

RAMA: A Rapid Multicut Algorithm on GPU

Ahmed Abbas Paul Swoboda

Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

We propose a highly parallel primal-dual algorithm for the multicut (a.k.a. correlation clustering) problem, a classical graph clustering problem widely used in machine learning and computer vision. Our algorithm consists of three steps executed recursively: (1) Finding conflicted cycles that correspond to violated inequalities of the underlying multicut relaxation, (2) Performing message passing between the edges and cycles to optimize the Lagrange relaxation coming from the found violated cycles producing reduced costs and (3) Contracting edges with high reduced costs through matrix-matrix multiplications.

Our algorithm produces primal solutions and lower bounds that estimate the distance to optimum. We implement our algorithm on GPUs and show resulting one to two orders-of-magnitudes improvements in execution speed without sacrificing solution quality compared to traditional sequential algorithms that run on CPUs. We can solve very large scale benchmark problems with up to $\mathcal{O}(10^8)$ variables in a few seconds with small primal-dual gaps. Our code is available at <https://github.com/pawelswoboda/RAMA>.

1. Introduction

Decomposing a graph into meaningful clusters is a fundamental problem in combinatorial optimization. The multicut problem [15] (also known as correlation clustering [10]) is a popular approach to decompose a graph into an arbitrary number of clusters based on affinities between nodes.

The multicut problem and its extensions such as higher order multicut [27, 32], lifted multicut [30], (asymmetric) multiway cut [14, 36], lifted disjoint paths [21] and joint multicut and node labeling [41] have found numerous applications in machine learning, computer vision, biomedical image analysis, data mining and beyond. Examples include unsupervised image segmentation [4, 5, 7, 58], instance-separating semantic segmentation [2, 33], multiple object tracking [21, 53], cell tracking [25], articulated human body pose estimation [22], motion segmentation [31], image and mesh segmentation [30], connectomics [6, 13, 48] and many more.

Multicut and its extensions are NP-hard to solve [10, 18]. Since large problem instances with millions or even billions of variables typically occur, powerful approximative algorithms have been developed [11, 12, 30, 39, 52]. However, even simple heuristics such as [30] require very large running times for very large instances. In particular, some instances, such as those investigated in [48] could not be solved in acceptable time (hence ad-hoc decomposition techniques were used). In other scenarios very fast running times are essential, e.g. when multicut is used in end-to-end training [2, 49]. Hence, the need for parallelization arises, preferably on GPUs. The parallelism offered by GPUs is typically difficult to exploit due to irregular data structures and the inherently sequential nature of most combinatorial optimization algorithms. This makes design of combinatorial optimization algorithms challenging on GPUs. An additional benefit of running our algorithms on GPU is that memory transfers between CPU and GPU are avoided when used in a deep learning pipeline.

Our contribution is a new primal-dual method that can be massively parallelized and run on GPU. This results in faster runtimes than previous multicut solvers while still computing solutions which are similar or better than CPU based solvers in terms of objective. Yet, our approach is rooted in solving a principled polyhedral relaxation and yields both a primal solution and a dual lower bound. In particular, finding primal solutions and approximate dual solving is interleaved such that both components of our algorithm can profit from each other. In more detail, our algorithmic contribution can be categorized as follows

Primal: Edge Contraction: Finding a primal solution depends similarly as in [30] on contracting edges that are highly likely to end up in the same component of the final clustering. To this end we propose to use a linear algebra approach by expressing edge contractions as sparse matrix-matrix multiplications. This allows us to accelerate edge contraction by exploiting highly parallel matrix-matrix multiplication GPU primitives.

Dual: Lagrange Relaxation & Message Passing: To find good edge contraction candidates, we consider approximately solving a polyhedral relaxation by searching for conflicting cycles, adding them to a Lagrange relaxation

and updating the resulting Lagrange multipliers iteratively by message passing. We propose a new message passing scheme that is both massively parallelizable yet yields monotonic increases w.r.t. the dual objective, speeding up the scheme of [52] by orders of magnitude.

Recursive Primal-Dual: We interleave the above operations of finding and solving a Lagrange relaxation and contracting edges, yielding the final graph decomposition. Hence, our algorithm goes beyond classical polyhedral approaches [26, 44, 52] that only consider the original graph.

On the experimental side we obtain primal solutions that are of comparable or better quality to those obtained by established high-quality heuristics [30, 38] in a fraction of the execution time but with additional dual lower bounds that help in estimating the quality of the solutions. We perform experiments on 2D and 3D instance segmentation problems for scene understanding [17] and connectomics [48] containing up to $\mathcal{O}(10^8)$ variables.

2. Related Work

Preprocessing and Inprocessing: For fixing variables to their optimal values and shrinking the problem before or during optimization, persistency or partial optimality methods have been proposed in [3, 37, 38]. These methods apply a family of criteria that, when passed, prove that any solution can be improved if its values do not coincide with the persistently fixed variables.

Primal heuristics: For obtaining primal solutions without optimality guarantees or estimates on the distance to optimum, a large number of methods have been proposed with different execution time/solution quality trade-offs. The first heuristic for multicut, the classical Kernighan&Lin move-making algorithm was originally proposed in [29] and slightly generalized in [30]. The algorithm consists of trying various moves such as joining two components, moving a node from one component to the next *etc.* and performing sequences of moves that decrease the objective. The faster but simpler greedy additive edge contraction (GAEC) heuristic, a move making algorithm that only can join individual components, was proposed in [30]. It is used in [30] to initialize the more complex Kernighan&Lin algorithm. Variants involving different join selection strategies were proposed in [28]. The greedy edge fixation algorithm [30] generalizes GAEC in that it can additionally mark edges as cut, constraining their endpoints to be in distinct components. The more involved Cut, Glue & Cut (CGC) move-making heuristic [12] works by alternating bipartitioning of the graph and exchanging nodes in pairs of clusters. The latter operation is performed by computing a max-cut on a planar subgraph via reduction to perfect matching. CGC was extended to a more

general class of possible “fusion moves” in [11]. A parallel algorithm for the simpler problem of unweighted correlation clustering problem was given in [46]. A comparative survey of some of the above primal heuristics is given in [40].

