(Cost-) Optimal Correlation Clustering via Max-SAT

Jeremias Berg, Matti Järvisalo

HIIT & Dept. of Computer Science University of Helsinki, Finland

July 6, 2015

Contributions of the paper

- Framework for optimally solving correlation clustering.
  - Combinatorial optimization problem.
  - Maximum Satisfiability (MaxSAT)

Benefits of our method

- Global optimum w.r.t any cost function.
- Easily adaptable to different variants of the problem.
- Constrained correlation clustering.
- Solvers improving.

Drawbacks

- Scalability.
- Extends to hundreds of points.

Contributions of the paper

- Framework for optimally solving correlation clustering.
  - Combinatorial optimization problem.
  - Maximum Satisfiability (MaxSAT)

Benefits of our method

- Global optimum w.r.t any cost function.
- Easily adaptable to different variants of the problem.
- Constrained correlation clustering.
- Solvers improving.

Drawbacks

- Scalability.
- Extends to hundreds of points.

Contributions of the paper

- Framework for optimally solving correlation clustering.
  - Combinatorial optimization problem.
  - Maximum Satisfiability (MaxSAT)

Benefits of our method

- Global optimum w.r.t any cost function.
- Easily adaptable to different variants of the problem.
- Constrained correlation clustering.
- Solvers improving.

Drawbacks

- Scalability.
- Extends to hundreds of points.

Outline of the talk

1. Background on correlation clustering
2. Background on MaxSAT
3. Correlation Clustering as Integer Programming
4. Correlation Clustering as MaxSAT
5. Experimental Evaluation
6. Future work
7. Conclusions

Background on correlation clustering

- NP-hard optimization problem.
- Categorical information.
- Studied both from theoretical and experimental p.o.v.
- Generalizations and variations.

Problem statement

INPUT	Undirected graph Edges labeled by "+" or "-"
OBJECTIVE:	Partition (cluster) the nodes s.t the number of positive edges between clusters and negative edges within clusters is minimized.

- Note! Number of clusters not part of input.

Background on correlation clustering

- NP-hard optimization problem.
- Categorical information.
- Studied both from theoretical and experimental p.o.v.
- Generalizations and variations.

Problem statement

INPUT	Undirected graph Edges labeled by "+" or "-"
OBJECTIVE:	Partition (cluster) the nodes s.t the number of positive edges between clusters and negative edges within clusters is minimized.

- Note! Number of clusters not part of input.

Background on correlation clustering

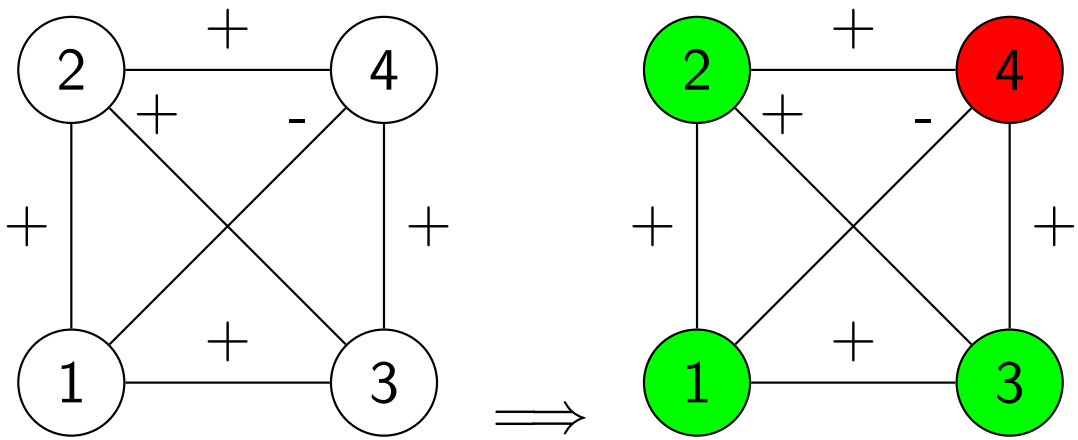
- NP-hard optimization problem.
- Categorical information.
- Studied both from theoretical and experimental p.o.v.
- Generalizations and variations.

Problem statement

INPUT	Undirected graph Edges labeled by "+" or "-"
OBJECTIVE:	Partition (cluster) the nodes s.t the number of positive edges between clusters and negative edges within clusters is minimized.

- Note! Number of clusters not part of input.

Example



Precise statement of problem

Given an undirected graph $G = (V, E)$ and a similarity function $s\colon E \to \{1, 0\}$ find a clustering $cl\colon V \to \{1 \ldots |V|\}$ minimizing:

$$\sum_{\substack{\{v_i, v_j\} \in E \\ cl(v_i) = cl(v_j)}} (1 - s(v_i, v_j)) + \sum_{\substack{\{v_i, v_j\} \in E \\ cl(v_i) \neq cl(v_j)}} s(v_i, v_j).$$

- Empty clusters allowed.
- Not necessarily complete similarity info.

Precise statement of problem

Given an undirected graph $G = (V, E)$ and a similarity function $s\colon E \to \{1, 0\}$ find a clustering $cl\colon V \to \{1 \ldots |V|\}$ minimizing:

$$\sum_{\substack{\{v_i, v_j\} \in E \\ cl(v_i) = cl(v_j)}} (1 - s(v_i, v_j)) + \sum_{\substack{\{v_i, v_j\} \in E \\ cl(v_i) \neq cl(v_j)}} s(v_i, v_j).$$

- Empty clusters allowed.
- Not necessarily complete similarity info.

Correlation Clustering as an Integer Program

[??]

