Accelerating Column Generation via Flexible Dual Optimal Inequalities with Application to Entity Resolution

Vishnu Suresh Lokhande,[1,3] Shaofei Wang,[2] Maneesh Singh,[3] Julian Yarkony[3]
[1]University of Wisconsin-Madison
[2]University of Pennsylvania
[3]Verisk Computational and Human Intelligence Laboratory

Abstract

In this paper, we introduce a new optimization approach to Entity Resolution. Traditional approaches tackle entity resolution with hierarchical clustering, which does not benefit from a formal optimization formulation. In contrast, we model entity resolution as correlation-clustering, which we treat as a weighted set-packing problem and write as an integer linear program (ILP). In this case, sources in the input data correspond to elements and entities in output data correspond to sets/clusters. We tackle optimization of weighted set packing by relaxing integrality in our ILP formulation. The set of potential sets/clusters can not be explicitly enumerated, thus motivating optimization via column generation. In addition to the novel formulation, we also introduce new dual optimal inequalities (DOI), that we call flexible dual optimal inequalities, which tightly lower-bound dual variables during optimization and accelerate column generation. We apply our formulation to entity resolution (also called de-duplication of records), and achieve state-of-the-art accuracy on two popular benchmark datasets. Our F-DOI can be extended to other weighted set-packing problems.

1 Introduction

In this paper we study the problem of entity resolution. Entity resolution aims to eliminate redundant information from multiple sources of data. This task plays a key role in information integration, natural language understanding, information processing on the World-Wide Web all of which are core areas of AI (Konda et al. 2016).

The Problem Setup Given a dataset of observations each associated with up to one object, entity resolution aims to pack (or partition) the observations into groups called hypothesis (or entities) such that there is a bijection from hypotheses to unique entities in the dataset. We are provided a set of observations called records, where each record is associated with a subset of fields (for example: name, social security number, phone number etc). We seek to partition the observations into hypothesis so that: **(1)** all observations of any real world entity are associated with exactly one selected hypothesis; **(2)** each selected hypothesis is associated with observations of exactly one real world entity.

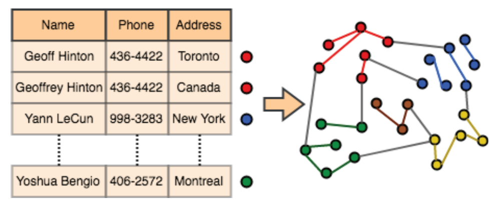

Figure 1: Entity Resolution posed as MWSP: Given a structured (tabular) data, we study the problem of entity resolution where we group rows representing the same real-world entity into the same cluster. Entity Resolution can be viewed as a node-clustering problem over a graph. In the example used in this figure, we identify similar rows with identical colored dots and then tranform it into a graph with nodes representing the entities/rows and edges represented some measure of similarity between between the nodes. Nodes that are not connected are understood to represent distict rows in the table.

Traditional Approaches Entity resolution has been studied using different clustering approaches (Saeedi, Peukert, and Rahm 2017). It is common to transform entity resolution to a graph problem and run a clustering algorithms on top of it as depicted in Figure 1. The popular clustering algorithms developed to attack entity resolution are ConCom, where the algorithm is based on computing the connected components of the input graph. Center clustering sequentially adds edges from a priority queue and either assigns the nodes to a cluster or tags them as a center (Hassanzadeh and Miller 2009). Star clustering (Aslam, Pelekhov, and Rus 2004), in a similar way, prioritizes in adding those nodes to a cluster that have the highest degree. Correlation Clustering (Bansal, Blum, and Chawla 2004), which forms the backbone of our method, has also been studied for entity resolution problem. However, the lengthy and numerous iterations to converge made it difficult for entity resolution problems (Saeedi, Peukert, and Rahm 2017).

Entity Resolution as MWSP Contrary to previous works, we propose to tackle entity resolution as an optimization problem, formulating it as a minimum weight set packing (MWSP) problem. The set of all possible hypotheses is the power set of the set of the observations. The real valued cost of a hypothesis, is a second order function of the observations that compose the hypothesis. The cost of a hypothesis decreases as the similarity among the observations in the hypothesis increases. Any non-overlapping subset of all possible hypotheses corresponds to a partition; we treat each observation not in any element in the subset as being in a hypothesis by itself. We model the quality of a packing as the total cost of the hypothesis in the packing. The lowest total cost packing is empirically a good approximation to the ground truth.

Efficient MWSP using Column Generation Enumerating the power set of the observations is often not possible in practice, thus motivating us to tackle MWSP using column generation (CG) (Gilmore and Gomory 1961; Barnhart et al. 1996; Desrosiers and Lübbecke 2005; Lübbecke 2010; Yarkony et al. 2019). CG solves a linear programming (LP) relaxation of MWSP by constructing a small sufficient subset of the power set, such that solving the LP relaxation over the sufficient subset provably provides the same solution as solving the LP relaxation over the entire power set. CG can often be accelerated using dual optimal inequalities (DOIs) (Ben Amor, Desrosiers, and Valério de Carvalho 2006), which bound the otherwise unbounded dual variables of the LP-relaxation, drastically reducing the search space of the LP problem. The use of DOI provably does not alter the solution produced at termination of CG.

Core Contribution We make the following contributions to the scientific literature. **(1).** Introduce a novel MWSP formulation for entity resolution, that achieves efficient exact/approximate optimization using CG. **(2).** Introduce novel DOIs called Flexible DOIs (F-DOI), which can be applied to broad classes of MWSP problems.

Paper Organization The paper is structured as follows. In Section 2 we review the integer linear programming (ILP) formulation of MWSP, and its solution via CG. In Section 3 we introduce F-DOIs. In Section 4 we devise optimization algorithms to solve entity resolution problem via CG and F-DOIs. In Section 5 we demonstrate the effectiveness of our approach on benchmark entity resolution datasets. In Section 6 we conclude.

2 Preliminaries

In this section we review the MWSP formulation and CG solution of (Yarkony et al. 2019). We outline this section as follows. In Section 2.1 we review the ILP formulation of MWSP. In Section 2.2 we review the CG algorithm that solves an LP relaxation of the ILP formulation. In Section 2.3 we review the varying DOIs introduced in (Yarkony et al. 2019). To be consistent with the notation used in the operations research community, we use the notation of (Yarkony et al. 2019) throughout this paper.

2.1 An ILP Formulation of MWSP

Observations An observation corresponds to an element in the traditional set-packing context and a data source in the entity resolution context. We use $\mathcal{D}$ to denote the set of observations, which we index by d.

Hypotheses A hypothesis corresponds to a set in the traditional set-packing context, and an entity in the entity resolution context. Given a set of observations $\mathcal{D}$, the set of all hypotheses is the power set of $\mathcal{D}$, which we denote as $\mathcal{G}$ and index by g.

We describe $\mathcal{G}$ using matrix $G \in \{0,1\}^{|\mathcal{D}|\times|\mathcal{G}|}$. Here $G_{dg} = 1$ if and only if hypothesis g includes observation d, and otherwise $G_{dg} = 0$. A real valued cost Γ_g is associated to each $g \in \mathcal{G}$, where Γ_g is the cost of including g in our packing. The hypothesis g containing no observations is defined to have cost $\Gamma_g = 0$. Γ_g is instantiated as a function of G_{dg}. For example, let $d_1 \in \mathcal{D}, d_2 \in \mathcal{D}$ and let $\theta_{d_1 d_2} \in \mathbb{R}$ denote the cost of putting d_1, d_2 in a single hypothesis g, then we could write $\Gamma_g = \sum_{d_1,d_2} \theta_{d_1 d_2} G_{d_1 g} G_{d_2 g}$. We will discuss more on the cost (Γ_g) formulation, in the light of entity resolution, in Section 4.2.

