A GPU-accelerated local search algorithm for the Correlation Clustering problem

Mario Levorato, Lúcia Drummond, Yuri Y. Frota, Rosa Figueiredo

A GPU-accelerated local search algorithm for the Correlation Clustering problem

Mario Levorato
Department of Computer Science, Fluminense Federal University
24210-240 Niterói, RJ – Brasil
mlevorato@ic.uff.br

Lúcia Drummond and Yuri Frota
Department of Computer Science, Fluminense Federal University
24210-240 Niterói, RJ – Brasil
{lucia, yuri}@ic.uff.br

Rosa Figueiredo
Laboratoire d'Informatique d'Avignon, University of Avignon
84911 Avignon, France
rosa.figueiredo@univ-avignon.fr

RESUMO

A solução ótima para o problema de Correlação de Clusters (*Correlation Clustering* ou CC) pode ser utilizada como medida do nível de equilíbrio em redes sociais de sinais, onde interações positivas (amizade) e negativas (antagonismo) estão presentes. Metaheurísticas têm sido utilizadas com sucesso para resolver não apenas este, como também outros problemas difícies de otimização combinatória, por serem capazes de fornecer soluções sub-ótimas em um tempo razoável. Este trabalho propõe uma implementação alternativa de busca local baseada em GPGPUs, a qual pode ser utilizada em conjunto com as metaheurísticas GRASP e ILS existentes para o problema CC. Esta nova abordagem supera, em tempo de execução, o procedimento de busca local até então aplicado, com a mesma qualidade de solução.

PALAVRAS CHAVE. CUDA, GPGPU, VND, GRASP, ILS, Correlação de Clusters.

Área Principal: MH - Metaheurísticas

ABSTRACT

The solution of the Correlation Clustering (CC) problem can be used as a criterion to measure the amount of balance in signed social networks, where positive (friendly) and negative (antagonistic) interactions take place. Metaheuristics have been used successfully for solving not only this problem, as well as other hard combinatorial optimization problems, since they can provide sub-optimal solutions in a reasonable time. In this work, we present an alternative local search implementation based on GPGPUs, which can be used with existing GRASP and ILS metaheuristics for the CC problem. This new approach outperforms the existing local search procedure in execution time, with the same solution quality.

KEYWORDS. CUDA, GPGPU, VND, GRASP, ILS, Correlation Clustering.

Main Area: MH - Metaheuristics

1. Introduction

Structural (or social) balance is considered a fundamental social process. It has been used to explain how the feelings, attitudes and beliefs, which the social actors have towards each other, can promote the formation of stable (but not necessarily conflict-free) social groups. The balance of a social system tends to follow the human tendency to preserve a cognitive consistency of hostility and friendship. The principle is simple: "my friend's friend is my friend, my friend's enemy is my enemy, my enemy's friend is my enemy, my enemy's enemy is my friend" (Heider, 1946). Absence of balance creates a kind of tension in the group members' minds that can eventually lead to changes in their opinions. Once balance is achieved, it tends to be stable, since no cognitive dissonance could change the state (Hummon and Doreian, 2003).

Determining the structural balance of a signed social network has been a key aspect in the study of the structure and origin of tensions and conflicts in a network of individuals whose mutual relationships are characterizable in terms of friendship and hostility. Structural balance theory was first formulated by Heider (1946) with the purpose of describing sentiment relations between people pertaining to a same social group (like/dislike, love/hate, trust/distrust). Signed graphs were then introduced by Cartwright and Harary (1956), who formalized Heider's theory stating that a balanced social group could be partitioned into two mutually hostile subgroups each having internal solidarity. In the last decades, signed graphs have shown to be a very attractive discrete structure for social network researchers (Doreian and Mrvar, 1996; Inohara, 1998; Yang et al., 2007; Abell and Ludwig, 2009; Doreian and Mrvar, 2009; Facchetti et al., 2011). Different criteria and solution approaches have been used in the literature so as to quantify and evaluate balance in a signed social network (Doreian and Mrvar, 2009; Leskovec et al., 2010; Facchetti et al., 2011; Srinivasan, 2011).

Clustering is the action of partitioning individual elements into groups based on their similarity. Clustering problems defined on signed graphs arise in many scientific areas (Bansal et al., 2002; Gülpinar et al., 2004; DasGupta et al., 2007; Traag and Bruggeman, 2009; Huffner et al., 2010; Macon et al., 2012; Figueiredo and Frota, 2014). The common element among these applications is the collaborative *vs.* conflicting environment in which they are defined. The solution of clustering problems defined on signed graphs can be used as a criteria to measure the degree of balance in social networks (Doreian and Mrvar, 1996, 2009; Figueiredo and Moura, 2013). By considering the original definition (Heider, 1946) of structural balance, the optimal solution of the very known Correlation Clustering (CC) Problem (Bansal et al., 2002) arises as a measure for the degree of balance in a social network. Alternative measures to the structural balance and the clustering problems associated with them have also been recently discussed (Doreian and Mrvar, 2009; Figueiredo and Moura, 2013).

From a practical point of view, in solving the clustering problem treated in this paper, heuristic approaches are primarily of interest, since large social networks may have to be analyzed (Kunegis et al., 2009; Leskovec et al., 2010; Facchetti et al., 2011). For example, online networks with two opposite kinds of relationships are nowadays very common. Slashdot, a technology-related news website, includes a feature which allows users to tag each other as friends or foes, thus allowing users to rate other users negatively. On online review websites such as Epinions users can either like or dislike other people's reviews. This behavior can be modeled as a signed network, where edge weights can be either greater or less than 0, representing positive or negative relationships respectively. The definition of a measure to represent the imbalance of a social network adds itself a degree of approximation to the task of evaluating balance in a social network. Thus, it is imperative that the clustering problem associated with this measure be solved efficiently.

To our knowledge, there are three metaheuristic approaches applied to the CC problem. Zhang et al. (2008) proposes genetic algorithms to the CC problem, with an application to document clustering. This strategy was impossible to reproduce though, for the absence of information about how the genetic operators are applied. Drummond et al. (2013) presents a Greedy Random-

ized Adaptive Search Procedure (GRASP) (Feo and Resende, 1995) implementation capable of efficiently solving the problem in networks of up to 800 vertices. Later, based on this work, Levorato et al. (2014) introduced an Iterated Local Search (ILS) (Lourenço et al., 2003) metaheuristic for the CC problem, which outperformed, in processing time, the GRASP metaheuristic proposed earlier, with similar or improved solution quality. By observing the great amount of time spent on the processing of larger graphs, we saw an opportunity to extend the aforementioned GRASP and ILS algorithms with a new implementation of local search that can solve the problem faster.

