UNIVERSITY OF SÃO PAULO
SÃO CARLOS SCHOOL OF ENGINEERING

Pedro Virgilio Basilio Jeronymo

Markov Blanket Discovery without Causal Sufficiency: Application in Credit Data

São Carlos

2021

Pedro Virgilio Basilio Jeronymo

Markov Blanket Discovery without Causal Sufficiency: Application in Credit Data

Master Dissertation submitted to the São Carlos School of Engineering, in partial fulfillment of the requirements for the degree of Master of Science.

Concentration Area: Dynamical Systems

Advisor: Carlos Dias Maciel, Ph.D.

Corrected Version

São Carlos

2021

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Virgilio Basilio Jeronymo, Pedro
V372d Descoberta de Markov Blankets Sem Suficiência Causal: Aplicação em Dados de Crédito / Pedro Virgilio Basilio Jeronymo; orientador Carlos Dias Maciel. São Carlos, 2021.

Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia Elétrica e Área de Concentração em Sistemas Dinâmicos -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2021.

1. Descoberta Causal. 2. Markov Blanket. 3. Redes Bayesianas. 4. Crédito. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidato: Engenheiro **PEDRO VIRGILIO BASÍLIO JERONYMO.**

Título da dissertação: “Descoberta de *Markov Blankets* sem suficiência causal: aplicação em dados de crédito”.

Data da defesa: 15/12/2021.

Comissão Julgadora	**Resultado**
Prof. Associado **Carlos Dias Maciel** **(Orientador)** (Escola de Engenharia de São Carlos – EESC/USP)	Aprovado
Prof. Dr. **Américo Barbosa da Cunha Júnior** (Universidade do Estado do Rio de Janeiro/UERJ)	Aprovado
Prof. Titular **Jorge Alberto Achcar** (Instituto de Ciências Matemáticas e de Computação/ICMC-USP)	Aprovado

Coordenador do Programa de Pós-Graduação em Engenharia Elétrica:
Prof. Associado **João Bosco Augusto London Junior**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Murilo Araujo Romero**

To my family. Thank you for everything.

ABSTRACT

JERONYMO, P. V. B. **Markov Blanket Discovery without Causal Sufficiency: Application in Credit Data**. 2021. 79p. Master Dissertation - São Carlos School of Engineering, University of São Paulo, São Carlos, 2021.

Faster feature selection algorithms become a necessity as Big Data dictates the *zeitgeist*. An important class of feature selectors are Markov Blanket (MB) learning algorithms. They are Causal Discovery algorithms that learn the local causal structure of a target variable. A common assumption in their theoretical basis, yet often violated in practice, is causal sufficiency. The M3B algorithm was proposed as the first to directly learn the MB without demanding causal sufficiency. The main drawback of M3B is that it is time inefficient, being intractable for high-dimensional inputs. Intending a faster method, we derive the Fast Markov Blanket Discovery Algorithm (FMMB). Empirical results that compare FMMB to M3B on the structural learning task show that FMMB outperforms M3B in terms of time efficiency, while preserving structural accuracy given a large enough sample size. Moreover, we introduce a new technique to aggregate bootstrapped MB structures, that first extracts a consensus MB, than constructs the aggregated structure as the union of the most probable path between each feature in the MB and the target. Comparisons with the state of the art shows that the proposed aggregation has a smaller loss of information. The analysis was conducted by using Credit-related data, with special focus on Peer-to-Peer lending platforms. Our results validate the credit scoring models used by these platforms as effective in identifying bad borrowers, yet still have room for improvement. Finally, we propose an ensemble of Bayesian Network Classifiers trained using the Cross-Entropy method. The ensemble performs better in credit scoring than Logistic Regression and Random Forests in the selected datasets.

Keywords: Causal Discovery. Markov Blanket. Bayesian Networks. Credit.

RESUMO

JERONYMO, P. V. B. **Descoberta de Markov Blankets Sem Suficiência Causal: Aplicação em Dados de Crédito**. 2021. 79p. Master Dissertation - São Carlos School of Engineering, University of São Paulo, São Carlos, 2021.

Seleção de *features* com maior velocidade se torna uma necessidade conforme Big Data dita o *zeitgeist*. Uma classe importante de seletores de *features* são algoritmos de descoberta de Markov Blanket (MB).São algoritmos de descoberta causal que aprendem a estrutura causal local de uma variável alvo. Uma suposição comum em sua base teórica, frequentemente violada na prática, é a de suficiência causal: a crença de que todas as causas em comum das variáveis que foram medidas, compondo o conjunto de dados, também estão no conjunto de dados. Recentemente, o algoritmo M3B foi proposto. É o primeiro a aprender diretamente o MB sem demandar suficiência causal. A maior desvantagem do M3B é sua ineficiência de tempo, sendo intratável para entradas muito grandes. Aqui, nós derivamos o Fast Markov Blanket Discovery Algorithm (FMMB). Resultados empíricos comparando o FMMB com o M3B em termos de aprendizado estrutural mostram que o FMMB tem melhor desempenho em termos de tempo, enquanto preservando a acurácia da estrutura causal dado um tamanho amostral grande o suficiente. Além disso, nós introduzimos uma nova técnica para agregar resultados de estruturas de MB que advém de bootstrap, que primeiro extrai um consenso de qual é o MB, então constrói a estrutura agregada como a união do caminho mais provável entre o alvo e as *features* que compõem o MB. Comparações com o estado da arte mostram que a agregação proposta perde menos informação. As analises foram conduzidas usando dados de crédito, com atenção especial à plataformas de empréstimos interpessoais. Nossos resultados validam os modelos de crédito usados por essas plataformas como efetivos na identificação de maus pagadores. Por fim, propomos um *ensemble* de Classificadores Baseados em Redes Bayesianas treinado usando o Método da Entropia Cruzada. O *ensemble* performou melhor em *Credit Scoring* do que Regressão Linear e *Random Forests* nos conjuntos de dados selecionados.

