Constraint Programming for an Efficient and Flexible Block Modeling Solver

Alex Lucía Mattenet, Ian Davidson, Siegfried Nijssen, and Pierre Schaus

Plan of the presentation

1. Motivations
2. Problem Statement
3. CP Model for Block Modeling
4. Experiments and Results
5. Conclusions, Future Work

Motivations

Motivations

- interesting graph data
- Identifying the structure is hard
- One approach: **Block Modeling**
  - Large body of work on the subject
  - Many variations,
    including those that incorporate
    expert knowledge:
    knowledge about **core-periphery** structure, **must-link** relationships, ...

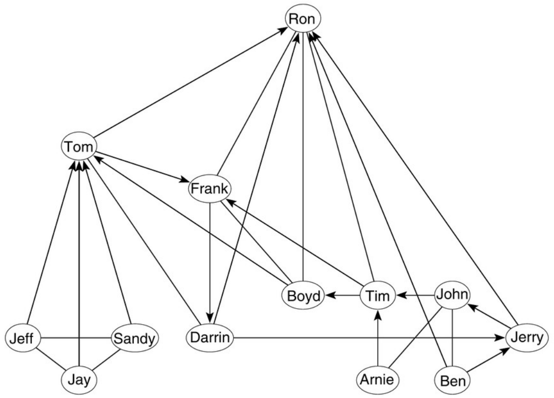
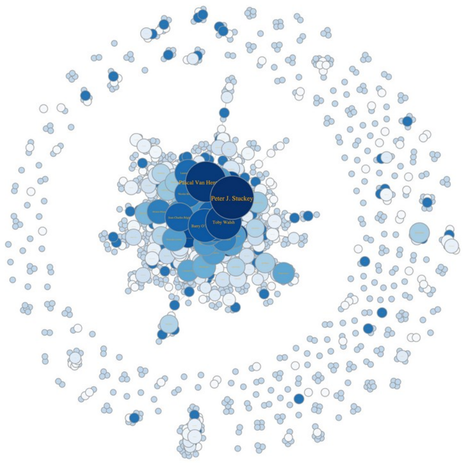
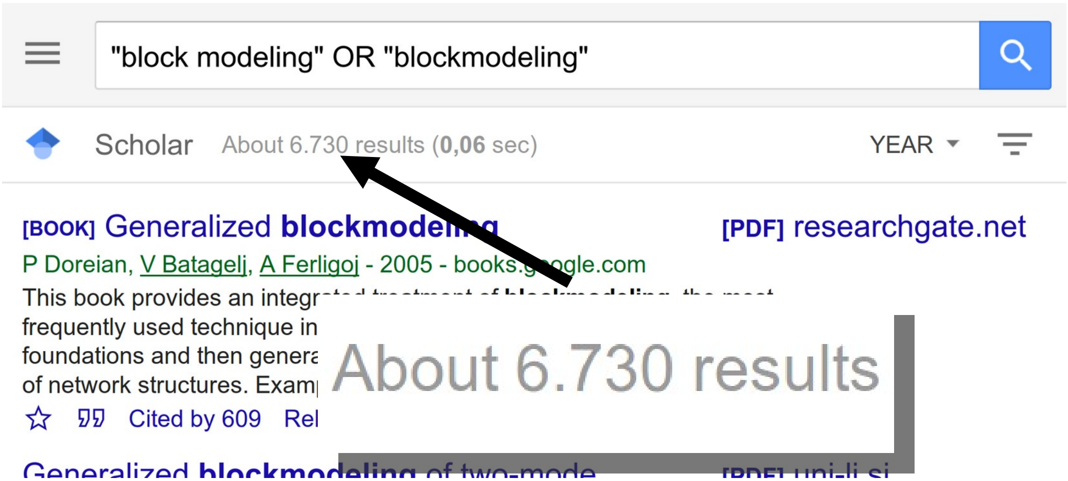


Figure 1.1. Transatlantic industries friendship ties.

Doreian, Patrick, Vladimir Batagelj, and Anuska Ferligoj. *Generalized Blockmodeling*. Cambridge University Press, 2005.



