# A Short Note on Discrete Representability of Independence Models

Petr Šimeček
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4,
182 08 Prague, Czech Republic

**Abstract**

The paper discusses the problem to characterize collections of conditional independence triples (independence model) that are representable by a discrete distribution. The known results are summarized and the number of representable models over four elements set, often mistakenly claimed to be 18300, is corrected. In the second part, the bounds for the number of positively representable models over four elements set are derived.

## 1 Introduction

Conditional independence relationships occur naturally among components of highly structured stochastic systems. In a field of graphical Markov models, a graph (nodes connected by edges) is used to represent the CI structure of a set of probability distributions.

Given the joint probability distribution of a collection of random variables it is easy to construct the list of all CIs among them. On the other hand, given the list (here called an independence model) of CIs an interesting question arises whether there exists a collection of discrete random variables meeting these and only these CIs, i.e. representing that model.

The problem of probabilistic representability comes originally from J. Pearl, cf. (Pearl, 1998). It was proved by M. Studený in (Studený, 1992) that there is no finite characterization (≡finite set of inference rules) of the set of representable independence models.

The interesting point is that for the fixed number of variables (or vertices) the number of representable independence models is much higher than the number of graphs. Therefore, even partial characterization of representable independence models may help to improve and understand the limits of learning of Bayesian networks and Markov models.

## 2 Independence models

For the reader's convenience, auxiliary results related to independence models are recalled in this section.

Throughout the paper, the singleton {a} will be shorten by $a$ and the union of sets $A \cup B$ will be written simply as juxtaposition $AB$. A random vector $\boldsymbol{\xi} = (\xi_a)_{a \in N}$ is a collection of random variables indexed by a finite set $N$. For $A \subseteq N$, a subvector $(\xi_a)_{a \in A}$ is denoted by $\boldsymbol{\xi}_A$; $\boldsymbol{\xi}_\emptyset$ is presumed to be a constant. Analogously, if $\boldsymbol{x} = (x_a)_{a \in N}$ is a constant vector then $\boldsymbol{x}_A$ is an appropriate coordinate projection.

Provided $A, B, C$ are pairwise disjoint subsets of $N$, "$\boldsymbol{\xi}_A \perp\!\!\!\perp \boldsymbol{\xi}_B | \boldsymbol{\xi}_C$" stands for a statement $\boldsymbol{\xi}_A$ and $\boldsymbol{\xi}_B$ are conditionally independent given $\boldsymbol{\xi}_C$. In particular, unconditional independence ($C = \emptyset$) is abbreviated as $\boldsymbol{\xi}_A \perp\!\!\!\perp \boldsymbol{\xi}_B$.

A random vector $\boldsymbol{\xi} = (\xi_a)_{a \in N}$ is called **discrete** if each $\xi_a$ takes values in a state space $X_a$ such that $1 < |X_a| < \infty$. A discrete random vector $\boldsymbol{\xi}$ is called **positive** if for any appropriate constant vector $\boldsymbol{x}$

$$0 < P(\boldsymbol{\xi} = \boldsymbol{x}) < 1.$$

In the case of discretely distributed random vector, variables $\xi_a$ and $\xi_b$ are independent[1] given

[1]The independence relation between random vectors

$\boldsymbol{\xi}_C$ iff for any appropriate constant vector $\boldsymbol{x}_{abC}$

$$P(\boldsymbol{\xi}_{abC} = \boldsymbol{x}_{abC})P(\boldsymbol{\xi}_C = \boldsymbol{x}_C) \quad = \\ P(\boldsymbol{\xi}_{aC} = \boldsymbol{x}_{aC})P(\boldsymbol{\xi}_{bC} = \boldsymbol{x}_{bC}).$$

Let $N$ be a finite set and $\mathcal{T}_N$ denotes the set of all pairs $\langle ab|C\rangle$ such that $ab$ is an (unordered) couple of distinct elements of $N$ and $C \subseteq N \setminus ab$.