A packing is described using $\gamma \in \{0,1\}^{|\mathcal{G}|}$ where $\gamma_g = 1$ indicates that hypothesis g is included in the solution, and otherwise $\gamma_g = 0$. MWSP is written as an ILP below.

$$\min_{\gamma\in\{0,1\}^{|\mathcal{G}|}} \sum_{g\in\mathcal{G}} \Gamma_g \gamma_g \quad \text{s.t.} \sum_{g\in\mathcal{G}} G_{dg}\gamma_g \leq 1 \quad \forall d \in \mathcal{D} \tag{1}$$

The constraints in Eq 1 enforce that no observation is included in more than one selected hypothesis in the packing.

2.2 Solving MWSP via Column Generation

Column Generation Algorithm Solving Eq 1 is challenging for two key reasons: **(1)** MWSP is NP-hard (Karp 1972); **(2)** $\mathcal{G}$ is too large to be considered in optimization for our problems. To tackle **(1)**, the integrality constraints on γ are relaxed, resulting in an LP:

$$\text{Eq } 1 \geq \min_{\gamma\geq 0} \sum_{g\in\mathcal{G}} \Gamma_g \gamma_g \quad \text{s.t.} \sum_{g\in\mathcal{G}} G_{dg}\gamma_g \leq 1 \quad \forall d \in \mathcal{D} \tag{2}$$

(Yarkony et al. 2019) demonstrates that **(2)** can be circumvented by using column generation (CG). Specifically, the CG algorithm constructs a small sufficient subset of $\mathcal{G}$, (which is denoted $\hat{\mathcal{G}}$ and initialized empty) s.t. an optimal solution to Eq 2 exists for which only hypothesis in $\hat{\mathcal{G}}$ are used. Thus CG avoids explicitly enumerating $\mathcal{G}$, which grows exponentially in $|\mathcal{D}|$. The primal-dual optimization over $\hat{\mathcal{G}}$, which is referred to as the restricted master problem (RMP),

Algorithm 1 MWSP via Column Generation

```
1: Ĝ ← ∅
2: repeat
3:     γ, λ ← Solve the RMP in Eq 3–4
4:     g* ← Solve the pricing problem in Eq 5
5:     Ĝ ← Ĝ ∪ {g*}
6: until Γ_{g*} − Σ_{d∈D} λ_d G_{dg*} ≥ 0
7: γ ← Solve MWSP in Eq (1) over Ĝ instead of G
8: Return γ
```

is written as:

$$\min_{\gamma \geq 0} \quad \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g \tag{3}$$
$$\text{s.t.} \quad \sum_{g \in \hat{\mathcal{G}}} G_{dg}\gamma_g \leq 1 \quad \forall d \in \mathcal{D}$$
$$= \max_{\lambda \leq 0} \quad \sum_{d \in \mathcal{D}} \lambda_d \tag{4}$$
$$\text{s.t.} \quad \Gamma_g - \sum_{d \in \mathcal{D}} G_{dg}\lambda_d \geq 0 \quad \forall g \in \hat{\mathcal{G}}$$

The CG algorithm is described in Alg 1. CG solves the MWSP problem by alternating between: **(1)** solving the RMP in Eq 4 given $\hat{\mathcal{G}}$ (Alg 1, line 3) and **(2)** Adding hypothesis in $\mathcal{G}$ to $\hat{\mathcal{G}}$, that have negative reduced cost given dual variables λ (Alg 1,line 4). The selection of the lowest reduced cost hypothesis in $\mathcal{G}$ is referred to as pricing, and is formally defined as:

$$\min_{g \in \mathcal{G}} \Gamma_g - \sum_{d \in \mathcal{D}} \lambda_d G_{dg} \tag{5}$$

Solving Eq 5 is typically tackled using a specialized solver exploiting specific structural properties of the problem domain (Gilmore and Gomory 1961; Wang et al. 2018; Zhang et al. 2017). In many problem domains pricing algorithms return multiple negative reduced cost hypothesis in $\mathcal{G}$. In these cases some or all returned hypotheses with negative reduced cost are added to $\hat{\mathcal{G}}$.

Convergence of Column Generation CG terminates when no negative reduced cost hypotheses remain in $\mathcal{G}$ (Alg 1,line 6). CG does not require that the lowest reduced cost hypothesis is identified during pricing to ensure that Eq 2 is solved exactly; instead, Eq 2 is solved exactly as long as a $g \in \mathcal{G}$ with negative reduced cost is produced at each iteration of CG if one exists.

If Eq 3 produces a binary valued γ at termination of CG (*i.e.* the LP-relaxation is tight) then γ is provably the optimal solution to Eq 1. However if γ is fractional at termination of CG, an approximate solution to Eq 1 can still be obtained by replacing $\mathcal{G}$ in Eq 1 with $\hat{\mathcal{G}}$ (Alg 1,line 7). (Yarkony et al. 2019) shows that Eq 2 describes a tight relaxation in practice; We refer readers interested in tightening Eq 2 to (Yarkony et al. 2019), which achieve this using subset-row inequalities (Jepsen et al. 2008).

2.3 Dual Optimal Inequalities (DOIs)

The convergence of Alg 1 often can be accelerated by providing bounds on the dual variables in Eq 4 without altering the final solution of Alg 1, thus limiting the dual space that Alg 1 searches over. We define DOI with Ξ_d which lower bounds dual variables in Eq 4 as $-\Xi_d \leq \lambda_d, \forall d \in \mathcal{D}$. The primal RMP in Eq 3 is thus augmented with new primal variables ξ, where primal variable ξ_d corresponds to the dual constraint $-\Xi_d \leq \lambda_d$.

$$\min_{\substack{\gamma \geq 0 \\ \xi \geq 0}} \quad \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g + \sum_{d \in \mathcal{D}} \Xi_d \xi_d \tag{6}$$
$$\text{s.t.} \quad -\xi_d + \sum_{g \in \hat{\mathcal{G}}} G_{dg}\gamma_g \leq 1$$
$$= \max_{-\Xi_d \leq \lambda_d \leq 0} \quad \sum_{d \in \mathcal{D}} \lambda_d \tag{7}$$
$$\text{s.t.} \quad \Gamma_g - \sum_{d \in \mathcal{D}} G_{dg}\lambda_d \geq 0 \quad \forall g \in \hat{\mathcal{G}}$$

Varying DOIs of (Yarkony et al. 2019) In the applications of (Yarkony et al. 2019), the authors observed that the removal of a small number of observations rarely causes a significant change to the cost of a hypothesis in $\hat{\mathcal{G}}$. This fact motivates the following DOIs, which are called varying DOIs.

Let $\bar{g}(g, \mathcal{D}_s)$ be the hypothesis consisting of g with all observations in $\mathcal{D}_s \subseteq \mathcal{D}$ removed. Formally, $G_{d\bar{g}(g,\mathcal{D}_s)} = G_{dg}[d \notin \mathcal{D}_s], \forall d \in \mathcal{D}$, where [] is the binary indicator function. Let ϵ be a tiny positive number. Varying DOI are computed as:

$$\Xi_d = \epsilon + \max_{g \in \hat{\mathcal{G}}} \Xi^*_{dg} \quad \forall d \in \mathcal{D} \tag{8}$$
$$\Xi^*_{dg} \geq \max_{\substack{\hat{g} \in \mathcal{G} \\ G_{\hat{d}\hat{g}} \leq G_{\hat{d}g} \forall \hat{d} \in \mathcal{D}}} \Gamma_{\bar{g}(\hat{g},\{d\})} - \Gamma_{\hat{g}}$$

Observe that Ξ_d may increase (but never decrease) over the course of CG as $\hat{\mathcal{G}}$ grows. In (Yarkony et al. 2019) the computation of Ξ^*_{dg} is done using problem specific worst case analysis for each g upon addition to $\hat{\mathcal{G}}$.

Related Approaches We now contrast DOIs from other dual stabilization methods, which also aim at accelerating CG. Dual stabilization approaches (excluding DOI) can all be understood as imposing a norm on the dual variables to prevent them from becoming extreme or leaving the area around a well established dual solution. DOI based methods, in contrast, are based on providing provable bounds on the optimal point in the dual space.
(Du Merle et al. 1999): This work optimizes the RMP with an ℓ_1 penalty on the distance from a box around the best dual solution found thus far. Here best is defined as the maximum lower bound identified thus far over the course of column generation. Variants on this approach are available and provide different schedules for weakening the ℓ_1 penalty. Other variants can replace the best solution found thus far with the most recent solution. The similar work of (Marsten, Hogan,

and Blankenship 1975) binds the dual variables to lie in a box around the previous dual solution.