LP-based algorithms: For obtaining dual lower bounds that estimate the distance to the optimum or even certify optimality of a solution a number of LP-relaxation based algorithms have been proposed. These algorithms can be used inside branch and bound and their computational results can be used to guide primal heuristics to provide increasingly better solutions. Quite surprisingly, it has been shown by [26, 32] that multicut problems of moderately large sizes can be solved with commercial integer linear programming (ILP) solvers like Gurobi [19] in a cutting plane framework in reasonable time to global optimality. Column generation based on solving perfect matching subproblems has been proposed in [42, 58]. Still, the above approaches break down when truly large scale problems need to be solved, since the underlying LP-relaxations are still solved by traditional LP-solvers that do not scale linearly with problem size and are not explicitly adapted to the multicut problem. Additionally, violated inequality separation (cutting planes) requires solving weighted shortest path problems which is not possible in linear time. The message passing algorithm [50] approximately solves a dual LP-relaxation faster than traditional LP-solvers and has faster separation routines than those of primal LP-solvers as well, thereby scaling to larger problems. An even faster, but less powerful, approximate cycle packing algorithm was proposed in [38].

Other efficient clustering Methods: The mutex watershed [57] and its generalizations [9] are closely related to the greedy additive edge fixation heuristic for multicut [40]. The corresponding algorithms can be executed faster than their multicut counterparts on CPU, but are sequential. Fast GPU schemes [8] were proposed for agglomerative clustering. Last, spectral clustering can be implemented on GPU with runtime gains [24, 43]. All these approaches however are not based on any energy minimization problem, hence do not come with the theoretical benefits that an optimization formulation offers.

3. Method

A *decomposition (or clustering)* of a graph $G = (V, E)$ is a partition $\{V_1, \dots, V_k\}$ of the node set such that $V_1 \cup \dots \cup V_k = V$ and $V_i \cap V_j = \emptyset \forall i \neq j$. The *cut* $\delta(V_1, \dots, V_k)$ induced by a decomposition is the subset of edges that straddle distinct clusters. Such edges are said to be *cut*. See Figure 1 for an illustration of a cut into three components.

The space of all multicuts is

$$\mathcal{M}_G = \left\{ \delta(V_1, \dots, V_k) : \begin{array}{l} k \in \mathbb{N} \\ V_1 \dot{\cup} \dots \dot{\cup} V_k = V \end{array} \right\}. \quad (1)$$

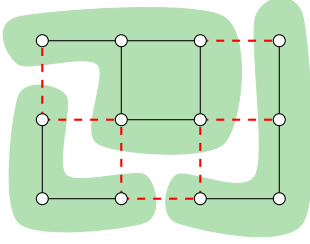


Figure 1. Decomposition of a graph into three components (green). The corresponding cut consists of the dashed edges straddling distinct components (red).

The associated minimum cost multicut problem is defined by an additional edge cost vector $c \in \mathbb{R}^E$. For any edge $uv \in E$, negative costs $c_{uv} < 0$ favour the nodes u and v to be in distinct components. Positive costs $c_{uv} > 0$ favour these nodes to lie in the same component. The minimum cost multicut problem is

$$\min_{y \in \mathcal{M}_G} \langle c, y \rangle, \quad (2)$$

where y_{uv} for edge $uv \in E$ is 1 (resp. 0) if u and v belong to distinct (resp. same) components.

Below we detail the key components of our algorithm: Starting from a graph where each node is a cluster, primal updates consist of edge contractions that iteratively merge clusters by join operations. Dual updates optimize a Lagrange relaxation via message passing to obtain better edge costs and lower bound. Primal and dual updates are interleaved to yield our primal-dual multicut algorithm. We additionally detail how each operation can be done in a highly parallel manner.

3.1. Primal: Parallel Edge Contraction

The idea of edge contraction algorithms is to iteratively choose edges with large positive costs. Such edges prefer their endpoints to be in the same component, hence they are contracted and end up in the same cluster. Edge contraction is performed until no contraction candidates are found. The special case of greedy additive edge contraction (GAEC) [30] chooses in each iteration an edge with maximum edge weight for contraction and stops if each edge in the contracted graph has negative weight. The following Lemma describes the operation of edge contraction.

Lemma 1. *Let an undirected graph $G = (V, E, c)$ and a set of edges $S \subseteq E$ to contract be given. Also let $G' = (V', E', c')$ be the graph obtained after edge contraction.*

- (a) *The corresponding surjective contraction mapping $f : V \rightarrow V'$ mapping node set V onto the contracted node set V' is up to isomorphism uniquely defined by $f(u) = f(v) \iff \exists uv\text{-path}(V, S)$. The contracted edge set is given by $E' = \{f(u)f(v) : f(u) \neq f(v), uv \in E\}$.*

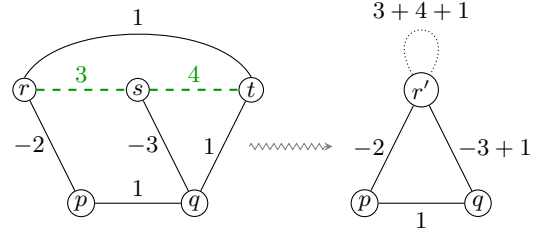


Figure 2. Contraction of a graph with contraction set $S = \{rs, st\}$ where vertices r, s and t are merged to form a cluster r' . The corresponding contraction mapping is $f(p) = p, f(q) = q, f(r) = f(s) = f(t) = r'$. Notice that edges qs and qt become parallel edges after contraction and their costs are added. Also notice the presence of self-loop in the contracted graph with cost indicating intra-cluster similarity.

- (b) *The edge weights for contracted edges are $c'_{ij} = \sum_{uv \in E: f(u)=i, f(v)=j} c_{uv}, \forall ij \in E'$.*

Lemma 1(a) relates the contraction mapping f with the set S of edges to contract. If two nodes have the same value in f then there must be a path between them in graph (V, S) . Moreover, the edges whose end points are not contracted are preserved in E' . Lemma 1(b) provides the costs of contracted edges that are obtained by summing the costs of parallel edges. An illustration of the lemma is given in Figure 2.