Teng Huang, Saharnaz Mehrani and David Bergman



UCLouvain

Artificial Intelligence & Algorithms

icteam

Motivations

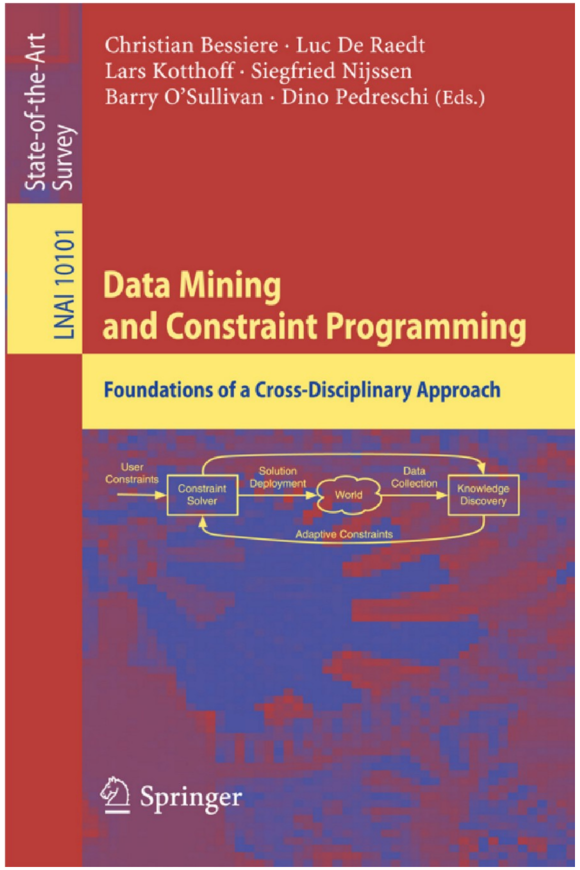
- Many approaches to solve the problem: local search,[1] gradient descent,[2] MIP,[3] …
- Many dedicated solvers with additional constraints: connected clusters,[4] must-link/cannot link,[5] image structure,[6] …
- Every solver is ad-hoc, no framework to combine constraints!

1. Batagelj, Vladimir, Anuška Ferligoj, and Patrick Doreian. "Direct and Indirect Methods for Structural Equivalence." *Social Networks*, Special Issue on Blockmodels, 14, no. 1 (March 1, 1992): 63–90.
2. Zhang, Yu, and Dit-Yan Yeung. "Overlapping Community Detection via Bounded Nonnegative Matrix Tri-Factorization." In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 606–614. KDD '12. New York, NY, USA: ACM, 2012.
3. Brusco, Michael J., and Douglas Steinley. "Integer Programs for One- and Two-Mode Blockmodeling Based on Prespecified Image Matrices for Structural and Regular Equivalence." *Journal of Mathematical Psychology* 53, no. 6 (December 1, 2009): 577–85.
4. Bai, Z., Qian, B., Davidson, I.: Discovering models from structural and behavioral brain imaging data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1128–1137. ACM (2018).
5. Ganji, M., Chan, J., Stuckey, P.J., Bailey, J., Leckie, C., Ramamohanarao, K., Park, L.: Semi-supervised blockmodelling with pairwise guidance. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 158–174. Springer (2018).
6. Ganji, M., Chan, J., Stuckey, P., Bailey, J., Leckie, C., Ramamohanarao, K., Davidson, I.: Image Constrained Blockmodelling: A Constraint Programming Approach. In: Proceedings of the 2018 SIAM International Conference on Data Mining, pp. 19–27. Proceedings, Society for Industrial and Applied Mathematics (May 2018).

Motivations

- Trend for use of CP in data mining
- Potential advantages:
  - More flexibility
  - Efficient

There is work on incorporating CP into existing iterative block modeling algorithms,[1] but **no efficient framework entirely in CP!**



2

1. Ganji, M., J. Chan, P. Stuckey, J. Bailey, C. Leckie, K. Ramamohanarao, and I. Davidson. "Image Constrained Blockmodelling: A Constraint Programming Approach," 19–27. Society for Industrial and Applied Mathematics, 2018.

