A Benders Decomposition Approach to Correlation Clustering

SC20 Workshop: Machine Learning in HPC Environments


Jovita Lukasik, Margret Keuper
Data and Web Science Group
University of Mannheim


Maneesh Singh, Julian Yarkony
Verisk

Agenda

Motivation

Related Work

Background on Correlation Clustering

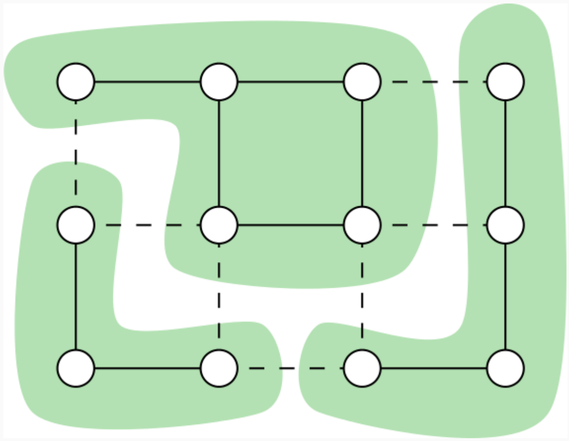
Benders Decomposition for Correlation Clustering

Experiments

Motivation

Motivation

- Many computer vision tasks involve partitioning a set of observations into unique entities
- Correlation clustering as a powerful formulation for this long-standing problem
- CC can be applied in a wide range of different applications



Advantages:

- No need to determine the amount of segments beforehand
- Size of segments does not matter

Disadvantages:

- Optimization based on LP with cutting planes approaches, which do not scale easily to large problems

We propose a Benders Decomposition Approach to Correlation Clustering, which allows for efficient optimization and possible massively parallel computations

Superpixels

Correlation clustering used in image segmentation on superpixel graphs

Figure 1: An image preprocessed into a set of superpixels (Kappes et al., 2011)