(Gschwind and Irnich 2016): This work derives bounds on dual variables corresponding to swapping elements in hypotheses with other elements in the primal problem. It is appropriate for tasks such as bin packing and cutting stock where cost terms are not defined in terms of the elements that make up a set. In contrast, the varying DOI and F-DOI describe bounds on the dual variables corresponding to removing elements from hypotheses in the primal problem.

3 Flexible Dual Optimal Inequalities

A major drawback of varying DOI is that Ξ_d depends on all hypotheses in $\hat{\mathcal{G}}$ (as defined in Eq 8), while often only a small subset of $\hat{\mathcal{G}}$ are active (selected) in an optimal solution to Eq 3. Thus during Alg 1, the presence of a hypothesis in $\hat{\mathcal{G}}$ may increase the cost of the optimal solution found in current iteration, making exploration of solution space slower. This motivates us to design new DOIs that circumvent this difficulty, which we name Flexible DOIs (F-DOIs).

We outline this section as follows. In Section 3.1 we introduce a MWSP formulation using CG featuring our F-DOIs. In Section 3.2 we consider pricing under this MWSP formulation.

3.1 Formulation with F-DOIs

Given any $g \in \mathcal{G}$, let Ξ_{dg} be positive if $G_{dg} = 1$ and otherwise $\Xi_{dg} = 0$, and defined such that for all non-empty $\mathcal{D}_s \subseteq \mathcal{D}$ the following bound is satisfied.

$$\sum_{d \in \mathcal{D}_s} \Xi_{dg} \geq \epsilon + \Gamma_{\bar{g}(g,\mathcal{D}_s)} - \Gamma_g \tag{9}$$

Let $\mathcal{Z}_d$ be the set of unique positive values of Ξ_{dg} over all $g \in \hat{\mathcal{G}}$, which we index by z. We order the values in $\mathcal{Z}_d$ from smallest to largest as $[\omega_{d1}, \omega_{d2}, \omega_{d3}...]$. We describe Ξ_{dg} using $Z_{dzg} \in \{0,1\}$ where $Z_{dzg} = 1$ if and only if $\Xi_{dg} \geq \omega_{dz}$. We describe Ξ_{dg} using Ξ_{dz} as follows: $\Xi_{dz} = \omega_{dz} - \omega_{d(z-1)} \quad \forall z \in \mathcal{Z}_d, z \geq 2$; $\Xi_{d1} = \omega_{d1}$. Below we use Z to model MWSP as a primal/dual LP.

$$\min_{\substack{\gamma \geq 0 \\ \xi \geq 0}} \quad \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g + \sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} \Xi_{dz} \xi_{dz} \tag{10}$$

$$\text{s.t.} \quad -\xi_{dz} + \sum_{g \in \hat{\mathcal{G}}} Z_{dzg} \gamma_g \leq 1 \quad \forall d \in \mathcal{D}, z \in \mathcal{Z}_d$$

$$= \max_{\substack{-\Xi_{dz} \leq \lambda_{dz} \leq 0 \\ \forall d \in \mathcal{D}, z \in \mathcal{Z}_d}} \quad \sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} \lambda_{dz} \tag{11}$$

$$\text{s.t.} \quad \Gamma_g - \sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} Z_{dzg} \lambda_{dz} \geq 0 \quad \forall g \in \hat{\mathcal{G}}$$

F-DOIs are the inequalities $-\Xi_{dz} \leq \lambda_{dz}$ in Eq 11. We now prove that at termination of CG that $\xi_{dz} = 0 \, \forall d \in \mathcal{D}, z \in \mathcal{Z}_d$ and hence Eq 10=Eq 2.

Proposition: Let ζ^* and ζ^*_{DOI} be the optimal values of Eq 2- Eq 10, at termination of CG respectively. If Ξ satisfies Eq 9, then $\zeta^*_{DOI} = \zeta^*$.

Proof. Let (γ^*, ξ^*) be an optimal solution to Eq 10. If $\xi^*_{dz} = 0$ for all $d \in \mathcal{D}, z \in \mathcal{Z}_d$, then $\zeta^*_{DOI} = \zeta^*$ because γ^* is feasible and optimal for Eq 10. Otherwise, there exists an observation $d \in \mathcal{D}$, $g \in \mathcal{G}$ such that $\gamma^*_g > 0$, $G_{dg} = 1$ and $(Z_{dzg} = 1) \rightarrow (\xi^*_{dz} > 0) \quad \forall z \in \mathcal{Z}_d$. Let $z_* \leftarrow \max_{\substack{z \in \mathcal{Z}_d \\ Z_{dgz}=1}} z$. Let $\alpha = \min \{\gamma^*_g, \xi^*_{dz_*}\}$. Consider the solution obtained from (γ^*, ξ^*) by decreasing γ^*_g and ξ^*_{dz} for all $z \in \mathcal{Z}_d$ s.t. $(Z_{dgz} = 1)$ by α and increasing $\gamma^*_{\bar{g}(g,\{d\})}$ by α. We have increased the objective by $\alpha(\Gamma_{\bar{g}(g,\{d\})} - \Gamma_g - \Xi_{dg})$ which is non-positive since $\sum_{z \in \mathcal{Z}_d} \Xi_{dz} Z_{gzg} = \Xi_{dg} > \Gamma_{\bar{g}(g,\{d\})} - \Gamma_g$ and $\alpha > 0$. Thus (γ^*, ξ^*) is feasible for Eq 10 and has a cost that is less than ζ^*_{DOI}. This contradicts the optimality of (γ^*, ξ^*) and proves that there is no $d \in \mathcal{D}, z \in \mathcal{Z}_d$ such that $\xi^*_{dz} > 0$. □

We can produce a feasible binary solution when γ is fractional at termination of CG as follows. We solve Eq 10 over $\hat{\mathcal{G}}$, while enforcing γ_g to be binary for all $g \in \hat{\mathcal{G}}$. If the solution has active ξ terms, then we apply the procedure described in the proof above to decrease the cost of the solution and ensure feasibility to Eq 1.

As CG proceeds we can not consider all of $\mathcal{Z}_d$ since the cardinality $\mathcal{Z}_d$ may explode for some or all $d \in \mathcal{D}$. Thus we use a subset of $\mathcal{Z}_d$ consisting of, the largest element and K others selected uniformly across $\mathcal{Z}_d$ denoted $\hat{\mathcal{Z}}_d$ (where K is a user defined parameter; e.g. $K = 5$ works well). Thus $\hat{\mathcal{Z}}_d = \{z_{\lceil \frac{k|\mathcal{Z}_d|}{K+1} \rceil} \quad \forall 1 \leq k \leq K+1\}$. With some abuse of notation we have Z_{dzg} be defined over Ξ^+_{dg}, where $\Xi^+_{dg} = G_{dg} \min_{\substack{z \in \hat{\mathcal{Z}}_d \\ \omega_{dz} \geq \Xi_{dg}}} \omega_{dz}$.

3.2 Efficient Pricing

Pricing for Eq 10 is conducted as $\min_{g \in \mathcal{G}} \Gamma_g - \sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} Z_{dzg} \lambda_{dz}$. Current MWSP applications (as in (Yarkony et al. 2019)) are associated with mechanisms to solve Eq 5 instead of $\min_{g \in \mathcal{G}} \Gamma_g - \sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} Z_{dzg} \lambda_{dz}$. We now prove that doing pricing using Eq 5 where $\lambda_d \leftarrow \sum_{z \in \mathcal{Z}_d} \lambda_{dz}$ $\forall d \in \mathcal{D}$ ensures that Eq 2=Eq 10 at termination of CG.

Claim: If λ^* is a dual optimal solution to Eq 11 (defined over some $\hat{\mathcal{G}} \subseteq \mathcal{G}$) satisfying that Eq 5$\geq$ 0 then $\sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} \lambda^*_{dz} =$ Eq 2.

