# Multiple Optimal Solutions but Single Search: A Study of the Correlation Clustering Problem

Nejat Arinik, Rosa Figueiredo, Vincent Labatut

▶ **To cite this version:**

Nejat Arinik, Rosa Figueiredo, Vincent Labatut. Multiple Optimal Solutions but Single Search: A Study of the Correlation Clustering Problem. 20ème congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF), Société Française de Recherche Opérationnelle et d'Aide à la Décision, Feb 2019, Le Havre, France. hal-02051683

**HAL Id: hal-02051683**
**https://hal.science/hal-02051683v1**

Submitted on 27 Feb 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Multiple Optimal Solutions but Single Search: A Study of the Correlation Clustering Problem

Nejat Arinik, Rosa Figueiredo, Vincent Labatut

Laboratoire Informatique d'Avignon LIA EA 4128, Avignon France
{name.surname}@univ-avignon.fr

**Keywords** : *Signed Graph, Graph Partitioning, Correlation Clustering, Structural Balance.*

## Introduction

A signed graph, whose links are labeled as positive (+) and negative (−), is considered *structurally balanced* [2] if it can be partitioned into a number of clusters, such that *positive* (*negative*) links are located *inside* (*in-between*) the clusters. Due to the imbalanced nature of real-world networks, various measures have been defined to quantify the amount of imbalance. Such measures are expressed relatively to a graph partition, so processing the graph balance amounts to identifying the partition corresponding to the lowest imbalance measure. A well-known measure among them corresponds to the definition of the *Correlation Clustering problem (CC)*, and it consists in counting the numbers of misplaced links [2].

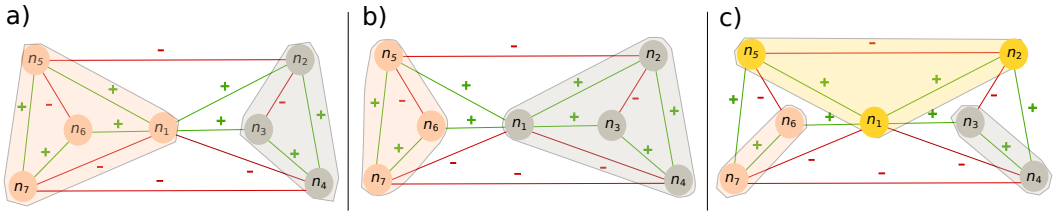


FIG. 1: 3 (out of 22) different optimal CC solutions obtained for the same network: a) $P_a = \{\{n_1,n_5,n_6,n_7\}, \{n_2,n_3,n_4\}\}$; b) $P_b = \{\{n_5,n_6,n_7\}, \{n_1,n_2,n_3,n_4\}\}$; and c) $P_c = \{\{n_1,n_2,n_5\}, \{n_6,n_7\}, \{n_3,n_4\}\}$. Red and green lines represent negative and positive links, respectively. The graph is complete, but for clarity, some negative links between clusters are intentionally omitted.

In the literature, a large number of applied works solve the CC problem to get a better understanding of some studied real-world system (e.g vote analysis [1], etc.). Most of them rely on heuristic approaches, especially when dealing with large networks, due to the complexity of the problem. But a non-negligible number of studies are also concerned with optimality. In any case, the standard approach is to find a single partition, even if other optimal or high scoring solutions possibly exist. Figure 1 illustrates such situation: solving the CC problem for this network of 7 nodes yields 22 distinct optimal partitions. We show only a few of them to highlight how similar these can be. For instance, $P_a$ (left) and $P_b$ (middle) are very similar partition-wise, as they are both bisections differing only in the cluster assignment of $n1$. By comparison, $P_c$ (right) is less similar: it contains an extra cluster obtained by separating an element from each cluster of the previous solutions, in addition to $n1$.

This focus on a single solution raises several questions. First, *many* optimal partitions may coexist. If so, are they all equally relevant to the application problem at hand? If not, it may be necessary to design a more appropriate version of CC, in order to distinguish them, possibly based on some external criteria related to the application context. Second, how different are

these partitions? Application-wise, very similar partitions (e.g. single node differences as between $P_a$ and $P_b$) could be assimilated to the same solution, whereas significantly different partitions might correspond to dramatically different interpretations. Finally, if groups of similar partitions exist, how to identify a representative partition to summarize them?

To the best of our knowledge, no one has ever tried to answer these questions in the literature. Our goal here is to do so, by characterizing the optimal solution space associated with a given network for the CC problem. Put differently, we want to study the CC problem itself, and not some of its existing resolution methods. As this preliminary work, we focus on complete unweighted graphs, and consider only optimal solutions.

## Method

We propose a 4-stepped method. The first step is to obtain all optimal partitions. For this purpose, we use an integer programming-based method able to solve the CC problem exactly. The second step consists in computing the similarity between the partitions through the measure *Variation of Information (VI)*[1], in order to perform the third step, which is a cluster analysis. This leads to a set of clusters (or only one), each one gathering similar partitions. The fourth step is to process what we call the *representative partition* for each cluster, which is supposed to be the most similar over all the partitions belonging to the cluster.

## Preliminary Experimental Results

We investigate into the space of optimal solutions based on synthetic signed networks, which are complete and unweighted. Network generation relies on three parameters: $n$ (graph size), $k$ (number of cluster) and $q$ (proportion of misplaced links). Once perfectly balanced networks are created with $n$ and $k$, the sign of links are inverted based on $q$. This process is repeated 10 times for replication. Figure 2 shows the number of optimal solutions by varying $n$ and $q$, while fixing $k = 4$. Interestingly, a large number of optimal solutions (1093) are obtained for $n = 28$, $q = 0.3$. However, there is no direct link between the number of optimal solutions and $n$, since increasing $n$ can also increase the number of cluster assignment combinations.

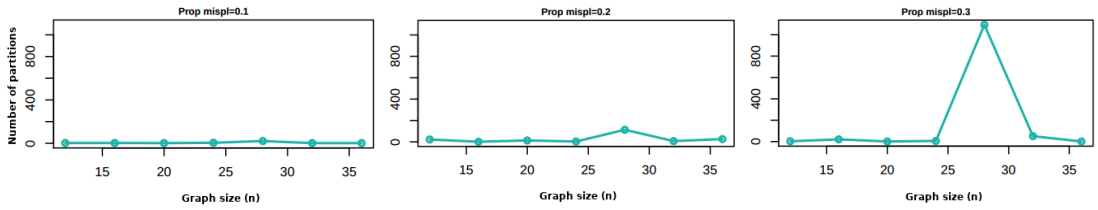


FIG. 2: Evolution of the number of optimal partitions found in synthetic signed networks generated with parameters $n \in \{12, 16, 20, 24, 28, 32, 36\}$, $q \in \{0.1, 0.2, 0.3\}$ and $k$=4.

## References

[1] N. Arinik, R. Figueiredo, and V. Labatut. Signed graph analysis for the interpretation of voting behavior. In *International Conference on Knowledge Technologies and Data-driven Business - International Workshop on Social Network Analysis and Digital Humanities*, Graz, AT, 2017.

[2] J. A. Davis. Clustering and structural balance in graphs. *Human Relations*, 20(2):181–187, 1967.

[1] Selecting an appropriate measure to compare partitions is currently dealing in an ongoing research