Correlation Clustering for Image Segmentation

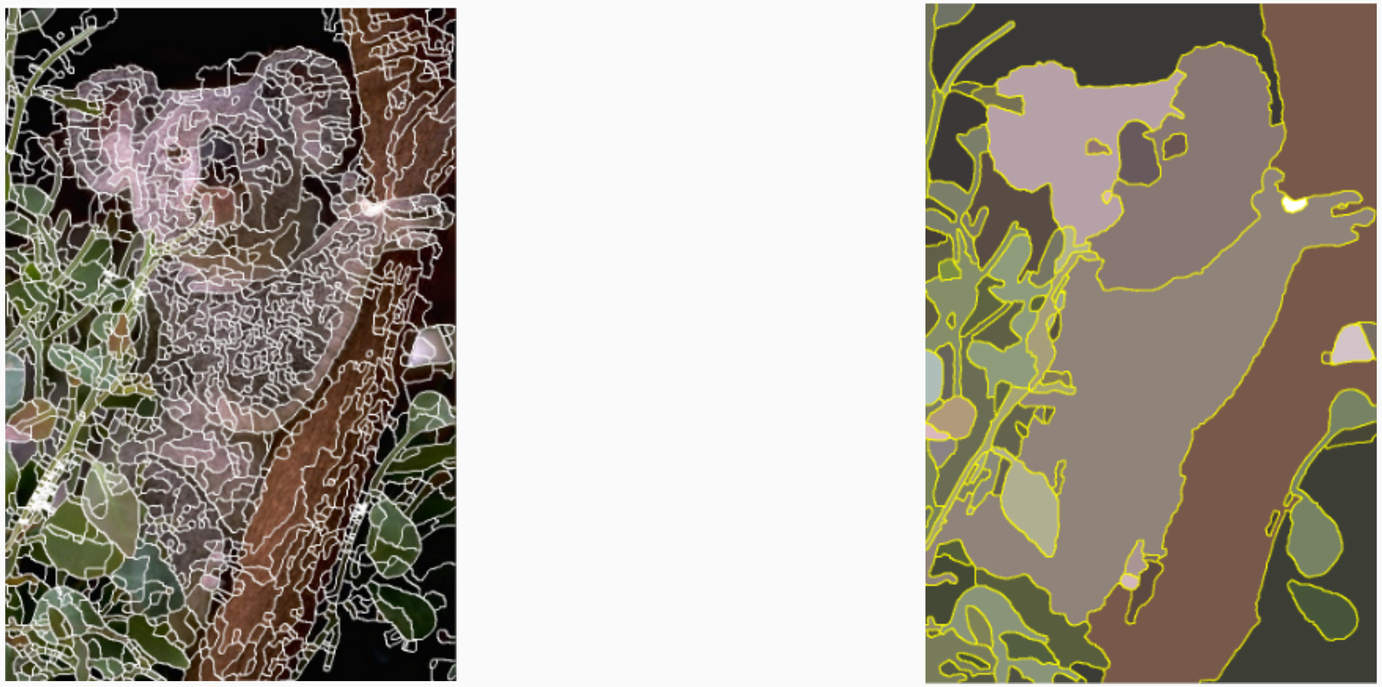
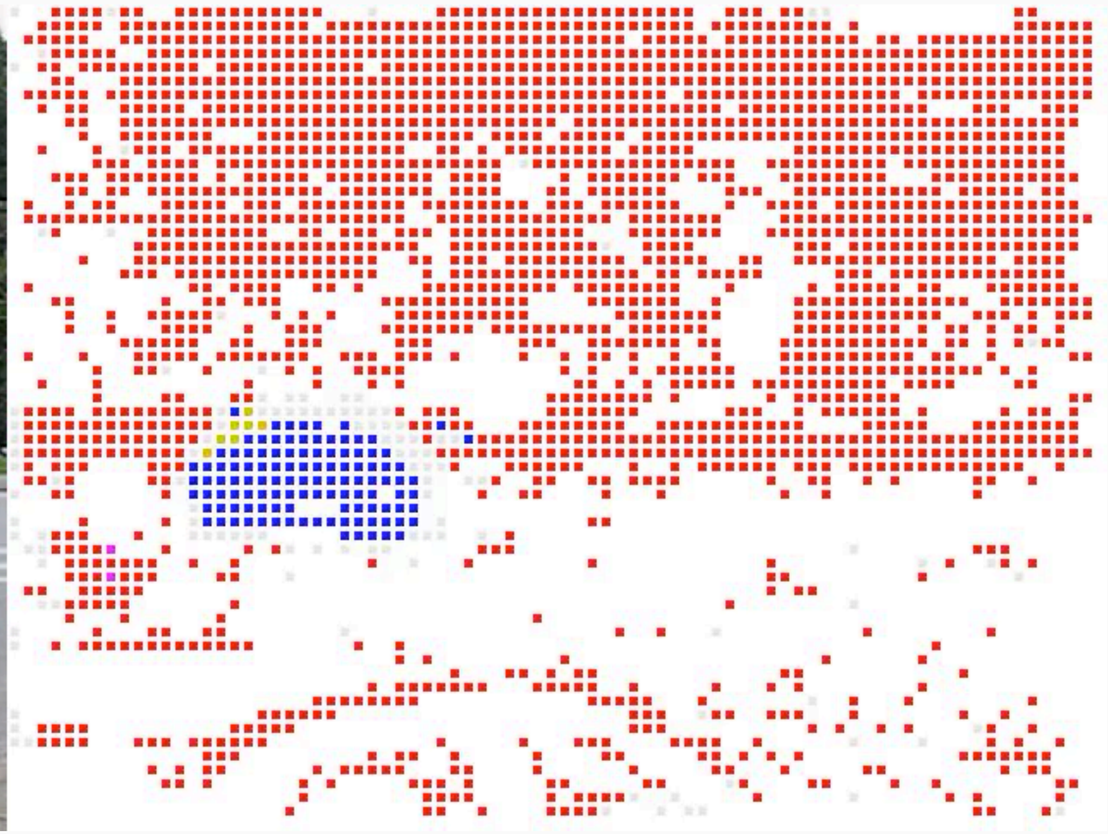


Figure 2: The curves that separate superpixels are shown in white (left). Image segmentation (right) (Andres et al., 2011).

Correlation clustering on superpixels uses ILP solver to find the optimal solution. This approach does not scale easily and is not easy paralellizable.