Proof: Since $\hat{\mathcal{G}} \subseteq \mathcal{G}$ then Eq 2 $\leq \sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} \lambda^*_{dz}$. Let λ^+ be defined as $\lambda^+_d = (\sum_{z \in \mathcal{Z}_d} \lambda^*_{dz}) \, \forall d \in \mathcal{D}$. Since Eq 5$\geq$ 0 then λ^+ is a dual feasible solution to Eq 4 where $\mathcal{G} = \hat{\mathcal{G}}$; thus $\sum_{d \in \mathcal{D}} \lambda^+_d \leq$ Eq 2. Since $\sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} \lambda^*_{dz} = \sum_{d \in \mathcal{D}} \lambda^+_d$ then we have lower and upper bounded $\sum_{\substack{d \in \mathcal{D} \\ z \in \mathcal{Z}_d}} \lambda^*_{dz}$ by Eq 2 establishing the claim.

4 Application: Entity Resolution

In this section we apply the MWSP formulation in Section 3 to entity resolution resulting in our approach, which we call F-MWSP. This section is structured as follows. In Section

4.1 we describe the problem domain of entity resolution, and outline our pipeline for solving such problems. In Section 4.2 we define problem specific cost function for evaluating a single hypothesis in entity resolution. In Section 4.3 we devise efficient pricing algorithms (*i.e.* finding hypotheses with negative reduced costs) that exploit structural properties of entity resolution. In Section 4.4 we describe the production of Ξ_{dg} terms that satisfy Eq 9, thus defining the F-DOIs for entity resolution problems. In Section 4.5 we provide a numerical example demonstrating the value of F-DOI in this application.

4.1 Pipeline for Entity Resolution

Entity resolution seeks to construct a surjection from observations in input dataset to real world entities. The observations in the dataset are denoted $\mathcal{D}$, as defined in Section 2.1. Specifically, the dataset consists of a structured table where each row (or tuple) represents an observation of a real world entity. We rely on the attributes of the table to determine if two observations represent the same real world entity.

A naive way of doing entity resolution is to compare every pair of observations in the input dataset and decide whether they belong to the same entity or not; this will result in $\binom{|\mathcal{D}|}{2}$ comparisons, which is often prohibitively large for real-world applications. We instead employ a technique called blocking (Konda et al. 2016), in which we use a set of pre-defined, fast-to-run predicates to identify the subset of pairs of observations which could conceivably correspond to common entities (thus blocking operates in the high recall regime).

We first use blocking to filter out majority of pairs of observations, which leaves only a small proportion of pairs for further processing. Next, we generate a score for each pair of observations returned by the blocking step. The probability score defined over a given pair of observations is the probability that the pair are associated with a common entity. The classifier that generates probability scores is trained by any learning algorithm on the annotated data. We take negative of probability scores and add a bias to them, forming the cost terms used in our MWSP algorithm. Finally, based on the cost terms, the MWSP algorithm packs the observations into hypothesis with the goal of creating a bijection from hypothesis in the packing to real world entities. We refer to the combination of the blocker and the scorer as the *classifier*. Our entire pipeline for solving the entity resolution problem is described in Fig 2.

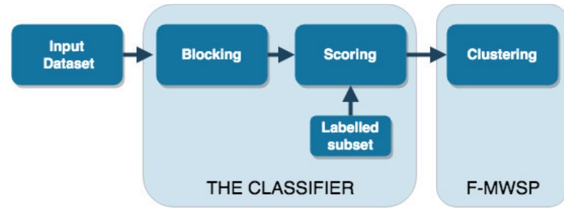

Figure 2: Entity Resolution Pipeline The stages of our pipeline are written in the following order. Given our input dataset we apply blocking to produce a limited set of pairs of observations that may be co-associated in a common hypotheses. Next we provide a probability score for each such pair using a classifier trained to distinguish between pairs that are/are not part of a common entity in the ground truth. Finally we convert the output of the probability scores to cost terms and treat the entity resolution as a MWSP problem as described in Section 3.

4.2 Cost Function for Entity Resolution

Consider a set of observations $\mathcal{D}$, where for any $d_1 \in \mathcal{D}, d_2 \in \mathcal{D}$ that $\theta_{d_1 d_2} \in \mathbb{R}$ is the cost associating with including d_1, d_2 in a common hypothesis. Here positive/negative values of $\theta_{d_1 d_2}$ discourage/encourage d_1, d_2 to be associated with a common hypothesis. The magnitude of $\theta_{d_1 d_2}$ describes the degree of discouragement/encouragement. We assume without loss of generality $\theta_{d_1 d_2} = \theta_{d_2 d_1}$. We construct $\theta_{d_1 d_2}$ from the output classifier as $(0.5 - p_{d_1 d_2})$ where $p_{d_1 d_2}$ is the probability provided by the classifier that d_1, d_2 are associated with a common hypothesis in the ground truth. It is a structural property of our problem domain that most pairs of observations can not be part of a common hypothesis. For such pairs d_1, d_2 then $\theta_{d_1 d_2} = \infty$. These are the pairs not identified by the blocker as being feasible. We use $\theta_{dd} = 0$ for all $d \in \mathcal{D}$. We define the cost of a hypothesis $g \in \mathcal{G}$ as follows.

$$\Gamma_g = \sum_{\substack{d_1 \in \mathcal{D} \\ d_2 \in \mathcal{D}}} \theta_{d_1 d_2} G_{d_1 g} G_{d_2 g} \tag{12}$$

With the cost of a hypothesis defined, we can now treat entity resolution as a MWSP problem, and use CG to solve it. Any observation not associated with any selected hypothesis in the solution to MWSP is defined to be in a hypothesis by itself of zero cost. Our formulation of entity resolution can also be rewritten as correlation clustering (Bansal, Blum, and Chawla 2004), which is usually tackled via LP relaxations with cycle inequalities and odd wheel inequalities (Nowozin and Jegelka 2009) in the machine learning literature. In the appendix we prove Eq 2 is no looser than (Nowozin and Jegelka 2009).

4.3 Pricing

With hypothesis cost Γ_g defined in Eq 12, we can now proceed to solve Eq 5. However, solving Eq 5 would be exceedingly challenging if we had to consider all $d \in \mathcal{D}$ at once. Fortunately, we can circumvent this difficulty using the following observation inspired by (Zhang et al. 2017), which studies biological cell instance segmentation. For any fixed $d^* \in \mathcal{D}$, solving for the lowest reduced cost hypothesis that includes d^* is much less challenging than solving Eq 5. This is because given d^* all $d \in \mathcal{D}$ for which $\theta_{d^* d} = \infty$ can be removed from consideration. Solving Eq 5 thus consists of solving many parallel pricing sub-problems, one for each $d^* \in \mathcal{D}$. All negative reduced cost solutions are then added to $\hat{\mathcal{G}}$. In this subsection we expand on this approach.

First we produce a small set of sub-problems each defined over a small subset of $\mathcal{D}$. Then we study *exact* optimization of those sub-problems, followed by *heuristic* optimization.

Pricing Formulation of (Zhang et al. 2017) We write pricing sub-problem adapted from (Zhang et al. 2017) given $d^* \in \mathcal{D}$ as follows:

$$\min_{\substack{g \in \mathcal{G} \\ G_{dg}=0 \quad \forall d \notin \mathcal{D}_{d^*} \\ G_{d^*g}=1}} \Gamma_g - \sum_{d \in \mathcal{D}} \lambda_d G_{dg} \tag{13}$$

$$\mathcal{D}_{d^*} = \{d \in \mathcal{D}; \theta_{dd^*} < \infty\}$$

Here $\mathcal{D}_{d^*}$ is the set of observations that may be grouped with observation d^*, which we call its neighborhood. Since the lowest reduced cost hypothesis must contain some $d^* \in \mathcal{D}$ by solving Eq 13 for each $d^* \in \mathcal{D}$ we solve Eq 5.

