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

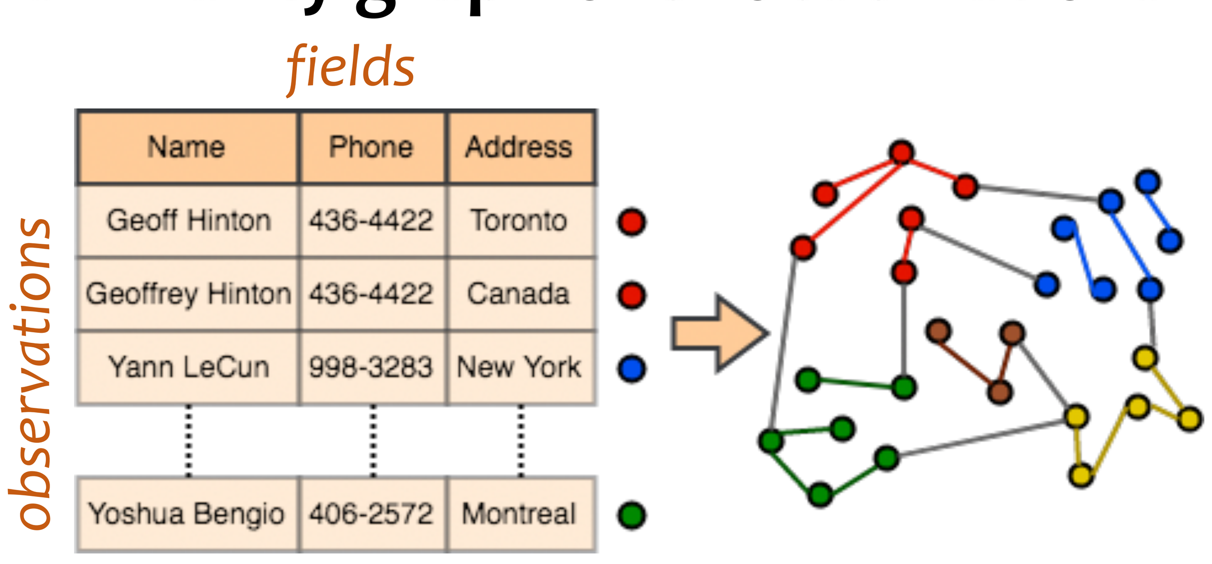
Vishnu Lokhande, Shaofei Wang, Maneesh Singh, Julian Yarkony

ENTITY RESOLUTION AS CORRELATION CLUSTERING

We present a **new optimization approach** to entity resolution (ER) modeled as **correlation-clustering**.

Correlation-clustering exploits relations between objects to cluster them without knowing the optimum number of clusters in advance.

Entity resolution can be viewed as a **node-clustering problem** on a **similarity graph over observations**.



ENTITY RESOLUTION AS SET PACKING

Traditional approaches for ER include hierarchical clustering and star clustering. They are often **slow** and produce **overlapping clusters**.

Our approach formulates ER as **minimum weight set packing**.

- novel formulation ensures **non-overlapping clusters** and **superior performance**
- novel & efficient optimization provides **significant speed-up**

MINIMUM WEIGHT SET PACKING (MWSP) FOR ER

$d \in \mathcal{D}$: observations

$g \in \mathcal{G}$: entities (or hypotheses) each g is a cluster of observations

Yikes, $\mathcal{G}$ is exponentially large!

$G \in \{0,1\}^{|\mathcal{D}|\times|\mathcal{G}|}$, $G_{dg} = 1$, if entity g includes observation d

$$\min_{\gamma\in\{0,1\}^{|\mathcal{G}|}} \sum_{g\in\mathcal{G}} \Gamma_g\gamma_g$$

objective: including more entities during resolution incurs a cost

$$\text{s.t.} \quad \sum_{g\in\mathcal{G}} G_{dg}\gamma_g \leq 1 \quad \forall d \in \mathcal{D}$$

constraints: each observation can only be associated with a single entity (non-overlap)

Yikes, ILP!
(NP-Hard)

SOLVING MWSP WITH COLUMN GENERATION

Column generation seeks to avoid explicit enumeration of $\mathcal{G}$ by iteratively growing a subset $\hat{\mathcal{G}}$.