In order to perform edge contraction fast we will use a linear algebraic representation that will allow to use highly parallel (sparse) matrix-matrix multiplication.

Definition 2 (Adjacency Matrix). *Given a weighted graph $G = (V, E, c)$ its (symmetric) adjacency matrix $A \in \mathbb{R}^{V \times V}$*

$$\text{is defined by } A_{uv} = \begin{cases} c_{uv}, & uv \in E \\ 0, & \text{otherwise} \end{cases}$$

We will perform edge contraction with the help of an edge contraction matrix defined as follows.

Definition 3 (Edge Contraction Matrix). *Given a weighted graph $G = (V, E, c)$ and an edge set $S \subseteq E$ to contract, let f be the contraction mapping and V' the contracted node set. The edge contraction matrix $K_S \in \{0, 1\}^{V \times V'}$ is defined as*

$$(K_S)_{uu'} = \begin{cases} 1, & f(u) = u' \\ 0, & \text{otherwise} \end{cases}$$

Lemma 4. *Given a weighted graph $G = (V, E, c)$, an edge set $S \subseteq E$ to contract and an associated edge contraction mapping f*

- (a) *the adjacency matrix of the contracted graph is equal to $K_S^\top A K_S - \text{diag}(K_S^\top A K_S)$, where $\text{diag}(\cdot)$ is the diagonal part of a matrix,*
- (b) *it holds for the diagonal entry $(K_S^\top A K_S)_{u'u'} = \sum_{uv \in E: u'=f(u)=f(v)} c_{uv}$.*

Lemma 4(a) provides a way to compute the contracted graph in parallel by sparse matrix-matrix multiplication. Lemma 4(b) allows to efficiently judge whether the newly formed clusters decrease the multicut objective. Specifically if the diagonal contains all positive terms then the corresponding multicut objective will also decrease after contraction.

A primal update iteration is given in Algorithm 1 that performs edge contraction as in Lemma 4(a).

Algorithm 1: Parallel-Edge-Contraction

Data: Graph $G = (V, E, c)$

Result: Contracted Graph $G' = (V', E', c')$,
contraction mapping $f : V \rightarrow V'$

- 1 Compute contraction set $S \subseteq E$
 - 2 Compute adjacency matrix A from G
 - 3 Construct contraction mapping $f : V \rightarrow V'$
 - 4 Construct contraction matrix K_S
 - 5 $A' = K_S^\top A K_S - \text{diag}(K_S^\top A K_S)$
 - 6 Compute contracted graph $G' = (V', E', c')$ from A'
-

Finding contraction edge set S : A vital step for ensuring a good primal update is selecting the edge set S for contraction in Algorithm 1. On one hand, we would like to choose edges in a conservative manner to avoid erroneous contractions. On the other hand, we need to contract as much edges as possible for efficiency. We propose two approaches allowing us to be at the sweet spot for both criterion as follows.

Maximum matching: Perform a fast maximum matching on the positive edges in using a GPU version of the Luby-Jones handshaking algorithm [16] and select the matched edges for contraction.

Maximum spanning forest without conflicts: Compute a maximum spanning forest on the positive edges with a fast GPU version of Borůvka’s algorithm [55] to find initial contraction candidates. Afterwards, iterate over all negative edges ij , find the unique path between i and j in the forest (if it exists) and remove the smallest positive edge. We make use of GPU connected components [23] to check for presence of these paths and to compute the final contraction mapping.

Both of the above strategies ensure that the resulting join operation decreases the multicut objective. We first find contraction edges via maximum matching. If not enough edges are found (*i.e.* fewer than $0.1|V|$), we switch to the spanning forest based approach. Note that if we chose only one largest positive edge for contraction, Algorithm 1 specializes to GAEC [30]. Since our algorithm depends upon many simultaneous edge contractions for efficiency, we do not use this strategy.

3.2. Dual: Conflicted Cycles & Message passing

Solving a dual of multicut problem (2) can help in obtaining a lower bound on the objective value and also yields a reparametrization of the edge costs which can help in better primal updates. Our dual algorithm works on the cycle relaxation for the multicut problem [15]. We present for its solution massively parallel inequality separation routines to search for the most useful violated constraints and efficient dual block coordinate ascent procedure for optimizing the resulting relaxation.

3.2.1 Cycle Inequalities & Lagrange Relaxation

Since the multicut problem is NP-hard [10, 18], we cannot hope to obtain a feasible polyhedral description of $\text{conv}(\mathcal{M}_G)$. A good relaxation for most practical problems is given in terms of cycle inequalities. Given a cycle $C = \{e_1, \dots, e_l\} \subseteq E$, a feasible multicut must either not contain any cut edge or should contain at least two cut edges. This constraint is expressed as

$$\forall C \in \text{cycles}(G) : \forall e \in C : y_e \leq \sum_{e' \in C \setminus \{e\}} y_{e'}. \quad (3)$$

Cycle inequalities together with the binary constraints $y_e \in \{0, 1\}$ actually define \mathcal{M}_G [15]. In other words, when relaxing $y_e \in [0, 1]$ we obtain a linear program relaxation to $\text{conv}(\mathcal{M}_G)$ with all integral points being valid multicuts.