Subsets of $\mathcal{T}_N$ will be referred as formal **independence models** over $N$. Independence models $\emptyset$ and $\mathcal{T}_N$ are called trivial.

The independence model $\mathcal{I}(\boldsymbol{\xi})$ induced by a random vector $\boldsymbol{\xi}$ indexed by $N$ is the independence model over $N$ defined as follows

$$\mathcal{I}(\boldsymbol{\xi}) = \{\langle ab|C\rangle;\ \xi_a \perp\!\!\!\perp \xi_b | \boldsymbol{\xi}_C\}.$$

Let us emphasize that an independence model $\mathcal{I}(\boldsymbol{\xi})$ uniquely determines also all other conditional independences among subvectors of $\boldsymbol{\xi}$, cf. (Matúš, 1992).

Diagrams proposed by R. Lněnička will be used for a visualisation of independence model $I$ over $N$ such that $|N| \leq 4$. Each element of $N$ is plotted as a dot. If $\langle ab|\emptyset\rangle \in I$ then dots corresponding to $a$ and $b$ are joined by a line. If $\langle ab|c\rangle \in I$ then we put a line between dots corresponding to $a$ and $b$ and add small line in the middle pointing in $c$–direction. If both $\langle ab|c\rangle$ and $\langle ab|d\rangle$ are elements of $I$, then only one line with two small lines in the middle is plotted. Finally, if $\langle ab|cd\rangle \in I$ is visualised by a brace between $a$ and $b$. See example in Figure 1.

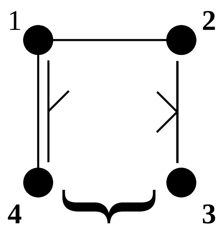


Figure 1: Diagram of the independence model $I = \{\langle 12|\emptyset\rangle, \langle 23|1\rangle, \langle 23|4\rangle, \langle 34|12\rangle, \langle 14|\emptyset\rangle, \langle 14|2\rangle\}$.

Independence models $I$ and $I^*$ over $N$ will be called **isomorphic** if there exists a permutation $\pi$ on $N$ such that

$$\langle ab|C\rangle \in I \Longleftrightarrow \langle \pi(a)\pi(b)|\pi(C)\rangle \in I^*,$$

where $\pi(C)$ stands for $\{\pi(c);\ c \in C\}$. See Figure 2 for an example of three isomorphic models.

An equivalence class of independence models with respect to the isomorphic relation will be referred as **type**.

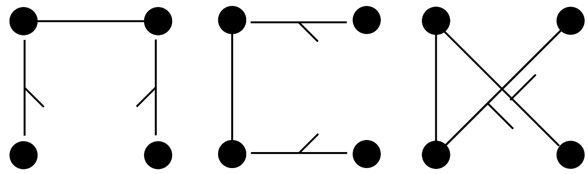

Figure 2: Example of three isomorphic models.

An independence model $I$ is said to be **representable**[2] if there exists a discretely distributed random vector $\boldsymbol{\xi}$ such that $I = \mathcal{I}(\boldsymbol{\xi})$. In addition, a special attention will be devoted to **positive representations**, i.e. representations by a positive discrete distribution.

Let us note that isomorphic models are either all representable or non-representable. Consequently, we can classify types as representable and non-representable.

**Lemma 1.** *If $I = \mathcal{I}(\boldsymbol{\xi})$ and $I^* = \mathcal{I}(\boldsymbol{\xi}^*)$ are representable independence models then the independence model $I \cap I^*$ is also representable. In particular, if they have positive representations then there exists a positive representation of $I \cap I^*$, too.*

*Proof.* Let $X = \prod X_a$ and $X' = \prod X'_a$ be state spaces of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, respectively. The required representation $\hat{\boldsymbol{\xi}}$ takes place in

$$\widehat{X} = \prod_{a \in N} X_a \times X'_a$$

and it is distributed as follows

$$P(\hat{\boldsymbol{\xi}} = (x_a, x'_a)_{a\in N}) = \\ P(\boldsymbol{\xi} = (x_a)_{a\in N}) \cdot P(\boldsymbol{\xi}' = (x'_a)_{a\in N}).$$

See (Studený and Vejnarová, 1998), pp. 5, for more details. □

$\boldsymbol{\xi}_A$, $\boldsymbol{\xi}_B$ given $\boldsymbol{\xi}_C$ might be defined analogously. However, we will see that such relationships are uniquely determined by the elementary ones ($|A| = |B| = 1$).