2. Nijssen, Siegfried ; O'Sullivan, Barry ; et. al. *Data Mining and Constraint Programming - Foundations of a Cross-Disciplinary Approach.* Springer (2016) (ISBN:978-3-319-50136-9) 349 pages

Generic Constraint-based Block Modeling using Constraint Programming

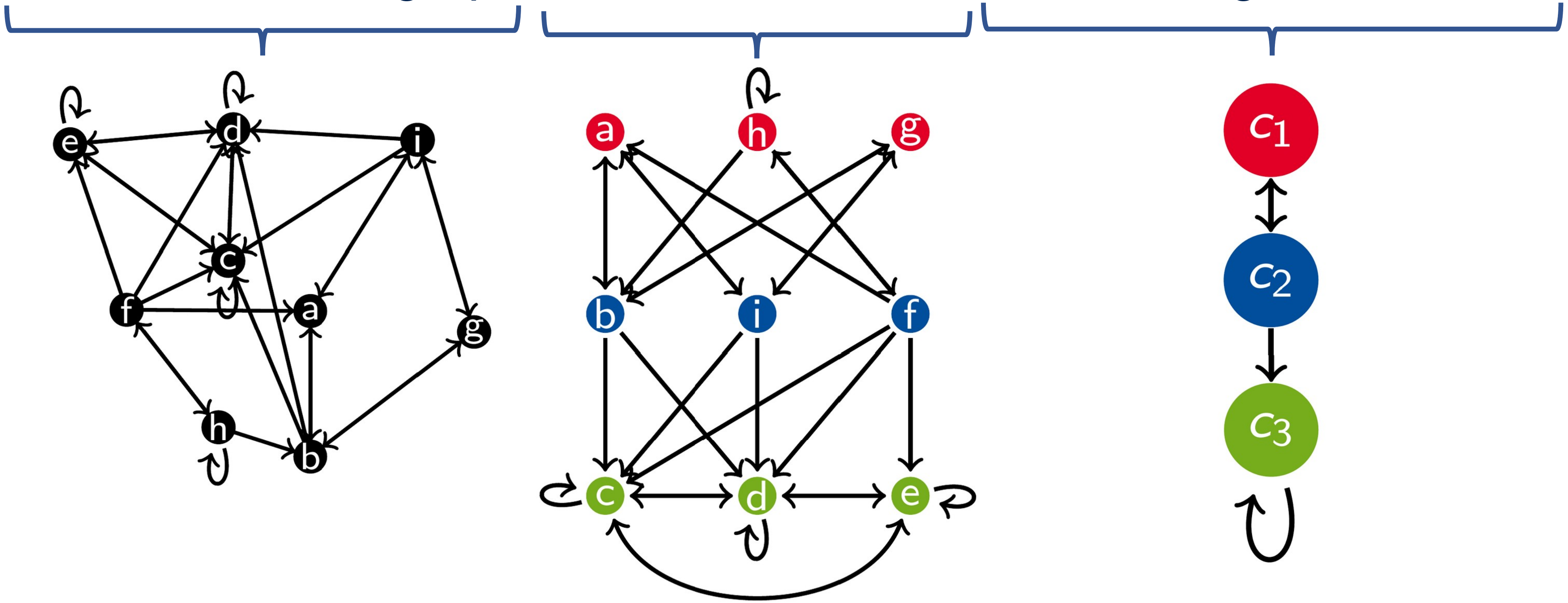
- Main objective of our work:
  practical efficiency and genericity
- Big focus on complexity of the filtering and on the search heuristics
- No focus on (arc-)consistency, idempotency, etc. if they hurt the practical performance.

https://www.ea.com/games/need-for-speed/need-for-speed-no-limits

Problem Statement

Block Modeling problem:

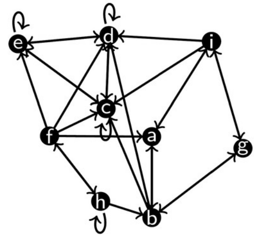
Given a directed graph, find vertex clusters, and their global structure



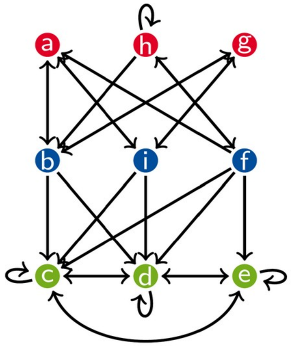
Block Modeling problem, formally :