Step 1 (solve): **Relax** ILP and **restrict** it to a subset of entities $\hat{\mathcal{G}} \subset \mathcal{G}$

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

restricted primal problem: LP relaxation over primal variables ($\gamma \geq 0$)

$$\max_{\lambda\leq 0} \sum_{d\in\mathcal{D}} \lambda_d \qquad \text{s.t.} \quad \Gamma_g - \sum_{d\in\mathcal{D}} G_{dg}\lambda_d \geq 0 \quad \forall g \in \hat{\mathcal{G}}$$

restricted dual problem: over dual variables ($\lambda \leq 0$)

Step 2 (grow): Find the hypothesis g with the smallest reduced cost to add to $\hat{\mathcal{G}}$ using the **pricing problem.**

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

FASTER CONVERGENCE WITH DUAL-OPTIMAL INEQUALITIES

Convergence can be accelerated by **bounding dual variables:** $-\Xi_d \leq \lambda_d$

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

Idea: dynamically identify a small number of tight lower bounds to effectively limit search space

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

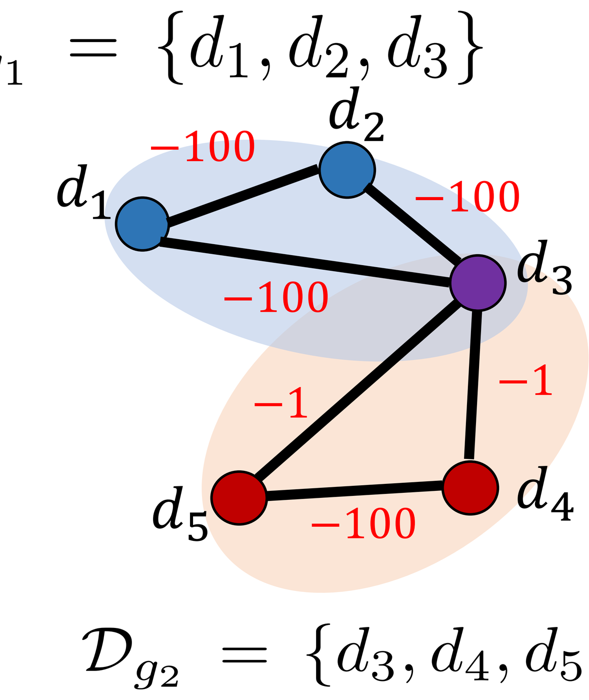
identify lower bounds by searching across thresholds of Ξ_{dz}

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

thresholds of Ξ_{dz} are enumerated by considering all the unique positive values of Ξ_{dg}