Correlation Clustering for Motion Segmentation



[0]Acknowledgement to Amirhossein Kardoost for providing this video.

Correlation Clustering for Tracking

[0]Acknowledgement to Kalun Ho for providing this video.

Related Work

Related Work Correlation Clustering

Correlation clustering for problems in computer vision

- Classical work of Andres et al. (2011) models image segmentation as CC, where nodes correspond to superpixels. Optimization by means of ILP solver
- Kim et al., 2011 extends CC to include higher-order cost terms over node sets
- Yarkony, Ihler, and Fowlkes (2012) tackles CC for planar graph structures problems by introducing a column generation approach, where the pricing problem corresponds to finding the lowest 2-colorable partition of the graph
- Tang et al. (2015) tackles multi-object tracking using formulation closely related to CC
- Insafutdinov et al. (2016) and Pishchulin et al. (2016) build on Tang et al., 2015 to formulate multi-person pose estimation using CC

Related Work Benders decomposition

- Classical work in operation research (Benders, 1962)
- We base our approach on the work of Cordeau et al. (2001), which solves a MILP over a set of fixed charge variables and a larger set of fractional variables associated with constraints
- MWR (Magnanti and Wong, 1981) are used in our approach instead of standard Benders rows
- BD in computer vision for the purpose of multi-person pose estimation (Wang, Kording, and Yarkony, 2017; Wang et al., 2018; Yarkony and Wang, 2018).

Background on Correlation Clustering

Correlation Clustering

Definition

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $v \in \mathcal{V}$ and undirected edges $(v_i, v_j) \in \mathcal{E}$. Given a label $x_{v_i v_j} \in \{0, 1\}$, with $x_{v_i v_j} = 1$ indicating nodes v_i, v_j are in separate components, zero otherwise. Given an edge weight $\phi_{v_i v_j} \in \mathbb{R}$, the binary edge labeling problem is to find an edge label $\mathbf{x} = (x_{v_i v_j}) \in \{0, 1\}^{|\mathcal{E}|}$, for which the total weight of the cut edges is minimized:

$$\min_{\mathbf{x} \in \{0,1\}^{|\mathcal{E}|}} \sum_{(v_i, v_j) \in \mathcal{E}^-} -\phi_{v_i v_j}(1 - x_{v_i v_j}) + \sum_{(v_i, v_j) \in \mathcal{E}^+} \phi_{v_i v_j} x_{v_i v_j} \tag{CC$_1$}$$

$$\text{s.t.} \sum_{(v_i, v_j) \in \mathcal{E}_c^+} x_{v_i v_j} \geqslant x_{v_i^c v_j^c} \quad \forall c \in \mathcal{C}, \tag{1}$$

where $\mathcal{E}^-, \mathcal{E}^+$ denote the subsets of $\mathcal{E}$, for which the weight $\phi_{v_i v_j}$ is negative and non-negative, respectively, $\mathcal{C}$ is the set of undirected cycles in $\mathcal{E}$ containing exactly one member of $\mathcal{E}^-$, (v_i^c, v_j^c) is the edge in $\mathcal{E}^-$ associated with cycle c and $\mathcal{E}_c^+ \subseteq \mathcal{E}^+$ associated with cycle c.

Standard Correlation Clustering

- Objective in (CC_1) is to minimize the total weight of the cut edges
- (1) referred as *cycle inequalities*, ensure that within every cycle of $\mathcal{G}$, the number of cut edges cannot be exactly one.
- Solving (CC_1) isintractable $\rightarrow$ Andres et al. (2011) generate solution by alternating btw. solving the ILP and adding new constraints from the set of currently violated cycle inequalities.