Palavras-chave: Descoberta Causal. Markov Blanket. Redes Bayesianas. Crédito.

LIST OF FIGURES

Figure 1 – Example of G-squared test with $dof = 12$, significance level $\alpha = 0.5$ and $power = 80\%$. The orange shaded area is the probability of a Type II error $\beta = 1 - power$. The mean of the chi-squared distribution is equal to dof, and the mean of the noncentral chi-squared is equal to dof plus the effect size, or non-centrality parameter, λ. If the G^2 statistic calculated for the sample is greater then the critical value C_α, the null hypothesis is rejected. 25

Figure 2 – (a) Markov Blanket of *Disease A* shaded in blue and (b) induced Markov Blanket of *Disease A* when hiding *Disease B*, composed of blue and green shaded nodes. Because *Disease B* can no longer be observed, *Disease C* and *Symptom 2* influence *Disease A* through *Symptom 1* (c) If *Disease B* is omitted, e.g. because of a lack of knowledge of its presence, the relationship between *Symptom 1* and *Symptom 2* is represented by a bidirected connection, indicating a correlation, and hinting a possible hidden common cause. 26

Figure 3 – Example of MB. The target is in orange. The nodes in blue are discovered by algorithms that assume causal sufficiency, while the nodes in green are only discovered by algorithms that do not assume causal sufficiency. The nodes in red are not part of the MB. 28

Figure 4 – A Bayesian Network representing the factorization $P(\mathcal{U}) = P(A) \cdot P(B \mid A) \cdot P(C \mid A) \cdot P(D \mid B, C)$. Two of the four CPDs (with fictitious values) are displayed along with the structure. 31

Figure 5 – Varieties of Naive Bayes 33

Figure 6 – Depiction of the states a node passed during the `Discover-Att` algorithm: not visited ($\mathcal{N}$), candidate ($\mathcal{C}$), admitted (($\mathcal{A}$), or excluded ($\mathcal{E}$). 38

Figure 7 – Alarm, a well-known network designed for monitoring patients in intensive care unit, commonly used as a benchmark. From its structure, test cases (TCs) were elaborated for validating the proposed algorithm. For each test case, a node was selected as target and other nodes were hidden in order to violate causal sufficiency. The target nodes are shaded in orange and the hidden nodes are shaded in blue. The dashed arrow between "LVF" and "STKV" indicates that the connection is present in the original network, but was removed to create the test cases. 39

Figure 8 – Correct structure of each test case created from the Alarm Network. The nodes in green only become members of the MB when the mentioned variables are hidden, while the nodes in blue are also part of the MB when all the variables are observable. Likewise, the arrows in green indicate connections that only appear in the MAG. 40

Figure 9 – Flowchart of the experimental design for studying aggregation methods. 42

Figure 10 – Example of a mapping from a binary array to a BAN structure with five features. Assuming an ordering $\sigma = (X_1, X_2, X_3, X_4, X_5)$, children (ch) edges can be mapped to the initial positions of the array and augmenting (aug) edges to the remaining positions following an upper triangular matrix, e.g. $(1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1)$. 43

Figure 11 – Normalized Mutual Information between every pair of features in the Alarm Network.Entries in white indicate independence of the variables, given a G^2 test at a significance level of 0.01. 47

Figure 12 – TC1 – Estimation of the MB of "VTUB" when "INT" and "PMB" are hidden. Results are the median of 30 runs of the experiment. 49

Figure 13 – TC2 – Estimation of the MB of "HR" when "HYP" and "STKV" are hidden, and the connection "LVF" $\rightarrow$ "STKV" is removed. Results are the median of 30 runs of the experiment. 50

Figure 14 – Comparison of FMMB and M3B, as well as structure aggregation methods, in credit related datasets considering classification performance (AUC). Each boxplot summarizes the results of 30 bootstraps. 56

Figure 15 – Comparison of FMMB and M3B, as well as structure aggregation methods, in credit related datasets considering MB size and Runtime. Each boxplot summarizes the results of 30 bootstraps. The runtime of the aggregations is in the order of milliseconds. 57

Figure 16 – Normalized Mutual Information between every pair of features in the Lending Club dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01. Notice that “Accounts Now Delinquent” and “Delinquent Amount” are not associated with any other feature. An explanation is that these features have a very low percentage of nonzero samples, making it unfeasible to detect a connection if any. 58

