Distributed Spectral Clustering on the Coordinator Model

CNMAC 2018
XXXVIII Congresso Nacional de Matemática Aplicada e Computacional
17 a 21 de Setembro de 2018 - IMECC/UNICAMP - Campinas - SP

Fabricio Mendoza Granada, Marcos Villagra

Núcleo de Investigación y Desarrollo Tecnológico
Facultad Politécnica - Univervisad Nacional de Asunción
`fabromendoza95@gmail.com, mvillagra@pol.una.py`

Abstract

Spectral Clustering is a technique that partitions a set of vertices of a given graph $G = (V, E)$ using the spectral properties of G. The optimal partition is found by using the eigenvectors of the first k eigenvalues of the Graph Laplacian matrix. In a real world situation the edges of G may be distributed among different sites. In order to apply spectral clustering in a distributed setting, every site constructs a spectral specifier of its own data graph. This reduces the communication costs by applying the spectral clustering technique leaving optimal results. In this work, we will study the communication complexity in the more extreme case where the vertex set is completely partitioned.

1. Motivation

Machine learning tasks often use unsupervised lerning to discover properties of data. In that context, clustering is one of the most fundamental tasks. Given a sample of points the goal is to collect points that are similar in the same group or cluster. In clustering, the number of clusters that should be found is not known *a priori*. A well known technique to find the number of optimal clusters and to partition the sample set is called *spectral clustering*. Also spectral clustering reduces significantly the dimension of the data to the number of clusters. This offer an advantage because the dimension can be larger than the number of clusters. However, when the data is distributed among different databases around the world makes clustering a difficult task and expensive in terms of the communication costs. Therefore new methods are necessary for this particular situation.

2. Introduction to Spectral Clustering

Given a set of points $\{x_1, x_2, ..., x_n\} \subseteq \mathbb{R}^d$, a graph $G = (V, E)$ is constructed and every point represents a vertex and each edge represents similarity between points. The laplacian matrix is defined as

$$L_G = D_G - A_G, \tag{1}$$

where D_G is the degree weighted matrix and A_G is the adyacent matrix. Spectral clustering uses the conductance ϕ of a subset $S \subseteq V$ to find good clusters

$$\phi(S) = \frac{w(S, V-S)}{\mu(S)}, \tag{2}$$

where $w(S, V-S)$ is the total weight of the edges crossing the cut and $\mu(S) = \sum_{i \in S} d_i$. In order to find an optimal k-partition of V we use the k-way *expansion constant*

$$\rho(k) = \min_{A_1, A_2, ..., A_k} \max_{1 \leq i \leq k} \phi(A_i), \tag{3}$$

which is related to the k-th eigenvalue taken in non-decreasing order by the *Cheeger inequality*:

$$\frac{\lambda_k(\mathcal{L}_G)}{2} \leq \rho(k) \leq O(k^2)\sqrt{\lambda_k(\mathcal{L}_G)}, \tag{4}$$

where $\mathcal{L}_G = I - D_G^{-1/2} A_G D_G^{-1/2}$ is called the *normalized Lapacian* if G.
As an example of spectral clustering works we have the followin graph

