Sparsification for Total Variation Clustering

Kevin Costello

Ehsan Kamalinejad

Thomas Laurent

AMS Sectional Meeting, UCR

Overview

- Total Variation Clustering
- Sparsification Based on Spanning Trees
- Results and Future Work

Total Variation Clustering

Let $V = \{x_1, x_2, \ldots, x_n\}$ be a data set of points and $W = \{w_{ij}\}_{1\le i\le n\,,\,1\le j\le n}$ be the similarity weights.

Then Cheeger Balance Cut of $S \subset V$ is given by

$$C(S) = \frac{\sum_{x_i\in S}\sum_{x_j\in S^c} w_{ij}}{\min\{|S|,|S^c|\}}.$$

Minimizing this problem over all sets S, is considered to be one of the most popular clustering problems.

This problem is NP-hard and it is approximated by a continuous relaxation given by

$$E(f) = \frac{\sum_{i,j} w_{ij}\,|f_i - f_j|^2}{\sum_i |f_i - m_2(f)|}$$

where $f: V \to R$ is a real function and $m_2(f)$ is the median of f.

The l^2 minimization problem

$$\arg min\{E(f) = \frac{\sum_{i,j} w_{ij}\left|f_i - f_j\right|^2}{\sum_i |f_i - m_2(f)|}\}$$

is known as the standard spectral clustering.

This problem gives us only an approximated (smooth) solution of Cheeger cut problem.

The following total variation (l^1) minimization problem is an exact continuous relaxation of Cheeger cut

$$\arg min\{E(f) = \frac{\sum_{i,j} w_{ij}\left|f_i - f_j\right|}{\sum_i |f_i - m_1(f)|}\}$$

where $m_1(f)$ is the median of f.

A Steepest Descent Algorithm

- The steepest descent algorithm is based on the sub-differential of the energy.
- The Convergence of the algorithm was proved and an adaptive multicluster version was developed.
- Under appropriate conditions, the algorithm converges to a binary minimizer which is also a minimizer of the Cheeger cut problem.

---

f^0 nonzero function with $\operatorname{med}(f) = 0$ zero.
c positive constant.
while $E(f^k) - E(f^{k+1}) \geq \text{TOL}$ **do**
  $v^k \in \partial_0 B(f^k)$
  $g^k = f^k + c\, v^k$
  $\hat{h}^k = \arg\min_{u \in \mathbb{R}^n} \{T(u) + \frac{E(f^k)}{2c} \|u - g^k\|_2^2\}$
  $h^k = \hat{h}^k - \operatorname{med}(\hat{h}^k)\mathbf{1}$
  $f^{k+1} = \frac{h^k}{\|h^k\|_2}$
end while

B is the Balance term.
$\partial_0 B$ is the sub-differential of B.
T is the total variation.

Next Step: Sparsification of the Graph

- A sparsifier which is compatible with the total variation algorithm should approximately preserve the weights of cuts.

- We used a sampling sparsifier based on the spanning trees of the graph.

Sampling (Benczur and Karger)

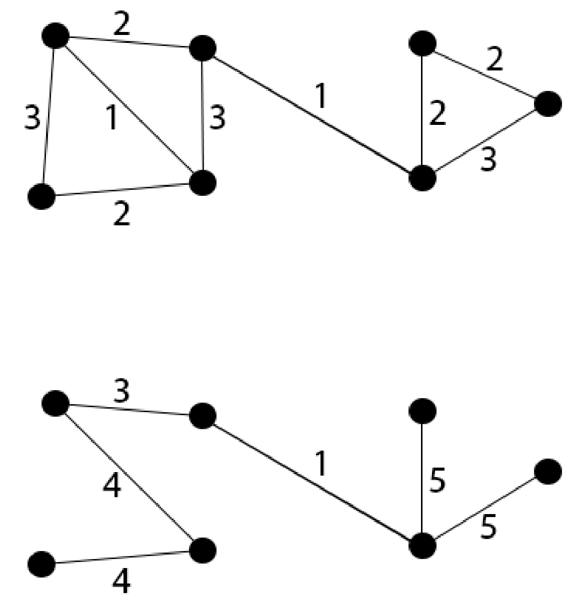
The connectivity con_{ij} of an edge e_{ij} is defined to be the minimal cut that separates v_i from v_j.