Figure 17 – Discovered structure of Lending Club using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue. 59
Figure 18 – Normalized Mutual Information between every pair of features in the Prosper dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01. 60
Figure 19 – Discovered structure of Prosper using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue. 61
Figure 20 – Normalized Mutual Information between every pair of features in the PAKDD2009 dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01. 62
Figure 21 – Discovered structure of PAKDD2009 using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue. 62
Figure 22 – Normalized Mutual Information between every pair of features in the Taiwan dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01. 63
Figure 23 – Discovered structure of Taiwan using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue. 64

Figure 24 – Normalized Mutual Information between every pair of features in the Polish dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01. 65
Figure 25 – Discovered structure of Polish using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue. 66

LIST OF TABLES

Table 1 – Edge types present in the graphical models learned by MB discovery algorithms. The arrowheads express our current knowledge about the relationship between the nodes connected by the edge. 27
Table 2 – Criteria for selecting variables as part of the MB that define each discovery algorithm. 38
Table 3 – Summary of datasets used in feature selection experiment. 41
Table 4 – Characteristics of credit datasets after preprocessing. 42
Table 5 – Summary of the preprocessed credit datasets used for comparing classifiers. 45
Table 6 – Feature selection experiment results for M3B. Values in bold indicate a better performance than it's counterpart (FMMB), and values in italic indicate equality. 10-fold cross-validation was used in the five smaller datasets – the results are in the format *mean±std*. For the five larger datasets, a single 70%/30% train/test split was measured. Computational time for the execution of the feature selection algorithms was limited to 15000*s*. Elapsed time only encompasses feature selection, i.e. classifier training and prediction times were excluded from the measurements. . . 51
Table 7 – Feature selection experiment results for FMMB. Values in bold indicate a better performance than it's counterpart (M3B), and values in italic indicate equality. 10-fold cross-validation was used in the five smaller datasets – the results are in the format *mean±std*. For the five larger datasets, a single 70%/30% train/test split was measured. Elapsed time only encompasses feature selection, i.e. classifier training and prediction times were excluded from the measurements. 51
Table 8 – Experimental results. Values in **bold** indicate a clear winner when considering each metric individually, whereas values in *italic* indicate a draw. 68

CONTENTS

1 INTRODUCTION 19

2 THEORY AND DEFINITIONS 23
2.1 Probability and Information Theory 23
2.1.1 Conditional Independence 23
2.1.2 Beta and Dirichlet Distributions 23
2.1.3 Mutual Information 23
2.2 Probabilistic Graphical Models 24
2.2.1 Structural Learning and Conditional Independence Testing 24
2.2.2 Intuition for the implications of the lack of causal sufficiency 24
2.2.3 Preceding work in MB discovery 26
2.2.4 Graphical models learned by CIT based MB discovery algorithms 27
2.2.5 MB discovery through the lens of Information Theory 29
2.3 Bayesian Networks 31
2.3.1 Parameter learning 32
2.3.2 Inference 32
2.3.3 Bayesian Network Classifiers 33

3 MATERIAL AND METHODS 35
3.1 MB discovery without causal sufficiency 35
3.1.1 The FMMB algorithm 35
3.1.2 Validation tests 37
3.1.3 Feature selection experiment 38
3.2 Aggregation of MB bootstraps and Application in Credit Data . . . 39
3.2.1 Proposed Aggregation Method 39
3.2.2 Experimental design 41
3.2.3 Data description 42
3.3 Credit Scoring with Bayesian Network Classifier Ensembles 43
3.3.1 Learning BNCs using CEM 43
3.3.2 Data preprocessing 44
3.3.3 Model training and evaluation 45

4 RESULTS AND DISCUSSION 47
4.1 MB discovery without causal sufficiency 47
4.2 Aggregation of MB bootstraps and Application in Credit Data . . . 52
4.3 Credit Scoring with Bayesian Network Classifier Ensembles 66

5 CONCLUSION 69

References 73

1 INTRODUCTION

Credit scoring is of vital importance to credit bureaus and credit providers such as banks, insurance companies and other financial institutions. A credit score is an estimation of the ability a borrower has to repay a loan. The models developed to tackle this task use data from previous completed loans and try to determine if a borrower has fully repaid the loan or has defaulted (BAESENS; ROESCH; SCHEULE, 2016). Recent efforts have been made to improve credit scoring models (ZHANG, W. et al., 2021; XIAO et al., 2020; ZHANG, T. et al., 2018; LESSMANN et al., 2015), specially in the Peer-to-peer (P2P) lending context, due to the public availability of data from P2P lending platforms such as Prosper and Lending Club (MOSCATO; PICARIELLO; SPERLÍ, 2021; MANCISIDOR et al., 2020; HOU et al., 2020; LI et al., 2017). Despite of technological advancements, the most commonly used models by credit bureaus like Experian are linear (EXPERIAN, 2021). A plausible argument for this preference is model interpretability. Traditional institutions might feel less compelled to adopt more complex models such as artificial neural networks due to their "black-box" reputation (LIPTON, 2018). An essential aspect of improving credit models is understanding the relative importance of the available features in predicting the causes of default. Previous research rely on linear or lasso regression (CROUX et al., 2020; POLENA; REGNER, 2018; SERRANO-CINCA; GUTIÉRREZ-NIETO; LÓPEZ-PALACIOS, 2015).