Benders Decomposition for Correlation Clustering

Benders Decomposition

- We apply Benders decomposition from operations research (Benders, 1962)
- Benders decomposition partitions the variables in the MILP between a master problem and a set of subproblems
- No row of the constraints in the MILP contains variables from more than one subproblem
- Variables are explicitly enforced to be integral lying only in the master problem
- optimization in Benders decomposition uses cutting plane algorithm:
  - master problem solves optimization over its variables by generating benders rows
  - solving subproblems can be done in parallel and provides primal/dual solutions
  - $\rightarrow$ solutions of subproblems give a tight lower bound for the master problem

Benders Decomposition for Correlation Clustering

Using Benders decomposition for CC, the following adaptions are needed:

- define a minimal vertex cover $\mathcal{E}^-$ with $\mathcal{S} \in \mathcal{V}$ indexed by v_s
- bender subproblem is associated with $s \in \mathcal{S}$, where v_s is referred as the root of s.
- edges in the subproblem: $\mathcal{E}_s^- \subset \mathcal{E}^-$ associated with subproblem s.
- cycle inequalities in the subproblem: $\mathcal{C}_s \subset \mathcal{C}$ containing edges in $\mathcal{E}_s^-$

We furthermore assume, $\mathcal{S}$ is given.

Benders Decomposition for Correlation Clustering

- Rewrite (CC_1) by means of an auxiliary fct. $Q(\phi, s, \mathbf{x})$

$$(\mathrm{CC}_1) \rightsquigarrow (\mathrm{CC}_2): \min_{\mathbf{x}\in\{0,1\}^{|\mathcal{E}|}} \sum_{(v_i,v_j)\in\mathcal{E}^-} -\phi_{v_iv_j}(1-x_{v_iv_j}) + \sum_{(v_i,v_j)\in\mathcal{E}^+} \phi_{v_iv_j}x_{v_iv_j} + \sum_{s\in\mathcal{S}} Q(\phi, s, \mathbf{x}), \tag{CC_2}$$

where $Q(\phi, s, \mathbf{x})$ is defined as follows.

$$\begin{aligned} Q(\phi, s, \mathbf{x}) = \min_{\mathbf{x}^s\in\{0,1\}^{|s|}} & \sum_{(v_i,v_j)\in\mathcal{E}_s^-} -\phi_{v_iv_j}(1-x^s_{v_iv_j}) + \sum_{(v_i,v_j)\in\mathcal{E}^+} \phi_{v_iv_j}x^s_{v_iv_j} \\ \text{s.t.} \quad & \sum_{(v_i,v_j)\in\mathcal{E}_c^+} x_{v_iv_j} + x^s_{v_iv_j} \geqslant x_{v_i^cv_j^c} - (1-x^s_{v_i^cv_j^c}) \quad \forall c \in \mathcal{C}_s. \end{aligned} \tag{2}$$

Duality Theory for Benders Subproblem

- Goal: find a lower bound for the master problem in the Bender decomposition approach (CC_2)
- rewriting (2) as a primal/dual LP

Let:

$$f^s_{v_i v_j} = \begin{cases} 1, & \text{for } (v_i, v_j) \in \mathcal{E}^+, \text{if } (v_i, v_j) \text{ is cut in } \mathbf{x}^s, \text{ but is not cut in } \mathbf{x} \\ 1, & \text{for } (v_i, v_j) \in \mathcal{E}^-_s, \text{if } (v_i, v_j) \text{ is not cut in } \mathbf{x}^s, \text{ but is cut in } \mathbf{x}. \end{cases} \tag{3}$$

$$m_v = \begin{cases} 1, & \text{node } v \in \mathcal{V} \text{ not in the component associated with } v_s \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Primal Subproblem

- Given binary $\mathbf{x}$, we only need to enforce that f, m are non-negative to ensure existence of optimizing solution with f, m being binary.

$$
\begin{aligned}
Q(\phi, s, \mathbf{x}) \quad = \quad & \min_{\substack{f^s_{v_i v_j} \geqslant 0 \\ m_v \geqslant 0}} \sum_{(v_i, v_j) \in \mathcal{E}^+} \phi_{v_i v_j} f^s_{v_i v_j} - \sum_{(v_s, v) \in \mathcal{E}^-_s} \phi_{v_s v} f^s_{v_s v} \qquad (5) \\
\lambda^-_{v_i v_j} \quad : \quad & m_{v_i} - m_{v_j} \leqslant x_{v_i v_j} + f^s_{v_i v_j} \quad \forall (v_i, v_j) \in (\mathcal{E}^+ \setminus \mathcal{E}^+_s), \\
\lambda^+_{v_i v_j} \quad : \quad & m_{v_j} - m_{v_i} \leqslant x_{v_i v_j} + f^s_{v_i v_j} \quad \forall (v_i, v_j) \in (\mathcal{E}^+ \setminus \mathcal{E}^+_s), \\
\psi^-_v \quad : \quad & x_{v_s v} - f^s_{v_s v} \leqslant m_v \quad \forall (v_s, v) \in \mathcal{E}^-_s, \\
\psi^+_v \quad : \quad & m_v \leqslant x_{v_s v} + f^s_{v_s v} \quad \forall (v_s, v) \in \mathcal{E}^+_s,
\end{aligned}
$$