MOTIVATING EXAMPLE

$$\mathcal{D}_{g_1} = \{d_1, d_2, d_3\}$$



$$\mathcal{D}_{g_2} = \{d_3, d_4, d_5\}$$

Without DOI

g_1 or g_2

With DOI

g_1 and g_2

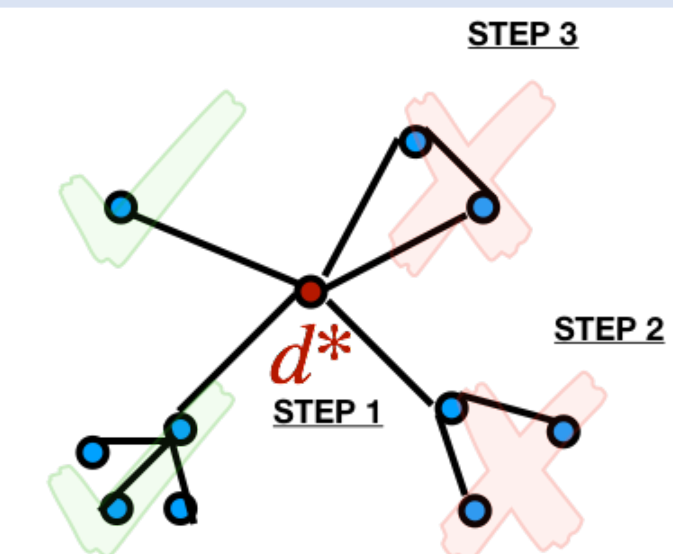
Penalty= $400 + \epsilon$

With F-DOI

g_1 and g_2

Penalty= $4 + \epsilon$

EFFICIENT PRICING



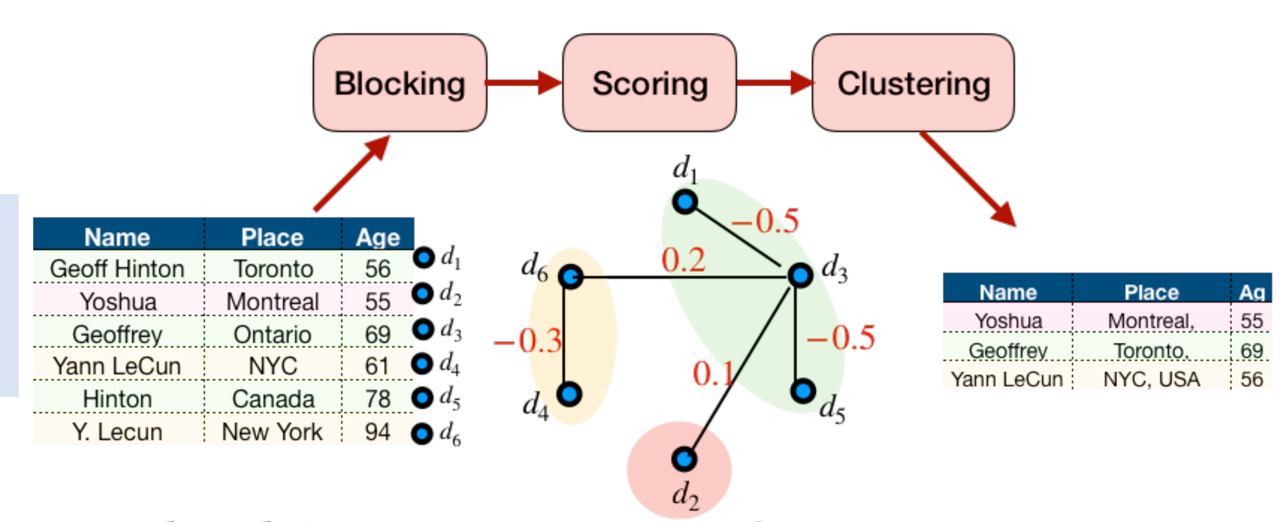
TWO VARIANTS
EXACT PRICING
HEURISTIC PRICING

Solve sub-problems with central node d* & at most one neighborhood

Enhancement 1: Remove leaves with higher relative neighborhood cardinality than the central node

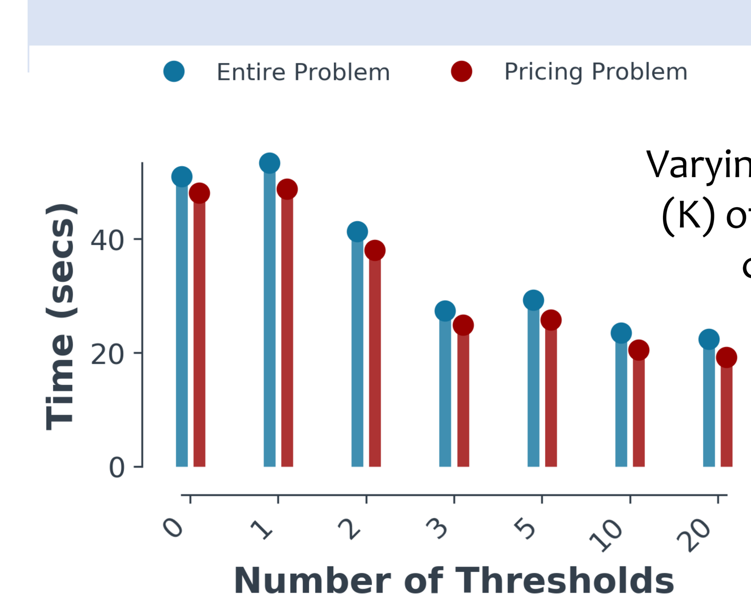
Enhancement 2: Remove leaves that form superfluous sub

OUR ENTITY RESOLUTION PIPELINE



Blocking: remove obvious non-matches by comparing key attribute across observations.
Scoring: similarity score between observations based on their fields
Clustering: with MWSP and FDOIs

RESULTS & CONCLUSIONS



Varying the number thresholds (K) of Flexible DOIs improves convergence speed.

F-MWSP is competitive on benchmark datasets.

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**

- **Superior performance** over hierarchical clustering baselines
- **Significant speed-ups** (at least 20%) owing to Flexible DOIs

*Work done during internship at Verisk

**Special thanks to Dr. Gautam Kunapuli for helping with the poster