KDD 2012 BEIJING
THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
Beijing, China
August 12-16, 2012

# Active Spectral Clustering via Iterative Uncertainty Reduction

**Fabian L. Wauthier**
UC Berkeley
flw@cs.berkeley.edu

**Nebojsa Jojic**
Microsoft Research
jojic@microsoft.com

**Michael I. Jordan**
UC Berkeley
jordan@cs.berkeley.edu

## One Minute Summary

Spectral clustering assumes *all* pairwise similarities are known *exactly*. This can render it unsuitable when similarities are expensive or noisy. We propose an iterative algorithm that uses intermediate clusterings to choose which similarity to measure next in order to remove clustering uncertainty. Our method relies on matrix perturbation theory and outperforms a related algorithm by Shamir and Tishby. An extension to noisy measurements is straightforward.

## Spectral clustering

- Let $W = \{w_{ij}\}$ be a matrix of pairwise similarities. Define $L = \mathrm{diag}(W\mathbf{1}) - W$. Find the embedding $v^*$

$$v^* = \mathrm{argmin}_v v^T L\, v = \mathrm{argmin}_v \sum_{ij} w_{ij}\big(v(i) - v(j)\big)^2$$
$$\text{s.t. } v^T v = 1, v^T\mathbf{1} = 0.$$

- Cluster by thresholding $v^*$ at 0: $c_i = \mathrm{sign}(v^*(i))$.
- Pairwise similarities are often expensive to acquire.
- Common to impute 0 for missing similarities $w_{ij}$, but performance can be poor if missing at random.
- Want to do better by actively choosing subset of similarities.

## Cluster Uncertainty

For $i$ if $\forall j, w_{ij} \approx c > c_0$ we have $v^*(i) \approx \frac{\sum_{j\neq i} v^*(j)}{n-1} \approx 0$.
Points that are sufficiently similar to all other points tend to lie near threshold. Clustering is most "uncertain" about points near threshold.

## Algorithm Sketch

- Try to push embedded clusters apart: Incrementally measure similarities so few points lie near threshold.

**Iterate**

**1)** Impute missing similarities in $W$ by 0; solve for $v^*$.

**2)** Let $k_{\min} = \mathrm{argmin}_i |v^*(i)|$, and find the similarity that would most change the embedding near the threshold, if it were perturbed:

$$(i^*, j^*) = \mathrm{argmax}_{(i,j)\in \mathrm{Unobserved}} \left|\frac{dv^*(k_{\min})}{dw_{ij}}\right|.$$

**3)** Measure $w_{i^*j^*}$ and add it to $W$.

- Embedding $v^*$ is an eigenvector of $L$. So $\left|\frac{dv^*(k_{\min})}{dw_{ij}}\right|$ can be computed using matrix perturbation theory.
- If $L = \sum_p \lambda_i v_i v_i^T$, with $0 = \lambda_1 \leq \cdots \leq \lambda_n$, then

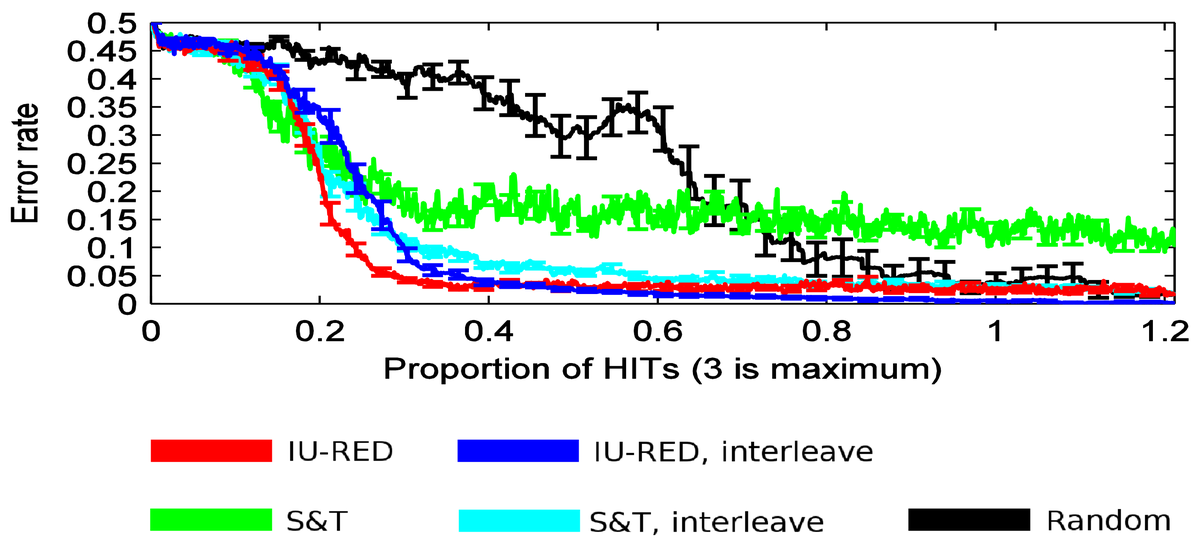
$$\left|\frac{dv^*(k_{\min})}{dw_{ij}}\right| = \left|\sum_{p>2}^{n} \frac{v_2^T[\partial L/\partial w_{ij}]v_p}{\lambda_2 - \lambda_p} v_p(k_{\min})\right|.$$