While cycle inequalities (3) give us a polyhedral relaxation of the multicut problem (2), our algorithm will operate on a Lagrangean decomposition that was proposed in [50]. It consists of two types of subproblems joined together via Lagrange variables: (i) edge subproblems for each edge $e \in E$ and (ii) triangle subproblems (*i.e.* cycles of length 3) for a subset of triangles $T \subset \binom{E}{3}$. Triangulation of cycles of length more than three is done to get triangles defining the same polyhedral relaxation as the one with all possible cycle inequalities (3) without loss of generality [15]. We define the set of feasible multicuts on triangle graphs as

$$\mathcal{M}_T = \{(0, 0, 0), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}, \quad (4)$$

which is a special case of (3) representing that either all edges are cut/joined or exactly two edges are cut. Given a set of edge and triangle subproblems our Lagrange decomposition is

$$\max_{\lambda} \underbrace{\sum_{uv \in E} \min_{y \in \{0, 1\}} c_{uv}^\lambda \cdot y + \sum_{t \in T} \min_{y \in \mathcal{M}_T} \langle c_t^\lambda, y \rangle}_{=: \text{LB}(\lambda)} \quad (5)$$

where the *reparametrized* edge costs $c_{uv}^\lambda \in \mathbb{R}$ and triangle

costs $c_t^\lambda \in \mathbb{R}^3$ for triangle $t = \{ij, jk, ki\} \in T$ are

$$c_{uv}^\lambda = c_{uv} + \sum_{t \in T: uv \in t} \lambda_{t,uv} \quad (6a)$$

$$c_t^\lambda = -(\lambda_{t,ij}, \lambda_{t,jk}, \lambda_{t,ki}) \quad (6b)$$

$\text{LB}(\lambda)$ in (5) is a lower bound on the cost of the optimum multicut for any λ . The optimum objective value of (5) equals that of the polyhedral relaxation [52].

3.2.2 Cycle Inequality Separation

For the dual problem (5) one would need to enumerate all possible cycle inequalities (3). However, as mentioned in [38] we can restrict only to conflicted cycles of G for efficiency without loosening the relaxation. A cycle is called a conflicted cycle if it contains exactly one repulsive edge.

Definition 5 (Conflicted cycles). *Let the set of attractive edges in E be $E^+ = \{c_e > 0 : \forall e \in E\}$ and repulsive edges $E^- = \{c_e < 0 : \forall e \in E\}$. Then conflicted cycles of G is the set $\{C \in \text{cycles}(G) : |C \cap E^-| = 1\}$.*

Lemma 6. *The search for conflicted cycles can be performed in parallel for each $ij \in E^-$ by finding shortest path w.r.t. hop distance between i and j in the graph $G = (V, E^+)$ making good use of parallelization capabilities of GPUs.*

3.2.3 Dual Block Coordinate Ascent (DBCA)

DBCA (a.k.a. message passing) was studied in [50] for multicut. However, the resulting message passing schemes are not easily parallelizable. The underlying reason for the inherent sequential nature of these schemes is that the effectiveness of the proposed message passing operations depend on the previous ones being executed. We propose a message passing scheme for multicut that is invariant to the message passing schedule, hence allowing parallel computation.

As in [50], our scheme iteratively improves the lower bound (5) by message passing between edges and triangles.

For each message passing operation we need to compute min-marginals, *i.e.* the difference of optimal costs on subproblems obtained by fixing a specified variable to 1 and 0. For edge costs the min-marginal is just the reparametrized edge cost. For triangle subproblems it is given as follows

Definition 7 (Marginalization for triangle subproblems). *Let $t \in T$ be a triangle containing an edge e .*

$$m_{t \rightarrow e}(c_t^\lambda) = \min_{\substack{y_e=1 \\ y \in \mathcal{M}_T}} \langle c_t^\lambda, y \rangle - \min_{\substack{y_e=0 \\ y \in \mathcal{M}_T}} \langle c_t^\lambda, y \rangle \quad (7)$$

is called min-marginal for triangle t and edge e .

The message passing algorithm iteratively sets min-marginal to zero first for edge subproblems and then for

triangles described in Algorithm 2. By sending back and forth messages the subproblems communicate their local optima and ultimately the min-marginals converge towards agreement (*i.e.* their corresponding edge labels y are consistent). In [52] it was shown that each such operation is non-decreasing in the dual objective value, yielding an overall monotonic convergence. Message are passed from edges to triangles in lines 2-5. After this step the reparametrized edge costs c_e^λ become zero. We perform multiple triangle to edge message passing updates (line 8-13) similar to the way it was done in [54] that distribute messages uniformly among all triangles which contain that edge. After this operation min-marginals for c_t^λ become zero.

Algorithm 2: Parallel-Message-Passing

Data: Graph $G = (V, E, c)$, triangles T , Lagrange multipliers λ .

Result: Updated Lagrange multipliers λ

// Messages from edges to triangles

1 **for** $e \in E$ in parallel **do**

2 $\alpha = c_e^\lambda$

3 **for** $t \in T : e \in t$ **do**

4 $\lambda_{t,e^-} = \frac{\alpha}{|t \in T : e \in t|}$

5 **end**

6 **end**

// Messages from triangles to edges

7 **for** $t = \{ij, jk, ki\} \in T$ in parallel **do**

8 $\lambda_{t,ij^+} = \frac{1}{3} m_{t \rightarrow ij}(c_t^\lambda)$

9 $\lambda_{t,ik^+} = \frac{1}{2} m_{t \rightarrow ik}(c_t^\lambda)$

10 $\lambda_{t,jk^+} = m_{t \rightarrow jk}(c_t^\lambda)$

11 $\lambda_{t,ij^+} = \frac{1}{2} m_{t \rightarrow ij}(c_t^\lambda)$

12 $\lambda_{t,ik^+} = m_{t \rightarrow ik}(c_t^\lambda)$

13 $\lambda_{t,ij^+} = m_{t \rightarrow ij}(c_t^\lambda)$

14 **end**

Convergence of Message Passing. Algorithm 2 converges towards fixed points, similar to other DBCA schemes for graphical models [34, 35, 54, 56]. These fixed points are characterized with the help of arc consistency and need not coincide with the optimal dual solution, but are typically close to them. Below, we characterize these fixed points.

