GRAPH CLUSTERING USING ONE-BIT COMPARISON DATA

Naveed Naimipour and Mojtaba Soltanalian*

Department of Electrical and Computer Engineering
University of Illinois at Chicago
Chicago, Illinois

ABSTRACT

Graph clustering with partial adjacency information has garnered significant interest in data processing and learning applications. In this paper, we propose an extension of a set of non-convex programs, generally referred to as Comparison-Aided 2-Hard and Comparison-Aided 2-Soft Clustering programs, that rely on matrix factorization formulations to process one-bit comparison data. Numerical results confirm not only the usefulness of the proposed methodologies, but also their potential to be used in future data processing applications. To the best of our knowledge, this is the first work in the area of graph clustering that relies only on comparison information to accomplish the clustering task.

***Index Terms*—** One-Bit Matrix Recovery, Non-Convex Methods, Graph Clustering, Soft/Hard Clustering, Matrix Factorization

1. INTRODUCTION

Identifying and categorizing data from pre-existing unlabeled useful information, otherwise known as clustering, has attracted interest due to its ability to structure data and give it *meaning* in a reasonable amount of time. With applications such as social media sites (e.g. Facebook) and online shopping companies (e.g. Amazon), numerous clustering methods have been devised despite the lack of consensus regarding the most efficient algorithm [1-4]. Two such popular graph clustering techniques are (i) the K-means methodology, owing to its simplicity in terms of implementation, as well as, (ii) spectral clustering, thanks to its usefulness in a vast array of applications [5-7]. However, such techniques have had numerous pitfalls (including a large computational cost) that have prevented them from effectively overcoming clustering's biggest obstacles; clustering algorithms need to have the ability to work regardless of missing data or initial knowledge of a cluster [8-9]. For example, although such problems can be handled via K-means methodologies, they will result in similar computational inefficiencies as their previously mentioned conventional counterparts [8-10]. Non-convex tools are proven to be more effective for data analysis in the aforementioned cases [11-12].

Note that research into one-bit sampling techniques has shown promise in many areas such as signal recovery and matrix completion [13-15]. Specifically, allowing the clustered matrix to represent the relationship between values, similar to that of a comparison matrix, has shown promise for matrix recovery [15]. Rather than sampling a signal using conventional techniques, one-bit formulations can capture a significant portion of the information available in the adjacency matrix.

The purpose of this work is to approach the problem of graph clustering through the lens of one-bit non-convex approaches stemming from the matrix factorization formulations in [11,15]. Specifically, we propose two extended non-convex clustering formulations, referred to as *Comparison-Aided 2-Hard Clustering (CA2-Hard Clustering)* and *Comparison-Aided 2-Soft Clustering (CA2-Soft Clustering)* programs. Based on such formulations, we show that the programs correctly identify a high percentage of clusters when used for randomized adjacency matrices. We also discuss a number of interesting observations that were made through the numerical simulations regarding the importance of the distance between clusters.

The rest of this paper is organized as follows: Section 2 introduces the definitions and clustering programs from [11] known as 2-Hard Clustering and 2-Soft Clustering formulations. Section 3 introduces our CA2-Hard Clustering and CA2-Soft Clustering programs stemming from [15]. This includes details regarding the important attributes of our programs. Section 4 discusses numerical results and multiple observations regarding the behavior of the programs. This includes adjustments to the simulation characteristics of the programs that led to better overall performance. Finally, we conclude with a summary of the work and possible future research directions.

2. CLUSTERING PROGRAMS

We begin with two definitions and clustering formulations from [11] that distinguish between hard and soft clustering:

* Corresponding author (e-mail: *nnaimi2@uic.edu*). This work was supported in part by U.S. National Science Foundation Grant CCF-1704401.