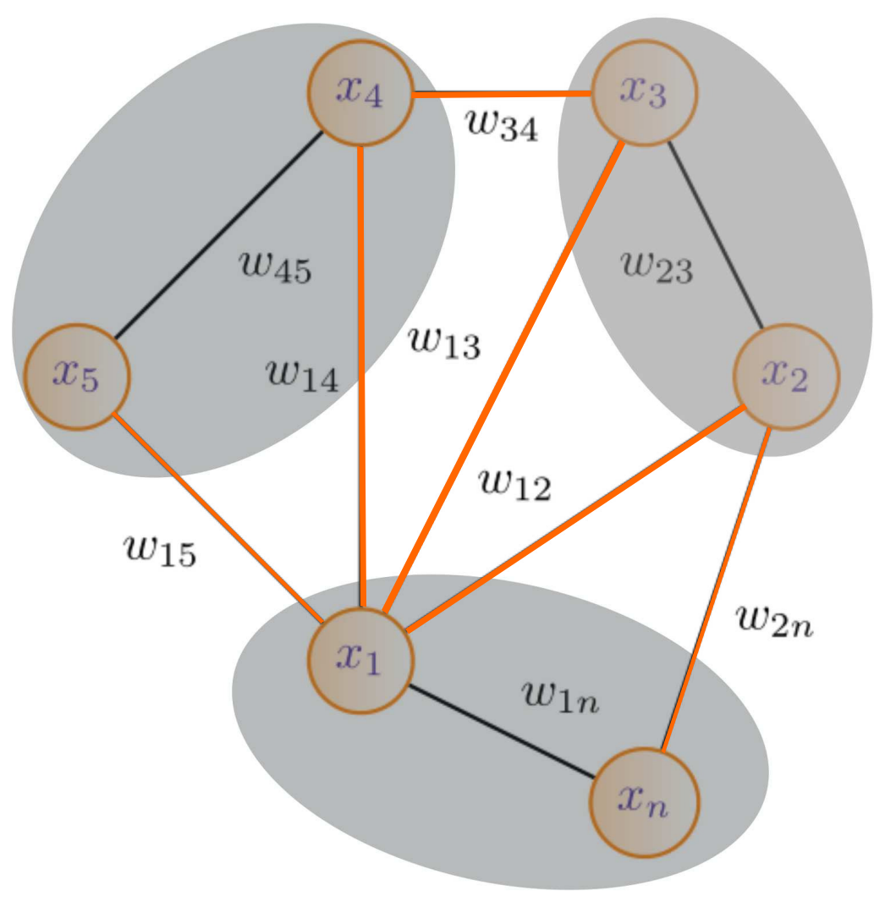

Figure 1: *A graph representing a data set. The shadows represent the partition.*

The following well-known algorithm finds an optimal partition over the vertex set of a given graph G

1. Compute the eigenvectors $f_1, f_2, ..., f_k$ associated with the first k eigenvalues $\lambda_1(\mathcal{L}_G), ..., \lambda_k(\mathcal{L}_G)$ taken in non-decreasen order.
2. Embed every vertex v to a point $\mathbb{R}^k$ through the embedding:

$$F(v) = \frac{1}{\sqrt{d_v}}.(f_1(v), ..., f_k(v))$$

3. run k-means on the embedded points $\{F(v)\}_{v \in V}$, and group the vertices of G into k clusters according to the output of k-means.

Acknowledgements. The authors are supported by Conacyt research grant PINV15-208.

3. Distributed Computing

When two (or more) computers are far from each other and they need to execute a joint task, a communication protocol needs to be defined.

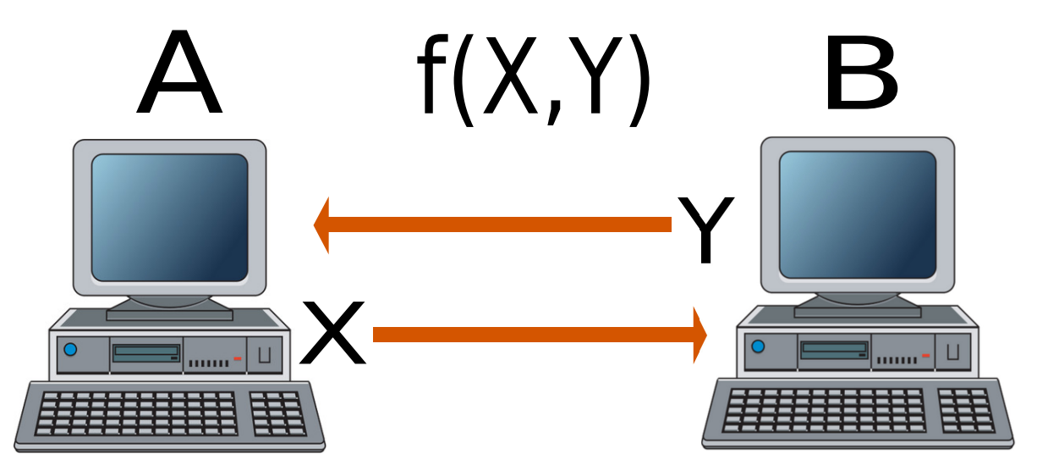

Figure 2: *Two computers trying to accomplish a task (function f) by sending bits to each other.*

In the most general case there are s players who wants to compute some function $f : X_1 \times ... \times X_s \to Z$ where X_i is the set of available inputs for player i. A *protocol* Π is defined as a sequence of binary strings sent by every player.

4. Clustering on the Coorinator model

In the Coordinator model there are s sites where every pair of them can communicate with each other through a special site called a *coordinator*.

