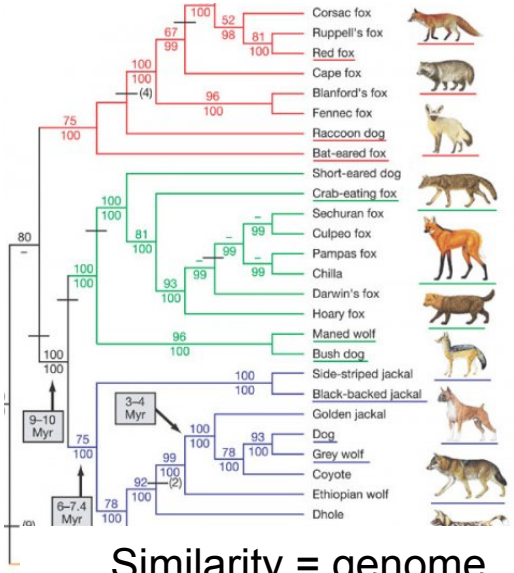
HS Hierarchical Spectral
SL Single Linkage
AL Average Linkage
CL Complete Linkage

Efficiency of Clustering

Similarities are costly to compute or measure

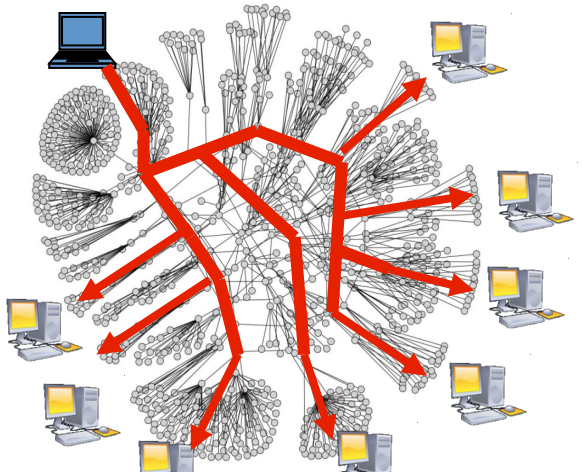


Phylogenetics



Similarity = genome sequence alignment

Identify network topology

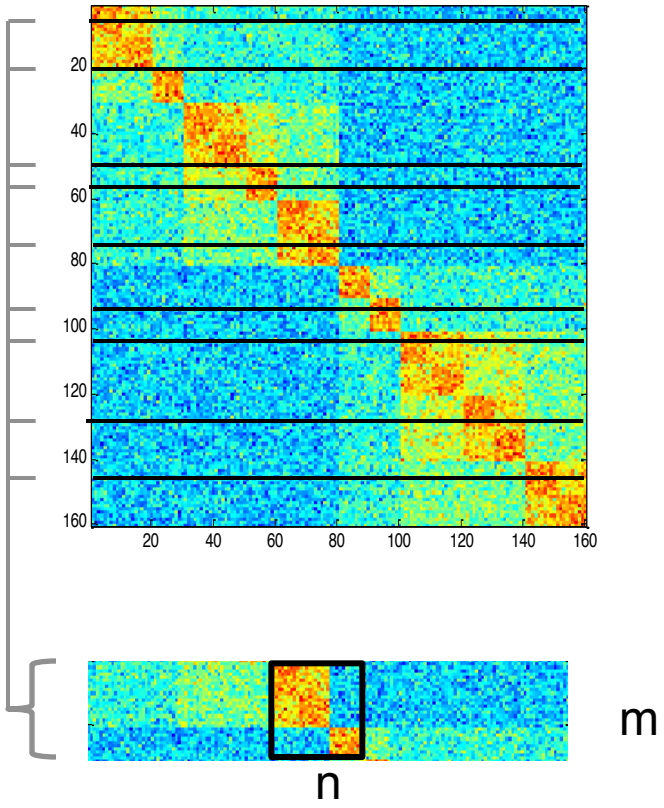


Similarity = delay covariance

Prior work: Can resolve clusters of size $\Omega(n)$ using $O(n \log n)$ **randomly chosen** similarities (fraction of misclusterings $\longrightarrow 0$) Hunter-Strohmer'10, Shamir-Tishby'11