In this dissertation, we advocate for graphical modelling (KOLLER; FRIEDMAN, 2009). In particular, for Markov Blanket (MB) discovery. A MB is the minimum set of features that carry all the information contained in a dataset about a target. The concept of a MB was introduced by Pearl (1988) as part of his work on reasoning under uncertainty, in which the theory of Bayesian Networks (BNs) is founded. BNs are probabilistic graphical models that factor the joint probability distribution of the variables in a system into a directed acyclic graph (DAG) (KOLLER; FRIEDMAN, 2009). MB discovery is relevant in feature selection and local-to-global structural learning of Bayesian Networks (RÍO; VILLANUEVA, 2021; YU; GUO, et al., 2020; WANG; LING, et al., 2020a). Earlier applications of graphical modelling in credit scoring are limited to a few papers (MASMOUDI; ABID; MASMOUDI, 2019; ANDERSON, 2019; LEONG, 2016; CHANG et al., 2000). Preceding work learned the whole structure, then extracted the MB of the target. We advocate for using algorithms that directly learn the MB. Using non-linear measurements of dependence between the variables, such as mutual information (BELGHAZI et al., 2018; COVER; THOMAS, 2006), we can represent the meaningful connections between the features and the target in a graphical, interpretable manner, while capturing facets of the data undetectable by linear methods (GLYMOUR; ZHANG;

SPIRTES, 2019).

A fundamental assumption for graphical modelling is Faithfulness. A dataset is faithful if its underlying probability distribution can be represented by a single graphical model that reliably represents all the dependencies present in the data. In other works, the faithfulness assumption is what enables graphical modelling (YU; WU, et al., 2016; SPIRTES et al., 2000). For MB discovery, an even stronger version of this assumptions is made. Its assumed that the underlying graphical model is a DAG ideally containing all of the common causes between variables, being Causal Sufficient. If X is a parent of Y, then X is a direct cause of Y. However, while the underlying DAG may be causally sufficient, there is no guarantee that every common cause of two or more variables present in the dataset is also in the dataset. Yet, this is often assumed by MB discovery algorithms (YU; GUO, et al., 2020; KOLLER; FRIEDMAN, 2009).

To withdraw the causal sufficiency assumption, often violated in practice, the M3B algorithm was proposed (YU; LIU, et al., 2018). To accommodate latent variables, it models the underlying distribution as a Maximal Ancestral Graph (MAG), an extension of DAGs that is capable of representing correlations and hidden common causes using bidirected connections. In contrast to previous MAG learning algorithms such as FCI (SPIRTES et al., 2000) and RFCI (COLOMBO et al., 2012) that required learning the complete MAG before extracting the MB, it is the first to directly discover the MB without assuming causal sufficiency. However, it is time demanding.

Recently, motivated by the desire for more time efficiency, we proposed the Fast MAG Markov Blanket (FMMB) algorithm in the XXIII Brazilian Congress of Automatics (CBA 2020) (JERONYMO; MACIEL, 2020). It adds features to the MB by directly testing the connection between the feature and the target, in contrast to M3B's indirect discovery, that consists of adding new features by adjacency, aiming to preserve statistical power. In this dissertation, a generalisation of M3B and FMMB is elaborated. We show that these algorithms can be seen as extensions of algorithms that assume causal sufficiency. The results demonstrate, using data from numerous areas that, given a large enough sample, FMMB is faster than M3B with equal structural accuracy. We apply both algorithms to analyse credit data, paying special attention to the P2P lending platforms.

We also explore the problem of combining the results of bootstrapping the algorithms on the data, a crucial step in structural learning (GLYMOUR; ZHANG; SPIRTES, 2019). The stability of an algorithm can be assessed by observing the diversity of the bootstrapped structures. Another byproduct is the estimation of probabilities for various properties of the structure, such as occurrence of edges and nodes. It is also common practice to aggregate the bootstrapped structures to extract a single structure with meaningful connections. Usually, it is assumed that the presence of an edge is independent of the presence of other edges. A threshold τ representing the confidence level in the existence of

an edge is chosen, and the aggregated structure is the union of all edges with probability $> \tau$. This method proposed by Friedman, Goldszmidt, and Wyner (2013) is ubiquitous in the literature (GROSS et al., 2019; RODGERS et al., 2019; MCNALLY; HEEREN; ROBINAUGH, 2017).

Here, we show that the structure aggregation approach universal in the literature does not work well for MB bootstraps. To the best of our knowledge, no previous work studied the implications of applying it to MB discovery nor proposed an alternative. The way that MB discovery algorithms consider edges breaks the edge independence assumption. Features included first affect the probability of inclusion of the remaining features. In particular, algorithms that not only identify the members of the MB, but also the structure, start by discovering the nodes adjacent to the target, then discover nodes at greater depths (YU; GUO, et al., 2020). As there may be multiple possible paths between a feature and the target, we experimentally show that the deeper the feature is, the more likely it is that the algorithm will reach the same feature through distinct paths given different bootstrap samples, diluting the frequency of occurrence of the edges. We show that disregarding these facts by assuming total independence is too restrictive and leads to aggregated structures with fewer nodes than the bootstraps, loosing information about the target. We propose a two step method of aggregation. First, we aggregate by nodes, forming a consensus of what is the most likely MB. Then, we form the aggregated structure as the union of the most likely path between each feature in the consensus MB and the target. In addition, we observe in the results the limitations of graphical modelling for some datasets. If a dataset is not faithful, then there is not a unique MB, but possibly many (YU; GUO, et al., 2020). Determining if a dataset is faithful in practice is difficult, because there is no method that can directly measure the degree of faithfulness of data. Most advancements are merely theoretical (WEINBERGER, 2018; SADEGHI, 2017). However we argue that it is possible to observe signs of an unfaithful distribution. An observable consequence is a high variability in the bootstrapped MBs, and the impossibility of reaching a consensus MB with similar size, because not many features appear consistently.