Definition 1. *Let $B = \{0, 1\}$. We call $\mathbf{X} \in B^{n\times k}$ a k-clustering matrix if and only if each row of $\mathbf{X}$ has exactly one 1. The subset of k-clustering matrices of $B^{n\times k}$ will be denoted by $\mathcal{H}_{n,k}$.*

Definition 2. *We call $\mathbf{X} \in \mathbb{R}^{n\times k}$ a soft k-clustering matrix if and only if the elements of $\mathbf{X}$ are nonnegative and the sum of the entries at each row is equal to one. The associated subset of $\mathbb{R}^{n\times k}$ will be denoted by $\Omega_{n,k}$.*

The above definitions result in the following non-convex clustering programs (see [11] for details):

(a) 2-Hard Clustering:

$$\min_{\mathbf{X}\in\mathcal{H}_{n,k},\mathbf{Q}\in\mathbb{C}^{n\times k}} \|\mathbf{XQ} - \mathbf{A}\|_F \text{ s.t. } \mathbf{QQ}^T = \mathbf{I}_k, \quad (1)$$

(b) 2-Soft Clustering:

$$\min_{\mathbf{X}\in\Omega_{n,k},\mathbf{Q}\in\mathbb{C}^{n\times k}} \|\mathbf{XQ} - \mathbf{A}\|_F \text{ s.t. } \mathbf{QQ}^T = \mathbf{I}_k. \quad (2)$$

In the above, $\mathbf{A}$ denotes the adjacency matrix of the graph. Herein, we are considering the case where the adjacency matrix is only partially available, not through a subset of its entries, but by capturing information on the comparisons of those entries.

3. CLUSTERING WITH ONE-BIT COMPARISON INFORMATION

We can utilize the above formulations in the case of partial information. Conventional clustering programs assume the entire adjacency matrix is known, but that is not always a valid assumption. The main obstacle stems from the development of formulations for recovering the adjacency matrix when the program is lacking information. We propose formulations which will utilize one-bit strategies and techniques to cluster the data effectively.

The expansion of one-bit techniques for matrix recovery involves the clustering programs having information regarding the relationships of each adjacency matrix component. Specifically, a one-bit comparison matrix can be used to store the comparison of matrix entries relative to each other. In combination with (1)-(2), such information will result in the programs detailed below.

To introduce the CA2-Hard and CA2-Soft Clustering formulations, we must first consider an $n \times n$ adjacency matrix, $\mathbf{A}$ with $[\mathbf{A}]_{i,j} = A_{i,j}$, and rank r. We assume that $A_{i,j}$ is bounded in $[0, \eta]$. Therefore, a one bit observation matrix can be constructed as $\mathbf{W} \in \{-1, 0, 1\}^{d\times n^2}$ with each row representing a comparison and d denoting the number of comparisons made. The comparison data are provided by the vector $\text{sgn}(\mathbf{W}\text{vec}(\mathbf{A}))$. Alternatively, the comparison data can be stored in the following matrix form:

$$\Phi = \text{Diag}(\text{sgn}(\mathbf{W}\text{vec}(\mathbf{A}))). \quad (3)$$

The task of clustering can be accomplished, along with recovering the matrix $\mathbf{A}$ as a side product, with the aid of the comparison information. The clustering programs CA2-Hard and CA2-Soft Clustering are as follows:

(a) CA2-Hard Clustering:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{X}\in\mathcal{H}_{n,k},\mathbf{Q}\in\mathbb{C}^{n\times k}} & \|\mathbf{XQ} - \mathbf{A}\|_F \quad (4)\\ \text{s.t.} \quad & \Phi \cdot \mathbf{W} \cdot \text{vec}(\mathbf{A}) \geq \mathbf{0},\\ & \mathbf{0} \leq \text{vec}(\mathbf{A}) \leq \eta \cdot \mathbb{1}_{n^2},\\ & \mathbf{QQ}^T = \mathbf{I}_k. \end{aligned}$$