[2] Of course, it is also possible to consider representability in other distribution frameworks that the discrete distributions, cf. (Lněnička, 2005).

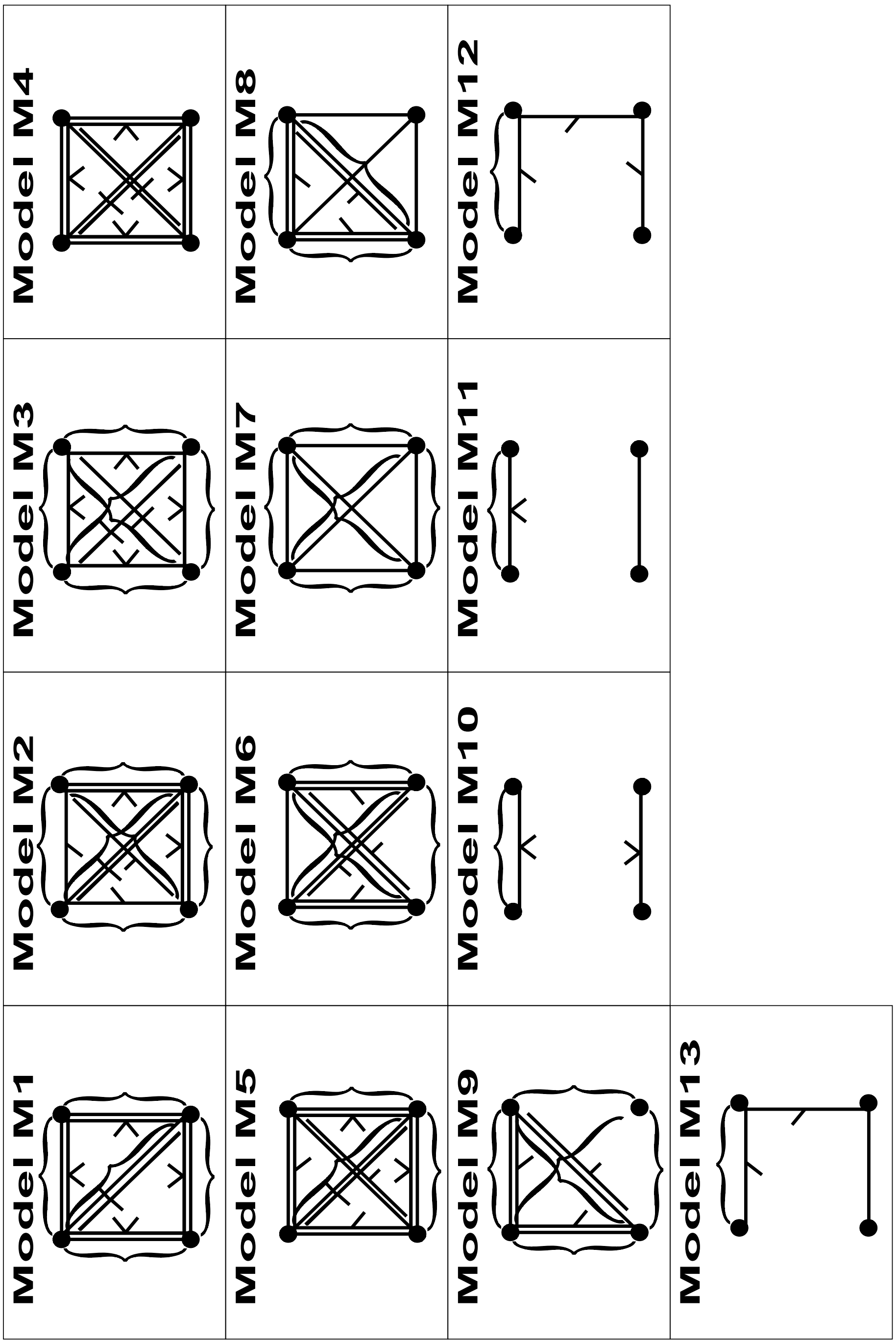


Figure 3: Irreducible models over $N = \{1, 2, 3, 4\}$.