Dual Subproblem

$$\max_{\substack{\lambda \geqslant 0 \\ \psi \geqslant 0}} \; - \sum_{(v_i, v_j) \in (\mathcal{E}^+ \setminus \mathcal{E}_s^+)} (\lambda^-_{v_i v_j} + \lambda^+_{v_i v_j}) x_{v_i v_j} + \sum_{(v_s, v) \in \mathcal{E}_s^-} \psi^-_v x_{v_s v} - \sum_{(v_s, v) \in \mathcal{E}_s^+} \psi^+_v x_{v_s v} \tag{6}$$

$$\text{s.t.} \quad \psi^+_{v_i} \, \mathbb{1}_{\mathcal{E}_s^+}(v_s, v_i) - \psi^-_{v_i} \, \mathbb{1}_{\mathcal{E}_s^-}(v_s, v_i) + \sum_{\substack{v_j \\ (v_i, v_j) \in (\mathcal{E}^+ \setminus \mathcal{E}_s^+)}} (\lambda^-_{v_i v_j} - \lambda^+_{v_i v_j}) + \sum_{\substack{v_j \\ (v_j, v_i) \in (\mathcal{E}^+ \setminus \mathcal{E}_s^+)}} (\lambda^+_{v_j v_i} - \lambda^-_{v_j v_i}) \geqslant 0 \qquad \forall v_i \in \mathcal{V} - v_s$$

$$\begin{aligned} -\phi_{v_s v} - \psi^-_v &\geqslant 0 \qquad \forall (v_s, v) \in \mathcal{E}_s^- \\ \phi_{v_s v} - \psi^+_v &\geqslant 0 \qquad \forall (v_s, v) \in \mathcal{E}_s^+ \\ \phi_{v_i v_j} - (\lambda^-_{v_i v_j} + \lambda^+_{v_i v_j}) &\geqslant 0 \qquad \forall (v_i, v_j) \in (\mathcal{E}^+ \setminus \mathcal{E}_s^+). \end{aligned}$$

Primal/Dual Linear Programming

- Any dual feasible solution for the dual problem describes an affine function of $\mathbf{x}$: tight lower bound on $Q(\phi, s, \mathbf{x})$
- Let $\mathcal{Z}$ denoted as the set of all dual feasible solutions across the subproblems $s \in \mathcal{S}$
- Rewrite (CC_2) using $\mathcal{Z}$

$$(\mathrm{CC}_2) \rightsquigarrow (\mathrm{CC}_3): \min_{\mathbf{x}\in\{0,1\}^{|\mathcal{E}|}} \sum_{(v_i,v_j)\in\mathcal{E}^+} \phi_{v_i v_j} x_{v_i v_j} - \sum_{(v_i,v_j)\in\mathcal{E}^-} (1 - x_{v_i v_j})\phi_{v_i v_j} \qquad (\mathrm{CC}_3)$$

$$\text{s.t.} \sum_{(v_i,v_j)\in\mathcal{E}} x_{v_i v_j} \omega^z_{v_i v_j} \leqslant 0 \qquad \forall z \in \mathcal{Z},$$

where $\omega^z_{v_i v_j}$ is associated with the dual variables λ, ψ for the $x_{v_i v_j}$ term,

Cutting Plane Method

- intractable numbers of constraints leading to an infinite set of feasible solutions, making optimizing (CC_3) intractable.
- $\rightarrow$ cutting plane method to construct a small sufficient set $\hat{\mathcal{Z}}$
- Iterate between solving LP relaxation of (CC_3) over $\hat{\mathcal{Z}}$ and generating new Benders rows until no violated constraints exists.