Given a digraph with $n \times n$ **adjacency matrix** $\boldsymbol{X}$, find an assignment into k **clusters** $\boldsymbol{F}$ and a $k \times$k **image matrix** $\boldsymbol{M}$ such that $X \simeq F\,M\,F^T$

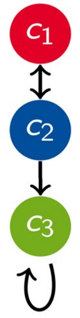
$$
X = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} a & b & c & d & e & f & g & h & i \end{array} \\
\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
\end{array}
$$



$$
F = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \end{array}
\begin{array}{c}
\begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0
\end{pmatrix}
\end{array}
$$



$$
M = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \end{array}
\begin{array}{c}
\begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 0 & 1
\end{pmatrix}
\end{array}
$$



The objective function to minimize is $||X - F\,M\,F^T||_1$.

UCLouvain

icteam

$$M = \begin{array}{c|ccc} & c_1 & c_2 & c_3 \\ \hline c_1 & 0 & 1 & 0 \\ c_2 & 1 & 0 & 1 \\ c_3 & \boxed{0} & 0 & \boxed{1} \end{array}$$

null block

complete block

Block Modeling problem, formally:

- Reordering the adjacency matrix to group the clusters together, blocks appear in the matrix.
- The (reordered) ideal matrix $F\,M\,F^T$ allows only **null blocks** (all 0) or **complete blocks** (all 1)

Cost of this solution = $||X - F\,M\,F^T||_1$ = #{*0* in complete blocks} + #{*1* in null blocks} = 9

$$\begin{array}{cc|ccc|ccc|ccc} & & & c_1 & & & c_2 & & & c_3 & \\ & & a & h & g & b & i & f & c & d & e \\ \hline c_1 & a & 0 & 0 & 0 & 1 & 1 & \mathit{0} & 0 & 0 & 0 \\ & h & 0 & \mathit{1} & 0 & 1 & \mathit{0} & 1 & 0 & 0 & 0 \\ & g & 0 & 0 & 0 & 1 & 1 & \mathit{0} & 0 & 0 & 0 \\ \hline c_2 & b & 1 & \mathit{0} & 1 & 0 & 0 & 0 & 1 & 1 & \mathit{0} \\ & i & 1 & \mathit{0} & 1 & 0 & 0 & 0 & 1 & 1 & \mathit{0} \\ & f & 1 & 1 & \mathit{0} & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline c_3 & c & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ & d & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ & e & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}$$

nullblock

complete block

CP Model for Block Modeling

UCLouvain

Artificial Intelligence & Algorithms

icteam

CP model: Variables

Assume k clusters and n vertices

Variable	Domain	Interpretation
C_i	$\{1..k\}$	$\mathsf{C}_i = c$ if vertex i is in cluster c
M_{cd}	$\{0,1\}$	$\mathsf{M}_{cd} = 0$ if the submatrix of rows in cluster c and columns in cluster d is a Null block, and $\mathsf{M}_{cd} = 1$ if it is a Complete Block
cost_{cd}	$\{0..n^2\}$	Number of entries in the submatrix c, d which do not match M_{cd}
$\mathsf{totalCost}$	$\{0..n^2\}$	The cost of the solution $\lVert X - FMF^T \rVert$

$$
\begin{array}{cc|c}
 & & \begin{array}{ccc|ccc|ccc} \multicolumn{3}{c|}{c_1} & \multicolumn{3}{c|}{c_2} & \multicolumn{3}{c}{c_3} \\ a & h & g & b & i & f & c & d & e \end{array} \\
c_1 & \begin{array}{c} a \\ h \\ g \end{array} & \\
c_2 & \begin{array}{c} b \\ i \\ f \end{array} & \\
c_3 & \begin{array}{c} c \\ d \\ e \end{array} &
\end{array}
\left(\begin{array}{ccc|ccc|ccc}
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
\hline
1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1
\end{array}\right)
$$

$$
M = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \end{array}
\begin{array}{c}
\begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\
\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}
\end{array}
$$

In our example:

- $C_a = 1$, $C_b = 2$, $C_c = 3$, …
- $M_{11} = 0$, $M_{12} = 1$, $M_{13} = 0$, …
- $\mathrm{cost}_{11} = 1$, $\mathrm{cost}_{12} = 3$, $\mathrm{cost}_{13} = 0$, …
- totalCost = 9

CP model: Constraints

- Sum(cost, totalCost) — cost and totalCost stay consistent
- $\forall$ i $\in$ {1..k}, atLeast(1, C, i) — all clusters contain at least 1 vertex
- BlockModelCost(C,M,cost,totalCost):
  A **new global constraint** we introduced in the paper.
  It filters (some) inconsistent values from the domains of C, M, cost, and totalCost.

Additional constraints can be added to C (constraints on the clusters) and on M (constraints on the image graph)

Experiments

UCLouvain

Artificial Intelligence & Algorithms

icteam

- Our CP model can find an exact **(optimal) solution** or approximation using **local search** (LNS)
- We compared the **optimal solution** setting with an existing MIP approach[1]
- We compared the **local search** with a well-known implementation in Pajek[2]

- CP system implemented in OscaR[3]
- MIP model written and solved in Java using Gurobi
- All experiments were run on a computer with Xeon Platinum 8160 24c/48t HyperThread processors

1. Dabkowski, M., Fan, N., Breiger, R.: Exploratory blockmodeling for one-mode, unsigned, deterministic networks using integer programming and structural equivalence. Social Networks 47, 93–106 (Oct 2016).
2. Batagelj, V., Mrvar, A., Ferligoj, A., Doreian, P.: Generalized blockmodeling with
pajek. Metodoloski zvezki 1(2), 455 (2004)

3. OscaR Team: OscaR: Scala in OR (2012), available from https://bitbucket.org/oscarlib/oscar

UCLouvain

Artificial Intelligence & Algorithms

icteam

We compared the MIP method with 3 approaches

1. **CP(bsl)**: A simple CP model without our global constraint
2. **CP(our)**: CP model with our BlockModelCost global constraint
3. **+heuris.**: CP model with our global contraint and a custom value heuristic to guide the search

Table 3. Run time of the MIP approach compared to a baseline CP approach (CP(bsl)), a CP approach with our global constraint (CP(our)), and our constraint + our value heuristic (+heuris.) for different number of clusters k. "–" indicates a timeout after 2 hours.

			CPU time (s)			
dataset	n	k	MIP	CP(bsl)	CP(our)	+heuris.
Transatlantic	13	2	1.73	0.80	0.45	0.28
		3	142.25	21.15	0.88	0.79
		4	–	386.20	2.94	2.07
Sharpstone	13	2	1.24	0.50	0.44	0.19
		3	62.46	13.57	1.17	0.85
		4	2952.13	221.41	2.78	1.82
		5	–	1102.68	2.31	1.30
Political Actor	14	2	2.14	1.13	0.62	0.31
		3	155.90	60.15	1.32	0.89
		4	2178.42	1936.43	2.68	2.20
		5	–	–	2.93	2.25
Search And Rescue	20	2	13.31	22.04	0.85	0.48
		3	–	–	6.01	5.18