Definition 8 (Locally Optimal Solutions). *Define the locally optimal solutions for edge $e \in E$ as*

$$\bar{c}_e := \{x \in \{0, 1\} : x \cdot c_e^\lambda = \min(0, c_e^\lambda)\} \quad (8)$$

and similarly for triangle $t \in T$ as

$$\bar{c}_t := \{x \in \mathcal{M}_T : \langle c_t^\lambda, x \rangle = \min_{x' \in \mathcal{M}_T} \langle c_t^\lambda, x' \rangle\} \quad (9)$$

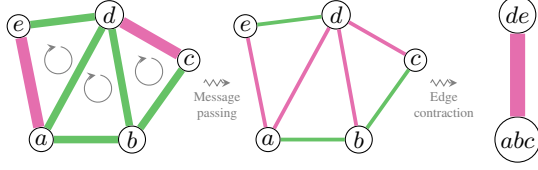


Figure 3. Example iteration of our primal-dual multicut solver on a graph with repulsive and attractive edges. Width of the edges indicate abs. cost. First we detect conflicted cycles and triangulate to get triangles (indicated by \triangle). Next, dual update reparametrizes edge costs which resolves the conflicted cycles. Lastly a primal update is done by contracting attractive edges.

Define the projection of triangle solutions onto one of its edges as

$$\Pi_e(\overline{c_t^\lambda}) := \{x \in \{0, 1\} : \exists x' \in \overline{c_t^\lambda} \text{ s.t. } x'_e = x\} \quad (10)$$

Definition 9 (Arc-Consistency). Lagrange multipliers λ are arc-consistent if $\Pi_e(\overline{c_t^\lambda}) = \overline{c_e^\lambda}$ for all $t \in T$ and $e \subset t$.

However, note that arc-consistency is not necessary for dual optimality. A necessary criterion is edge-triangle agreement.

Definition 10 (Edge-Triangle Agreement). Lagrange multipliers are in edge-triangle agreement if there exist non-empty subsets $\xi_e \subseteq \overline{c_e^\lambda}$ for all $e \in E$ and $\xi_t \subseteq \overline{c_t^\lambda}$ for all $t \in T$ such that ξ_e is arc-consistent, i.e. $\xi_e = \Pi_e(\xi_t)$ for all $t \in T$ and $e \subset t$.

In words, edge-triangle agreement signifies that there exists a subset (also called kernel in [56]) of locally optimal solutions that are arc-consistent.

Theorem 11. Algorithm 2 converges to edge-triangle agreement.

3.3. Primal-Dual Updates

While the two building blocks of our multicut solver *i.e.* edge contraction and cycle separation with message passing can be used in isolation to compute a primal solution and lower bound, we propose an interleaved primal-dual solver in Algorithm 3.

In each iteration we separate cycles and perform message passing to get reparameterized edge costs. We use these reparameterized edge costs to perform parallel edge contraction. This interleaved process continues until no edge contraction candidate can be found. Such scheme has the following benefits

Better edge contraction costs: The reparameterization in line 6 produces edge costs c^λ that are more indicative of whether an edge is contracted or not in the final solution thus yielding better primal updates in line 8. In case

Algorithm 3: Primal-Dual Multicut

Data: Graph $G = (V, E, c)$
Result: Contraction mapping $f : V \rightarrow V'$
// Init. each node as a separate cluster
1 $f = V \rightarrow V, f(v) = v \quad \forall v \in V$
2 **while** G has positive edges without conflicts **do**
 // Find conflicted cycles (Lemma. 6)
3 $T = \text{Cycle-Separation}(G)$
4 **for** $iter = 1, \dots, k$ **do**
5 $\lambda = \text{Parallel-Message-Pass.}(G, T)$
 // Reparametrize edge costs
6 $c_e = c_e^\lambda \quad \forall e \in E$ by (6a)
7 **end**
8 $G, f' = \text{Parallel-Edge-Contract.}(G)$
 // Update contraction mapping
9 $f(v) = f'(f(v)) \quad \forall v \in V$
10 **end**

the relaxation (5) is tight, the sign of c_e^λ perfectly predicts whether an edge e is separating two clusters or is inside one.

Better cycle separation: For fast execution times we stop cycle separation for cycles greater than a given length (5 in our case). Since cycle separation is performed again after edge contraction, this corresponds to finding longer cycles in the original graph. Such approach alleviates the need to perform a more exhaustive and time-consuming initial search.

Note that a valid lower bound can be obtained from Algorithm 3 by recording (5) after cycle separation and message passing on the original graph.

4. Experiments

We evaluate solvers on multicut problems for neuron segmentation for connectomics in the fruit-fly brain [48] and unsupervised image segmentation on Cityscapes [17]. We use a single NVIDIA Volta V100 (16GB) GPU for our solvers unless otherwise stated and an AMD EPYC 7702 for CPU solvers. Our solvers are implemented using the CUDA [45] and Thrust [20] GPU programming frameworks.

Datasets We have chosen three datasets containing the largest multicut problem instances we are aware of. The instances are available in [51].

Connectomics-SP: Contains neuron segmentation problems from the fruit-fly brain [48]. The raw data is taken from the CREMI-challenge [1] acquired by [59] and converted to multiple multicut instances by [48]. For this conversion [48] also reduced the problem size by creating super-pixels. The majority of these instances are different crops of one global problem. There are 3

small (400000 – 600000 edges), 3 medium (4 – 5 million edges) and 5 large (28 – 650 million edges) multicut instances. For the largest problem we use NVIDIA RTX 8000 (48GB) GPU.

Connectomics-Raw: We use the 3 test volumes (sample A+, B+, C+) from the CREMI-challenge [1] segmenting directly on the pixel level without conversion to superpixels. Conversion to multicut instances is carried out using [47]. We report results on two types of instances: (i) The three full problems where the underlying volumes have size $1250 \times 1250 \times 125$ with around 700 million edges and (ii) six cropped problems created by halving each volume and creating the corresponding multicut instances each containing almost 340 million edges. For all these instances we use NVIDIA RTX 8000 (48GB) GPU.

Cityscapes: Unsupervised image segmentation on 59 high resolution images (2048×1024) taken from the validation set [17]. Conversion to multicut instances is done by computing the edge affinities produced by [2] on a grid graph with 4-connectivity and additional coarsely sampled longer range edges. Each instance contains approximately 2 million nodes and 9 million edges.

Algorithms As baseline methods we have chosen, to our knowledge, the fastest primal heuristics from the literature.

GAEC [30]: The greedy additive edge contraction corresponds to Algorithm 2 with choosing a single highest edge to contract. We use our own CPU implementation that is around 1.5 times faster than the one provided by the authors.