**Lemma 2.** *Let $a, b, c$ be distinct elements of $N$ and $D \subseteq N \setminus abc$. If an independence models $I$ over $N$ is representable, then*

$$\big(\{\langle ab|cD\rangle, \langle ac|D\rangle\} \subseteq I\big) \iff \big(\{\langle ac|bD\rangle, \langle ab|D\rangle\} \subseteq I\big).$$

*Moreover, if $I$ is positively representable, then*

$$\big(\{\langle ab|cD\rangle, \langle ac|bD\rangle\} \subseteq I\big) \Longrightarrow \big(\{\langle ab|D\rangle, \langle ac|D\rangle\} \subseteq I\big).$$

*Proof.* These are so called "semigraphoid" and "graphoid" properties, cf. (Lauritzen, 1996) for the proof. □

# 3 Representability of Independence Models over $N = \{1, 2, 3, 4\}$

For $N$ consisting of three or less elements, all independence models not contradicting properties from Lemma 2 are representable, resp. positively representable, cf. (Studený, 2005). That is why we focus on $N = \{1, 2, 3, 4\}$ from now to the end of the paper.

The first subsection summarizes known results related to (general) representability of independence models. The second subsection is devoted to positive representability.

## 3.1 General Representability

The problem was solved in a brilliant series of papers (Matúš and Studený, 1995), (Matúš, 1995) and (Matúš, 1999) by F. Matúš and M. Studený. The final conclusions are clearly and comprehensibly presented in (Studený and Boček, 1994) and (Matúš, 1997)[3].

In brief, due to Lemma 1 an intersection of two representable models is also representable. Therefore, the class of all representable models over $N$ can be described by a set of so called **irreducible** models $\mathcal{C}$, i.e. nontrivial representable models that cannot be written as an intersection of two other representable models. It is not difficult to evidence that a nontrivial independence model $I$ is representable if and only if there exists $\mathcal{A} \subseteq \mathcal{C}$ such that

$$I = \bigcap_{C \in \mathcal{A}} C.$$

There are only only 13 types of irreducible models, see Figure 3 or (Studený and Boček, 1994), pp. 277–278. The problematic point is the total number of representable independence models over $N$. It has been believed that this number is 18300, cf. (Studený, 2002), (Lauritzen and Richardson, 2002), (Robins et al., 2003), (Šimeček, 2006)... However, working on this paper I have discovered that there actually exist 18478 different representable independence models over $N$ of 1098 types. The list of models has been checked by several programs including SG POKUS written by M. Studený and P. Boček. The list can be downloaded from the web page

`http://5r.matfyz.cz/skola/models`

## 3.2 Positive Representability

Only a little is known about positive representability. This paper would like to be the first step to the complete characterisation of positively representable models over $N = \{1, 2, 3, 4\}$.

Obviously, a set of positively representable models is a subset of the set of (generally) representable models. In addition, positively representable model must fulfill properties following the second part of Lemma 2 and Lemma 3 below.

**Lemma 3.** *Let $a, b, c, d$ be distinct elements of $N$. If $I$ is a positively representable independence model oven $N$ such that*

$$\{\langle ab|cd\rangle, \langle cd|ab\rangle, \langle cd|a\rangle\} \subseteq I,$$

*then*

$$\langle cd|b\rangle \in I \iff \langle cd|\emptyset\rangle \in I.$$

*Proof.* See (Spohn, 1994), pp. 15. □

There are 5547 (generally) representable models (356 types) meeting requirements on for positively representable models from Lemma 2 and Lemma 3. This is the upper bound to the set of all positively representable models.