(b) CA2-Soft Clustering:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{X}\in\Omega_{n,k},\mathbf{Q}\in\mathbb{C}^{n\times k}} & \|\mathbf{XQ} - \mathbf{A}\|_F \quad (5)\\ \text{s.t.} \quad & \Phi \cdot \mathbf{W} \cdot \text{vec}(\mathbf{A}) \geq \mathbf{0},\\ & \mathbf{0} \leq \text{vec}(\mathbf{A}) \leq \eta \cdot \mathbb{1}_{n^2},\\ & \mathbf{QQ}^T = \mathbf{I}_k. \end{aligned}$$

The above optimization problems can be tackled efficiently in a cyclic manner with respect to $\mathbf{X}$, $\mathbf{Q}$, and $\mathbf{A}$; see [11]. In particular, the optimization with respect to $\mathbf{A}$ is a convex linearly constrained quadratic program.

4. NUMERICAL RESULTS

The numerical results in Fig. 1 and Fig. 2 exhibit the ability of the proposed methods to produce the correct clusters for different sizes of an adjacency matrix. It was observed that that CA2-Soft and CA2-Hard Clustering perform best when there are larger distances between clusters.

Fig. 1 presents the case where 100% of the clusters are correctly identified via CA2-Hard Clustering, while Fig. 2 exemplifies the case where the close proximity of the clusters makes it difficult to cluster perfectly using only the comparison information. Despite the close proximity of the clusters, CA2-Hard Clustering led to 75% of the clusters being correctly identified.

Furthermore, the formulations were tested with completely randomized adjacency matrices, meaning the one-bit comparison matrix parameters were the only captured information of the original adjacency matrix. In other words, there are no matrix values that are known prior to applying our clustering programs. Thus, the presented results are more significant than potential simulations where parts of the matrix are known beforehand.

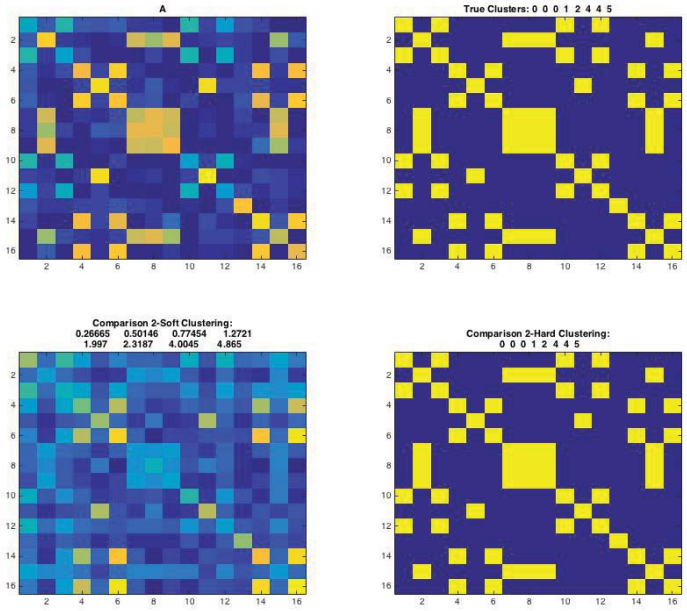

Fig. 1: Graph clustering with partial comparison data via the CA2-Hard and CA2-Soft Clustering formulations ($n = 16$). CA2-Hard Clustering correctly identifies all the clusters.