KLj [30]: The Kernighan&Lin with joins algorithm performs local move operations which can improve the objective. To avoid large runtimes the output of GAEC is used for initialization.

GEF [40]: The greedy edge fixation algorithm is similar to GAEC but additionally visits negative valued (repulsive) edges and adds non-link constraints between their endpoints.

BEC [28]: Balanced edge contraction, a variant of GAEC which chooses edges to contract based on their cost normalized by the size of the two endpoints.

ICP [38]: The iterated cycle packing algorithm searches for cycles and greedily solves a packing problem that approximately solves the multicut dual (5).

P: Our purely primal Algorithm 1 using the maximum matching and spanning forest based edge contraction strategy.

PD: Our primal-dual Algorithm 3 which additionally makes use of the dual information. We find conflicted cycles up to length 5 on original graph and up to a length of 3 for later iterations on contracted graphs.

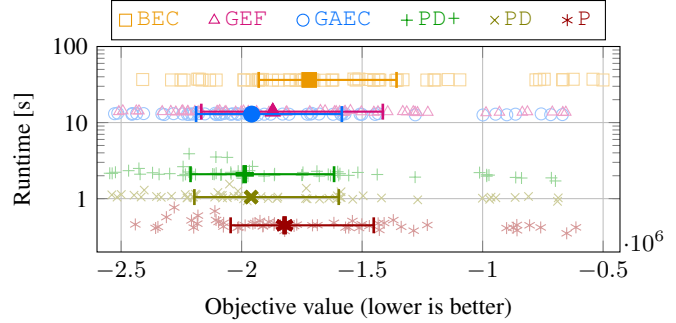


Figure 4. Comparison of primal solutions from *Cityscapes* dataset. Our purely primal algorithm (P) is $30\times$ faster than GAEC [30] and GEF [40], although with worse objective values. Incorporating dual information enables our solvers (PD, PD+) to even surpass the sequential solvers in objective while being faster by an order of magnitude. Error bars mark the 0.25, 0.75-quantile. (KLj not shown due to high runtime).

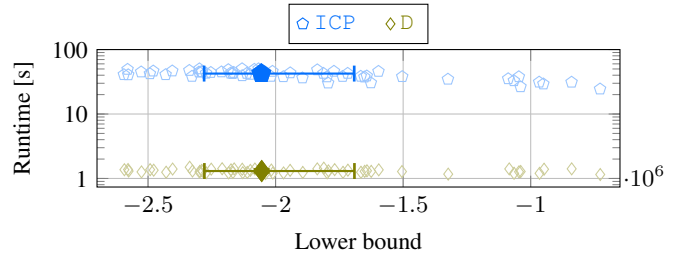


Figure 5. Comparison of lower bounds from *Cityscapes* dataset. Our parallel message passing scheme (D) is more than an order of magnitude faster than ICP [38] and gives slightly better lower bounds. Error bars mark the 0.25, 0.75-quantile.

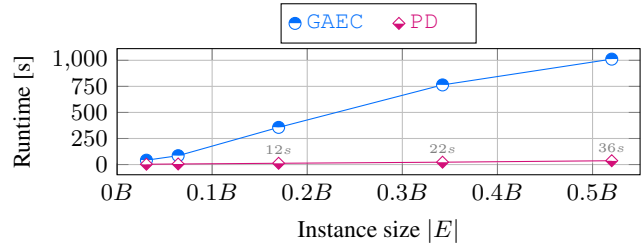


Figure 6. Runtime scaling comparison computed on different crops of CREMI test data showing that RAMA scales very well as compared to GAEC [30] w.r.t. increasing problem sizes

PD+: Variant of PD which always considers conflicted cycles up to a length 5 for reparametrization which can lead to even better primal solutions although with higher runtime.

D: Our dual cycle separation algorithm followed by message passing on the original graph via Algorithm 2 producing lower bounds.

Method	<i>Connectomics-SP</i>						<i>Connectomics-Raw</i>				<i>Cityscapes</i>	
	Small (3)		Med. (3)		Large (5)		Crops (6)		Full (3)		(59)	
	C($\times 10^5$)	t(s)	C($\times 10^5$)	t(s)	C($\times 10^5$)	t(s)	C($\times 10^8$)	t(s)	C($\times 10^8$)	t(s)	C($\times 10^6$)	t(s)
Primal												
KLj [30]	-1.794	3.8	-9.225	125	†	†	†	†	†	†	-1.858	5e4
GAEC [30]	-1.794	0.4	-9.224	4.7	-1.512	280	-1.464	570	-2.963	1140	-1.826	13
GEF [40]	-1.793	0.7	-9.223	9.0	-1.511	699	-1.458	582	-2.949	1762	-1.743	14
BEC [28]	-1.787	0.5	-9.199	5.6	-1.507	309	-1.402	1688	-2.838	4150	-1.613	36
P	-1.780	0.1	-9.173	0.6	-1.505	6	-1.430	9	-2.895	19	-1.711	0.4
PD	-1.791	0.2	-9.217	1.0	-1.509	13	-1.477	24	-2.981	32	-1.846	1
PD+	-1.791	0.3	-9.219	1.4	-1.509	20	-1.480	115	-2.995	224	-1.862	2.2
Dual												
ICP [38]	-1.798	0.8	-9.246	11.3	-1.518	1235	-1.507	513	-3.053	1091	-1.930	41.1
D	-1.797	0.2	-9.241	0.8	-1.517	13	-1.499	34	*	*	-1.928	1.3

Table 1. Comparison of results on all datasets. (C: cost, t(s): time in seconds, †: timed out, *: out of GPU memory). We report average primal and dual costs and runtime over instances within each category. In terms of primal solutions our primal-dual solvers (PD, PD+) achieve objectives close to or better than sequential solvers while being substantially faster especially on larger instances. Moreover, our parallel message passing approach (D) gives better lower bounds than ICP with up to two orders of magnitude reduction in runtime.

Discussion Results on all datasets are given in Table 1. On the *Connectomics-SP* dataset we attain primal objectives very close to those produced by GAEC [30] but faster by more than an order of magnitude on large instances.