In this work, we present a parallel local search procedure for the CC problem, accelerated by General Purpose Graphics Processing Units (GPGPUs). Then, by applying the proposed local seach in the GRASP and ILS metaheuristics, we show the improvements over the existing sequential local search procedure. The paper is organized as follows. Section 2 presents the Correlation Clustering problem, including a mathematical formulation and a literature review of it. Section 3 describes the parallel local search algorithm for the CC problem that runs on the GPU, while Section 4 lists the experimental results of it as well as a comparison with other available solution approaches. Finally, Section 5 presents our concluding remarks.

2. The CC problem

Correlation Clustering (Bansal et al., 2002) is a clustering technique motivated by the problem of document clustering, in which given a large corpus of documents such as web pages, one wants to find their optimal partition into clusters. The problem consists of minimizing the number of unrelated pairs that are clustered together, plus the number of related pairs that are separate. In this section, we formally describe the CC problem and present a mathematical formulation of it, followed by a literature review.

2.1. Mathematical Formulation

Let $G = (V, E)$ be an undirected graph where V is the set of n vertices and E is the set of edges. In this text, a signed graph is allowed to have parallel edges but no loops. Also, we assume that parallel edges always have opposite signs. For a vertex set $S \subseteq V$, let $E[S] = \{(i,j) \in E \mid i,j \in S\}$ denote the *subset of edges induced by* S. For two vertex sets $S, W \subseteq V$, let $E[S : W] = \{(i,j) \in E \mid i \in S, j \in W\}$. One observes that, by definition, $E[S : S] = E[S]$. Consider a function $s : E \to \{+, -\}$ that assigns a sign to each edge in E. An undirected graph G together with a function s is called a *signed graph*. An edge $e \in E$ is called *negative* if $s(e) = -$ and *positive* if $s(e) = +$. Let E^- and E^+ denote, respectively, the set of negative and positive edges in a signed graph.

A *partition* of V is a division of V into non-overlapping and non-empty subsets. Consider a partition $P = \{S_1,S_2,\ldots,S_l\}$ of V. The *cut edges* and the *uncut edges* related with this partition are defined, respectively, as the edges in sets
$\cup_{1\leq i<j\leq l} E[S_i : S_j]$ and $\cup_{1\leq i\leq l} E[S_i]$. Let w_e be a nonnegative edge weight associated with edge $e \in E$. Also, for $1 \leq i, j \leq l$, let

$$\Omega^+(S_i,S_j) = \sum_{e\in E^+\cap E[S_i:S_j]} w_e \quad \text{and} \quad \Omega^-(S_i,S_j) = \sum_{e\in E^-\cap E[S_i:S_j]} w_e.$$

The *imbalance* $I(P)$ of a partition P is defined as the total weight of negative uncut edges and positive cut edges, i.e.,

$$I(P) = \sum_{1\leq i\leq l} \Omega^-(S_i,S_i) + \sum_{1\leq i<j\leq l} \Omega^+(S_i,S_j). \tag{1}$$

Likewise, the *balance* $B(P)$ of a partition P can be defined as the total weight of positive uncut edges and negative cut edges. Clearly, $B(P) + I(P) = \sum_{e\in E} w_e$. That being said, we are ready to give a formal definition to the CC problem.

Problem 2.1 (CC problem) *Let $G = (V,E,s)$ be a signed graph and w_e be a nonnegative edge weight associated with each edge $e \in E$. The correlation clustering problem is the problem of finding a partition P of V such that the imbalance $I(P)$ is minimized or, equivalently, the balance $B(P)$ is maximized.*

Observe that the given definition comprises a weighted version of the problem. To obtain a non-weighted version, it suffices to make $w_e = 1$, for each $e \in E$.

The classical mathematical formulation for the CC problem is an integer linear programming (ILP) model proposed to uncapacitated clustering problems (Mehrotra and Trick, 1998). In this formulation a binary decision variable x_{ij} is assigned to each pair of vertices $i,j \in V$, $i \neq j$, and defined as follows: $x_{ij} = 0$ if i and j are in a common set; $x_{ij} = 1$ otherwise. The model minimizes the total imbalance.

$$\text{minimize} \sum_{(i,j)\in E^-} w_{ij}(1 - x_{ij}) + \sum_{(i,j)\in E^+} w_{ij}x_{ij} \tag{2}$$

$$\text{subject to } x_{ip} + x_{pj} \geq x_{ij}, \qquad \forall\, i,p,j \in V, \tag{3}$$

$$x_{ij} = x_{ji}, \qquad \forall\, i,j \in V, \tag{4}$$

$$x_{ij} \in \{0,1\}, \qquad \forall\, i,j \in V. \tag{5}$$

The triangle inequalities (3) say that if i and p are in a same cluster as well as p and j, then vertices i and j are also in a same cluster. Constraint (4) written to $i,j \in V$ establishes that variables x_{ij} and x_{ji} assume always the same value in this formulation. Constraints (5) impose binary restrictions to the variables while the objective function (2) minimizes the total imbalance defined by equation (1). Even though this formulation is polynomial-sized, having $n(n-1)$ variables and n^3+n^2 constraints, notice that, according to constraints (4), half of the variables can be eliminated, which reduces both the number of variables and constraints of the formulation.

A set partitioning formulation (Mehrotra and Trick, 1998) is proposed in the literature to uncapacitated clustering problems and could also be used in the solution of the CC problem. As we can expect, these two formulations are not appropriate solution approaches when time limit is a constraint in the solution process. The authors in Figueiredo and Moura (2013) report that the classical formulation starts to fail (time limit set to 1h) with random instances of 40 vertices and negative density equal to 0.5.

2.2. Literature Review

To the best of our knowledge, the CC problem, as defined in the previous section, was addressed for the first time in Doreian and Mrvar (1996) (not under this name) where its heuristic solution was used as a criteria for analyzing structural balance in social networks. The heuristic approach proposed by the authors is a simple greedy neighborhood search procedure that assumes a prior knowledge of the number of clusters in the solution. This heuristic is implemented in software Pajek (Batagelj and Mrvar, 2008). Lately, motivated by the solution of a document clustering problem, the unweighted version of the CC problem was formalized in Bansal et al. (2002). The weighted version of the problem was addressed in Demaine et al. (2006). The CC problem has been largely investigated from the point of view of constant factor approximation algorithms and has been applied in the solution of many applications, including portfolio analysis in risk management (Huffner et al., 2010), biological systems (DasGupta et al., 2007; Huffner et al., 2010), efficient document classification (Bansal et al., 2002), detection of embedded matrix structures (Gülpinar et al., 2004) and community structure (Traag and Bruggeman, 2009; Macon et al., 2012).