[3] To avoid confusion, note that (Matúš, 1997) contains a minor typo in Figure 14 on pp. 21. Over the upper line in the first two diagrams should be $\emptyset$ instead of $*$.

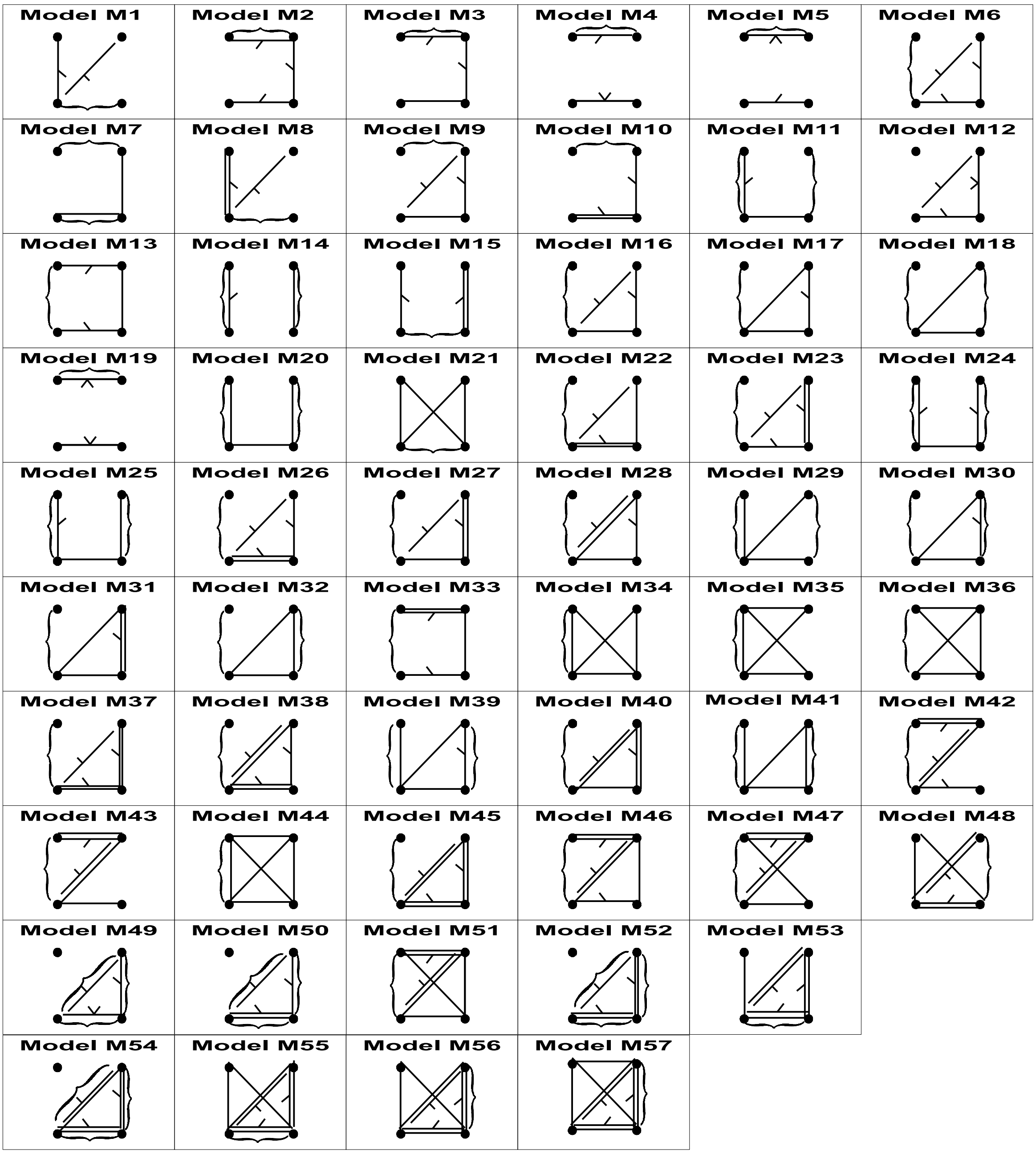


Figure 4: Undecided types.