For the *Cityscapes* and *Connectomics-Raw* datasets we achieve even better primal solutions than sequential algorithms by incorporating dual information while also being substantially faster. Our best solver (PD+) is more than 10^4 times faster than KLj [30] and produces better solutions. Distributions of runtimes and primal resp. dual objectives for all instances of *Cityscapes* are shown in Figures 4 and 5. We compare the scaling behaviour of our solver w.r.t increasing instance sizes in Figure 6 showing that RAMA scales much more efficiently than GAEC. An example visual comparison of results is given in Figure 7 in Appendix.

Lastly, our dual algorithm (D) produces speedups of up to two orders of magnitude and better lower bounds compared to the serial ICP [38], except on the full instances of *Connectomics-Raw* where we run out of GPU memory.

Runtime breakdown Runtime breakdown of our PD algorithm is given in Table 2. Most of the time is spent in finding conflicted cycles which we found to be challenging to implement on GPU while keeping runtime and memory consumption low. Future improvements offer a potential for even better results and speedups by finding longer cycles more efficiently.

5. Conclusion

We have demonstrated that multicut, an important combinatorial optimization problem for machine learning and

Finding S	Contract.	Conf. cycles	Message passing
30%	7%	43%	20%

Table 2. Runtime breakdown for PD algorithm on *Cityscapes*

computer vision, can be effectively parallelized on GPU. Our approach produces better solutions than state of the art efficient heuristics on grid graphs and comparable ones on super-pixel graphs while being faster by one to two orders-of-magnitude. We believe that performance gap on super-pixel graphs is due to a graph structure containing much more (and longer) conflicted cycles. Since our implementation can only find cycles of length up to 5, better implementations that can efficiently handle longer cycles might yield further improvements.

In contrast to CPU algorithms, where execution speed is the limiting factor, for our GPU algorithm, comparatively smaller amount of GPU-memory limits application to even larger instances. We hope that our work will enable more compute intensive applications of multicut, where until now the slower serial CPU codepath has hindered its adoption.

6. Acknowledgments

We would like to thank Shweta Mahajan and Jan-Hendrik Lange for insightful discussions and Constantin Pape, Adrian Wolny and Anna Kreshuk for their suggestions regarding the experiments. We also thank all anonymous reviewers for their valuable feedback.

References

- [1] CREMI MICCAI Challenge on circuit reconstruction from Electron Microscopy Images. <https://cremi.org>. 6, 7
- [2] Ahmed Abbas and Paul Swoboda. Combinatorial Optimization for Panoptic Segmentation: A Fully Differentiable Approach. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 7
- [3] Amir Alush and Jacob Goldberger. Ensemble segmentation using efficient integer linear programming. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1966–1977, 2012. 2
- [4] Amir Alush and Jacob Goldberger. Break and conquer: Efficient correlation clustering for image segmentation. In *International Workshop on Similarity-Based Pattern Recognition*, pages 134–147. Springer, 2013. 1
- [5] Bjoern Andres, Jörg H. Kappes, Thorsten Beier, Ullrich Köthe, and Fred A. Hamprecht. Probabilistic image segmentation with closedness constraints. In *ICCV*, 2011. 1
- [6] Bjoern Andres, Thorben Kröger, Kevin L. Briggman, Winfried Denk, Natalya Korogod, Graham Knott, Ullrich Köthe, and Fred A. Hamprecht. Globally optimal closed-surface segmentation for connectomics. In *ECCV*, 2012. 1
- [7] Bjoern Andres, Julian Yarkony, BS Manjunath, Steffen Kirchhoff, Engin Turetken, Charless C Fowlkes, and Hanspeter Pfister. Segmenting planar superpixel adjacency graphs wrt non-planar superpixel affinity graphs. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 266–279. Springer, 2013. 1
- [8] Bas Fagginger Auer and Rob H Bisseling. Graph coarsening and clustering on the GPU. *Graph Partitioning and Graph Clustering*, 588:223, 2012. 2
- [9] Alberto Bailoni, Constantin Pape, Steffen Wolf, Thorsten Beier, Anna Kreshuk, and Fred A Hamprecht. A generalized framework for agglomerative clustering of signed graphs applied to instance segmentation. *arXiv preprint arXiv:1906.11713*, 2019. 2
- [10] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56(1-3):89–113, 2004. 1, 4
- [11] Thorsten Beier, Björn Andres, Ullrich Köthe, and Fred A Hamprecht. An efficient fusion move algorithm for the minimum cost lifted multicut problem. In *European Conference on Computer Vision*. Springer, 2016. 1, 2
- [12] Thorsten Beier, Thorben Kroeger, Jorg H Kappes, Ullrich Kothe, and Fred A Hamprecht. Cut, glue & cut: A fast, approximate solver for multicut partitioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1, 2
- [13] Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101, 2017. 1
- [14] Sunil Chopra and Mendu R Rao. On the multiway cut polyhedron. *Networks*, 21(1):51–89, 1991. 1
- [15] Sunil Chopra and Mendu R Rao. The partition problem. *Mathematical Programming*, 59(1-3):87–115, 1993. 1, 4
- [16] Jonathan Cohen and Patrice Castonguay. Efficient graph matching and coloring on the gpu. In *GTC*. NVIDIA, 2012. 4
- [17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 2, 6, 7
- [18] Erik D Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2-3):172–187, 2006. 1, 4
- [19] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2019. 2
- [20] Jared Hoberock and Nathan Bell. Thrust: A parallel template library, 2010. Version 1.7.0. 6, 13
- [21] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted disjoint paths with application in multiple object tracking. In *International Conference on Machine Learning*, pages 4364–4375. PMLR, 2020. 1
- [22] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. Arttrack: Articulated multi-person tracking in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1
- [23] Jayadharini Jaiganesh and Martin Burtcher. A high-performance connected components implementation for GPUs. In *Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing*, pages 92–104, 2018. 4
- [24] Yu Jin and Joseph F JaJa. A high performance implementation of spectral clustering on CPU-GPU platforms. In *Parallel and Distributed Processing Symposium Workshops, 2016 IEEE International*, pages 825–834. IEEE, 2016. 2
- [25] Florian Jug, Evgeny Levinkov, Corinna Blasse, Eugene W Myers, and Bjoern Andres. Moral lineage tracing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1
- [26] Jörg Hendrik Kappes, Markus Speth, Björn Andres, Gerhard Reinelt, and Christoph Schn. Globally optimal image partitioning by multicuts. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2011. 2
- [27] Jörg Hendrik Kappes, Markus Speth, Gerhard Reinelt, and Christoph Schnörr. Higher-order segmentation via multicuts. *Computer Vision and Image Understanding*, 143:104–119, 2016. 1
- [28] Amirhossein Kardoost and Margret Keuper. Solving minimum cost lifted multicut problems by node agglomeration. In *Asian Conference on Computer Vision*, pages 74–89. Springer, 2018. 2, 7, 8
- [29] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2):291–307, 1970. 2
- [30] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjorn Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 1, 2, 3, 4, 7, 8, 14