A comparison of several heuristic strategies (greedy and local search methods) for the problem is presented in Elsner and Schudy (2009) and applied to document clustering and natural language processing (instances of $n = 1000$), to which ILP does not scale. In this context, the authors' recommended strategy for solving the CC Problem is a greedy algorithm called $VOTE/BOEM$, which can quickly achieve good objective values with tight bounds.

In Yang et al. (2007), the CC problem is called *community mining* and an agent-based heuristic is proposed to its solution. As far as we know, there are three metaheuristic approaches applied to the CC problem. A solution based on genetic algorithms has been proposed in Zhang et al. (2008) for the CC problem and applied to document clustering, but unfortunately there is no explanation about how the genetic operators are applied, making it difficult to understand and reproduce the proposed algorithm. Recently, Drummond et al. (2013) presented a GRASP (Feo and Resende, 1995) implementation that provides an efficient solution to the CC problem in networks of up to 8000 vertices. Later on, Levorato et al. (2014) introduced an ILS (Lourenço et al., 2003) metaheuristic for the CC problem, which outperformed, in processing time, the GRASP metaheuristic proposed earlier, with similar or improved solution quality.

3. Parallelizing local search for the CC problem in the GPU

Our work started with an analysis of two existing metaheuristics for the CC problem. Drummond et al. (2013) report the results obtained with sequential and parallel GRASP procedures. The algorithm was implemented in C++ with MPI for message passing (Gropp et al., 1999). Then, based on this work, Levorato et al. (2014) later introduced an ILS metaheuristic for the CC problem, which outperformed, in processing time, the GRASP algorithm proposed earlier, with similar or improved solution quality.

By observing the great amount of time spent on the local search phase of the aforementioned algorithms (Figure 1), we saw an opportunity to improve their performance by extending both of them with a new implementation of local search, which is capable of solving the problem faster, without altering the behavior of the metaheuristic. In this section we present a local search procedure for the CC problem that uses the parallelism offered by GPGPUs.

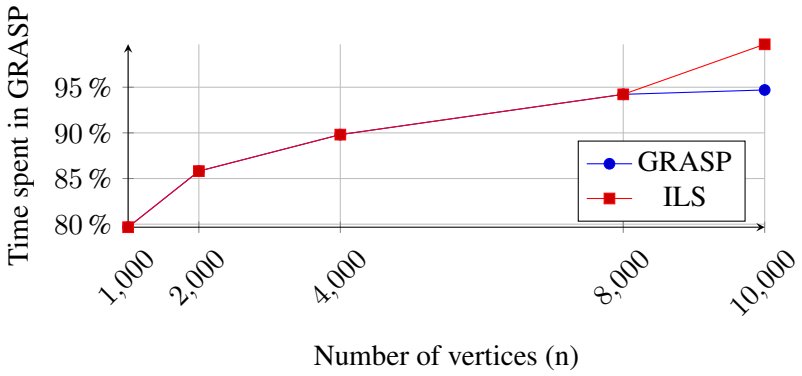

Figure 1: Time spent by GRASP and ILS algorithms on sequential 1-opt local search on Slashdot-based signed graphs.

3.1. Using General Purpose GPUs to solve optimization problems

The use of Graphics Processing Units (GPUs) has been extended to a wide range of application domains (e.g. computational science) thanks to the publication of the CUDA (Compute Unified Device Architecture) development toolkit (NVIDIA, 2015), which allows GPU programming in C-like language. When used as general-purpose computing devices, GPUs can efficiently accelerate many non-graphics programs, especially vector-and matrix-based codes that exhibit lots of parallelism with low synchronization requirements. Because their hardware is primarily designed to perform complex computations on blocks of pixels at high speed and with wide parallelism, GPU architectures differ substantially from conventional CPU hardware. Therefore, writing efficient programs to solve combinatorial optimization problems on GPUs is not a straightforward task and requires a huge effort not only at design but also at implementation level. Indeed, several challenges mainly related to the hierarchical memory management have to be dealt with. The major issues consist of efficient distribution of data processing between CPU and GPU, thread synchronization,

optimization of data transfer between the different memories, as well as the capacity constraints of these memories (Van Luong et al., 2013).

Whenever parallel algorithms are applied to solve optimization problems, it is worth noticing that, in general, for distributed architectures (like MPI), the global performance in metaheuristics is limited by high communication latencies. However, in GPU architectures, performance is bounded by memory access latencies. This being said, several works have already demonstrated the potential speedups when using GPUs to accelerate metaheuristics. For example, GRASP, ILS and EA algorithms have been already adapted to use local search procedures implemented in GPGPU. Table 1 lists some results available in the literature.

Author	**Title**	**Speedup**
Fujimoto and Tsutsui (2011)	A highly-parallel TSP solver for a GPU computing platform	Up to x24.2
Rocki and Suda (2012)	Accelerating 2-opt and 3-opt local search using GPU in the travelling salesman problem	From x3 to x26 speedup compared to parallel CPU w/ 32 cores in ILS for TSP
Van Luong et al. (2013)	GPU computing for parallel local search metaheuristic algorithms	From x0.5 up to x73.3 in local search metaheuristics in GPU
Krüger et al. (2010)	Generic local search (memetic) algorithm on a single GPGPU chip	Between x70 and x120
Pena et al.	Local search for the observer positioning over terrain problem	?
Santos et al.	Parallel GRASP for the p-median problem	Between x1.14 and x13.89

Table 1: Speedups obtained when using GPGPUs to accelerate metaheuristics.

3.2. GPGPU architecture and the CUDA programming model

CUDA has made possible the development of algorithms to solve time-consuming problems using the large number of parallel multiprocessors as well as the high memory bandwidth provided by GPUs. To accomplish high-performance computing, it is necessary to develop parallel algorithms that are partially or totally executed on the GPU. The CUDA-enabled graphics cards are composed of multiple processors, more specifically, Single Instruction Multiple Data (SIMD) processors called Stream Multiprocessors (SMs), which allow the execution of multiple parallel threads. Thus, GPU processors can efficiently execute instructions involving operations with data parallelism, when the same operation is applied to different data.