Comparison with Pajek's local search

Pajek has a limit of 256 vertices

We plot the cost of the best solution as a function of time

LNS starts with worse solutions but finds better solutions fast

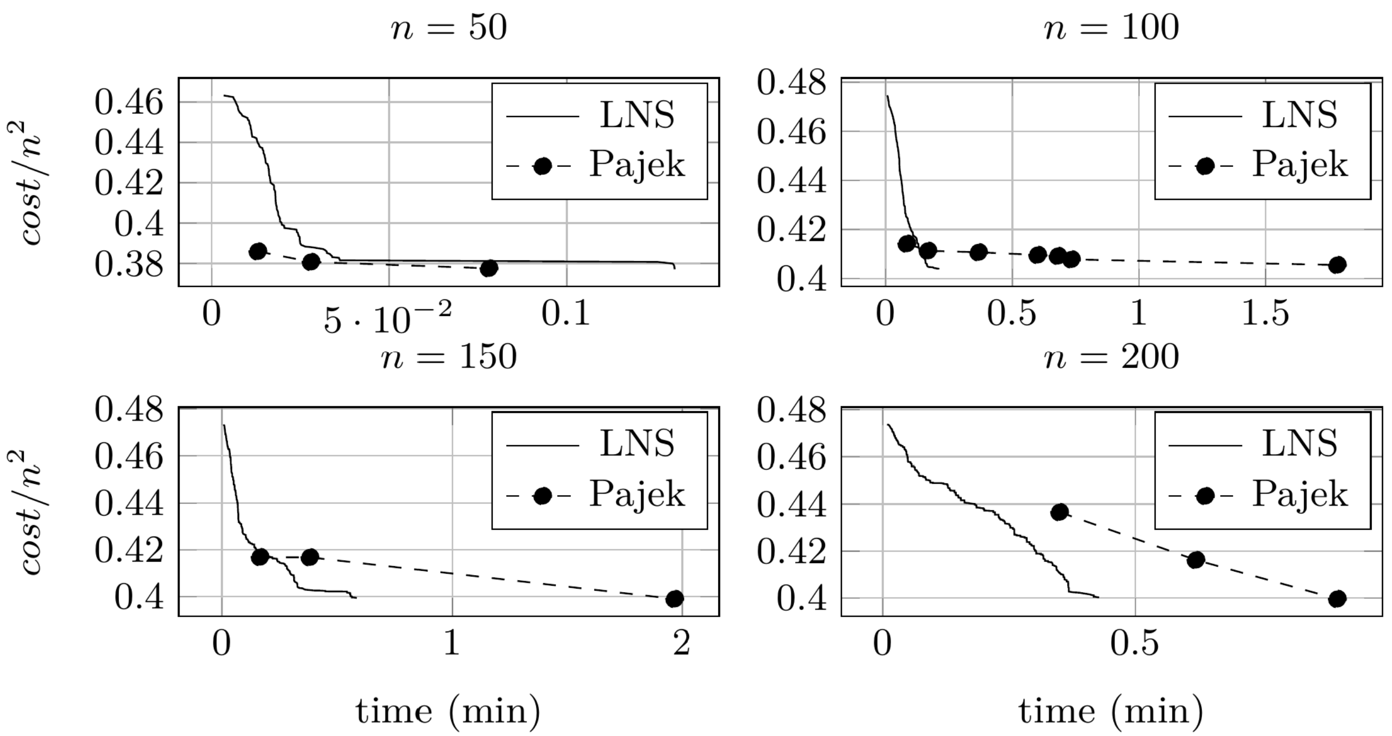
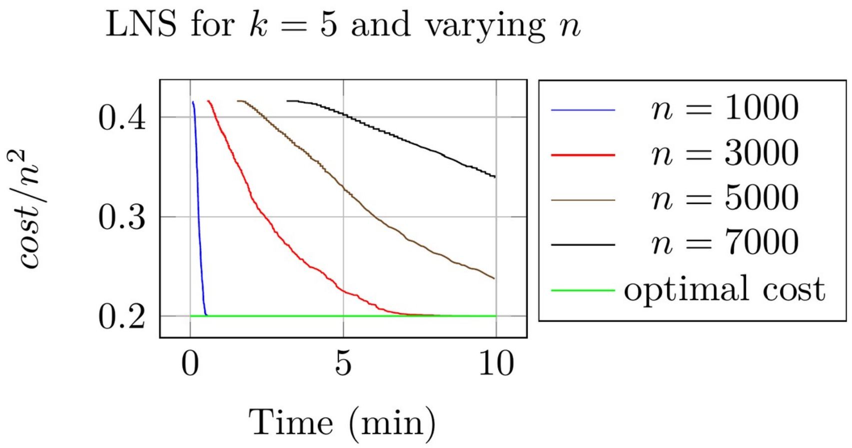
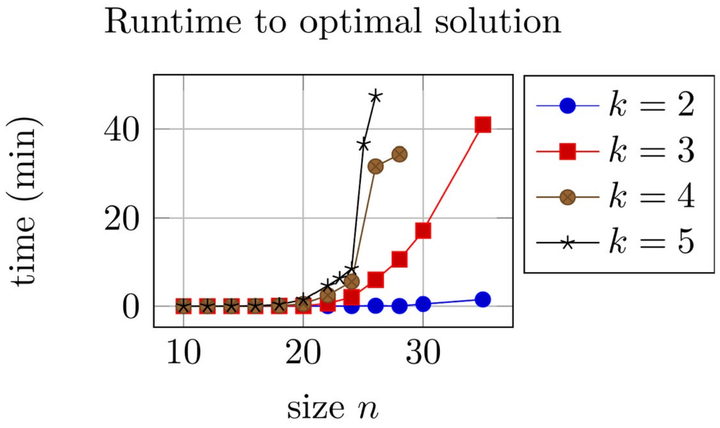