- [31] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):140–153, 2018. 1
- [32] Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang D Yoo. Higher-order correlation clustering for image segmentation. In *Advances in neural information processing systems*, 2011. 1, 2
- [33] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1
- [34] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. In *International Workshop on Artificial Intelligence and Statistics*, pages 182–189. PMLR, 2005. 5, 11
- [35] Vladimir Kolmogorov. A new look at reweighted message passing. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):919–930, 2014. 5, 11
- [36] Thorben Kroeger, Jörg H Kappes, Thorsten Beier, Ullrich Koethe, and Fred A Hamprecht. Asymmetric cuts: Joint image labeling and partitioning. In *German Conference on Pattern Recognition*. Springer, 2014. 1
- [37] Jan-Hendrik Lange, Bjoern Andres, and Paul Swoboda. Combinatorial persistency criteria for multicut and max-cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6093–6102, 2019. 2
- [38] Jan-Hendrik Lange, Andreas Karrenbauer, and Bjoern Andres. Partial optimality and fast lower bounds for weighted correlation clustering. In *International Conference on Machine Learning*, 2018. 2, 5, 7, 8
- [39] Evgeny Levinkov, Alexander Kirillov, and Bjoern Andres. A comparative study of local search algorithms for correlation clustering. In *German Conference on Pattern Recognition*. Springer, 2017. 1
- [40] Evgeny Levinkov, Alexander Kirillov, and Bjoern Andres. A comparative study of local search algorithms for correlation clustering. In *GCPR*, 2017. 2, 7, 8
- [41] Evgeny Levinkov, Jonas Uhrig, Siyu Tang, Mohamed Omran, Eldar Insafutdinov, Alexander Kirillov, Carsten Rother, Thomas Brox, Bernt Schiele, and Bjoern Andres. Joint graph decomposition & node labeling: Problem, algorithms, applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1
- [42] Jovita Lukasik, Margret Keuper, Maneesh Singh, and Julian Yarkony. A benders decomposition approach to correlation clustering. In *2020 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC) and Workshop on Artificial Intelligence and Machine Learning for Scientific Applications (AI4S)*. IEEE, 2020. 2
- [43] Maxim Naumov and Timothy Moon. Parallel spectral graph partitioning. *tech. rep., NVIDIA tech. rep*, 2016. 2
- [44] Sebastian Nowozin and Stefanie Jegelka. Solution stability in linear programming relaxations: Graph partitioning and unsupervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 769–776, 2009. 2
- [45] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. CUDA, release: 11.2, 2021. 6
- [46] Xinghao Pan, Dimitris Papailiopoulos, Samet Oymak, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Parallel correlation clustering on big graphs. In *Advances in Neural Information Processing Systems*, 2015. 2
- [47] Constantin Pape. torch-em. <https://github.com/constantinpape/torch-em>, 2021. 7
- [48] Constantin Pape, Thorsten Beier, Peter Li, Viren Jain, Davi D Bock, and Anna Kreshuk. Solving large multicut problems for connectomics via domain decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2, 6
- [49] Jie Song, Bjoern Andres, Michael J Black, Otmar Hilliges, and Siyu Tang. End-to-end learning for graph decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1
- [50] Paul Swoboda and Bjoern Andres. A message passing algorithm for the minimum cost multicut problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4, 5
- [51] Paul Swoboda, Andrea Hornakova, Paul Roetzer, and Ahmed Abbas. Structured Prediction Problem Archive. *arXiv preprint arXiv:2202.03574*, 2022. 6
- [52] Paul Swoboda, Jan Kuske, and Bogdan Savchynskyy. A dual ascent framework for lagrangean decomposition of combinatorial problems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5
- [53] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1
- [54] Siddharth Tourani, Alexander Shekhovtsov, Carsten Rother, and Bogdan Savchynskyy. MPLP++: Fast, parallel dual block-coordinate ascent for dense graphical models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5, 11
- [55] W Hwu Wen-mei. *GPU Computing Gems Jade Edition*. Elsevier, 2011. 4
- [56] Tomas Werner. A linear programming approach to max-sum problem: A review. *IEEE transactions on pattern analysis and machine intelligence*, 29(7):1165–1179, 2007. 5, 6
- [57] Steffen Wolf, Alberto Bailoni, Constantin Pape, Nasim Rahaman, Anna Kreshuk, Ullrich Köthe, and Fred A Hamprecht. The mutex watershed and its objective: Efficient, parameter-free graph partitioning. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2
- [58] Julian Yarkony, Alexander Ihler, and Charless C Fowlkes. Fast planar correlation clustering for image segmentation. In *European Conference on Computer Vision*. Springer, 2012. 1, 2
- [59] Zhihao Zheng, J Scott Lauritzen, Eric Perlman, Camenzind G Robinson, Matthew Nichols, Daniel Milkie, Omar Torrens, John Price, Corey B Fisher, Nadiya Sharifi, et al. A complete electron microscopy volume of the brain of adult *Drosophila melanogaster*. *Cell*, 174(3):730–743, 2018. 6