Depending on the algorithm, GPUs can provide greater processing power than CPUs because they are specialized in performing parallel tasks involving many calculations. On the other hand, the CPUs are designed to perform tasks involving execution flow control and data cache. The physical difference between both architectures can be visualized in Figure 2: GPUs dedicate most of their area for processing units (in green), while CPUs dedicate most of their area for execution control and data cache (in yellow and orange, respectively).

A CUDA application consists in code that is executed on CPU and functions (called kernels) that are executed on GPU. The GPU is able to do parallel processing by creating threads such that each thread may execute the kernel operations on different data. This way, the GPU is used as a coprocessor to perform certain tasks more efficiently than the CPU.

In CUDA, the processing units (cores) are grouped to share a single instruction unit, so that threads mapped on these cores execute the same instruction each cycle, but on different data. Each logical group of threads sharing instructions is called a warp. Moreover, threads belonging to different warps can execute different instructions on the same cores, but in a different time slot. In practice, CUDA cores are time-shared between warps, and a group of threads in a warp performs as a SIMD unit.

That said, modern GPU architectures relax SIMD constraints by allowing threads in a given warp to execute different instructions. However, these varying instructions cannot be executed concurrently, since each SIMD unit must execute the same instruction on all cores. This way, the instructions are serialized in time, which severely degrades performance. This advanced feature is called SIMT (Single Instruction Multiple Threads) and provides increased programming flexibility by deviating from SIMD at the cost of performance. Threads executing different instructions in a warp are said to diverge; if-then-else statements and loop-termination conditions are common sources of divergence.

Another major concern about CUDA implementation which greatly impacts performance is memory access. Bottlenecks can appear not only during data transfer between host (CPU) and device (GPU) memory, but also during memory access on the device; namely, data locality is very important. Depending on the accessed addresses, concurrent memory requests from multiple threads from a warp can exhibit three possible behaviors:

- Requests targeting the same address are merged to be one unless they are atomic operations. In the case of write operations, the value actually written to memory is nondeterministically chosen from among merged requests;
- Requests exhibiting spatial locality are maximally coalesced. For example, accesses to addresses i and $i + 1$ are served by a single memory fetch, as long as they are aligned;
- All other memory requests (including atomic ones) are serialized in a nondeterministic order.

This last behavior, often called the scattering access pattern, greatly reduces memory throughput, since each memory request utilizes only a few bytes from each memory fetch.

The CUDA programming model includes the notion of shared memory and thread blocks, a reflection of the underlying hardware architecture as shown in Figure 2. All threads in a thread block can access the same shared memory, which provides lower latency and higher bandwidth access than global GPU memory but is limited in size. Threads in a thread block may also communicate with each other via this shared memory.

3.2.1. Modifying the search algorithm to run in the GPU

Our approach to parallelize the local search procedure followed the Iteration-level Parallel Model (Van Luong et al., 2013). As can be seen on Figure 3, the evaluation of the neighborhood is made in parallel. At the beginning of each iteration, the master thread, that runs on the CPU, duplicates the current solution, which is made available to all threads of the GPU. Each of them evaluates a specific movement in the neighborhood of candidates, and the results are returned back to the master.

At this point, it is important to list some optimizations in the Correlation Clustering local search algorithm that have been applied for the code to run efficiently in the GPU. First of all, the graph had to be stored in Compressed Sparse Row format (Figure 4), in order to save space and avoid unnecessary data transfers between host (CPU) and device (GPU) memory. This representation consists of two arrays: column indices and row offsets. The column indices array is a concatenation of each vertex's adjacency list into an array of m elements. The row offsets array is an $n + 1$ element array that points at where each vertex's adjacency list begins and ends within the column indices array. For example, the adjacency list of vertex v starts at $C[R[v]]$ and

Memory type	Access latency	Size
Global	Medium	Big
Registers	Very fast	Very small
Local	Medium	Medium
Shared	Fast	Small
Constant	Fast (cached)	Medium
Texture	Fast (cached)	Medium

- Highly parallel multi-threaded many core
- High memory bandwidth compared to CPU
- Different levels of memory (different latencies)

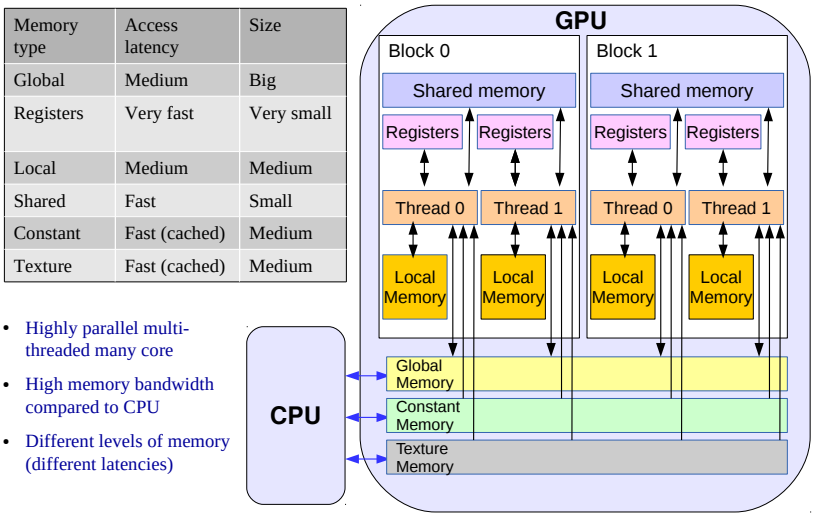

Figure 2: GPU Memory Hierarchy

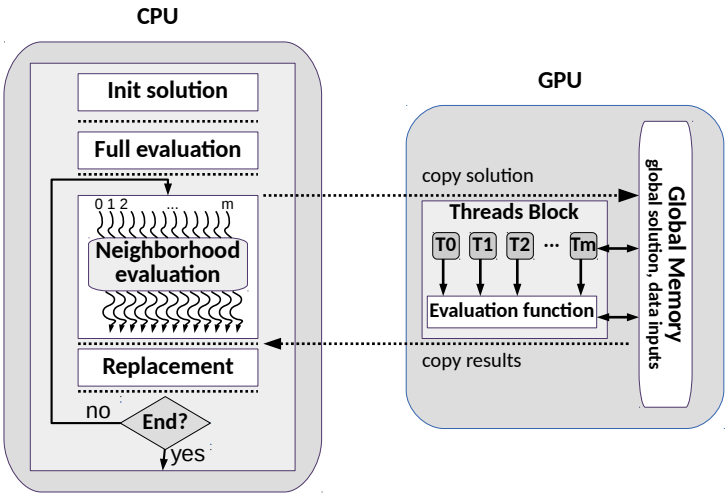

Figure 3: CUDA local search parallelization scheme

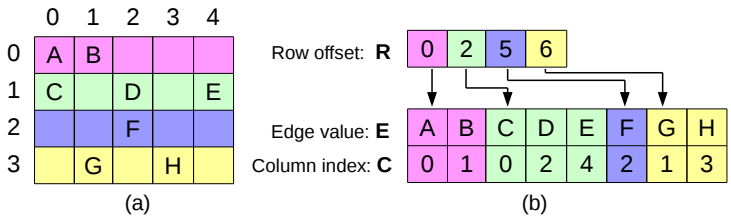

Figure 4: Using the Compressed Sparse Row (CSR) format to store graph's adjacency matrix.