Improving on (Zhang et al. 2017) by decreasing sub-problem size We improve on (Zhang et al. 2017) by decreasing the number of observations considered in sub-problems, particularly those with large numbers of observations. We achieve this by associating a unique rank r_d to each observation $d \in \mathcal{D}$, such that r_d increases with $|\mathcal{D}_d|$, *i.e.* the more neighbors an observation has, the higher rank it is assigned. To ensure that each observation has unique rank we break ties arbitrarily.

Given that d^* is the lowest ranking observation in the hypothesis we need only consider the set of observations s.t. $d \in \{\mathcal{D}_{d^*} \cap \{r_d \geq r_{d^*}\}\}$, which we define to be $\mathcal{D}^*_{d^*}$. We write the resultant pricing sub-problem as follows.

$$\min_{\substack{g \in \mathcal{G} \\ G_{dg}=0 \quad \forall d \notin \mathcal{D}^*_{d^*} \\ G_{d^*g}=1}} \Gamma_g - \sum_{d \in \mathcal{D}} \lambda_d G_{dg} \tag{14}$$

Further improving on (Zhang et al. 2017) by removing superflous sub-problems We can also decrease the number of sub-problems considered as follows. First we relax the constraint $G_{d^*g} = 1$ in Eq 14. Now observe that for any $d_2 \in \mathcal{D}, d \in \mathcal{D}$ s.t. $\mathcal{D}^*_d \subset \mathcal{D}^*_{d_2}$ that the lowest reduced cost hypothesis over $\mathcal{D}^*_{d_2}$ has no greater reduced cost than that over $\mathcal{D}^*_d$. We refer a neighborhood $\mathcal{D}^*_{d^*}$ as being non-dominated if no $d_2 \in \mathcal{D}$ exists s.t. $\mathcal{D}^*_d \subset \mathcal{D}^*_{d_2}$.

During pricing we iterate over non-dominated neighborhoods. For a given non-dominated neighborhood $\mathcal{D}^*_{d^*}$ we write the pricing sub-problem below.

$$\min_{\substack{g \in \mathcal{G} \\ G_{dg}=0 \quad \forall d \notin \mathcal{D}^*_{d^*}}} \Gamma_g - \sum_{d \in \mathcal{D}} \lambda_d G_{dg} \tag{15}$$

(A) Exact Pricing We now consider the exact solution of Eq 15. We frame Eq 15 as a ILP, which we solve using a mixed integer linear programming (MILP) solver. We use decision variables x, y as follows. We set binary variable $x_d = 1$ to indicate that d is included in the hypothesis being generated and otherwise set $x_d = 0$. We set $y_{d_1d_2} = 1$ to indicate that both d_1, d_2 are included in the hypothesis being generated and otherwise set $y_{d_1d_2} = 0$. Defining $\mathcal{E}^- = \{(d_1, d_2) : \theta_{d_1d_2} = \infty\}$ as the set containing pairs of observations that cannot be grouped together, and $\mathcal{E}^+ = \{(d_1, d_2) : \theta_{d_1d_2} < \infty\}$ as the set containing pairs of observations that can be grouped together, we write the solution to Eq 15 as a MILP, which we annotate below.

$$\min_{\substack{x_d \in \{0,1\} \\ \forall d \in \mathcal{D}^*_{d^*} \\ y \geq 0}} \sum_{d \in \mathcal{D}^*_{d^*}} -\lambda_d x_d + \sum_{\substack{d_1 \in \mathcal{D}^*_{d^*} \\ d_2 \in \mathcal{D}^*_{d^*} \\ (d_1,d_2) \in \mathcal{E}^+}} \theta_{d_1d_2} y_{d_1d_2} \tag{16}$$

$$\text{s.t.} \quad x_{d_1} + x_{d_2} \leq 1 \quad \forall (d_1, d_2) \in \mathcal{E}^- \tag{17}$$

$$y_{d_1d_2} \leq x_{d_1} \quad \forall (d_1, d_2) \in \mathcal{E}^+ \tag{18}$$

$$y_{d_1d_2} \leq x_{d_2} \quad \forall (d_1, d_2) \in \mathcal{E}^+ \tag{19}$$

$$x_{d_1} + x_{d_2} - y_{d_1d_2} \leq 1 \quad \forall (d_1, d_2) \in \mathcal{E}^+ \tag{20}$$

Eq 16: Defines the reduced cost of the hypothesis being constructed.**Eq 17**: Enforce that pairs for which $\theta_{d_1d_2} = \infty$ are not include in a common hypothesis. **Eq 18-Eq 20**: Enforce that $y_{d_1d_2} = x_{d_1} x_{d_2}$. Observe that given that x is binary, that y must also be binary so as to obey Eq 18-Eq 20. Thus we need not explicitly enforce y to be binary.

(B) Heuristic Pricing Solving Eq 15 exactly using Eq 16-Eq 20 for each non-dominated neighborhood can be too time intensive for some scenarios. In fact Eq 15 generalizes max-cut, which is NP-hard (Karp 1972). This motivates the use of heuristic methods to solve Eq 15. Heuristic pricing is commonly used in operations research, however *we are the first paper in machine learning/entity resolution to employ this strategy*. Thus we decrease the computation time of pricing by decreasing the number of sub-problems solved, and solving those that are solved heuristically.

- **Early termination of pricing:** Observe that solving pricing (exactly or heuristically) over a limited subset of the sub-problems produces an approximate minimizer of Eq 5. We decrease the number of sub-problems solved during a given iteration of CG as follows. We terminate pricing in a given iteration when M negative reduced cost hypothesis have been added to $\hat{\mathcal{G}}$ in that iteration of CG (M is a user defined constant; $M = 50$ in our experiments). This strategy is called partial pricing (Lübbecke and Desrosiers 2005)
- **Solving sub-problems approximately:** We found empirical success solving Eq 16-Eq 20 using the quadratic pseudo-Boolean optimization with the improve option used (QPBO-I) (Rother et al. 2007).

The use of heuristic pricing does not prohibit the exact solution of Eq 2. One can switch to exact pricing after heuristic pricing fails to find a negative reduced cost hypothesis in $\mathcal{G}$.

4.4 Computing Ξ_{dg} for Entity Resolution

In this section, for any given $g \in \hat{\mathcal{G}}$ we construct Ξ_{dg} to satisfy Eq 9, which in practice leads to efficient optimization. We rewrite $\epsilon + \Gamma_{\bar{g}(g,\mathcal{D}_s)} - \Gamma_g$ by plugging in the expressions for Γ_g in Eq 12. We use $\mathcal{D}_g$ to denote the subset of $\mathcal{D}$ for which $G_{dg} = 1$.

$$\epsilon + \sum_{\substack{d_1 \in \mathcal{D}_g \\ d_2 \in \mathcal{D}_g}} -\theta_{d_1d_2} \max([d_1 \in \mathcal{D}_s], [d_2 \in \mathcal{D}_s]) \tag{21}$$

We now bound components of Eq 21 as follows. For $\theta_{d_1d_2} < 0$ we upper bound $-\theta_{d_1d_2} \max([d_1 \in \mathcal{D}_s], [d_2 \in \mathcal{D}_s])$

with: $-\theta_{d_1 d_2}([d_1 \in \mathcal{D}_s] + [d_2 \in \mathcal{D}_s])$ For $\theta_{d_1 d_2} > 0$ we upper bound $-\theta_{d_1 d_2} \max([d_1 \in \mathcal{D}_s], [d_2 \in \mathcal{D}_s])$ with: $-\frac{\theta_{d_1 d_2}}{2}([d_1 \in \mathcal{D}_s] + [d_2 \in \mathcal{D}_s])$ Below we plug the upper bounds into Eq 21; group by $[d \in \mathcal{D}_s]$; and enforce non-negativity of the result. Eq 21 $\leq \sum_{d \in \mathcal{D}}[d \in \mathcal{D}_s]\Xi_{dg}$ where $\Xi_{dg} = 0$ for $d \notin \mathcal{D}_g$, and is otherwise defined below.

$$\Xi_{dg} = \epsilon + \max(0, -\sum_{d_1 \in \mathcal{D}_g} \theta_{dd_1}(1 + [\theta_{dd_1} < 0])) \quad \forall d \in \mathcal{D}_g$$

The analysis in in this section can be applied directly for other pairwise cost functions such as multi-person pose estimation and multi-cell segmentation as described in (Yarkony et al. 2019).