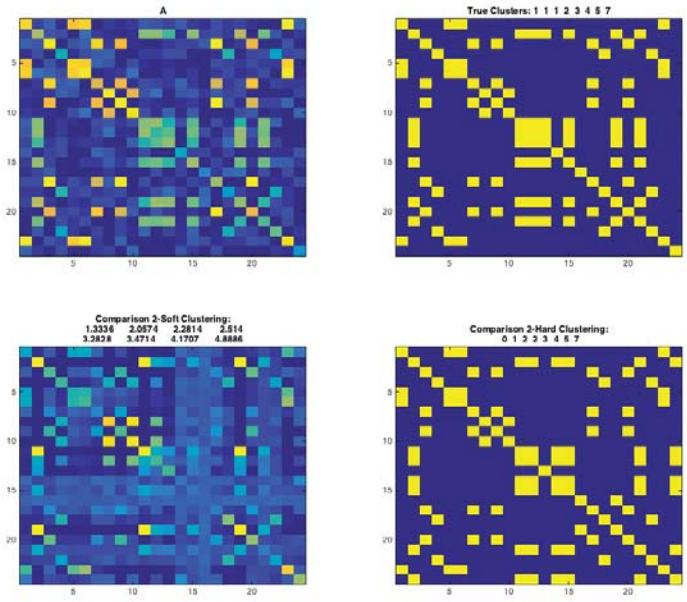

Fig. 2: Graph clustering with partial comparison data via the CA2-Hard and CA2-Soft Clustering formulations ($n = 24$). CA2-Hard Clustering correctly identifies 75% of the clusters.

5. CONCLUSION

The fusion of non-convex clustering formulations with one-bit comparison matrix information results in an alternative methodology to cluster matrices with missing information. Inspired by matrix factorization formulations, we presented a set of non-convex CA2-Hard and CA2-Soft Clustering programs that have shown great potential for partial information clustering applications. The proposed methods were shown to satisfactorily identify the clusters when tested with randomized adjacency matrices. Furthermore, the ability of the proposed methods to handle clustering efficiently in smaller cases shows promise in regards to using similar techniques for efficient big data clustering applications.

6. REFERENCES

[1] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert Tarjan. "Clustering Social Networks," *Algorithms and Models for the Web-Graph, volume 4863 of Lecture Notes in Computer Science*, chapter 5, pages 56-67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[2] H. Wu, K. Liu and C. Trappey, "Understanding customers using Facebook Pages: Data mining users feedback using text analysis," *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Hsinchu, 2014, pp. 346-350.

[3] C. Fry and S. Manna, "Can We Group Similar Amazon Reviews: A Case Study with Different Clustering Algorithms," *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, 2016, pp. 374-377.

[4] G. Linden, B. Smith and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," in *IEEE Internet Computing*, vol. 7, no. 1, pp. 76-80, Jan.-Feb. 2003.

[5] C. Boutsidis, A. Zouzias, M. Mahoney, and P. Drineas, "Randomized dimensionality reduction for K-means clustering," *Comput. Res. Repos.*, 2011.

[6] G. A. Wilkin and X. Huang, "K-Means Clustering Algorithms: Implementation and Comparison," *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, Iowa City, IA, 2007, pp. 133-136.

[7] D. Hamad and P. Biela, "Introduction to spectral clustering," *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, Damascus, 2008, pp. 1-6.

[8] M. Pattanodom, N. Iam-On and T. Boongoen, "Clustering data with the presence of missing values by ensemble approach," *2016 Second Asian Conference on Defence Technology (ACDT)*, Chiang Mai, 2016, pp. 151-156.

[9] C. Gautam and V. Ravi, "Evolving clustering based data imputation," *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]*, Nagercoil, 2014, pp. 1763-1769.

[10] K. Wagstaff, S. Rogers, and S. Schroedl, "Constrained K-means clustering with background knowledge ," *Proc. 8th Int. Conf. Machine Learning*, pp. 577-584, 2001.

[11] N. Naimipour and M. Soltanalian, "Efficient Non-Convex Graph Clustering for Big Data," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 2896-2900.

[12] Poddar, Sunrita and Jacob, Mathews. "Clustering of Data with Missing Entries using Non-convex Fusion Penalties," *arXiv preprint arXiv:1709.01870*, 2017.

[13] S. Khobahi, and M. Soltanalian, "Signal Recovery from 1-Bit Quantized Noisy Samples via Adaptive Thresholding," *2018 IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 2018.

[14] S. A. Bhaskar and A. Javanmard, "1-bit matrix completion under exact low-rank constraint," *2015 49th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, 2015, pp. 1-6.

[15] A. Bose, A. Ameri, M. Klug and M. Soltanalian, "Low-Rank Matrix Recovery from One-Bit Comparison Information," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 4734-4738.