- Use indicator variables $x_{ij} \in \{0, 1\}$.
- $x_{ij} = 1$ if $cl(v_i) = cl(v_j)$.

IP formulation:

$$\text{MINIMIZE:} \quad \sum_{s(v_i,v_j)=0} x_{ij} - \sum_{s(v_i,v_j)=1} x_{ij}$$

where: $x_{ij} + x_{jk} \leq 1 + x_{ik}$ transitivity

$x_{ij} \in \{0, 1\}$ for all i, j.

- $\mathcal{O}(n^2)$ variables and $\mathcal{O}(n^3)$ constraints (very large).

Correlation Clustering as an Integer Program

[??]

- Use indicator variables $x_{ij} \in \{0, 1\}$.
- $x_{ij} = 1$ if $cl(v_i) = cl(v_j)$.

IP formulation:

$$\text{MINIMIZE:} \quad \sum_{s(v_i,v_j)=0} x_{ij} - \sum_{s(v_i,v_j)=1} x_{ij}$$

where: $x_{ij} + x_{jk} \leq 1 + x_{ik}$ transitivity

$x_{ij} \in \{0, 1\}$ for all i, j.

- $\mathcal{O}(n^2)$ variables and $\mathcal{O}(n^3)$ constraints (very large).

Maximum Satisfiability

- Simple constraint language.
- Literal: a boolean variable x or $\neg x$.
- Clause C: a disjunction ($\vee$) of literals. e.g $(x \vee y \vee \neg z)$
- Truth assignment τ: a function from boolean variables to $\{0, 1\}$.
- $\tau(C) = 1$ if
  $\tau(x) = 1$ for a literal $x \in C$,
  $\tau(x) = 0$ for a literal $\neg x \in C$.

Problem statement

Given two sets of clauses F_h, F_s find τ s.t.

1. $\tau(C) = 1$ for all $C \in F_h$.
2. $\sum_{C \in F_s} \tau(C)$ is maximized.

- NP-hard, hence usable for solving cost optimal correlation clustering.

Maximum Satisfiability

- Simple constraint language.
- Literal: a boolean variable x or $\neg x$.
- Clause C: a disjunction ($\vee$) of literals. e.g $(x \vee y \vee \neg z)$
- Truth assignment τ: a function from boolean variables to $\{0, 1\}$.
- $\tau(C) = 1$ if
  $\tau(x) = 1$ for a literal $x \in C$,
  $\tau(x) = 0$ for a literal $\neg x \in C$.

Problem statement

Given two sets of clauses F_h, F_s find τ s.t.

1. $\tau(C) = 1$ for all $C \in F_h$.
2. $\sum_{C \in F_s} \tau(C)$ is maximized.

- NP-hard, hence usable for solving cost optimal correlation clustering.

Maximum Satisfiability

- Simple constraint language.
- Literal: a boolean variable x or $\neg x$.
- Clause C: a disjunction ($\vee$) of literals. e.g $(x \vee y \vee \neg z)$
- Truth assignment τ: a function from boolean variables to $\{0, 1\}$.
- $\tau(C) = 1$ if
  $\tau(x) = 1$ for a literal $x \in C$,
  $\tau(x) = 0$ for a literal $\neg x \in C$.

Problem statement

Given two sets of clauses F_h, F_s find τ s.t.

1. $\tau(C) = 1$ for all $C \in F_h$.
2. $\sum_{C \in F_s} \tau(C)$ is maximized.

- NP-hard, hence usable for solving cost optimal correlation clustering.

MaxSAT Encoding 1

- $(\{v_1, \ldots v_n\}, E)$ undirected graph and $s\colon E \to \{0, 1\}$.
- Reformulation of IP.
- Hard clauses: encode well defined clustering.
- Soft clauses: encode cost function.
- $\mathcal{O}(n^2)$ variables and $\mathcal{O}(n^3)$ clauses.
  - Same as IP.
  - However, more memory efficient.

Encoding 1

Variables

- $x_{ij} = 1$ if $cl(v_i) = cl(v_j)$

Hard clauses

- For all distinct i, j, k : $(x_{ij} \wedge x_{jk}) \rightarrow x_{ik}$.
- As clause: $(\neg x_{ij} \vee \neg x_{jk} \vee x_{ik})$.

Soft clauses

- Capture cost function.
- As clauses:

$$(x_{ij}) \quad \text{for all } i, j \quad s(v_i, v_j) = 1$$
$$(\neg x_{ij}) \quad \text{for all } i, j \quad s(v_i, v_j) = 0$$

Encoding 1

Variables

- $x_{ij} = 1$ if $cl(v_i) = cl(v_j)$

Hard clauses

- For all distinct i, j, k : $(x_{ij} \wedge x_{jk}) \rightarrow x_{ik}$.
- As clause: $(\neg x_{ij} \vee \neg x_{jk} \vee x_{ik})$.

Soft clauses

- Capture cost function.
- As clauses:

$$(x_{ij}) \quad \text{for all } i, j \quad s(v_i, v_j) = 1$$
$$(\neg x_{ij}) \quad \text{for all } i, j \quad s(v_i, v_j) = 0$$

Encoding 1

Variables

- $x_{ij} = 1$ if $cl(v_i) = cl(v_j)$

Hard clauses

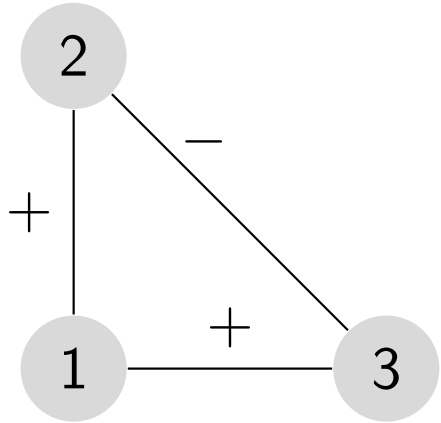
- For all distinct i, j, k : $(x_{ij} \wedge x_{jk}) \rightarrow x_{ik}$.
- As clause: $(\neg x_{ij} \vee \neg x_{jk} \vee x_{ik})$.