Our work: Can resolve all clusters of size up to $\Omega(\log n)$ in a hierarchy using $O(n \log^2 n)$ **selective** similarities

Active Hierarchical Clustering



- Pick m objects at random
- Split into two (or k) clusters

e.g. Compute $\mathbf{L}_m = \mathbf{D}_m - \mathbf{W}_m$

$\mathbf{v}_2 \leftarrow$ second smallest evec

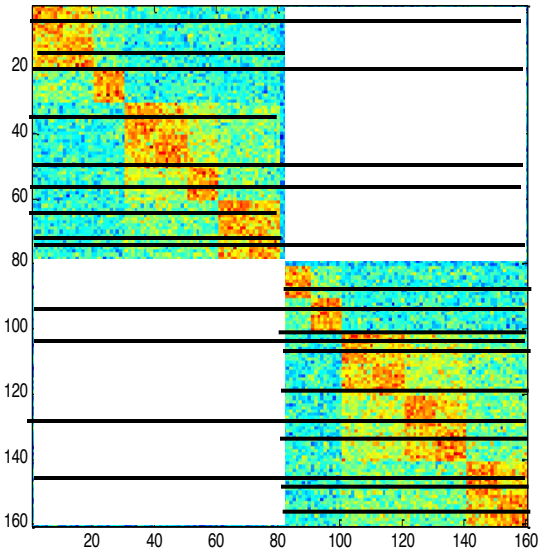
$$C_1 = \{i : \mathbf{v}_2(i) \geq 0\}$$

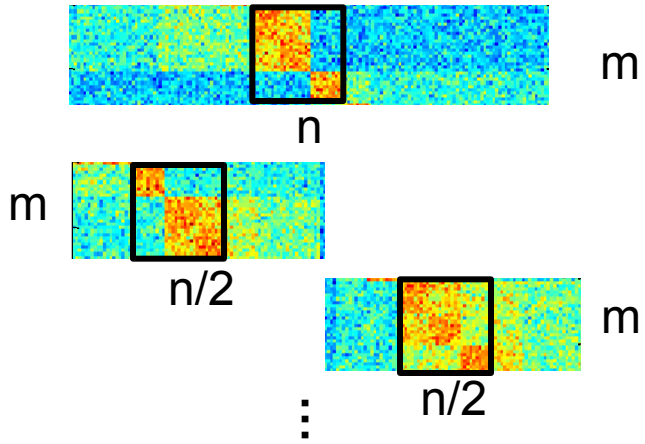
$$C_2 = \{i : \mathbf{v}_2(i) < 0\}$$

- Assign each remaining object to the cluster with higher average similarity
- *Repeat* on each cluster

Efficient Active Algorithms for Hierarchical Clustering, ICML 2012 (*To appear*).

Active Hierarchical Clustering





- Pick m objects at random
- Split into two (or k) clusters

e.g. Compute $\mathbf{L}_m = \mathbf{D}_m - \mathbf{W}_m$

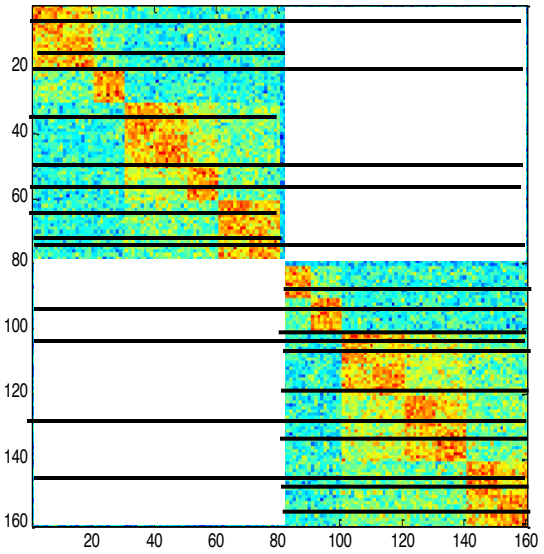
$\mathbf{v}_2 \leftarrow$ second smallest evec