Sample the edges independently with probability

$$P_{ij} = max\left\{\frac{c\,w_{ij}}{con_{ij}}, 1\right\},$$

and if an edge is sampled, then set its weight in the new graph equal to $\frac{w_{ij}}{P_{ij}}$.

(c is a constant, $c \sim \theta(log^2 n)$)



Connectivity Sampling Theorems

Sample the edges of a graph G independently with probability

$$P_{ij} = max\left\{\frac{c\, w_{ij}}{con_{ij}}, 1\right\},$$

and if an edge is sampled, then set its weight in the new graph equal to $\frac{w_{ij}}{P_{ij}}$.

This sampling applied to a dense graph G results in a sparse graph G_s which has $O(\frac{n \log n}{\varepsilon})$ edges in expectation, where for every cut $c_G(S, S^c)$ in G we have

$$\left|c_{G_s}(S, S^c) - c_G(S, S^c)\right| < \varepsilon$$

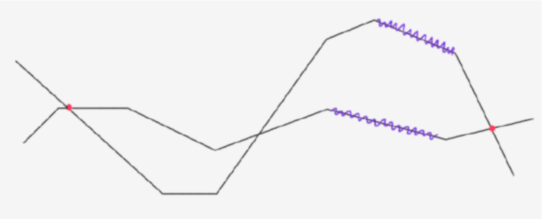
With high probability (1 − inverse polynomial in n).

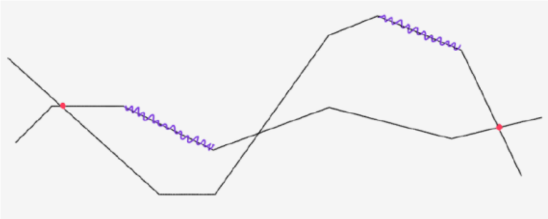
Approximated Connectivity

con_{ij} is very expensive to calculate.