Soft clauses

- Capture cost function.
- As clauses:

$$\begin{array}{rll} (x_{ij}) & \text{for all } i,j & s(v_i, v_j) = 1 \\ (\neg x_{ij}) & \text{for all } i,j & s(v_i, v_j) = 0 \end{array}$$

Example of Enc 1



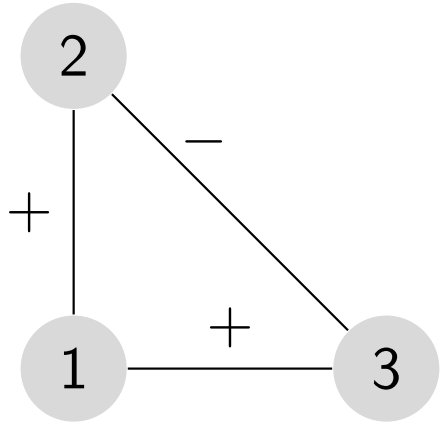
- Hard Clauses:

$$\{(\neg x_{12} \vee \neg x_{23} \vee x_{13}), (\neg x_{12} \vee \neg x_{13} \vee x_{23}), (\neg x_{23} \vee \neg x_{13} \vee x_{12})\}$$

- Soft Clauses:

$$\{(x_{12}), (x_{13}), (\neg x_{23})\}$$

Example of Enc 1



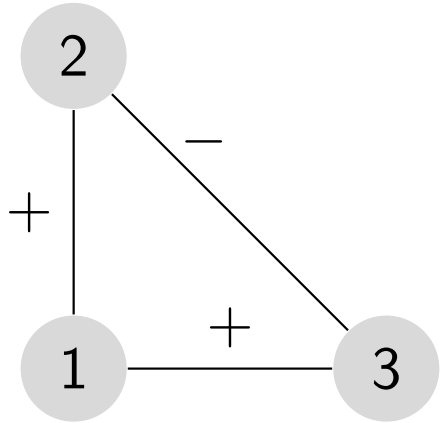
- Hard Clauses:

$$\{(\neg x_{12} \vee \neg x_{23} \vee x_{13}), (\neg x_{12} \vee \neg x_{13} \vee x_{23}), (\neg x_{23} \vee \neg x_{13} \vee x_{12})\}$$

- Soft Clauses:

$$\{(x_{12}), (x_{13}), (\neg x_{23})\}$$

Example of Enc 1



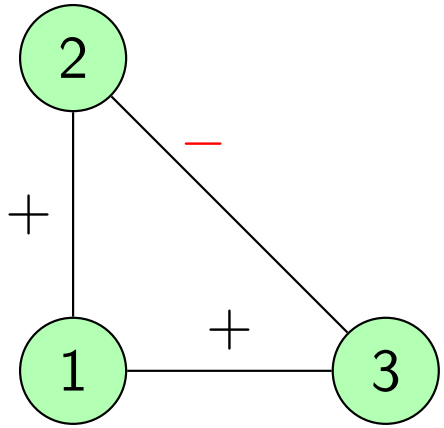
- Hard Clauses:

$$\{(\neg x_{12} \vee \neg x_{23} \vee x_{13}), (\neg x_{12} \vee \neg x_{13} \vee x_{23}), (\neg x_{23} \vee \neg x_{13} \vee x_{12})\}$$

- Soft Clauses:

$$\{(x_{12}), (x_{13}), (\neg x_{23})\}$$

Example of Enc 1



- Hard Clauses:

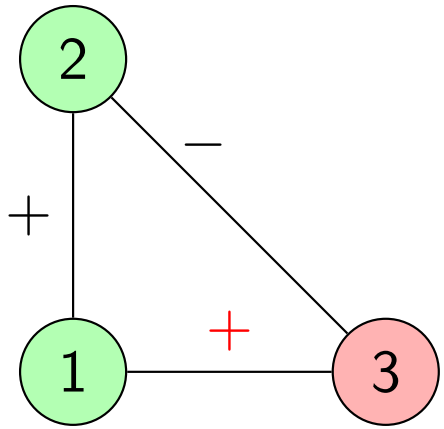
$$\{(\neg x_{12} \vee \neg x_{23} \vee x_{13}), (\neg x_{12} \vee \neg x_{13} \vee x_{23}), (\neg x_{23} \vee \neg x_{13} \vee x_{12})\}$$

- Soft Clauses:

$$\{(x_{12}), (x_{13}), (\neg x_{23})\}$$

- 1 Cluster. Cost: 1.

Example of Enc 1



- Hard Clauses:

$$\{(\neg x_{12} \vee \neg x_{23} \vee x_{13}), (\neg x_{12} \vee \neg x_{13} \vee x_{23}), (\neg x_{23} \vee \neg x_{13} \vee x_{12})\}$$

- Soft Clauses:

$$\{(x_{12}), (x_{13}), (\neg x_{23})\}$$

- 2 Clusters. Cost: 1.

Constructing a clustering from a MaxSAT solution

- Given MaxSAT solution τ.
- Assign $cl_\tau(v_1) = 1$ and $cl_\tau(v_i) = 1$ when $\tau(x_{1i}) = 1$.
- Iterate until all points assigned.