4.5 Numerical Example

We provide a motivating example demonstrating how the use of F-DOI provides a lower valued LP relaxation given fixed $\hat{\mathcal{G}}$ than using varying DOI or no DOI. Consider a correlation clustering problem instance over $\mathcal{D} = \{d_1, d_2, d_3, d_4, d_5$ where θ is defined as follows. Let $\theta_{d_1,d_2} = \theta_{d_2,d_3} = \theta_{d_1,d_3} = \theta_{d_4 d_5} = -100; \theta_{d_3,d_4} = \theta_{d_3 d_5} = -1$ (note $\theta_{d\hat{d}} = \theta_{\hat{d}d}$). All other θ terms not defined above take on value ∞ . Consider that $\hat{\mathcal{G}} = g_1, g_2$ where $\mathcal{D}_{g_1} = \{d_1, d_2, d_3\}$ and $\mathcal{D}_{g_2} = \{d_3, d_4, d_5\}$. The optimal solution to the set packing RMP using no DOI is limited to selecting either g_1 or g_2. Solving the RMP using varying DOI both g_1, g_2 can be selected but a penalty of $400+\epsilon$ must be paid, resulting in higher objective than selecting only g_1. Solving the RMP using F-DOI both g_1, g_2 can be selected with a penalty of $4 + \epsilon$, resulting in a lower objective than selecting either g_1 or g_2.

5 Experiments

In this section, we study the different properties of the F-MWSP clustering algorithm and evaluate the performance scores on certain benchmark datasets. The classifier, which encompasses the blocker and the scorer, is a crucial component of the entity resolution pipeline (see Figure 2 and Section 4.1). We leverage the methods provided in a popular and open source entity resolution library called Dedupe (Gregg and Eder) to handle the blocking and scoring functionalities for us. Dedupe offers attribute type specific blocking rules and a ridge logistic regression algorithm as a default for scoring. Certainly, a more powerful classifier, especially if designed keeping the domain of the dataset in mind, can significantly boost the performance of the clustering outcome. As the focus of this paper has been F-MWSP clustering algorithm, an intuitive and reasonably good classifier such as Dedupe suits our setting.
In the following sections, we first demonstrate the different properties of F-MWSP algorithm on a single dataset and then compare its performance with other methods on benchmark datasets.

5.1 Characteristics of F-MWSP algorithm

The Setting To understand the benefits of F-MWSP clustering, it will be helpful to first conduct ablation study on a single dataset. The dataset that we choose in this section is called *patent_example* and is publicly available on Dedupe. *patent_example* is a labelled dataset listing the patent statistics of the Dutch innovators. It has has 2379 entities and 102 clusters where the mean size of the cluster is 23. We split the dataset into two halves and set aside the second half only to report the accuracies. The first half of the dataset that is visible to the learning algorithm from which we randomly sample about 1% of the total matches and provide it to the classifier as a labelled data.

Performance Metric	Hierarchical Clustering	F-MWSP Clustering
Precision / Recall	95.5% / 89.1%	95.4% / 94.3%
F1 measure	92.2%	94.8%
Homogn. / Compltn.	94.6% / 94%	94.4% / 96.5%
V measure	94.3%	95.4%
Adjusted Rand Index	91.3%	94.2%
Fowlkes Mallows	92.2%	94.8%

Table 1: F-MWSP algorithm performs better than the baseline hierarchical clustering algorithm.

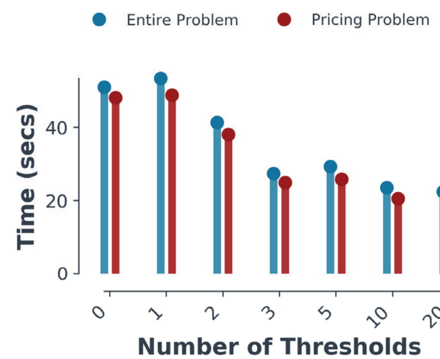

Figure 3: Speedups with Flexible DOIs Varying the number thresholds (value of K) of the Flexible DOIs improves the convergence speed. Threshold value 0 corresponds to the Varying DOIs used in (Yarkony et al. 2019).

(A) Superior performance over hierarchical clustering Table 5.1 shows that F-MWSP clusters offers better performance over hierarchical clustering, a standard method of choice for clustering problems (Hastie et al. 2005). The performance has been evaluated against standard clustering metrics.

(B) Significant speed-ups owing to Flexible DOIs We obtain at least 20% speed up with our proposed Flexible DOIs over Varying DOIs (Yarkony et al. 2019) as indicated in Figure 3. Moreover, we also observe that the computation time of the problem decreases as the number of thresholds (value of K) increases, with up to 60% speedup.

(C) Tractable solutions to the pricing problem Recall the strategies discussed to solve the pricing problem from Section 4.3, namely, exact and heuristic. Exact pricing is often not feasible in entity resolution owing to the large neighborhoods of some sub-problems. Fortunately, the heuristic

Dataset	Entities	Matches	Clusters	Mean	Max
patent_example	2379	293785	102	23	676
csv_example	3337	6608	1162	3	18
Affiliations	2260	16795	330	7	47
Settlements	3054	4388	820	3.7	4
Music 20K	19375	16250	10000	2	5

Table 2: Dataset statistics The statistics of all the datasets used in the paper are presented here. *Mean* and *Max* denote the respective statistics over the cluster sizes.

Method	Settlements	Music 20K
ConCom	0.65	0.26
CCPivot	0.90	0.74
Center	0.88	0.66
MergCenter	0.68	0.39
Star1	0.82	0.62
Star2	0.92	0.69
F-MWSP	**0.96**	**0.81**

Table 3: F-MWSP on benchmark datasets We obtain higher F1 score over the methods reported in (Saeedi, Peukert, and Rahm 2017). The F1 scores for other methods are extracted from the paper's bar plot.

solver helps cut down the computation time by a large fraction. For instance, *patent_example* experiment takes atleast **1 hour** for completion with the exact solver while with the heuristic solver it takes about **20 seconds**.

5.2 F-MWSP algorithm on benchmark datasets

To make sure that our findings are broadly applicable, we conducted experiments with more entity resolution benchmark datasets. (Saeedi, Peukert, and Rahm 2017) provides us with some interesting entity resolution datasets which we also include in this section. The statistics of all the datasets used in the paper are available in Table 2.

The Setting We make our setting consistent with (Saeedi, Peukert, and Rahm 2017) to be able to compare against their clustering algorithms. (Saeedi, Peukert, and Rahm 2017) leverages hand-crafted rules designed on the entire dataset to generate the cost terms. The costs are then fed into various clustering algorithms and the performance is evaluated over the whole dataset. We use dedupe classifier which is trained on a small percentage of matches from a split half of the dataset similar to Section 5.1. F-MWSP is then evaluated based on the dedupe cost terms over the entire dataset.

F-MWSP is competitive In this section, we report the performance of F-MWSP clustering on different datasets and compare against the baselines available for them. We start with *csv_example* dataset which is publically available on Dedupe akin to *patent_example*. On *csv_example*, F-MWSP achieves a higher F1 score of **95.2 %** against hierarchical clustering 94.4%, the default in Dedupe. In Table 3, we compare the performance of our entity resolution pipeline against algorithms in (Saeedi, Peukert, and Rahm 2017). Table 3 demonstrates that our pipeline, with F-MWSP clustering, is as powerful as the recognized entity resolution algorithms. The *affiliations* dataset used in (Aumueller and Rahm 2009) which is unique in the sense that the lack of structure in the data generates poorer cost terms. Despite this, F-MWSP gives us an F1 score of 63%, however, we note that a well handcrafted rule-based classifier improve the F1 score as demonstrated in (Aumueller and Rahm 2009).