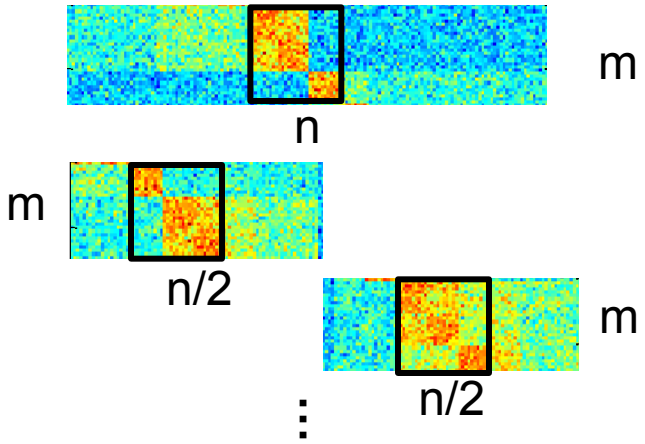
$$C_1 = \{i : \mathbf{v}_2(i) \geq 0\}$$

$$C_2 = \{i : \mathbf{v}_2(i) < 0\}$$

- Assign each remaining object to the cluster with higher average similarity
- *Repeat* on each cluster

Active Hierarchical Spectral Clustering





Measurement Efficiency:

#similarities needed on each iteration = nm

If clusters approx balanced,

total # similarities = $nm + 2\frac{nm}{2} + 4\frac{nm}{4} + \ldots$

$= O(nm \log n)$

$= O(n \log^2 n)$ if $m = \log n$

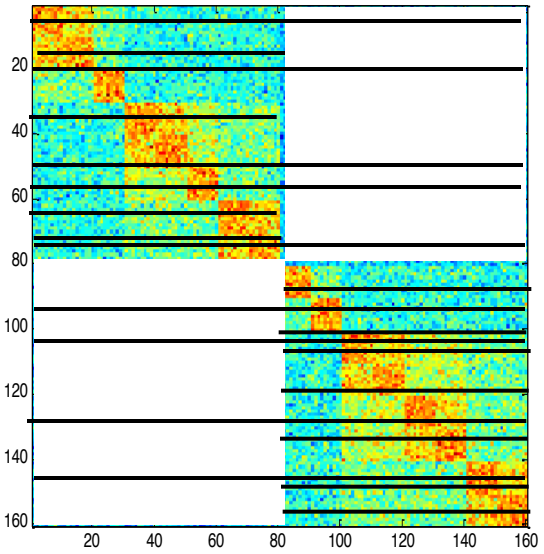
minimum possible

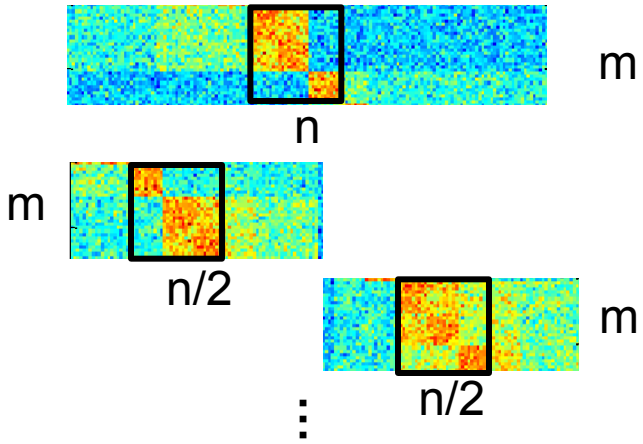
Computational Efficiency:

only need to compute eigenvectors of m x m matrices (log n x log n)

$O(nm^2)$ $= O(n \log^2 n)$ if $m = \log n$

Active Hierarchical Spectral Clustering





Robustness Analysis:

Let m = log n

- Each split succeeds with high probability

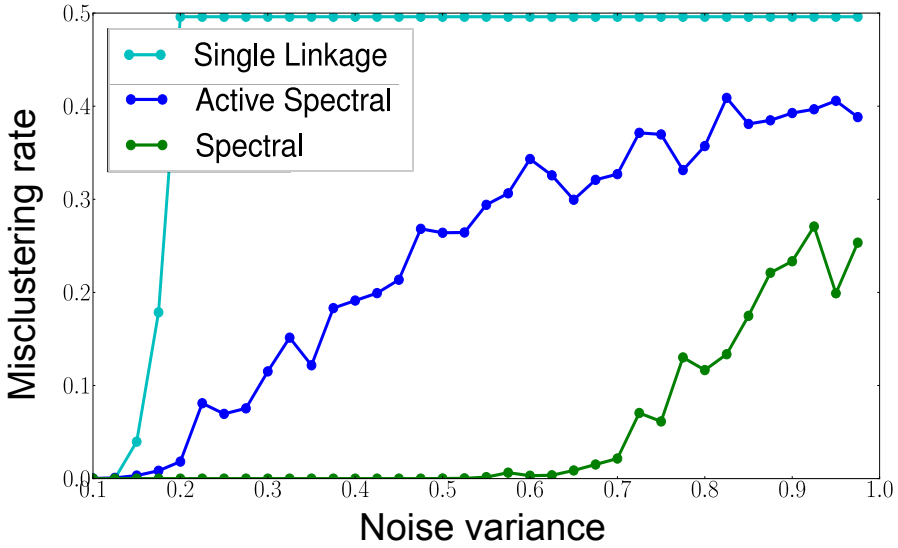
  if $\left(\frac{\mu}{\sigma}\right)^2 \sim \sqrt{\frac{\log n}{m}} \quad \Rightarrow \frac{\mu}{\sigma} = \text{constant}$

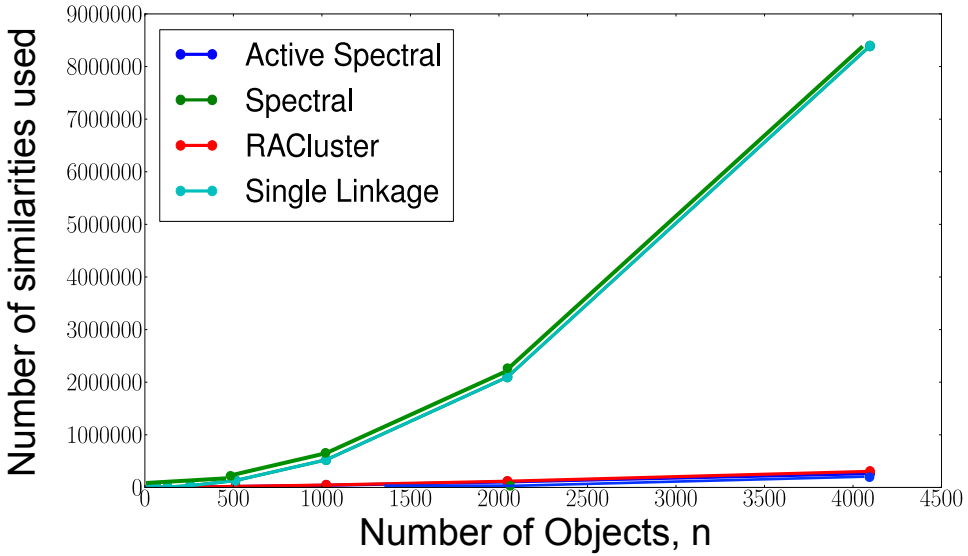
- Each round of object assignments succeeds with high probability

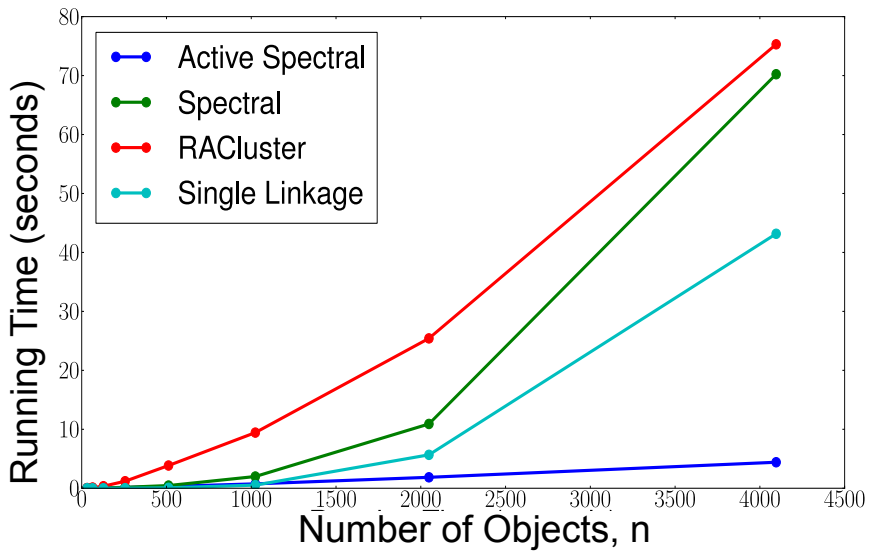
  if $\underbrace{\sigma\sqrt{\frac{\log n}{m}}}_{\text{max of n subgaussians with scale factor } \sigma/\sqrt{m}} < \mu \quad \Rightarrow \frac{\mu}{\sigma} = \text{constant}$