As stated above, previous applications of graphical modelling in credit scoring learned BNs using methods focused on general structural learning. The approach is not only inefficient for MB discovery, but also is not optimized for the task of classifying a single variable. We advocate for the use of Bayesian Network Classifiers (BNCs) which are a subset of BN models specialized in classification. Here, a novel method for constructing BNCs using the Cross-Entropy Method (CEM) combined with Bootstrap Aggregation (Bagging) is proposed (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). CEM is a Monte-Carlo program for importance sampling and optimization (RUBINSTEIN; KROESE, 2013; DE BOER et al., 2005). We used it to randomly generate and evaluate network structures. To the best of our knowledge, CEM has not been used previously for learn BNs. It was chosen due to it's fast convergence and simplicity. This combined Monte-Carlo approach is employed to

avoid a finer exploration of the super-exponential domain of possible structures (KLEITER, 1999). The core idea is to build an ensemble of simpler structures that together attempt to form a good approximation of the underlying distribution. Additionally, Bagging improves the average performance of noisy estimators (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). We applied the algorithm in credit data for the task of credit scoring, and the results show its potential. The algorithm will be applied to more datasets and compared with other models. The final results will be submitted in January to the IEEE World Congress on Computational Intelligence 2022.

Our main objective is introducing FMMB as faster method of local causal discovery, aiming to mitigate the curse of dimensionality, imposed by the super-exponential nature of the search space (GROSS et al., 2019; CAMPOS, 2006; KOLLER; FRIEDMAN, 2009). The secondary goal is analysing the applicability of causal discovery to credit data. A minor, yet related, objective is to explore the applicability of Bayesian Networks in the task of credit scoring, realized by introducing the CEM trained BNC.

This dissertation is organized as follows. Ch. 2 covers concepts from Probability and Statistics, Information Theory, Bayesian Networks, and other Probabilistic Graphical Models. We mix concepts from the literature with our own definitions. Ch. 3 introduces the FMMB algorithm and compares it with M3B, also bringing the proposed aggregation method. Furthermore, it presents the method for creating a BNC emsemble using CEM. Ch. 4 discusses the results obtained by applying the methods to credit-related datasets. Finally, Ch. 5 concludes the dissertation, highlighting the main contributions to Bayesian Networks, MB Discovery, and Credit Scoring.

2 THEORY AND DEFINITIONS

2.1 Probability and Information Theory

2.1.1 Conditional Independence

The universe, i.e. the set of all variable in the system, is denoted as $\mathcal{U}$. Two variables $X, Y \in \mathcal{U}$ are conditionally independent given a set of observed variables $\mathbf{Z}$, denoted $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, if the joint probability distribution $p(x, y \mid \mathbf{z})$ of X and Y given $\mathbf{Z}$ can be factored as $p(x \mid \mathbf{z}) \cdot p(y \mid \mathbf{z})$. On the other hand, if X and Y are not conditionally independent, we denote $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$.

2.1.2 Beta and Dirichlet Distributions

The Dirichlet distribution generalizes a Beta distribution for $K \geq 2$ categories, being the conjugate prior of the Multinomial. It has support over the open (K-1)–simplex $S = \{(\Theta_1, ..., \Theta_K) \mid \Theta_i > 0 \wedge \sum_{i=1}^{K} \Theta_i = 1\}$ and probability density function (PDF)

$$f(\Theta_1, ..., \Theta_K; m_i, ..., m_K) = \frac{1}{B(\mathbf{m})} \prod_{i=1}^{K} \Theta_i^{m_i - 1} \tag{2.1}$$

where $B(\mathbf{m}) = \frac{\prod_{i=1}^{K} \Gamma(m_i)}{\Gamma(\sum_{i=1}^{K} m_i)}$ and $\mathbf{m} = [m_1, ..., m_K], m_i > 0$ are the concentration parameters. If $\{m_i = 1\}$, the distribution generates uniform samples over its support, and $\{m_i = 1/2\}$ is the Jeffreys prior (AGRESTI, 2003). Let $m = \sum_{i=1}^{K} m_i$ be called the *effective sample size* (COOPER; HERSKOVITS, 1992). An alternative parameterization for the Dirichlet distribution is

$$\Theta = \langle \Theta_1, ..., \Theta_K \rangle \sim Dir(m; \bar{\Theta}_1, ..., \bar{\Theta}_K) \tag{2.2}$$

where $\bar{\Theta}_i = E(\Theta_i) = m_i/m$. If $K = 2$, since $\Theta_1 + \Theta_2 = 1$, let $\Theta_1 = \Theta$ and $\Theta_2 = 1 - \Theta$. The Dirichlet reduces to a Beta distribution over Θ with parameters $\nu = m_1 + m_2$ and $\mu = m_1/\nu$. Moreover, the variance can be expressed as

$$var(\Theta) = \frac{\mu(1 - \mu)}{1 + \nu} \tag{2.3}$$

2.1.3 Mutual Information

Mutual information (MI) is an information theoretical quantity that quantifies the degree of dependence between two variables X and Y, denoted as $I(X, Y)$. If the variables are independent, then $I(X, Y) = 0$, whereas if the value of Y is completely determined if X is known, then $I(X, Y)$ is maximal. The normalized mutual information (NMI) is defined in Eq. 2.4, where $H(\cdot)$ is the Shannon entropy. It has values in the interval $[0, 1]$ (COVER; THOMAS, 2006).