6 Conclusions

In this paper we formulate entity resolution as MWSP problem. To solve such a problem, we devise a novel CG formulation that employs flexible dual optimal inequalities which use hypothesis specific information when constructing dual bounds. Our formulation exploits the fact that most pairs of observations can not be in a common hypothesis to produce pricing subproblems over small subsets of the observations that can be easily solved for some datasets, and for others can be solved to high quality heuristically. We demonstrate superior performance to the baseline hierarchical clustering formulation to entity resolution.

Acknowledgement We thank Dr. Vikas Singh from UW Madison and Dr. Konrad Kording from UPenn for the discussions on the project. Shaofei Wang is supported by NIH grant R01NS063399. Julian Yarkony acknowledges the help provided by Verisk.

7 Appendix

In this section we describe Correlation Clustering (CC) (Bansal, Blum, and Chawla 2004), its standard relaxation and finally prove that our MWSP relaxation of CC is tighter than the standard relaxations in the literature (Nowozin and Jegelka 2009; Andres et al. 2011; Kim et al. 2011).

Given a graph with node set $\mathcal{D}$ we use $f \in \{0,1\}^{|\mathcal{D}|\times|\mathcal{D}|}$ which we index by d_1, d_2 to describe a partition of $\mathcal{D}$. We set $f_{d_1d_2} = 1$ if and only if d_1, d_2 are in a common component in our solution. We use $\theta_{d_1d_2}$ to denote the cost of including d_1, d_2 in a common component. The objective of CC is written below.

$$\min_{f\in\{0,1\}} \sum_{\substack{d_1\in\mathcal{D}\\ d_2\in\mathcal{D}}} \theta_{d_1d_2} f_{d_1d_2} \tag{22}$$

CC uses cycle inequalities to enforce that f describes a valid partitioning of the vertices. Cycle inequalities state that for every cycle of nodes, that the number of "cut" edges (meaning edges where the connected observations are in separate components) is a number other than one. Let $\mathcal{H}$ be the set of cycles of vertices, which we index by h. We treat h as the set of edges on the associated cycle. The cycle inequality associated with any $h \in \mathcal{H}$, $f_{d_ad_b} \in h$ is written below.

$$\sum_{(d_1,d_2)\in h-(d_a,d_b)} (1 - f_{d_1d_2}) \geq (1 - f_{d_ad_b}) \tag{23}$$

A solution f must satisfy Eq 23 for all $h \in \mathcal{H}$, $(d_a, d_b) \in h$ to be a feasible partition of the observations (Nowozin and Jegelka 2009). It is sufficient to enforce only the cycle inequalities over each cycle of three nodes in order to enforce all cycle inequalities (Chopra and Rao 1993). We write the cycle inequality over observations d_1, d_2, d_3 below.

$$(1 - f_{d_1d_3}) + (1 - f_{d_2d_3}) \geq (1 - f_{d_1d_2}) \tag{24}$$

Odd wheel inequalities (OWI) are a common valid inequality in CC used to tighten the LP relaxation of CC. OWI are defined over a cycle of edges of odd cardinality h_b with a single additional node d_b connected to all other nodes in the center. We define OWI below for any $b \in \mathcal{B}$ where $\mathcal{B}$ is the set of OWI. We use d^b_m to denote the m'th observation in the cycle h_b; and note that $d^b_{|h_b|+1} = d^b_1$.

$$\sum_{m=1}^{|h_b|} f_{d^b_m d_b} - f_{d^b_m d^b_{m+1}} \leq \lfloor \frac{|h_b|}{2} \rfloor \tag{25}$$

We produce an LP relaxation of CC by relaxing f to lie in $[0, 1]$ and enforcing Eq 24- Eq 25.

$$\min_{1 \geq f \geq 0} \sum_{\substack{d_1 \in \mathcal{D} \\ d_2 \in \mathcal{D}}} \theta_{d_1 d_2} f_{d_1 d_2} \tag{26}$$
$$(1 - f_{d_1 d_3}) + (1 - f_{d_2 d_3}) \geq (1 - f_{d_1 d_2})$$
$$\sum_{m=1}^{|h_b|} f_{d^b_m d_b} - f_{d^b_m d^b_{m+1}} \leq \lfloor \frac{|h_b|}{2} \rfloor$$

For any $d_1 \in \mathcal{D}, d_2 \in \mathcal{D}$ the variable $f_{d_1 d_2}$ corresponds to the following in our MWSP formulation for EC.

$$f_{d_1 d_2} = \sum_{g \in \mathcal{G}} G_{d_1 g} G_{d_2 g} \gamma_g \tag{27}$$

The goal of the remainder of this section is to show that Eq 2 is no looser a relaxation of CC than Eq 26. We outline this section as follows. In Section 7.1 we show that any feasible solution to Eq 2 satisfies $0 \leq f \leq 1$. In Section 7.2 we show that any feasible solution to Eq 2 satisfies Eq 24. In Section 7.3 that that any feasible solution to Eq 2 satisfies Eq 25. In Section 7.4 we establish that Eq 2 $\geq$ Eq 26.

7.1 Proof Bound Obeyed

Eq 26 enforces that for each $1 \geq f_{d_1 d_2} \geq 0$ and we now establish that this holds for Eq 2. Using Eq 27 we observe that the following must hold for any γ satisfying Eq 2.

$$0 \leq \sum_{g \in \mathcal{G}} G_{d_1 g} G_{d_2 g} \gamma_g = f_{d_1 d_2} \tag{28}$$

$$\sum_{g \in \mathcal{G}} G_{d_1 g} G_{d_2 g} \gamma_g = f_{d_1 d_2} \leq 1 \tag{29}$$

Eq 28 is satisfied since G is a binary matrix and γ is non-negative. Eq 29 is satisfied since $\sum_{g \in \mathcal{G}} G_{d_1 g} G_{d_2 g} \gamma_g \leq \sum_{g \in \mathcal{G}} G_{d_1 g} \gamma_g$ and Eq 2 ensures that $\sum_{g \in \mathcal{G}} G_{d_1 g} \gamma_g \leq 1$ for each $d_1 \in \mathcal{D}$.

7.2 MWSP Satisfies All Cycle Inequalities

In this section we establish that any feasible solution γ to Eq 2 satisfies Eq 24. To assist in our discussion we use the notation $j^{\mathcal{D}^-}_{\mathcal{D}^+}$ to denote the sum of the γ terms associated with hypothesis that include all elements in $\mathcal{D}^+$ and no elements in $\mathcal{D}^-$ as follows.

$$j^{\mathcal{D}^-}_{\mathcal{D}^+} = \sum_{g \in \mathcal{G}} \gamma_g (\prod_{d \in \mathcal{D}^+} G_{dg})(\prod_{d \in \mathcal{D}^-} (1 - G_{dg})) \tag{30}$$

We now use proof by contradiction to establish that γ obeys Eq 24.

Claim:

All γ satisfying Eq 2 satisfy all inequalities of the form in Eq 24.

Proof: Suppose the claim is false. Thus there exists a γ that is feasible to Eq 2 for which there exists a d_1, d_2, d_3 that does not satisfy Eq 24. We re-write Eq 24 for the violated cycle inequality using j.

$$(1 - j_{d_1 d_3}) + (1 - j_{d_2 d_3}) < (1 - j_{d_1 d_2}) \tag{31}$$
$$1 + j_{d_1 d_2} < j_{d_1 d_3} + j_{d_2 d_3}$$
$$1 + j_{d_1 d_2 d_3} + j^{d_3}_{d_1 d_2} < j^{d_2}_{d_1 d_3} + j^{d_1}_{d_2 d_3} + 2 j_{d_1 d_2 d_3}$$
$$1 + j^{d_3}_{d_1 d_2} < j^{d_2}_{d_1 d_3} + j^{d_1}_{d_2 d_3} + j_{d_1 d_2 d_3}$$

We now bound the RHS by j_{d3}, which we in turn bound by 1.

$$1 + j^{d_3}_{d_1 d_2} < j^{d_2}_{d_1 d_3} + j^{d_1}_{d_2 d_3} + j_{d_1 d_2 d_3} \leq j_{d_3} \leq 1 \tag{32}$$

Since $j^{d_3}_{d_1 d_2}$ is non negative it can not be less than zero thus establishing a contradiction.