ends at $C[R[v+1]-1]$ (inclusively) and the edge values are stored in elements from $E[R[v]]$ to $E[R[v+1]-1]$.

Also, since it is impossible to store the graph in shared memory (graph is big, shared memory is too small), the graph was copied to the (slower) GPU global memory. It was then used to calculate matrices that contain the sum of edge weights between vertex i and every cluster k in

the current solution. As we are processing a signed graph, there are 2 sum matrices: one for positive edges and the other for negative edges, following the layout depicted in Figure 5. These matrices, also stored in GPU global memory, contain all the information needed to evaluate the imbalance of a new clustering configuration, without the need to traverse the graph, thus saving GPU memory accesses and execution time.

Our parallel approach, although fast, is limited by the GPU's shared memory size. Access to the shared memory is very fast, therefore data stored in shared memory can be accessed with very low latency. However, due to the limitations of GPU architecture, avilable shared memory is limited to 48kB per MultiProcessor (NVIDIA, 2015)[Appendix G]. Since our local search algorithm stores the current clustering solution array (to which cluster number a vertex belongs) in shared memory, it is unable to solve graph instances larger than 12,888 vertices (48 kB/ 4 Bytes [int type]).

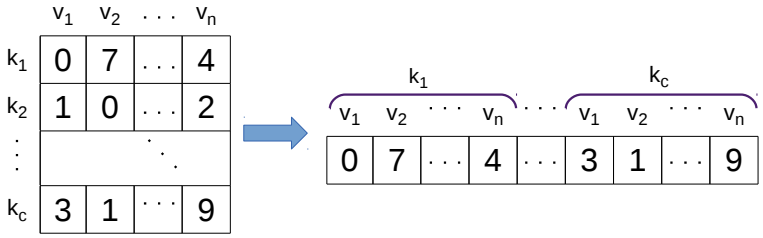

Figure 5: Layout of the matrices that store the sum of positive and negative edge weights between vertex v_i and each cluster k_c ($positiveSumArray$ and $negativeSumArray$, respectively).

3.3. CUDA local search kernel implementation

Algorithm 1: $1OptLocalSearchKernel$

```
1  Input: positiveSumArray, negativeSumArray, currentImbalance, clusterArray, number of clusters (c)
2  Output: destImbArray
3  idx = blockIdx.x * blockDim.x + threadIdx.x;
4  i = idx  mod n;            → Vertex i is in cluster k1
5  k2 = idx div n;            → Vertex i is being moved to cluster k2
6  if (i ≤ n and k2 ≤ c + 1) {
7     k1 = clusterArray[ i ];      → obtains the cluster number of vertex i
8     /* calculates only the difference in positive and negative imbalance */
9     positiveSum = - positiveSumArray[ i + k2 * n ] + positiveSumArray[ i + k1 * n ];
10    negativeSum = - negativeSumArray[ i + k1 * n ] + negativeSumArray[ i + k2 * n ];
11    destImbArray[ idx ] = currentImbalance + positiveSum + negativeSum;
12 }
```

Algorithm 1 presents the kernel pseudocode for CUDA CC $1-opt$ local search kernel and Figure 6 summarizes the work executed. Each thread running in the GPU (uniquely identified by idx) is responsible for calculating the delta of imbalance caused by moving a specific vertex i to a different cluster, for example, in the range k_1 to k_c. Afterwards, another kernel performs a reduction of the results, also in parallel, returning the best move for this specific local search.

Finally, whenever a vertex move is applied due to an improvement in imbalance, a third CUDA kernel is invoked to update the clustering configuration and the vertex-cluster edge-weight-sum arrays ($positiveSumArray$ and $negativeSumArray$) after a change in the clustering. This update is a necessary step to allow the execution of the variable neighborhood descent procedure, that is, invoking the $1-opt$ local search procedure again (new local search iteration), as long as the obtained clustering solution brings an improvement in imbalance.

Our initial implementation approach consisted of running only the local search algorithm ($1OptLocalSearchKernel$) inside the GPU, keeping the reduction of best result and update of the

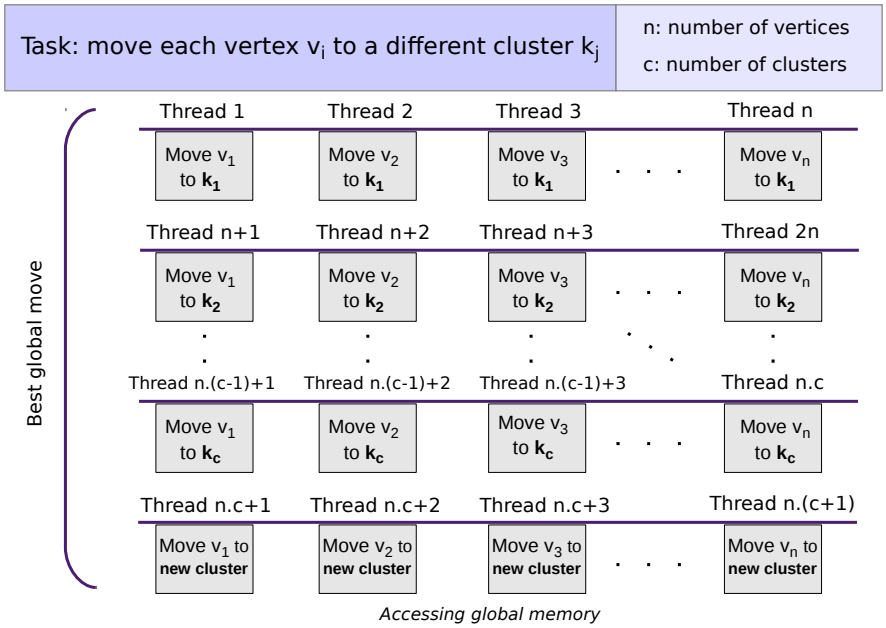

Figure 6: GPU thread work representation for 1-opt local search. Each thread idx is responsible for moving vertex i to a different cluster, from k_1 to k_c, and to a new cluster $(k_c + 1)$.

auxiliary matrices in the CPU. However, poor performance was obtained due to the overhead of memory transfers between host and device memory: 95% of time spent by local search procedure did not involve computation, only memory copies between host (CPU) and device (GPU) ($memcpy$ operation). In the final version of the algorithm, solution reduction and updates of data structures between local search iterations are also performed in the GPU, which led to the computational results that shall be presented in the next section.