- Union bound over all splits

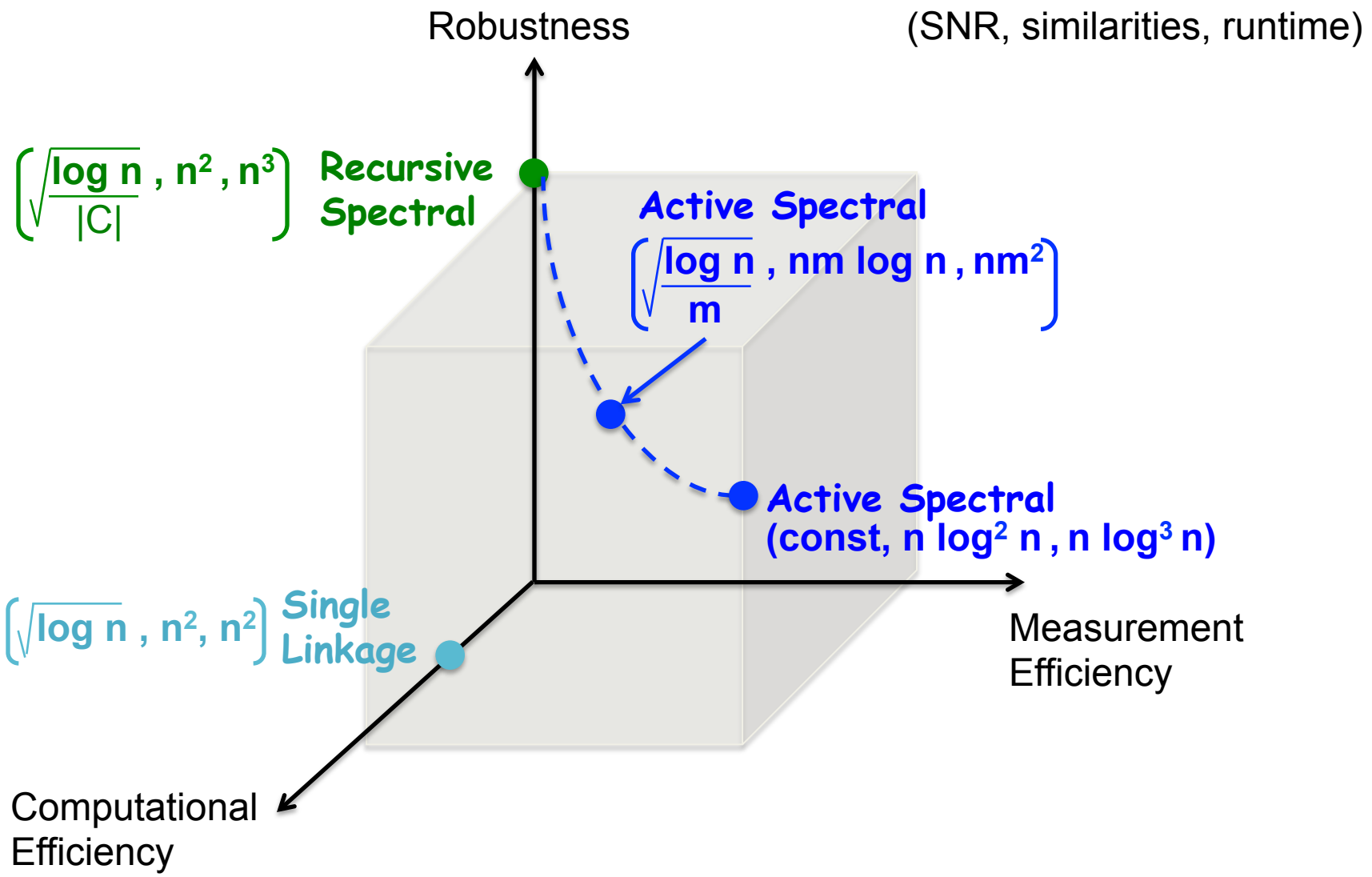
Simulation Results



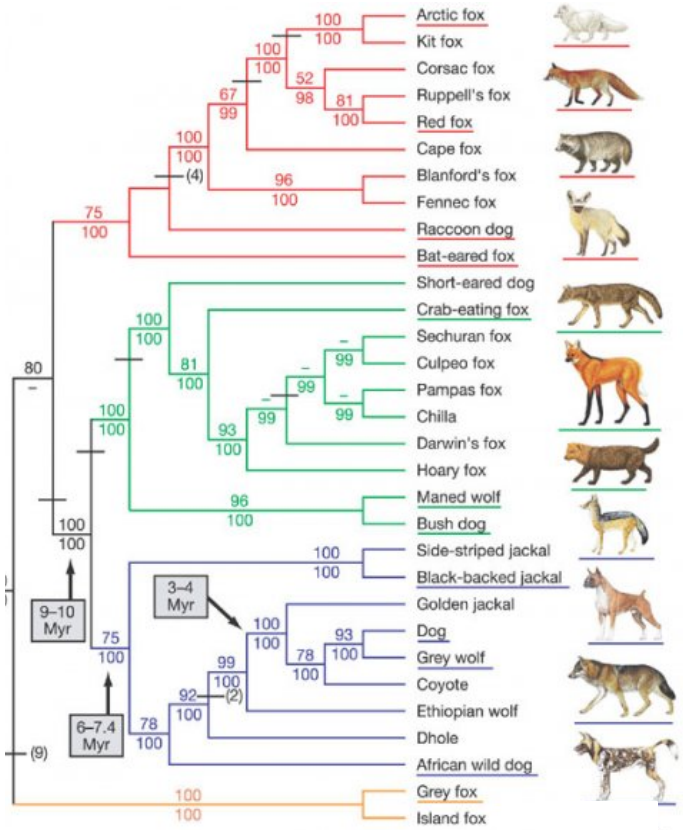




Robustness vs Efficiency Tradeoffs



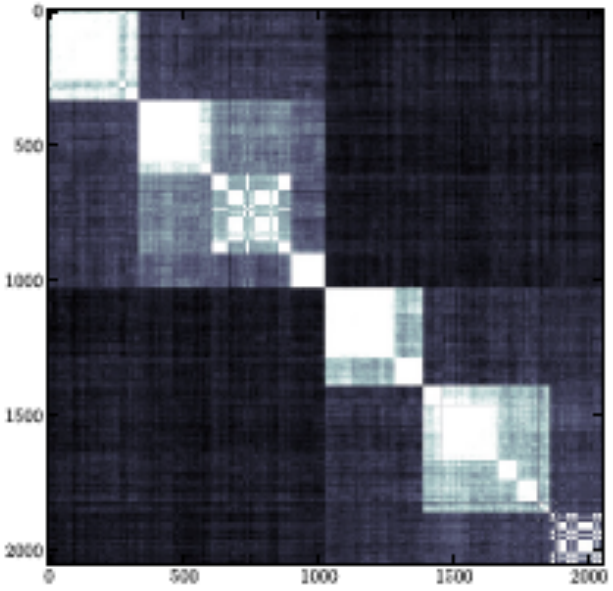
Application to Evolutionary tree reconstruction



Similarity = genome sequence alignment

2048 genome sequences
with 2000 base pairs (phyClust)

Non-active: Runtime 15000s, all similarities



Active Spectral Clustering

Runtime 600s, # similarities 3.5%

Thanks

Hierarchically-structured high-rank matrices can be completed using $O(n \log^2 n)$ selectively sampled entries!

References:

- Noise Thresholds for Spectral Clustering, NIPS 2011.
- Efficient Active algorithms for Hierarchical Clustering, ICML 2012.

Funding Acknowledgements:

National Science Foundation

Air Force Office of Scientific Research

NIH's MIDAS Center University of Pittsburgh