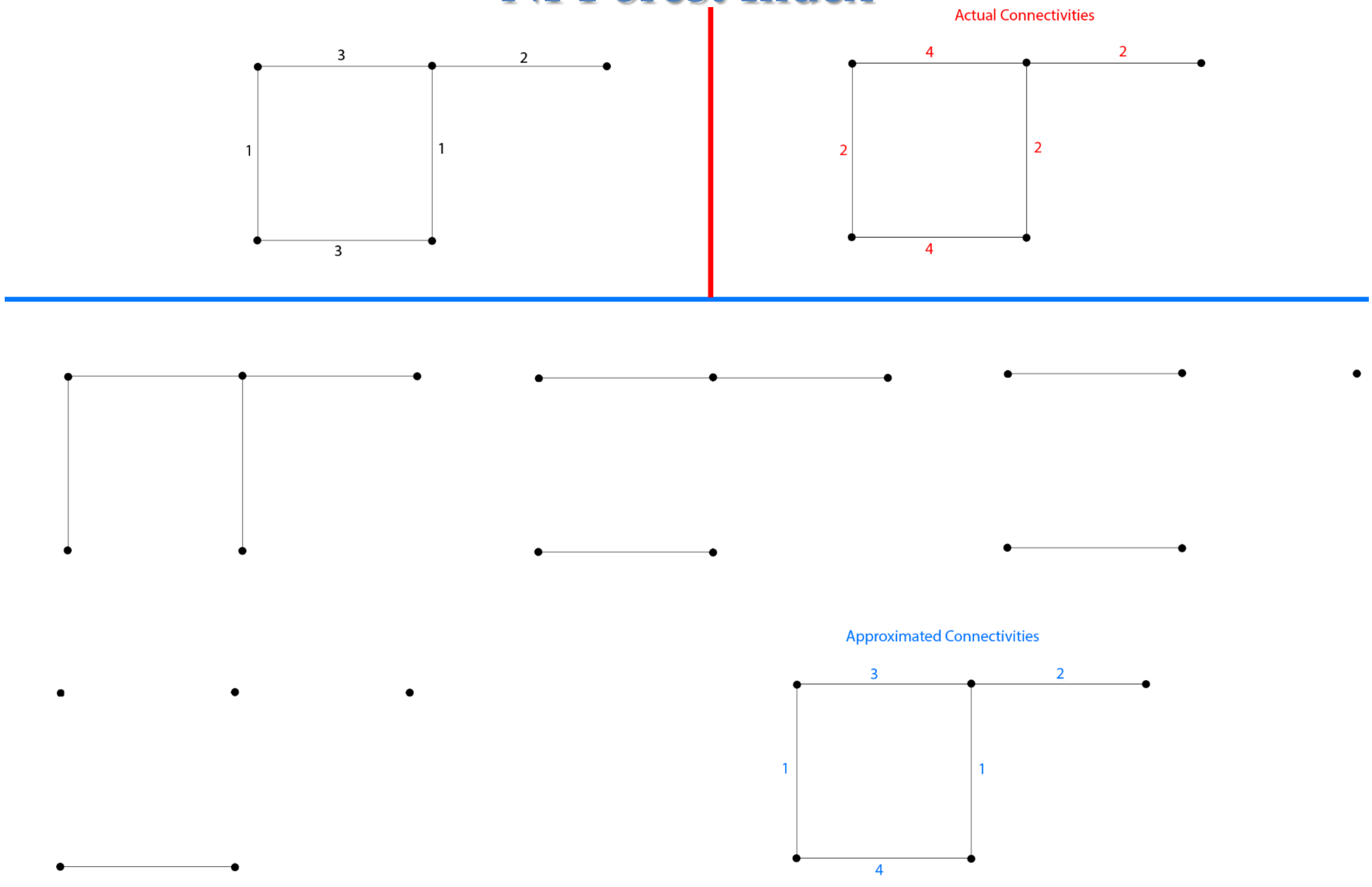
We use Nagamochi-Ibaraki index to approximate the connectivities.

This algorithm works based on the number of spanning forests that connect two vertices together.