4. Experimental results

The algorithms described in the previous section were implemented in ANSI C++ and CUDA. All experiments were performed (with exclusive access) on a workstation with an Intel Core i7 QuadCore processor @2.66GHz and 32Gb of RAM under Linux Mint 16 operating system. The workstation is also equipped with NVIDIA Fermi C2050 GPU containing 14 SMs, 32 SPs per SM and 48 KB of shared memory per SM. CUDA code was written in the "C for CUDA V6.5" (NVIDIA, 2015) programming environment.

All heuristic outcomes are average results of 5 independent executions. Speedups are computed by dividing the sequential CPU time with the parallel time, which is obtained with the same CPU and the GPU acting as a co-processor. The following configuration was used to run the local search CUDA kernels: (a) Block size of 256; (b) $(c + 1) \times n$ threads in $1 - opt$ search, where n is the number of vertices of the graph and c is the number of clusters of the current solution.

Computational experiments were carried out on (i) a set of 24 random instances, and (ii) a set of 5 social networks from the literature. Next, we describe briefly these instances[1].

(i) We generated random social networks with $n \in \{400, 600\}$, varying network density $d = 2 \times |E|/(n^2 - n)$ and negative graph density defined here as $d^- = |E^-|/|E|$. For each value of n, we considered a set of 12 random instances having d and d^- ranging, respectively, in sets $\{0.1, 0.2, 0.5, 0.8\}$ and $\{0.2, 0.5, 0.8\}$.

(ii) This set of instances is composed by 5 signed networks extracted from the large scale social network representing the technology-related news website Slashdot (Leskovec

[1] all instances are available in http://www.ic.uff.br/∼ yuri/files/CCinst.zip.

et al., 2010; Facchetti et al., 2011), containing the first n vertices, with $n \in \{1000, 2000, 4000, 8000, 10000\}$.

4.1. Sequential GRASP *vs.* Sequential GRASP with CUDA local search

In this section, we present the experiments performed with the sequential GRASP algorithm (SeqGRASP) available in Drummond et al. (2013) and the sequential GRASP with CUDA parallel Variable Neighborhood Descent (SeqGRASP/CUDAVND), when solving random instances (Table 2) and Slashdot instances (Table 4). Both experiments used the following set of GRASP parameters:

Time limit	Alpha	Neighborhood	Number of iterations without improvement
2 hours	$\alpha = 1.0$	$r = 1$	$iter = 400$

4.2. Sequential ILS *vs.* Sequential ILS with CUDA local search

Here we list the results of the experiments performed with the sequential ILS algorithm (SeqILS) available in Levorato et al. (2014) and the sequential ILS with CUDA parallel Variable Neighborhood Descent (SeqILS/CUDAVND), when solving random instances (Table 3) and Slashdot instances (Table 5). The following configuration was used in the ILS procedure:

Time limit	Alpha	Neighborhood	Iterations	ILS iterations	Perturbation level
2 hours	$\alpha = 1.0$	$r = 1$	$iter = 10$	$iterMaxILS = 5$	$perturbMax = 30$

5. Concluding remarks

The aim of this paper was to design an efficient parallelization strategy for the implementation of a parallel local search procedure for the Correlation Clustering problem on GPU. After applying the procedure, known as CUDAVND, in existing GRASP and ILS metaheuristics for the CC problem, our experimental results showed significant speedups, outperforming, in processing time, the local search available in the literature.

The GRASP/CUDAVND algorithm presented an average speedup of x43 (up to x121) on random instances and x2.9 (up to x4.21) on Slashdot instances, while the ILS/CUDAVND showed an average speedup of x14 (up to x33) on random instances and x3.5 (up to x5.6) on Slashdot instances. In both algorithms, the solution quality was equal or close to their sequential counterparts.

The next step of our work will focus on improving the analysis of larger signed social networks. The numerical experience indicates that, in order to handle instances like Epinions (131,828 vertices and 841,372 edges) or Slashdot (82,144 vertices and 549,202 edges) networks, we need to implement better parallelization strategies. One possible approach is implementing a hybrid application, using the parallelism available both in CPU (multicore) and GPU (CUDA).

References

Abell, P. and Ludwig, M. (2009). Structural balance: a dynamic perspective. *Journal of Mathematical Sociology*, 33:129–155.

Bansal, N., Blum, A., and Chawla, S. (2002). Correlation clustering. In *Proceedings of the 43rd annual IEEE symposium of foundations of computer science*, pages 238–250, Vancouver, Canada.

Batagelj, V. and Mrvar, A. (2008). Pajek wiki. http://pajek.imfm.si/. Accessed on 12.05.2014.

Cartwright, D. and Harary, F. (1956). Structural balance: A generalization of heiders theory. *Psychological Review*, 63:277–293.

DasGupta, B., Encisob, G. A., Sontag, E., and Zhanga, Y. (2007). Algorithmic and complexity results for decompositions of biological networks into monotone subsystems. *BioSystems*, 90:161–178.

Demaine, E. D., Emanuel, D., Fiat, A., and Immorlica, N. (2006). Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361:172–187.