Theorem

The clustering cl_{τ^*} constructed as above from an optimal τ^* is optimal.

Constructing a clustering from a MaxSAT solution

- Given MaxSAT solution τ.
- Assign $cl_\tau(v_1) = 1$ and $cl_\tau(v_i) = 1$ when $\tau(x_{1i}) = 1$.
- Iterate until all points assigned.

Theorem

The clustering cl_{τ^*} constructed as above from an optimal τ^* is optimal.

MaxSAT Encoding 2

- More compact than encoding 1.
- Allows an upper limit K on number of clusters.
- $\mathcal{O}(|E| \cdot K + n \cdot K)$ variables and $\mathcal{O}(|E| \cdot K)$ clauses.
  - In practice requires: $K < n$.

Variables

- y_{ik} for all points v_i and clusters k.
- $y_{ik} = 1$ if $cl(v_i) = k$.

Optimality

- Given a MaxSAT solution τ assign $cl_\tau(v_i) = k$ when $\tau(y_{ik}) = 1$.
- If τ is an optimal truth assignment, then cl_τ is an optimal clustering.

MaxSAT Encoding 2

- More compact than encoding 1.
- Allows an upper limit K on number of clusters.
- $\mathcal{O}(|E| \cdot K + n \cdot K)$ variables and $\mathcal{O}(|E| \cdot K)$ clauses.
  - In practice requires: $K < n$.

Variables

- y_{ik} for all points v_i and clusters k.
- $y_{ik} = 1$ if $cl(v_i) = k$.

Optimality

- Given a MaxSAT solution τ assign $cl_\tau(v_i) = k$ when $\tau(y_{ik}) = 1$.
- If τ is an optimal truth assignment, then cl_τ is an optimal clustering.

MaxSAT Encoding 2

- More compact than encoding 1.
- Allows an upper limit K on number of clusters.
- $\mathcal{O}(|E| \cdot K + n \cdot K)$ variables and $\mathcal{O}(|E| \cdot K)$ clauses.
  - In practice requires: $K < n$.

Variables

- y_{ik} for all points v_i and clusters k.
- $y_{ik} = 1$ if $cl(v_i) = k$.

Optimality

- Given a MaxSAT solution τ assign $cl_\tau(v_i) = k$ when $\tau(y_{ik}) = 1$.
- If τ is an optimal truth assignment, then cl_τ is an optimal clustering.

Hard Clauses

- Every point to exactly one cluster.

$$\sum_{k=1}^{K} y_{ik} = 1.$$

- In our work: *a sequential counter* [?]

Soft Clauses

- For all $s(v_i, v_j) = 1$:

$$\bigvee_{k=1}^{K} (y_{ik} \wedge y_{jk})$$

- For all $s(v_i, v_j) = 0$:

$$\bigwedge_{k=1}^{K} (\neg y_{ik} \vee \neg y_{jk})$$

Optimizations to Encoding 2

- Redundant clauses that speed up the MaxSAT solver.

1. Symmetries: Enough to search over clusterings where $cl(v_1) = 1$.
   - Add: (y_{11}) and $(\neg y_{1k})$ for all $k = 2..K$.
2. Erroneous triangles can be exploited.
   1. Guaranteed min cost of 1.
   2. Lower cost of MaxSAT solutions in a controlled fashion.

Optimizations to Encoding 2

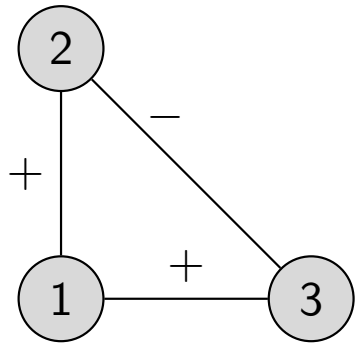
- Redundant clauses that speed up the MaxSAT solver.

1. Symmetries: Enough to search over clusterings where $cl(v_1) = 1$.
   - Add: (y_{11}) and $(\neg y_{1k})$ for all $k = 2..K$.
2. Erroneous triangles can be exploited.
   1. Guaranteed min cost of 1.
   2. Lower cost of MaxSAT solutions in a controlled fashion.

Optimizations to Encoding 2

- Redundant clauses that speed up the MaxSAT solver.

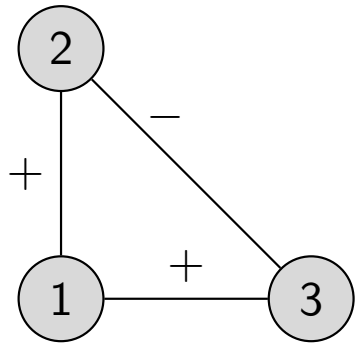
1. Symmetries: Enough to search over clusterings where $cl(v_1) = 1$.
   - Add: (y_{11}) and $(\neg y_{1k})$ for all $k = 2..K$.
2. Erroneous triangles can be exploited.
   1. Guaranteed min cost of 1.
   2. Lower cost of MaxSAT solutions in a controlled fashion.



Optimizations to Encoding 2

- Redundant clauses that speed up the MaxSAT solver.

1. Symmetries: Enough to search over clusterings where $cl(v_1) = 1$.
   - Add: (y_{11}) and $(\neg y_{1k})$ for all $k = 2..K$.
2. Erroneous triangles can be exploited.
   1. Guaranteed min cost of 1.
   2. Lower cost of MaxSAT solutions in a controlled fashion.