$$NMI(X,Y) = \frac{I(X,Y)}{\min\{H(X), H(Y)\}} \tag{2.4}$$

2.2 Probabilistic Graphical Models

2.2.1 Structural Learning and Conditional Independence Testing

The task of learning a graphical model that represents the conditional independence relationships in the data is known as *structural learning*. Two important classes of structural learning algorithms are those based on maximizing a *score*, i.e. the likelihood of the data given the structure, and those based on conditional independence tests (CITs) (CAMPOS, 2006). A complementary approach is expert elicitation (CHEN; POLLINO, 2012). Here, we focus on CIT based discovery. A commonly adopted CIT for discrete variables is the G-squared. The G^2 statistic is proportional to the empirical measure of MI: $G^2 = 2N\hat{I}(X,Y)$ where N is the sample size. Under the null hypothesis, the statistic has a asymptotic chi-squared distribution with degrees of freedom (dof) equal to $(r_X - 1)(r_Y - 1)$ where r_X is the number of states of the variable X. Under the alternative hypothesis, the statistic assumes a asymptotic non-central chi-squared distribution with the same dof and non-centrality parameter proportional to the actual MI between the two variables: $\lambda = 2NI(X,Y)$, i.e. the mutual information is calculated replacing the probabilities estimated from the sample with the population values. Fig. 1 gives an example of the test's workings (MCDONALD, 2009; AGRESTI, 2003; KULLBACK, 1997).

2.2.2 Intuition for the implications of the lack of causal sufficiency

Fig. 2 illustrates a hypothetical scenario useful for understanding the implications of assuming causal sufficiency of the data. Imagine a patient arriving at the doctor with *Symptom 1*. From the health history and genetic factors of the patient, the most likely cause is *Disease A*. However, given the intensity of *Symptom 1*, unusual for A, but commonly caused by *Disease B*, the doctor asks if the patient has *Symptom 2*, often manifested by patients with B, but not caused by A. The response is positive. However, to complicate matters, *Symptom 2* may also be caused by a wide range of seasonal diseases such as *Disease C*, common during the winter. The doctor asks if the patient was sick with C recently, also receiving a positive response. This vastly decreases the chance that the patient has B, and further increases the belief that A is the cause of *1*. A medical exam that tests for A was requested. The positive result made *Disease A* 99% likely. The fact that the prompt diagnose of *Disease B* is not possible, makes the knowledge about *Symptom 2* and *Disease C* affect the diagnose of *1*, consequently affecting the probability of A, as shown in Fig. 2b and 2c. If the answer to the question "Does the patient have *Disease B*?" was immediately available, the doctor would not have to ask about *Symptom 2* or *Disease C*, as in Fig. 2a.

