Graph-based Semi-supervised Local Clustering with Few Labeled Nodes (Supplement)

Zhaiming Shen[1] , **Ming-Jun Lai**[1] and **Sheng Li**[2]
[1]University of Georgia, Athens, GA, USA
[2]University of Virginia, Charlottesville, VA, USA
{zhaiming.shen, mjlai}@uga.edu, shengli@virginia.edu

1 Description of Geometric Data

Three Lines. The three lines are generated by sampling points uniformly at random in the two dimensional x-y plane where the x coordinate is between 0 and 6 and y coordinate equals to 0, 1, and 2 respectively. We draw 1200 points in each line to create three clusters. We then embed each data point into $\mathbb{R}^{100}$ by appending zeros and then adding Gaussian random noise to each coordinate with mean 0 and standard deviation 0.15.

Three Circles. The three circles are generated by sampling points uniformly at random from three concentric circles of radii 1, 2.4 and 3.8 respectively. We draw around 500 points from the smallest circle, around 1200 points from the middle circle and around 1900 points from the largest circle (the numbers are chosen so that the total number of points is 3600). We then embed each data point into $\mathbb{R}^{100}$ by appending zeros and then adding Gaussian random noise to each coordinate with mean 0 and standard deviation 0.15.

Three Moons. The three moons are generated by sampling points uniformly at random from the upper semicircle of radius 1 centered at (0,0), the lower semi-circle of radius 1.5 centered at (1.5, 0.4) and the upper semi-circle of radius 1 centered at (3,0). We draw 1200 points in each semi-circle to create three clusters. We then embed each data point into $\mathbb{R}^{100}$ by appending zeros and then adding Gaussian random noise to each coordinate with mean 0 and standard deviation 0.15.

2 Hyperparameters Setup.

For each cluster to be recovered, we sampled the seed vertices Γ_i uniformly from C_i for all of our implementations. We fix the rejection parameter $R = 0.1$, the random walk depth $t = 3$ and random walk threshold parameter $\epsilon = 0.8$ for all of our implementations. We fix the least squares threshold parameter with $\gamma = 0.2$ for all experiments.

All the numerical experiments are implemented in MATLAB and can be run on a local machine.

3 Image Data Preprocessing.

For our approach, the images data coming from each of the AT&T, OptDigits, MNIST, USPS dataset have to be firstly constructed into an auxiliary graph before feeding into the algorithm. We adopt the following way to build the auxiliary graphs.

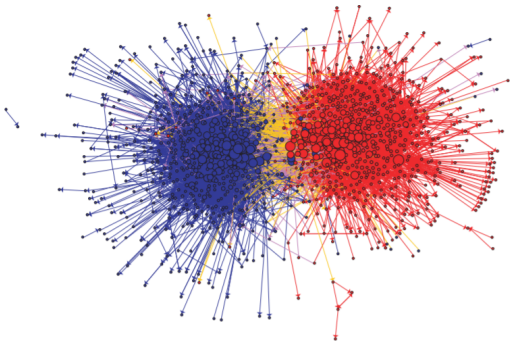

Figure 1: Community structure of political blogs. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it [Adamic and Glance, 2005].

Let $\mathbf{x}_i \in \mathbb{R}^n$ be the vectorization of an image from the original data set, we define the following affinity matrix of the K-NN auxiliary graph based on Gaussian kernel according to [Jacobs *et al.*, 2018] and [Zelnik-Manor and Perona, 2004],.

$$A_{ij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma_i\sigma_j} & \text{if} \quad \mathbf{x}_j \in NN(\mathbf{x}_i, K) \\ 0 & \text{otherwise} \end{cases}$$

The notation $NN(\mathbf{x}_i, K)$ indicates the set of K-nearest neighbours of $\mathbf{x}_i$, and $\sigma_i := \|\mathbf{x}_i - \mathbf{x}_i^{(r)}\|$ where $\mathbf{x}_i^{(r)}$ is the r-th closest point of $\mathbf{x}_i$. Note that the above A_{ij} is not necessary symmetric, so we consider $\tilde{A}_{ij} = A^T A$ for symmetrization. Alternatively, one may also want to consider $\tilde{A} = \max\{A_{ij}, A_{ji}\}$ or $\tilde{A} = (A_{ij} + A_{ji})/2$. We use $\tilde{A}$ as the input adjacency matrix for our algorithms. In our implementation, we choose the local scaling parameters $K = 5$, $r = 3$ for AT&T faces images, and $K = 15$, $r = 10$ for OptDigits, MNIST, USPS.

References

[Adamic and Glance, 2005] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 us election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43, 2005.

[Jacobs *et al.*, 2018] Matt Jacobs, Ekaterina Merkurjev, and Selim Esedoḡlu. Auction dynamics: A volume constrained mbo scheme. *Journal of Computational Physics*, 354:288–310, 2018.

[Zelnik-Manor and Perona, 2004] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, 17, 2004.