Magnanti-Wong Benders Rows

- Benders rows given by the dual subproblem provides a tight bound at $\mathbf{x}^*$, with $\mathbf{x}^*$ being the master problem's solution used to generate the Benders row.
- ideally, the Benders row can provide a good lower bound for a large set of $\mathbf{x}$, while being tight at $\mathbf{x}^*$.
- modifying the dual subproblem by replacing the objective and adding a term (including a tolerance τ) leads to the desired goal

$$\tau Q(\phi, s, \mathbf{x}) \leqslant -\sum_{(v_i,v_j)\in(\mathcal{E}^+\setminus\mathcal{E}_s^+)} (\lambda^-_{v_iv_j} + \lambda^+_{v_iv_j}) x_{v_iv_j} + \sum_{(v_s,v)\in\mathcal{E}_s^-} \psi_v^- x_{v_sv} - \sum_{(v_s,v)\in\mathcal{E}_s^+} \psi_v^+ x_{v_sv}, \quad (7)$$

Experiments

Experiments

Demonstrate value of BDCC on CC problem instances for image segmentation on the benchmark Berkeley Segmentation Data Set (BSDS) (Martin et al., 2001).

- BDCC solves CC instances for image segmentation
- BDCC successfully exploits parallelization

Experiments- Image Segmentation

- Demonstrate effectiveness of BDCC with various τ for different problem difficulties.
- presence of MWR accelerates optimization
- exact τ does not affect the speed of optimization dramatically.
- parallel processing time assumes one CPU for each subproblem.

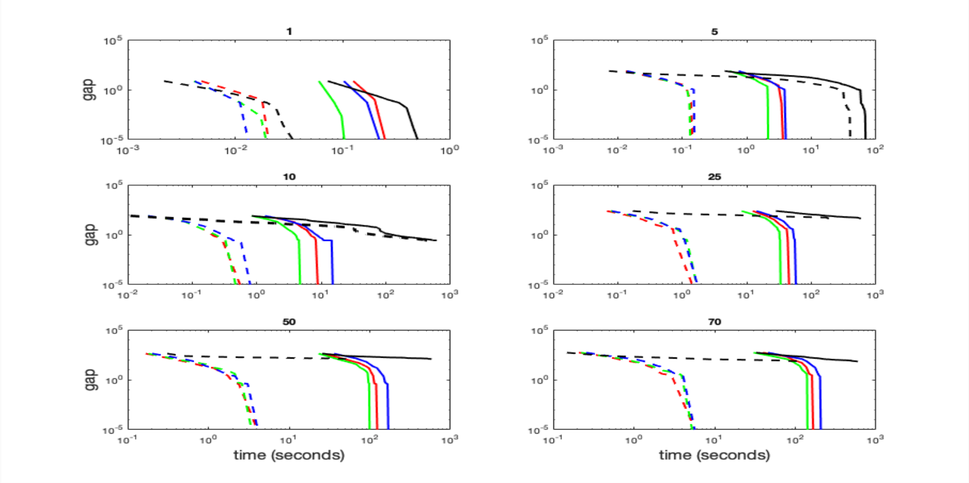


Figure 3: Gap between upper and lower bounds as a fct. of time for various values of τ. We use red,green,blue for $\tau = [0.5, 0.99, .01]$,black for not using MWR. We show both the computation time with and without exploiting parallelization of subproblems with dotted and solid lines, respectively. We use titles to indicate the approximate difficulty of the problem as ranked by input file size of 100 files.