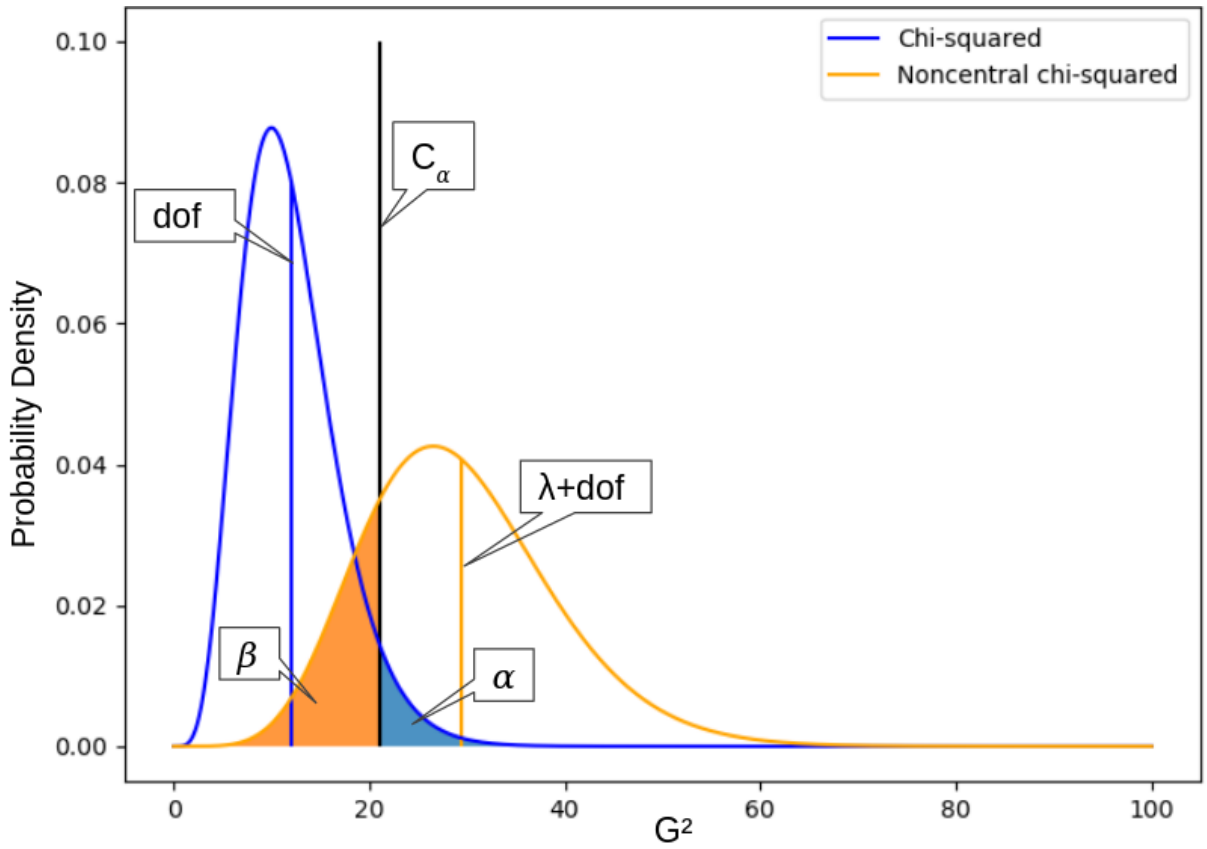


Figure 1 – Example of G-squared test with $dof = 12$, significance level $\alpha = 0.5$ and $power = 80\%$. The orange shaded area is the probability of a Type II error $\beta = 1 - power$. The mean of the chi-squared distribution is equal to dof, and the mean of the noncentral chi-squared is equal to dof plus the effect size, or non-centrality parameter, λ. If the G^2 statistic calculated for the sample is greater then the critical value C_α, the null hypothesis is rejected.

If presented with a dataset containing *i.i.d.* samples of the variables generated by the underlying DAG of Fig. 2a, but lacking a column that indicates the presence of *Disease B*, as in Fig. 2b, an algorithm assuming the causal sufficiency of the data will find the relationship between *Symptom 1* and *Symptom 2*, as represented in Fig. 2c, because it is not possible to separate *2* from *A* without observing *B*. Yet, because it is expecting causal sufficiency, it will treat *2* as a cause of *A*, and will never search for *Disease C*, thus missing variables from the MB. The root of the mistake is the collider connection. It is a type of connection that happens when there is variable X that is only connected to the target T when some other variable Y is observed. If the collider connection between X and T is detected in a causally sufficient dataset, the only possibility is that X and T are common causes of Y (KOLLER; FRIEDMAN, 2009). Because *Disease A* and *B* form a collider with *Symptom 1*, whether the patient has *B* affects the diagnose of *A*. But if *Symptom 1* was not present, since the diseases are independent from each other, knowing *B* would not bring any information about *A*. However, if *B* is not observable, the algorithm will think that *Symptom 2* and *Disease A* are common causes of *Symptom 1*, because *2* is collider