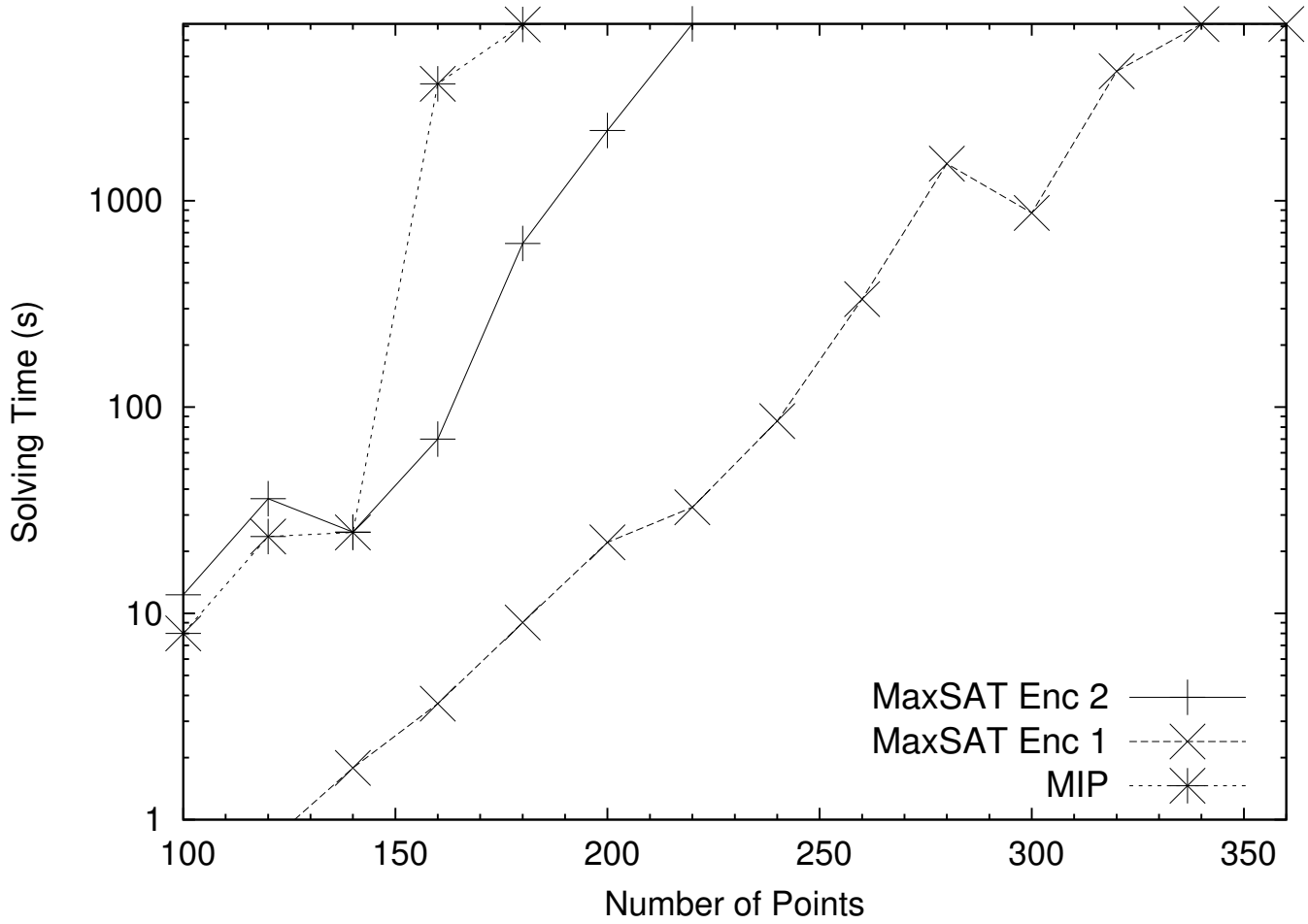


Preliminary experimental evaluation

- Clustering proteins given pairwise similarities (BLAST) of their amino acid sequences.
- 4 datasets and ground truth clusterings

  http://www.paccanarolab.org/software/scps/.
- Compare:
  - Academic MaxSAT solver (MaxHS)
  - Commercial MIP solver (Cplex)
  - Local search SCPS algorithm

Running time of the exact solvers



- Note: Cplex quickly runs out of memory.

Cost of obtained solutions

Limited number of points

Dataset	SCPS / Encoding 2 / GT	Enc 1
D1 300P	616/**491**/591	487
D2 320P	883/**786**/869	719
D3 260P	626/**470**/623	464
D4 200P	184/**106**/177	106

Whole datasets, edges pruned, K set according to GT

	SCPS Cost	Encoding 2 Cost
D1 669P	2452	**2042**
D2 586P	2167	**2153**
D3 567P	2596	**2146**
D4 654P	1881	**1643**

Cost of obtained solutions

Limited number of points

Dataset	SCPS / Encoding 2 / GT	Enc 1
D1 300P	616/**491**/591	487
D2 320P	883/**786**/869	719
D3 260P	626/**470**/623	464
D4 200P	184/**106**/177	106