Again, this class of models $\mathcal{A}$ is generated by its subset $\mathcal{C}$ by the operation of intersection. The elements of $\mathcal{C}$ can be found by starting with an empty set $\mathcal{C}$ and in each step adding the element of $\mathcal{A}$ not yet generated by $\mathcal{C}$ with the greatest size. Actually, $\mathcal{C}$ contains (nontrivial) models of 23 types (may be downloaded from the mentioned web page).

To find some lower bound to the set of positively representable models, large amount of positive binary ($\equiv$sample space $\{0,1\}^N$) random distributions have been randomly generated. The description of the generating process will be omitted here[4], see the above mentioned web page for the list of corresponding independence models and their binary representations. Using Lemma 1 we obtained 4555 (299 types) different positive representations of independence models. The remaining problematic 57 types are plotted in Figure 4.

### 3.3 Conclusion

Let us summarize the results into the concluding theorem.

**Theorem 1.** *There are* 18478 *different (generally) representable independence models (*1098 *types) over the set* $N = \{1, 2, 3, 4\}$*. There are between* 4555 *and* 5547 *different positively representable independence models (*299*–*356 *types) over the set* $N = \{1, 2, 3, 4\}$.

**Acknowledgments**

This work was supported by the grants GA ČR n. 201/04/0393 and GA ČR n. 201/05/H007.

## References

S. L. Lauritzen and T. S. Richardson. 2002. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:321–361.

S. L. Lauritzen. 1996. *Graphical Models*. Oxford University Press.

R. Lněnička. 2005. On gaussian conditional independence structures. Technical report, ÚTIA AV ČR.

F. Matúš and M. Studený. 1995. Conditional independences among four random variables I. *Combinatorics, Probability & Computing*, 4:269–278.

F. Matúš. 1992. On equivalence of markov properties over undirected graphs. *Journal of Applied Probability*, 29:745–749.

F. Matúš. 1995. Conditional independences among four random variables II. *Combinatorics, Probability & Computing*, 4:407–417.

F. Matúš. 1997. Conditional independence structures examined via minors. *The Annals of Mathematics and Artificial Intelligence*, 21:99–128.

F. Matúš. 1999. Conditional independences among four random variables III. *Combinatorics, Probability & Computing*, 8:269–276.

J. Pearl. 1998. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. 2003. Uniform consistency in causal inference. *Biometrika*, 90:491–515.

P. Šimeček. 2006. Gaussian representation of independence models over four random variables. Accepted to COMPSTAT conference.

W. Spohn. 1994. On the properties of conditional independence. In P. Humphreys, editor, *Patrick Suppes: Scientific Philosopher*, volume 1, pages 173–194. Kluwer.

M. Studený and P. Boček. 1994. CI-models arising among 4 random variables. In *Proceedings of the 3rd workshop WUPES*, pages 268–282.

M. Studený and J. Vejnarová. 1998. The multiinformation function as a tool for measuring stochastic dependence. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 261–298. Kluwer.

M. Studený. 1992. Conditional independence relations have no finite complete characterization. In S. Kubík and J.A. Víšek, editors, *Information Theory, Statistical Decision Functions and Random Processes. Transactions of the 11th Prague Conference*, volume B, pages 377–396. Kluwer.

M. Studený. 2002. On stochastic conditional independence: the problems of characterization and description. *Annals of Mathematics and Artificial Intelligence*, 35:241–323.

M. Studený. 2005. *Probabilistic Conditional Independence Structures*. Springer.

[4] Briefly, the parametrization of joint distribution by moments has been used. Further, both systematic and random search through parameter space was performed.