Experiments- Parallelization

- Demonstrate the speed-up induced by the use of parallelization
- if MWR are not used total CPU time prohibitively large

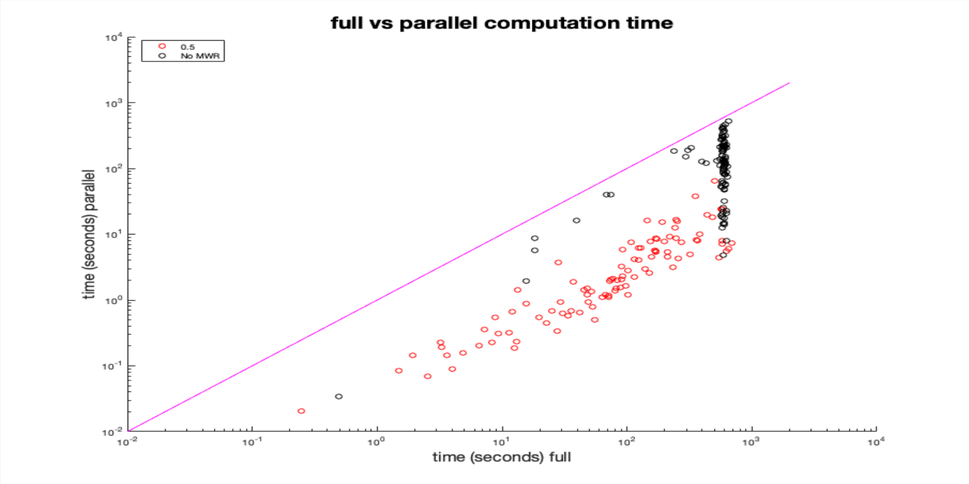


Figure 4: Comparison of benefits of parallelization and MWR across our data set. We scatter plot the total running time versus the total running time when solving each subproblem is done on its own CPU across problem instances. We use red to indicate $\tau = 0.5$ and black to indicate that MWR are not used. We draw a line with slope=1 in magenta to better enable appreciation of the red and black points. NOTE: The time spent generating Benders rows, in a given iteration of BDCC when using parallel processing, is the maximum time spent to solve any sub-problem for that iteration.

- Andres, B. et al. (2011). “Probabilistic image segmentation with closedness constraints”. In: *Proceedings of the Fifth International Conference on Computer Vision (ICCV-11)*, pp. 2611–2618.
- Benders, J. F. (1962). “Partitioning procedures for solving mixed-variables programming problems”. In: *Numerische mathematik* 4.1, pp. 238–252.
- Cordeau, J.-F. et al. (2001). “Benders decomposition for simultaneous aircraft routing and crew scheduling”. In: *Transportation science* 35.4, pp. 375–388.
- Insafutdinov, E. et al. (2016). “Deepercut: A deeper, stronger, and faster multi-person pose estimation model”. In: *European Conference on Computer Vision*. Springer, pp. 34–50.
- Kappes, J. H. et al. (2011). “Globally Optimal Image Partitioning by Multicuts”. In: *Energy Minimazation Methods in Computer Vision and Pattern Recognition - 8th International Conference, EMMCVPR 2011, St. Petersburg, Russia, July 25-27, 2011. Proceedings*. Vol. 6819. Lecture Notes in Computer Science. Springer, pp. 31–44.
- Kim, S. et al. (2011). “Higher-Order Correlation Clustering for Image Segmentation”. In: *Advances in Neural Information Processing Systems,25*, pp. 1530–1538.

- Magnanti, T. L. and R. T. Wong (1981). "Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria". In: *Operations research* 29.3, pp. 464–484.
- Martin, D. et al. (2001). "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics". In: *Proceedings of the Eighth International Conference on Computer Vision (ICCV-01)*, pp. 416–423.
- Pishchulin, L. et al. (2016). "Deepcut: Joint subset partition and labeling for multi person pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4929–4937.
- Tang, S. et al. (2015). "Subgraph Decomposition for Multi-Target Tracking". In: *CVPR*.
- Wang, S., K. Kording, and J. Yarkony (2017). "Exploiting skeletal structure in computer vision annotation with Benders decomposition". In: *arXiv preprint arXiv:1709.04411*.
- Wang, S. et al. (2018). "Accelerating Dynamic Programs via Nested Benders Decomposition with Application to Multi-Person Pose Estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 652–666.


Yarkony, J., A. Ihler, and C. Fowlkes (2012). "Fast Planar Correlation Clustering for Image Segmentation". In: *Proceedings of the 12th European Conference on Computer Vision(ECCV 2012)*.


Yarkony, J. and S. Wang (2018). "Accelerating Message Passing for MAP with Benders Decomposition". In: *arXiv preprint arXiv:1805.04958*.

Thank you!