Doreian, P. and Mrvar, A. (1996). A partitioning approach to structural balance. *Social Networks*, 18:149–168.

n	e	e+	e-	d	d-	SeqGRASP		SeqGRASP/CUDAVND			
						Avg I(P)	Avg Time	Avg I(P)	Gap %I(P)	Avg Time	Speedup
400	15960	12768	3192	0.1	0.2	3192	100.67	3192.0	0.00%	2.99	33.70
400	15960	7980	7980	0.1	0.5	5803.6	565.66	5796.7	-0.12%	21.86	25.88
400	15960	3192	12768	0.1	0.8	2338	1409.42	2357.0	0.81%	33.69	41.84
400	31920	25536	6384	0.2	0.2	6384	132.25	6384.0	0.00%	2.81	47.07
400	31920	15960	15960	0.2	0.5	12840.4	656.99	12833.2	-0.06%	28.46	23.08
400	31920	6384	25536	0.2	0.8	5324.4	3254.87	5355.6	0.59%	53.40	60.96
400	79800	63840	15960	0.5	0.2	15960	232.70	15960.0	0.00%	3.64	63.87
400	79800	39900	39900	0.5	0.5	34862.4	1219.75	34829.6	-0.09%	68.51	17.80
400	79800	15960	63840	0.5	0.8	14636	3606.43	14667.6	0.22%	100.32	35.95
400	127680	102144	25536	0.8	0.2	25536	346.12	25536.0	0.00%	4.71	73.49
400	127680	63840	63840	0.8	0.5	57466.8	2535.17	57437.0	-0.05%	132.01	19.20
400	127680	25536	102144	0.8	0.8	24086.8	3608.31	24122.0	0.15%	157.46	22.92
600	35940	28752	7188	0.1	0.2	7188	334.64	7188.0	0.00%	4.65	71.89
600	35940	17970	17970	0.1	0.5	13915.2	1414.86	13914.8	0.00%	44.62	31.71
600	35940	7188	28752	0.1	0.8	5730.4	3601.82	5771.6	0.72%	77.31	46.59
600	71880	57504	14376	0.2	0.2	14376	447.66	14376.0	0.00%	5.07	88.28
600	71880	35940	35940	0.2	0.5	30149.2	2670.39	30123.3	-0.09%	89.50	29.84
600	71880	14376	57504	0.2	0.8	12613.6	3605.18	12642.3	0.23%	139.35	25.87
600	179700	143760	35940	0.5	0.2	35940	905.51	35940.0	0.00%	8.13	111.39
600	179700	89850	89850	0.5	0.5	80670.8	3602.54	80654.4	-0.02%	198.65	18.13
600	179700	35940	143760	0.5	0.8	33808.8	3629.47	33854.7	0.14%	293.85	12.35
600	287520	230016	57504	0.8	0.2	57504	1423.95	57504.0	0.00%	11.68	121.95
600	287520	143760	143760	0.8	0.5	132176.8	3603.92	132108.4	-0.05%	304.84	11.82
600	287520	57504	230016	0.8	0.8	55182	3656.96	55203.2	0.04%	427.06	8.56
Average						**-**	**1940.22**	**-**	**0.10%**	**92.27**	**43.51**

Table 2: SeqGRASP and SeqGRASP/CUDAVND results for random instances in (i). Number of vertices: n; Avg I(P): average value of the best solution found within time limit; AvgTime: average time spent (in seconds) on 5 independent executions of each algorithm. Gap %I(P) is the % gap between SeqGRASP and SeqGRASP/CUDAVND solutions. Speedup measures the acceleration of the parallel algorithm over its sequential counterpart.

Doreian, P. and Mrvar, A. (2009). Partitioning signed social networks. *Social Networks*, 31:1–11.

Drummond, L., Figueiredo, R., Frota, Y., and Levorato, M. (2013). Efficient solution of the correlation clustering problem: An application to structural balance. In Demey, Y. and Panetto, H., editors, *On the Move to Meaningful Internet Systems: OTM 2013 Workshops*, volume 8186 of *Lecture Notes in Computer Science*, pages 674–683. Springer Berlin Heidelberg.

Elsner, M. and Schudy, W. (2009). Bounding and comparing methods for correlation clustering beyond ilp. In *ILP'09 Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 19–27.

Facchetti, G., Iacono, G., and Altafini, C. (2011). Computing global structural balance in large-scale signed social networks. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 108, pages 20953–20958.

Feo, T. A. and Resende, M. G. (1995). Greedy randomized adaptive search procedures. *Journal of global optimization*, 6(2):109–133.

Figueiredo, R. and Frota, Y. (2014). The maximum balanced subgraph of a signed graph: Applications and solution approaches. *European Journal of Operational Research*, 236(2):473 – 487.

Figueiredo, R. and Moura, G. (2013). Mixed integer programming formulations for clustering problems related to structural balance. *Social Networks*, 35(4):639–651.

Fujimoto, N. and Tsutsui, S. (2011). A highly-parallel tsp solver for a gpu computing platform. In *Numerical Methods and Applications*, pages 264–271. Springer.

Gropp, W., Lusk, E., and Skjellum, A. (1999). *Using MPI: Portable Parallel Programming with*

n	e	e+	e-	d	d-	SeqILS		SeqILS/CUDAVND			
						Avg I(P)	Avg Time	Avg I(P)	Gap %I(P)	Avg Time	Speedup
400	15960	12768	3192	0.1	0.2	3192	10.07	3192	0.00%	10.69	0.94
400	15960	7980	7980	0.1	0.5	5714.4	443.81	5750.4	0.63%	44.65	9.94
400	15960	3192	12768	0.1	0.8	2171.2	1675.43	2223.2	2.39%	78.06	21.46
400	31920	25536	6384	0.2	0.2	6384	14.18	6384	0.00%	12.95	1.10
400	31920	15960	15960	0.2	0.5	12734.8	563.41	12796.8	0.49%	49.54	11.37
400	31920	6384	25536	0.2	0.8	5152.8	2709.74	5222.8	1.36%	123.28	21.98
400	79800	63840	15960	0.5	0.2	15960	36.86	15960	0.00%	19.65	1.88
400	79800	39900	39900	0.5	0.5	34675.2	1302.08	34826.4	0.44%	71.14	18.30
400	79800	15960	63840	0.5	0.8	14438.4	5849.67	14527.2	0.62%	226.57	25.82
400	127680	102144	25536	0.8	0.2	25536	60.71	25536	0.00%	26.26	2.31
400	127680	63840	63840	0.8	0.5	57242.4	2101.79	57446	0.36%	97.54	21.55
400	127680	25536	102144	0.8	0.8	23886.4	7030.23	24002.4	0.49%	312.83	22.47
600	35940	28752	7188	0.1	0.2	7188	17.35	7188	0.00%	14.86	1.17
600	35940	17970	17970	0.1	0.5	13756	967.29	13827.6	0.52%	75.80	12.76
600	35940	7188	28752	0.1	0.8	5453.6	5078.13	5558	1.91%	162.17	31.31
600	71880	57504	14376	0.2	0.2	14376	32.92	14376	0.00%	18.89	1.74
600	71880	35940	35940	0.2	0.5	29958.8	1495.55	30057.2	0.33%	89.13	16.78
600	71880	14376	57504	0.2	0.8	12328.4	7182.95	12445.2	0.95%	269.33	26.67
600	179700	143760	35940	0.5	0.2	35940	83.27	35940	0.00%	31.46	2.65
600	179700	89850	89850	0.5	0.5	80393.2	3278.09	80694.4	0.37%	130.38	25.14
600	179700	35940	143760	0.5	0.8	33503.6	7200.54	33670.8	0.50%	594.54	12.11
600	287520	230016	57504	0.8	0.2	57504	133.27	57504	0.00%	43.70	3.05
600	287520	143760	143760	0.8	0.5	131710.4	5641.22	132099.6	0.30%	171.88	32.82
600	287520	57504	230016	0.8	0.8	54851.2	7200.52	55021.2	0.31%	758.44	9.49
Average						**-**	**2504.54**	**-**	**0.50%**	**143.07**	**13.95**