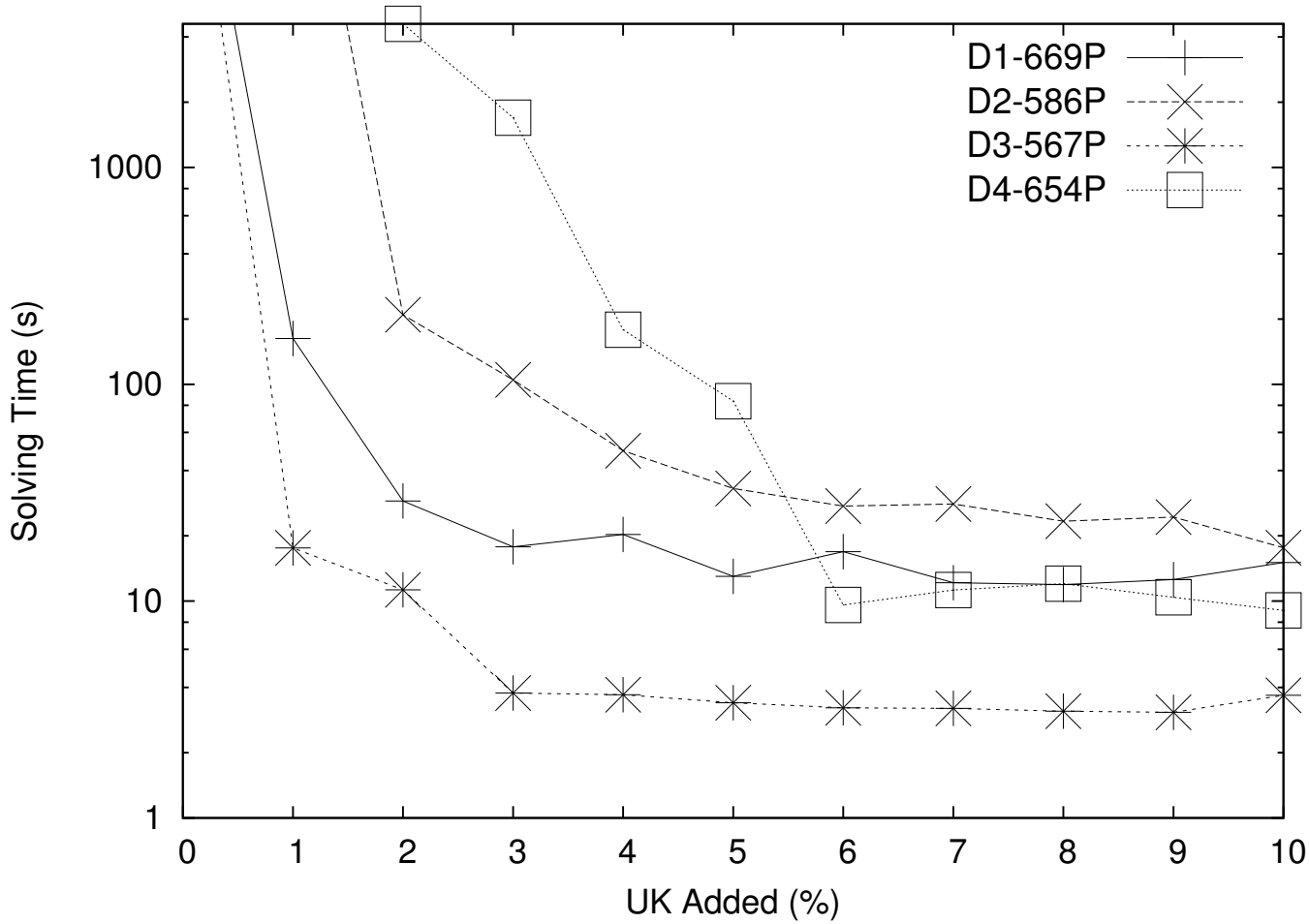
Whole datasets, edges pruned, K set according to GT

	SCPS Cost	Encoding 2 Cost
D1 669P	2452	**2042**
D2 586P	2167	**2153**
D3 567P	2596	**2146**
D4 654P	1881	**1643**

Constrained correlation clustering

- Guide the search towards clusterings of interest.
- Iteratively adding constraints based on previous solutions.
  - Can be non-trivial.
  - For example: "Only allow clusterings where at most 2 out of any 3 given points are assigned to the same cluster".
- Non trivial addition to problem.
  - MaxSAT running time improves.
- We simulated this setting by randomly sampling information from the ground truth clustering.

Running time in the presence of user knowledge



- Encoding 2 on the entire dataset with K set according to GT.

Precision and Recall

- Standard measures.
  - Measure similarity between a clustering and the GT, range from 0 to 1.
- F-Score: Harmonic mean of precision and recall.
- Optimizing F-score requires different choice of cost function.
  - Or a constrained setting.

F-scores with added user knowledge

Data Set	1% UK	2% UK	3% UK	4% UK	$\geq$ 5% UK
$D1$	0.954	0.997	0.997	1	1
$D2$	–	0.97	1	1	1
$D3$	0.986	0.996	1	1	1
$D4$	–	0.987	0.991	0.998	1

Precision and Recall

- Standard measures.
  - Measure similarity between a clustering and the GT, range from 0 to 1.
- F-Score: Harmonic mean of precision and recall.
- Optimizing F-score requires different choice of cost function.
  - Or a constrained setting.

F-scores with added user knowledge

Data Set	1% UK	2% UK	3% UK	4% UK	$\geq$ 5% UK
$D1$	0.954	0.997	0.997	1	1
$D2$	–	0.97	1	1	1
$D3$	0.986	0.996	1	1	1
$D4$	–	0.987	0.991	0.998	1

Conclusions

- A flexible framework for solving the correlation clustering problem cost-optimally.
- Works well for datasets up to hundreds of points.
- Extends to *constrained* correlation clustering and other variants.
  - Overlapping
  - Chromatic

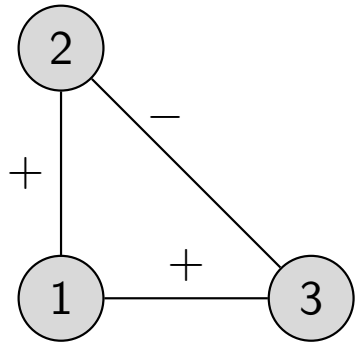
Further work

- Extension to weighted correlation clustering.
  - MaxSAT solver capable of handling non integral clause weights.
  - Allows use of any similarity measure directly.
- New, more compact encoding.
  - Allows both unrestricted and restricted number of clusters.
  - In a weighted setting, scales up to 669P (solved in under 3h).
- Possible other IP formulations.