- Exploits emerging structure in a partial clustering as a guide to active learning. Works well if dataset actually clusters.
- Shamir and Tishby do not assume dataset clusters and do not exploit emerging structure as a guide.

## Image Clustering

Sample images in an open kitchen and adjacent living room.

- Task: Cluster images taken in an open kitchen/living room (see above) into their location.
- Unconstrained images; very hard for computer vision.
- Collect similarities using HITs on Mechanical Turk: Query: **"How likely is it that two photos were taken in the same room?"**
- Use median response of three workers as similarity.

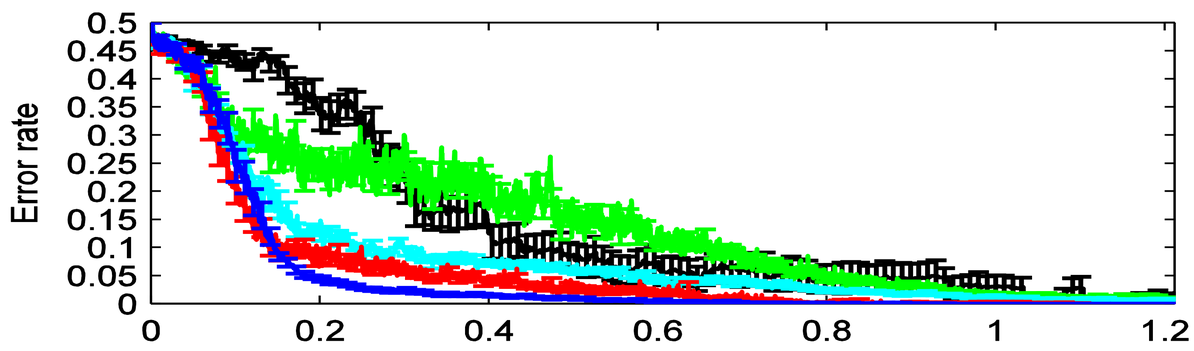



## Noisy Similarities

- Similarities are often noisy. E.g. Mechanical Turk, biological experiments, etc.
- Can reduce noise by computing median. Expensive!
- Gradient $\left|\frac{dv^*(k_{\min})}{dw_{ij}}\right|$ gives noise sensitivity of embedding. Tells us which similarities need to be known accurately.
- Augment algorithm:
  - Maintain running median $w_{ij}$ of each similarity using repeat measurements.
  - Estimate associated standard deviations $\sigma_{ij}$.
  - Measure similarity that is most uncertain *and* would most change the embedding, if it were perturbed:

$$(i^*, j^*) = \mathrm{argmax}_{(i,j)} \sigma_{ij} \left|\frac{dv^*(k_{\min})}{dw_{ij}}\right|.$$

## Image Clustering with Noisy Similarities




## References

[1] U. von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17:395-416, 2007.
[2] G.W. Stewart and J Sun. *Matrix Perturbation Theory.* Computer Science and Scientific Computing. Academic Press, 1990.
[3] O. Shamir and N. Tishby. Spectral Clustering on a Budget. *Proceedings of the International Conference on Artificial Intelligence and Statistics,* 2011.

**Paper ID: 607**