Table 3: SeqILS and SeqILS/VND results for random instances in (i).

Instance					SeqGRASP		SeqGRASP/CUDAVND			
n	$\|E^-\|$	$\|E^+\|$	$w(E^-)$	$w(E^+)$	$AvgI(P)$	$AvgTime$	$AvgI(P)$	$Gap\%I(P)$	$AvgTime$	$Speedup$
1000	859	5132	859	5132	600.0	23.69	600.0	0.00%	18.75	1.26
2000	3217	17598	3217	17598	2186.0	232.48	2187.4	0.06%	73.93	3.14
4000	8664	40868	8664	40868	6202.6	1415.45	6206.2	0.06%	335.88	4.21
8000	22789	86916	22789	86916	16082.6	7030.32	16087.2	0.03%	2189.18	3.21
10000	29805	109266	29805	109266	20594.6	7200.49	20596.6	0.01%	2680.88	2.69
Average						**3180.49**		**0.03%**	**1059.72**	**2.90**

Table 4: SeqGRASP and SeqGRASP/VND results for Slashdot instances in (ii).

Instance					SeqILS		SeqILS/CUDAVND			
n	$\|E^-\|$	$\|E^+\|$	$w(E^-)$	$w(E^+)$	$AvgI(P)$	$AvgTime$	$AvgI(P)$	$Gap\%I(P)$	$AvgTime$	$Speedup$
1000	859	5132	859	5132	600.2	33.64	600.2	0.00%	15.71	2.14
2000	3217	17598	3217	17598	2187.5	107.42	2201.6	0.64%	38.15	2.82
4000	8664	40868	8664	40868	6218.6	591.84	6213.2	-0.09%	105.96	5.59
8000	22789	86916	22789	86916	16072.1	2229.04	16082.6	0.07%	604.51	3.69
10000	29805	109266	29805	109266	20595.1	3618.14	20600.6	0.03%	1087.89	3.33
Average						**1316.01**		**0.13%**	**370.44**	**3.51**

Table 5: SeqILS and SeqILS/VND results for Slashdot instances in (ii).

the Message-passing Interface. Number v. 1 in Scientific and engineering computation. MIT Press.

Gülpinar, N., Gutin, G., Mitra, G., and Zverovitch, A. (2004). Extracting pure network submatrices in linear programs using signed graphs. *Discrete Applied Mathematics*, 137:359–372.

Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology*, 21:107–112.

Huffner, F., Betzler, N., and Niedermeier, R. (2010). Separator-based data reduction for signed graph balancing. *Journal of Combinatorial Optimization*, 20:335–360.

Hummon, N. P. and Doreian, P. (2003). Some dynamics of social balance processes: bringing heider back into balance theory. *Social Networks*, 25(1):17–49.

Inohara, T. (1998). On conditions for a meeting not to reach a deadlock. *Applied Mathematics and Computation*, 90:1–9.

Krüger, F., Maitre, O., Jiménez, S., Baumes, L., and Collet, P. (2010). Speedups between$\times$ 70 and$\times$ 120 for a generic local search (memetic) algorithm on a single gpgpu chip. In *Applications of Evolutionary Computation*, pages 501–511. Springer.

Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009). The slashdot zoo: mining a social network with negative edges. In *WWW'09 Proceedings of the 18th international conference on World wide web*, pages 741–750.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010). Signed networks in social media. In *CHI'10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370.

Levorato, M., Figueiredo, R., Drummond, L., and Frota, Y. (2014). Uma metaheurística iterated local search aplicada ao problema de correlação de clusters. In *Anais do XLVI Simpósio Brasileiro de Pesquisa Operacional (SBPO'14)*.

Lourenço, H. R., Martin, O. C., and Stützle, T. (2003). *Iterated local search*. Springer.

Macon, K., Mucha, P., and Porter, M. (2012). Community structure in the united nations general assembly. *Physica A: Statistical Mechanics and its Applications*, 391:343–361.

Mehrotra, A. and Trick, M. A. (1998). Cliques and clustering: A combinatorial approach. *Oper. Res. Lett.*, 22(1):1–12.

NVIDIA (2015). CUDA Toolkit. https://developer.nvidia.com/cuda-toolkit. Accessed on 27.02.2015.

Pena, G. C., Magalhaes, S. V., Andrade, M. V., and Ferreira, C. R.. Algoritmo paralelo usando gpu para o posicionamento de observadores em terrenos.

Rocki, K. and Suda, R. (2012). Accelerating 2-opt and 3-opt local search using gpu in the travelling salesman problem. In *High Performance Computing and Simulation (HPCS), 2012 International Conference on*, pages 489–495. IEEE.

Santos, L., Madeira, D., Clua, E., Martins, S., and Plastino, A.. A parallel grasp resolution for a gpu architecture.

Srinivasan, A. (2011). Local balancing influences global structure in social networks. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 108, pages 1751–1752.

Traag, V. and Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Physical Review E*, 80:036115.

Van Luong, T., Melab, N., and Talbi, E.-G. (2013). Gpu computing for parallel local search metaheuristic algorithms. *Computers, IEEE Transactions on*, 62(1):173–185.

Yang, B., Cheung, W., and Liu, J. (2007). Community mining from signed social networks. *IEEE Transactions on Knowledge and Data Engineering*, 19:1333–1348.

Zhang, Z., Cheng, H., Chen, W., Zhang, S., and Fang, Q. (2008). Correlation clustering based on genetic algorithm for documents clustering. In *IEEE Congress on Evolutionary Computation*, pages 3193–3198.