Bibliography I

Optimizations to Encoding 2

1. Erroneous triangles can be exploited.
   1. Guaranteed min cost of 1.
   2. Relax soft clauses: $(r_1 \vee \bigvee_{k=1}^{K} A_{12k})$, $(r_2 \vee \bigvee_{k=1}^{K} A_{13k})$ and $(r_3 \vee \neg D_{23})$
   3. Add $r_1 + r_2 + r_3 = 1$ as hard.
   4. Iterate for all edge disjoint triangles.



Example applications of correlation clustering

- Categorical information.
  - Social networks.
  - Trajectories/GPS-data. [?]
  - Protein sequences. [?]
  - Crosslingual link detection. [?]
  - Image segmentation. [?]
- Combining clusterings.
- Agnostic learning.

MaxSAT Encoding 3

- Not included in the paper.
- More compact than enc 2
  - Allows $K = n$.
- $\mathcal{O}(|E| + n \cdot \log n)$ variables and $\mathcal{O}(|E| \cdot \log n)$ clauses.
- For each point $v_i \in V$ introduce $b_1^i, \ldots, b_{\log n}^i$.
- $cl(v_i) = b_{\log n}^i \ldots b_1^i$ as binary.

Constraints

- $cl(v_i) = cl(v_j)$ iff $b_k^i = b_k^j \quad \forall k = 1..\log n$
- Similarly: $cl(v_i) \neq cl(v_j)$ iff $\exists k : b_k^i \neq b_k^j$

MaxSAT Encoding 3

- Not included in the paper.
- More compact than enc 2
  - Allows $K = n$.
- $\mathcal{O}(|E| + n \cdot \log n)$ variables and $\mathcal{O}(|E| \cdot \log n)$ clauses.
- For each point $v_i \in V$ introduce $b_1^i, \ldots, b_{\log n}^i$.
- $cl(v_i) = b_{\log n}^i \ldots b_1^i$ as binary.