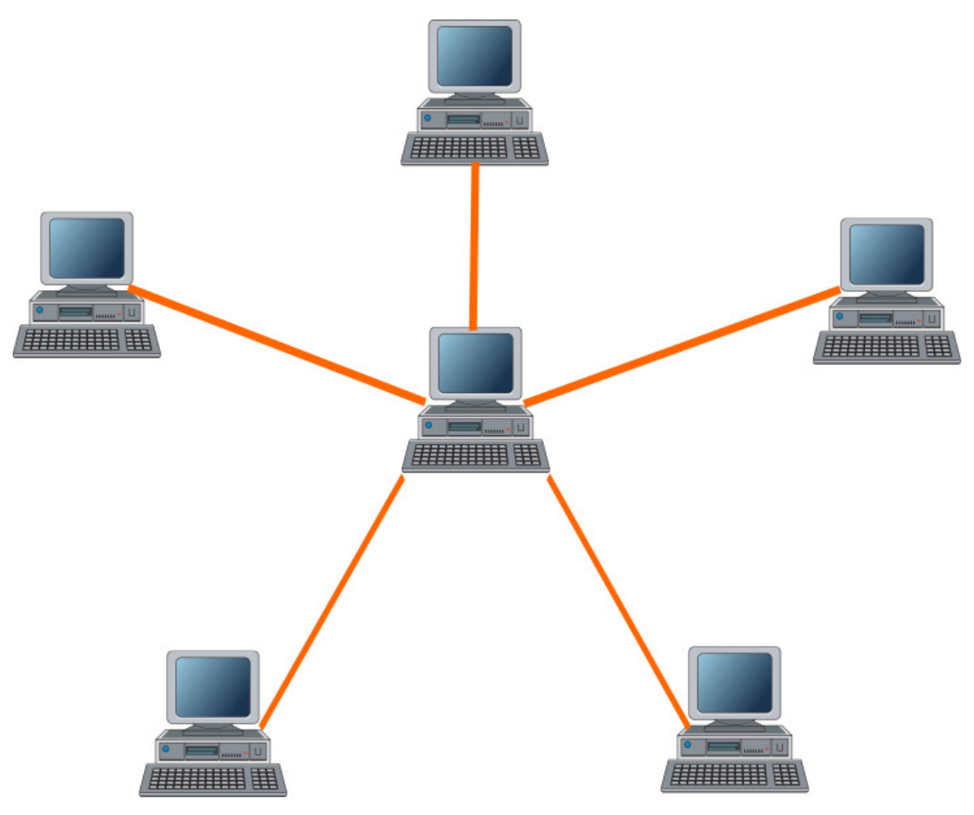

Figure 3: *Example of the Coordinator model with* 5 *sites.*

The idea behind the protocol is that every site sends a subgraph H with edges having new weights such that $(1-\epsilon)x^T L_G x \leq x^T L_H x \leq (1+\epsilon)x^T L_G x$. The graph H is called a *spectral sparsifier* of G.
Given a graph G, where its edge set is partitioned around s sites, the following protocol of [3] runs the spectral clustering on the coordinator model:

1. For every subgraph $G_i = (V, E_i)$ where $1 \leq i \leq s$ computes an spectral sparsifier $H_i = (V, E_i')$,
2. every site i sends its new edge set E_i' to the coordinator,
3. the coordinator constructs the graph $G' = (V, \bigcup_{i=1}^s E_i')$ and runs spectral clustering algorithm on that graph, and
4. the coordinator ouputs the clusters.

5. Proposal

The work of Chen et al. [3] studied the case where each player knows the vertices of the data graph G, but only a subset of the edges. In this work, we will study the communication complexity in the more extreme case where the vertex set is completely partitioned. Let $G = (V, E)$ be graph of the data points. In the coordinator model we have s sites where site i, with $1 \leq i \leq s$, knows a graph $G_i = (V_i, E_i)$, where $\{V_i\}$ and $\{E_i\}$ are partitions of the vertex set and the edge set of G, respectively. Each site can communicate with the coordinator with messages but no site can communicate with another site. Then after a finite number of rounds of communication, the coordinator computes an optimal partition of V. The goal is to find a protocol where an optimal partition of V can be computed using the minimum amount of communication.

References

[1] J. Batson, D. A. Spielman, N. Srivastava, and S. H. Teng, Spectral sparsification of graphs: theory and algorithms. *Communications of the ACM*, 2013. 56(8), 87-94.

[2] M. Braverman, F. Ellen, R. Oshman, T. Pitassi and V. Vaikuntanathan, A tight bound for set disjointness in the message-passing model. *Foundations of Computer Science (FOCS)*, 2013 IEEE 54th Annual Symposium on. IEEE, 2013.

[3] J. Chen, H. Sun, D. Woodruff and Q. Zhang, Communication-optimal distributed clustering.*In Advances in Neural Information Processing Systems*, 2016. (pp. 3727-3735).

[4] Z. Huang, B. Radunovic, M. Vojnovic and Q. Zhang, Communication complexity of approximate matching in distributed graphs. *In LIPIcs-Leibniz International Proceedings in Informatics*, volume 30, 2015. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[5] J. R. Lee, S. O. Gharan and L. Trevisan, Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 2014. 61(6), 37.

CONACYT

PROCIENCIA
PROGRAMA PARAGUAYO PARA EL DESARROLLO DE LA CIENCIA Y TECNOLOGÍA