NI Forest Index



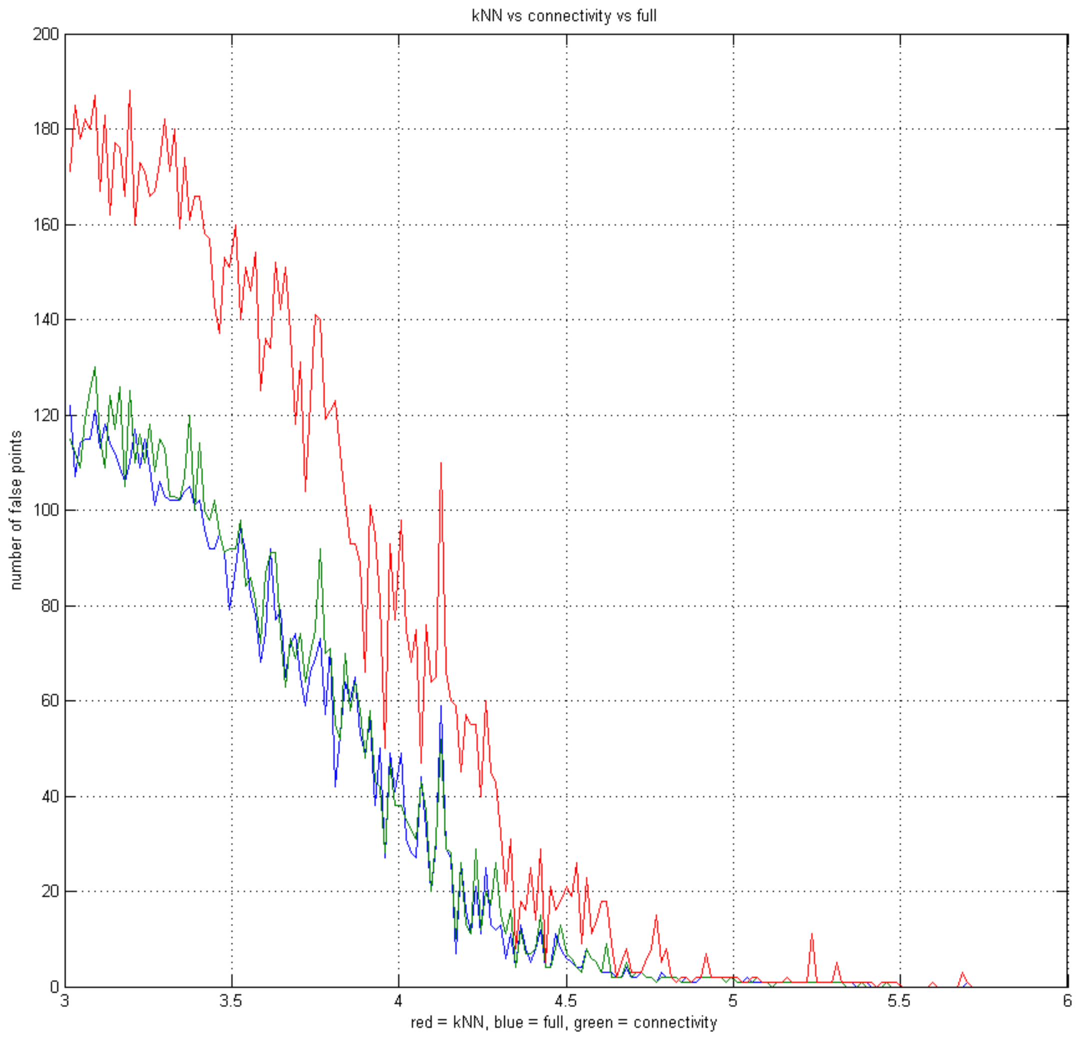
NI Forest Index Approximation for Connectivity

- NI index is a lower approximation of the connectivity hence the cuts are preserved.

- NI index algorithm is very sensitive to the starting vertex. This issue can be solved with randomly picking different starting vertices.

- The run-time of NI index algorithm is $O(n\ log^2 n)$.

Results



kNN vs connectivity vs full
number of false points
200
180
160
140
120
100
80
60
40
20
0
3
3.5
4
4.5
5
5.5
6
red = kNN, blue = full, green = connectivity

Future Work

- Data sets with non-local similarity functions.
- Combining kNN and smoothening algorithms with Connectivity sparsifier.
- Optimization and enhancement of the algorithm.

References

- X. Bresson, T. Laurent, D. Uminsky, J. Brecht ; “Convergence and Energy Landscape for Cheeger Cut Clustering”
- X. Bresson, T. Laurent, D. Uminsky, J. Brecht ; “An Adaptive Total Variation Algorithm for Computing the Balanced Cut of a Graph”
- X. Bresson, T. Laurent ; “Asymmetric Cheeger cut and application to multi-class unsupervised clustering”
- W. S. Fung, N. Harvey ; “Graph Sparsification by Edge-connectivity and Random Spanning Trees”
- D. Spielman, N. Srivastava ; “Graph Sparsification by Effective Resistances”
- D. Spielman, S. Teng ; “A Local Clustering Algorithm for Massive Graphs and its Application to Nearly-Linear Time Graph Partitioning”
- D. Spielman, S. Teng ; “Spectral Sparsification of Graphs”
- A. Benczúr, D. Karger ; “Randomized approximation schemes for cuts and flows in capacitated graphs
- H. Nagamochi, T. Ibaraki ; “Computing edge-connectivity in multigraphs and capacitated graphs”

Thank you