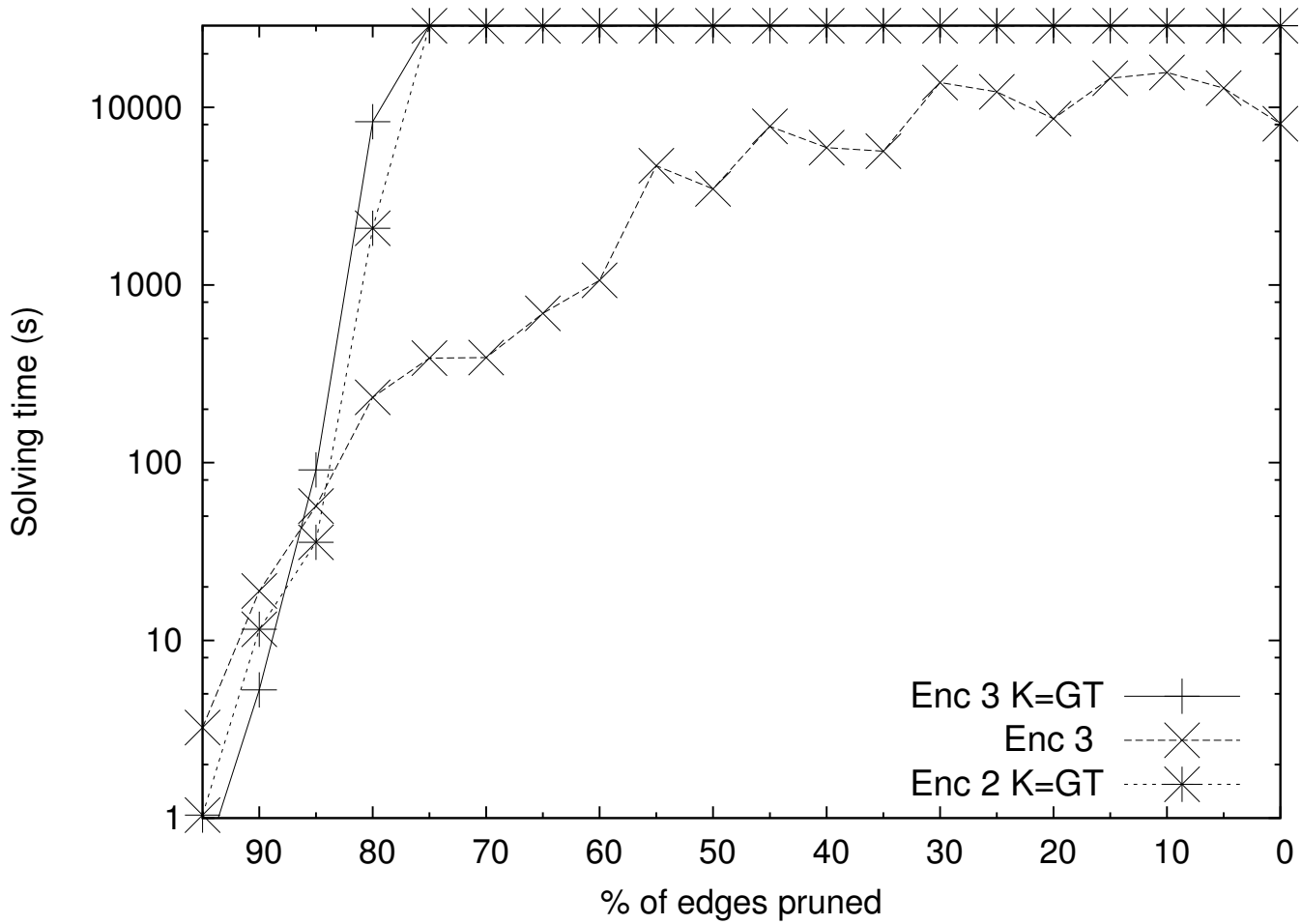
Constraints

- $cl(v_i) = cl(v_j)$ iff $b_k^i = b_k^j \quad \forall k = 1..\log n$
- Similarly: $cl(v_i) \neq cl(v_j)$ iff $\exists k : b_k^i \neq b_k^j$

Summary of encoding sizes

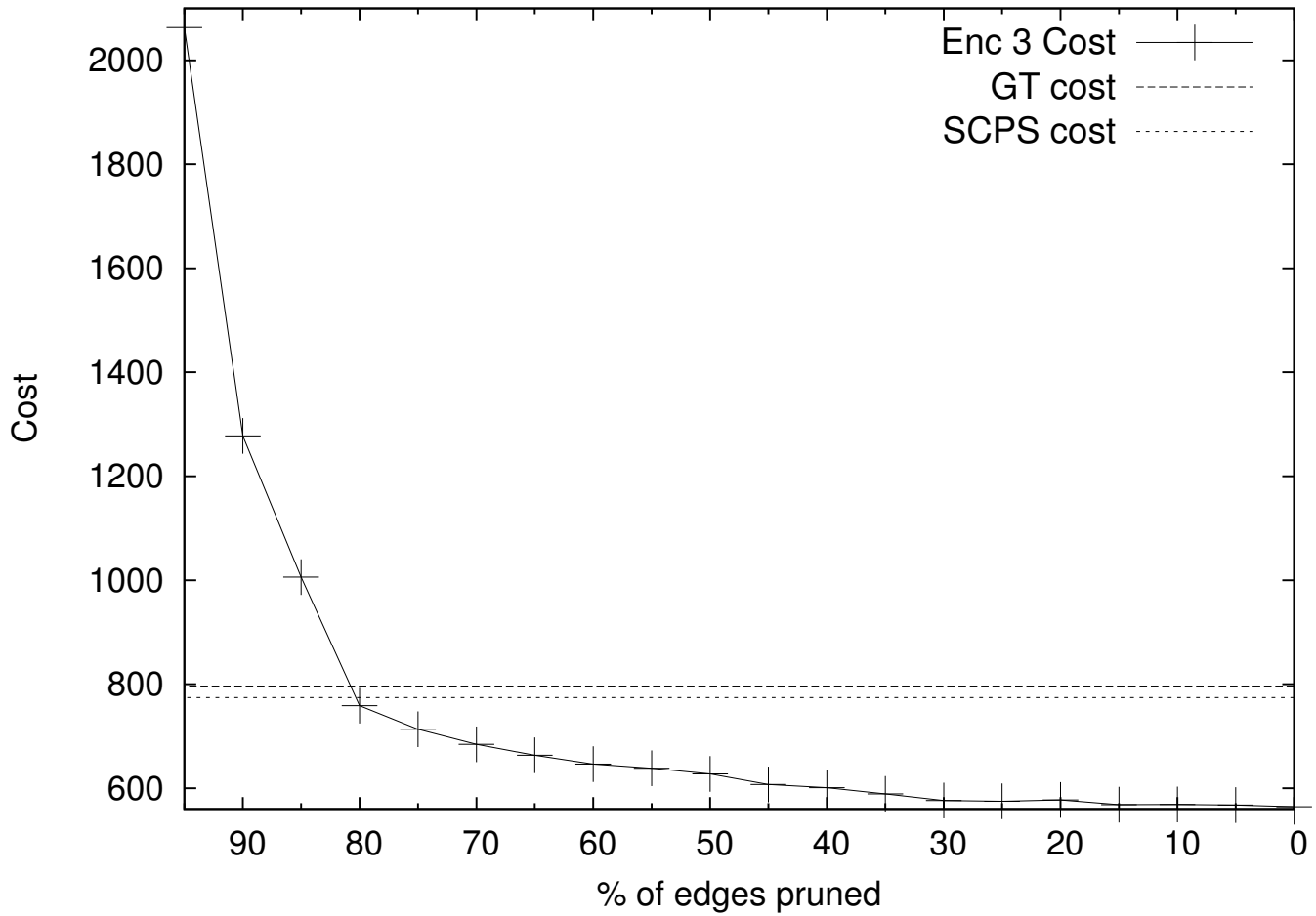
Encoding	Variables	Clauses
IP	$\mathcal{O}(n^2)$	$\mathcal{O}(n^3)$
1	$\mathcal{O}(n^2)$	$\mathcal{O}(n^3)$
2	$\mathcal{O}(\lvert E\rvert \cdot K + n \cdot K)$	$\mathcal{O}(\lvert E\rvert \cdot K)$
3	$\mathcal{O}(\lvert E\rvert + n \cdot \log n)$	$\mathcal{O}(\lvert E\rvert \cdot \log n)$

Preliminary results with Encoding 3



- Not comparable to results in paper, another computing environment and weighted setting.

Evolution of cost 3



- Not comparable to results in paper, another computing environment and weighted setting.