Fig. 3. Comparison of LNS search with the local search bundled in Pajek for synthetic dataset with 40% of noise. Each graph shows a different instance of the problem, for $n = 50$ to $n = 200$ by increments of 50.

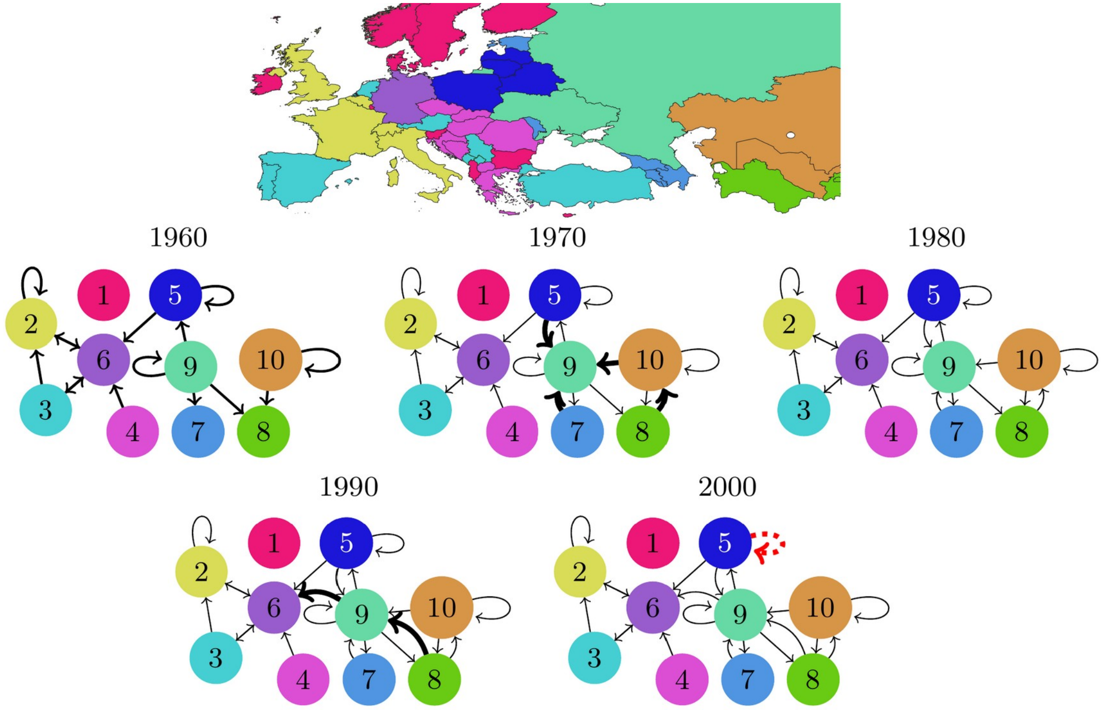
More experiments in the paper

Scalability





Application with additional Constraints (RESCAL like)



Bai, Z., Qian, B., Davidson, I.: Discovering models from structural and behavioral brain imaging data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1128–1137. ACM (2018)

Conclusions and Future Work

- Our new global constraint allows to solve constraint-based block modeling problems
  - Optimally, more efficiently than MIP
  - Heuristically, more efficiently than a specialized algorithm
- Many variations of block modeling can be studied using a similar approach:
  - Nonboolean image matrices
  - Attributes on the nodes/edges
  - For variable numbers of clusters
  - Weights for different types of errors
  - ...