7.3 MWSP Satisfies All Odd Wheel Inequalities

We now establish that all OWI are satisfied for any feasible solution to Eq 2 using proof by contradiction.

Claim

$$\sum_{m=1}^{|h_b|} j_{d^b_m d_b} - j_{d^b_m d^b_{m+1}} \leq \lfloor \frac{|h_b|}{2} \rfloor \quad \forall b \in \mathcal{B} \tag{33}$$

Proof: Consider a solution γ and $b \in \mathcal{B}$ violating the claim.

$$\sum_{m=1}^{|h_b|} j_{d^b_m d_b} - j_{d^b_m d^b_{m+1}} > \lfloor \frac{|h_b|}{2} \rfloor$$
$$\sum_{m=1}^{|h_b|} j_{d^b_m d_b d^b_{m+1}} + j^{d^b_{m+1}}_{d^b_m d_b} - j^{d_b}_{d^b_m d^b_{m+1}} - j_{d^b_m d^b_{m+1} d_b} > \lfloor \frac{|h_b|}{2} \rfloor$$
$$\sum_{m=1}^{|h_b|} j^{d^b_{m+1}}_{d^b_m d_b} - j^{d_b}_{d^b_m d^b_{m+1}} > \lfloor \frac{|h_b|}{2} \rfloor \tag{34}$$

We upper bound the LHS of Eq 34 by removing the $j^{d_b}_{d^b_m d^b_{m+1}}$ terms. $\sum_{m=1}^{|h_b|} j^{d^b_{m+1}}_{d^b_m d_b} > \lfloor \frac{|h_b|}{2} \rfloor$. We express j using Eq 30.

$$\sum_{m=1}^{|h_b|} \sum_{g \in \mathcal{G}} \gamma_g G_{d^b_m g}(1 - G_{d^b_{m+1} g}) G_{d_b g} > \lfloor \frac{|h_b|}{2} \rfloor$$
$$\sum_{g \in \mathcal{G}} \gamma_g G_{d_b g} \sum_{m=1}^{|h_b|} G_{d^b_m g}(1 - G_{d^b_{m+1} g}) > \lfloor \frac{|h_b|}{2} \rfloor$$

Observe that the term $\sum_{m=1}^{|h_b|} G_{d^b_m g}(1 - G_{d^b_{m+1} g})$ is bounded from above by $\lfloor \frac{|h_b|}{2} \rfloor$. This is because the largest independent set defined on a cycle graph contains half the nodes (rounded down). We apply this bound below.

$$\sum_{g\in\mathcal{G}}\gamma_g G_{d_bg}\lfloor\frac{|h_b|}{2}\rfloor > \lfloor\frac{|h_b|}{2}\rfloor$$
$$\sum_{g\in\mathcal{G}}\gamma_g G_{d_bg} > 1 \tag{35}$$

Eq 2 ensures that $\sum_{g\in\mathcal{G}} G_{dg}\gamma_g \leq 1$ for all $d \in \mathcal{D}$ which contradicts Eq 35 thus proving that the claim in Eq 33 true.

7.4 Eq 2 $\geq$ Eq 26

Since every feasible solution to Eq 2 obeys all constraints in Eq 26 then the minimal cost solution to Eq 2 obeys all constraints in Eq 26 thus Eq 2 $\geq$ Eq 26. We have not established the existence of cases for which Eq 2 $>$ Eq 26 and leave consideration of such cases to future research.

References

Andres, B.; Kappes, J. H.; Beier, T.; Kothe, U.; and Hamprecht, F. A. 2011. Probabilistic image segmentation with closedness constraints. In *Proc. 13th International Conference on Computer Vision*, 2611–2618.

Aslam, J. A.; Pelekhov, E.; and Rus, D. 2004. The star clustering algorithm for static and dynamic information organization. *J. Graph Algorithms Appl.* 8:95–129.

Aumueller, D., and Rahm, E. 2009. Web-based affiliation matching. In *ICIQ*, 246–256. Citeseer.

Bansal, N.; Blum, A.; and Chawla, S. 2004. Correlation clustering. *Journal of Machine Learning* 56(1-3):89–113.

Barnhart, C.; Johnson, E. L.; Nemhauser, G. L.; Savelsbergh, M. W. P.; and Vance, P. H. 1996. Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46:316–329.

Ben Amor, H.; Desrosiers, J.; and Valério de Carvalho, J. M. 2006. Dual-optimal inequalities for stabilized column generation. *Operations Research* 54(3):454–463.

Chopra, S., and Rao, M. R. 1993. The partition problem. *Mathematical Programming* 59(1-3):87–115.

Desrosiers, J., and Lübbecke, M. E. 2005. A primer in column generation. In Desaulniers, G.; Desrosiers, J.; and Solomon, M. M., eds., *Column Generation*. New York, NY: Springer. 1–32.

Du Merle, O.; Villeneuve, D.; Desrosiers, J.; and Hansen, P. 1999. Stabilized column generation. *Discrete Mathematics* 194(1-3):229–237.

Gilmore, P., and Gomory, R. 1961. A linear programming approach to the cutting-stock problem. *Operations Research* 9(6):849–859.

Gregg, F., and Eder, D. Dedupe. https://github.com/dedupeio/dedupe.

Gschwind, T., and Irnich, S. 2016. Dual inequalities for stabilized column generation revisited. *INFORMS Journal on Computing* 28(1):175–194.

Hassanzadeh, O., and Miller, R. J. 2009. Creating probabilistic databases from duplicated data. *The VLDB Journal—The International Journal on Very Large Data Bases* 18(5):1141–1166.

Hastie, T.; Tibshirani, R.; Friedman, J.; and Franklin, J. 2005. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27(2):83–85.

Jepsen, M.; Petersen, B.; Spoorendonk, S.; and Pisinger, D. 2008. Subset-row inequalities applied to the vehicle-routing problem with time windows. *Operations Research* 56(2):497–511.

Karp, R. M. 1972. Reducibility among combinatorial problems. In *Proc. Symposium on the Complexity of Computer Computations*, 85–103.

Kim, S.; Nowozin, S.; Kohli, P.; and Yoo, C. D. 2011. Higher-order correlation clustering for image segmentation. In *Advances in Neural Information Processing Systems,25*, 1530–1538.

Konda, P.; Das, S.; Suganthan GC, P.; Doan, A.; Ardalan, A.; Ballard, J. R.; Li, H.; Panahi, F.; Zhang, H.; Naughton, J.; et al. 2016. Magellan: Toward building entity matching management systems. *Proceedings of the VLDB Endowment* 9(12):1197–1208.

Lübbecke, M. E., and Desrosiers, J. 2005. Selected topics in column generation. *Operations Research* 53(6):1007–1023.

Lübbecke, M. E. 2010. Column generation. *Wiley Encyclopedia of Operations Research and Management Science*.

Marsten, R. E.; Hogan, W.; and Blankenship, J. W. 1975. The boxstep method for large-scale optimization. *Operations Research* 23(3):389–405.

Nowozin, S., and Jegelka, S. 2009. Solution stability in linear programming relaxations: Graph partitioning and unsupervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 769–776. ACM.

Rother, C.; Kolmogorov, V.; Lempitsky, V.; and Szummer, M. 2007. Optimizing binary mrfs via extended roof duality. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 1–8.

Saeedi, A.; Peukert, E.; and Rahm, E. 2017. Comparative evaluation of distributed clustering schemes for multi-source entity resolution. In *European Conference on Advances in Databases and Information Systems*, 278–293. Springer.

Wang, S.; Ihler, A.; Kording, K.; and Yarkony, J. 2018. Accelerating dynamic programs via nested benders decomposition with application to multi-person pose estimation. In *Proc. 15th European Conference on Computer Vision*, 652–666.

Yarkony, J.; Adulyasak, Y.; Singh, M.; and Desaulniers, G. 2019. Data association via set packing forcomputer vision applications. In *Les Cahiers du GERAD*.

Zhang, C.; Wang, S.; Gonzalez-Ballester, M. A.; and Yarkony, J. 2017. Efficient column generation for cell detection and segmentation. *arXiv preprint arXiv:1709.07337*.