FINDING MISSING DATA BY MODELLING AND SOLVING THE MCC PROBLEM

AUTHORS
Boukria Ali
Farouk Toumani
AFFILIATIONS
LIMOS - ISIMA

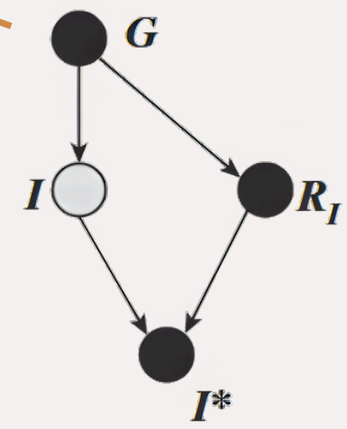
MISSING DATA

G	I
F	?
F	?
?	H
M	L

Extracts blocks

Compute tuple probabilities

MISSINGNESS GRAPHS [1]



(b)

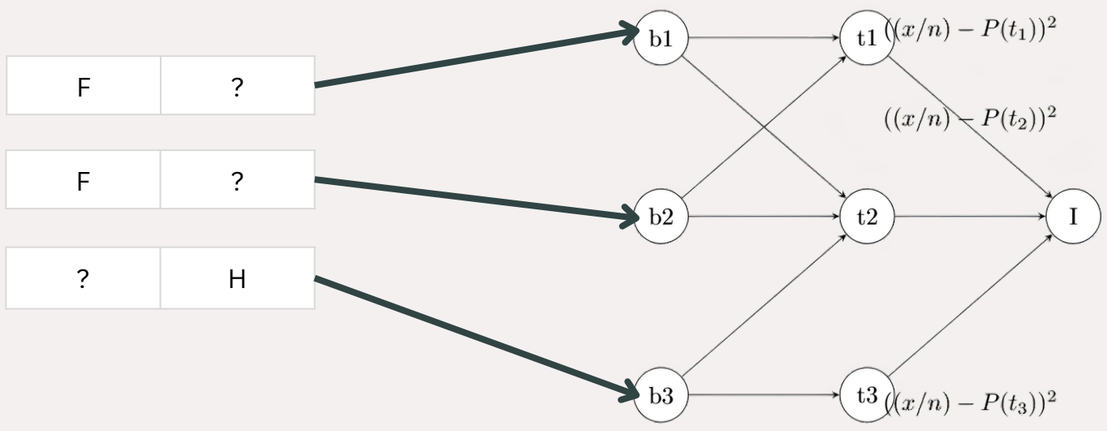
BLOCK INDEPENDENT DATABASES[2]

B	G	I	P(T)
B1	F	L	P(I=L|G=F)
B1	F	H	P(I=H|G=F)
B2	F	L	P(I=L|G=F)
B2	F	H	P(I=H|G=F)
B3	M	H	P(F=M|I=H)
B3	F	H	P(G=F|I=H)
B4	M	L	1

A MODELISATION OF THE MCC PROBLEM

The MCC problem can be reduced to a min cost flow problem with convex costs on the edge flows with integer flows.

1. The compliance of a class is a euclidian distance of the repetitions of tuples.
2. A set of quadratic equations can be extracted from said euclidian distance.
3. Each quadratic distance can be assigned to one of the final edges in a min cost flow graph.
4. These edge's flows indicate how many times a tuple was selected and thus the class.



An example of a class :

- A world is a possible imputation.
- A class is a set of imputations with equally repeated tuple.
- The probability of a class is the sum of the probability of its worlds

G	I
F	L
F	H
M	H
M	L

World 1
Probability :
P(I=H|G=F)*P(I=L|G=F)*P(F=M|I=H)

G	I
F	H
F	L
M	H
M	L

World 2
Probability :
P(I=H|G=F)*P(I=L|G=F)*P(F=M|I=H)

The compliance of a class with **K** unique tuples : $\sqrt{\sum_{i=0}^{k} (\#ti - Pmg(ti))^2}$

The total cost of the min cost flow : $\sum_{i=0}^{k} ((x(ti, I) + Kti) - Pmg(ti))^2$

CUTTING EDGE TECHNIQUES

There have been a variety of approaches and research projects done to explore methods and techniques for finding integer flow solutions to a min cost flow problem with convex as opposed to linear flow costs. We have prioritized exploring them, understanding them, and analyzing their implementations or theoretical performance to decide which might be most appropriate for our approach, if any are. These include :

SCIP[3]

- SCIP is a very fast non-commercial solver for MIP and MINLP problems.
- The minimum cost flow problem can be formulated as a MINLP.
- It also provides a flexible framework for constraint integer programming and branch-cut-and-price.

SOLVING THE CONVEX COST INTEGER DUAL NETWORK FLOW PROBLEM [4]

- This article by Ahuja, Hochbaum, and Orlin surveys insights and methods for solving the dual network flow problem.
- It synthesizes results from earlier work and contemporary studies under the previously defined parameters.
- The product is a theoretically polynomial time algorithm.

GENERAL CONCLUSIONS AND NOTES

- An initial analysis of methodologies for solving the convex cost integer network flow problem shows that it is a difficult problem to solve, with some publicly available implementations of solviers relying on branch and bound techniques (Exponential complexity in worst case scenario, though much faster in practice) .

REFERENCES :

1. Graphical Models for Inference with Missing Data by Mohan, Karthika and Pearl, Judea and Tian, Jin.
2. Query Answering in Incomplete Databases under Missingness Mechanisms by Leopoldo Bertossi∗ (Carleton Univ., Canada & IMFD, Chile), Maxime Buron, Idris Moulay, Farouk Toumani (LIMOS, CNRS, UCA, France).
3. The SCIP Optimization Suite 10.0 by Christopher Hojny, Mathieu Besançon, Ksenia Bestuzheva, Sander Borst, Antonia Chmiela, João Dionísio, Leon Eifler, Mohammed Ghannam, Ambros Gleixner, Adrian Göß, Alexander Hoen, Rolf van der Hulst, Dominik Kamp, Thorsten Koch, Kevin Kofler, Jurgen Lentz, Stephen J. Maher, Gioni Mexi, Erik Mühmer, Marc E. Pfetsch, Sebastian Pokutta, Felipe Serrano, Yuji Shinano, Mark Turner, Stefan Vigerske, Matthias Walter, Dieter Weninger, Liding Xu.
4. Ravindra K. Ahuja, Dorit S. Hochbaum, James B. Orlin, (2003) Solving the Convex Cost Integer Dual Network Flow Problem. Management Science 49(7):950-964.

FURTHER RELATED READINGS :

- A Strongly Polynomial Algorithm for a Class of Minimum-Cost Flow Problems with Separable Convex Objectives by László A. Végh.
- A polynomial algorithm for minimum quadratic cost flow problems by M. Minoux .