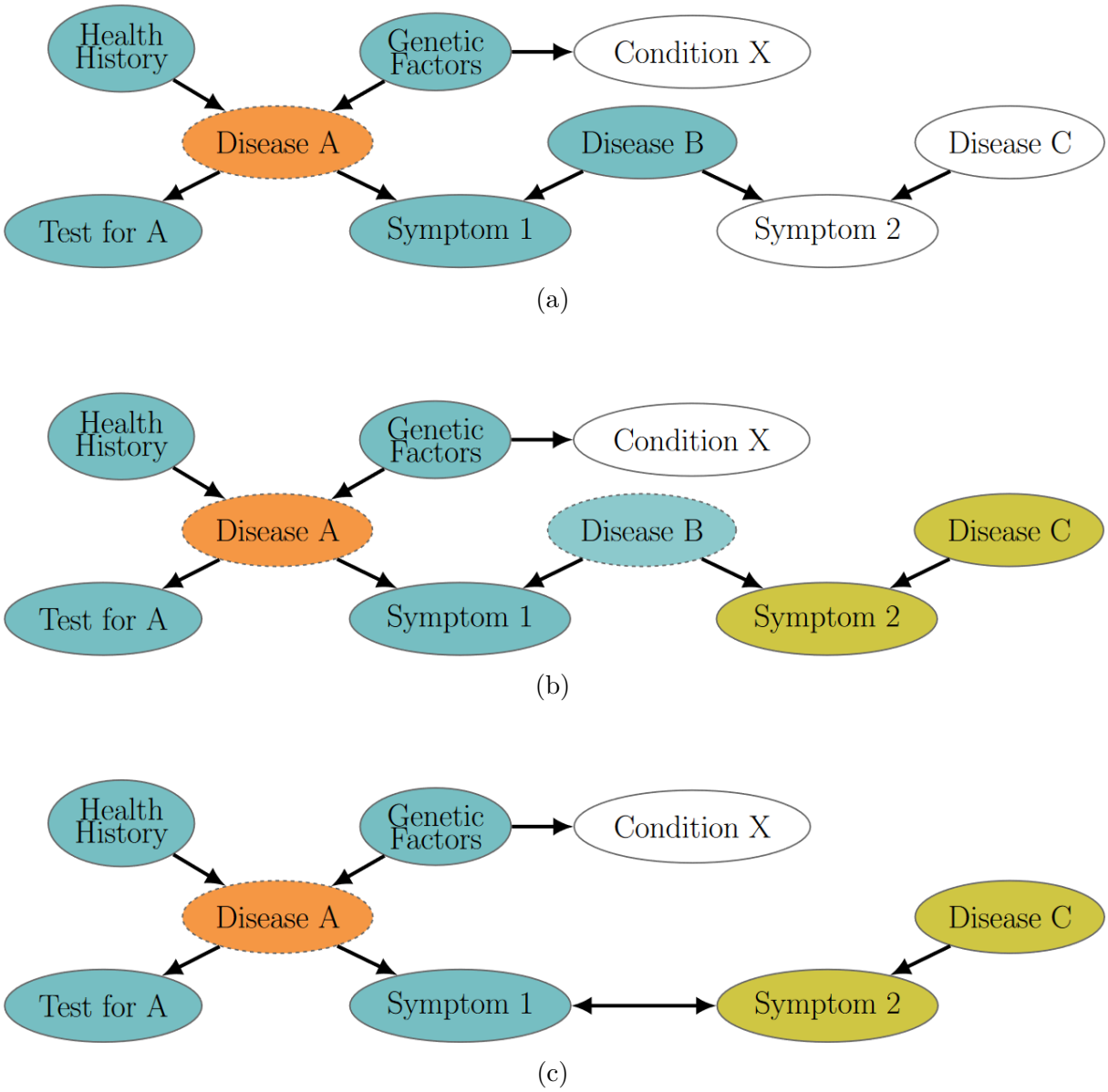


Figure 2 – (a) Markov Blanket of *Disease A* shaded in blue and (b) induced Markov Blanket of *Disease A* when hiding *Disease B*, composed of blue and green shaded nodes. Because *Disease B* can no longer be observed, *Disease C* and *Symptom 2* influence *Disease A* through *Symptom 1* (c) If *Disease B* is omitted, e.g. because of a lack of knowledge of its presence, the relationship between *Symptom 1* and *Symptom 2* is represented by a bidirected connection, indicating a correlation, and hinting a possible hidden common cause.

connected to A through 1.

2.2.3 Preceding work in MB discovery

The MB was established as the optimal subset of features in the seminal paper of Koller and Sahami (1996), gathering interest in the Feature Selection community. Estimating the MB of a target became feasible after Margaritis and Thrun (2000) proposed the

first tractable solution, the Grow-Shrink (GS) algorithm, using CITs for variable selection. First, it grows a candidate set of features by testing association with the target. Then, it shrinks the candidates removing false positives. Since GS, IAMB (TSAMARDINOS; ALIFERIS, 2003) and variants (Tsamardinos, Aliferis, Statnikov, and Statnikov (2003), Yaramakala and Margaritis (2005) and Yang et al. (2019) among others) improved on these ideas, forming the family of simultaneous algorithms.

Pena et al. (2007) showed that simultaneous algorithms are not data efficient. The amount of data required to maintain the power of the CITs can sometimes surpass the size of the dataset, as the conditioning set grows proportional to the candidate set. Another drawback is that they do not differentiate immediate causes from variables that are connected through colliders. To overcome these limitations, divide-and-conquer algorithms were proposed, the pioneer being the MMMB algorithm (TSAMARDINOS; ALIFERIS; STATNIKOV, 2003). They first learn the adjacent variables to the target T, then repeat the procedure, assigning each variable X adjacent to T as the new target. The collider connection with T is tested for each candidate Y that is adjacent to X, eliminating false positives, such as any descendant of X. This strategy reduces the sample size requirements at the expense of a greater time complexity. A theoretical framework and broad experimental evaluation for this family of algorithms was given in Aliferis et al. (2010a) and Aliferis et al. (2010b).

Combining the strategies of simultaneous and divide-and-conquer algorithms to derive a method capable of simultaneously learning the MB while distinguishing the structure, the EEMB algorithm is among the latest developments in this line of research (WANG; LING, et al., 2020b). An extensive review on MB learning algorithms is found in Yu, Guo, et al. (2020).

2.2.4 Graphical models learned by CIT based MB discovery algorithms

Table 1 – Edge types present in the graphical models learned by MB discovery algorithms. The arrowheads express our current knowledge about the relationship between the nodes connected by the edge.

Edge Type	**Interpretation**
$X \rightarrow Y$	X is a cause of Y;
$X \leftrightarrow Y$	X and Y are correlated, and may share a hidden common cause;
$X \circ\!\!\rightarrow Y$	Either X is a cause of Y, or X and Y are correlated, and may share a hidden common cause;
$X \circ\!\!-\!\!\circ Y$	Any of the above may apply.

Tab. 1 shows the symbols used to represent connections in graphical models learned by MB discovery and their interpretations. In practice, edges of the type $X \rightarrow Y$, that express total knowledge about the relationship, are not possible to learn from pure observational data, needing the incorporation of domain knowledge or the realization of

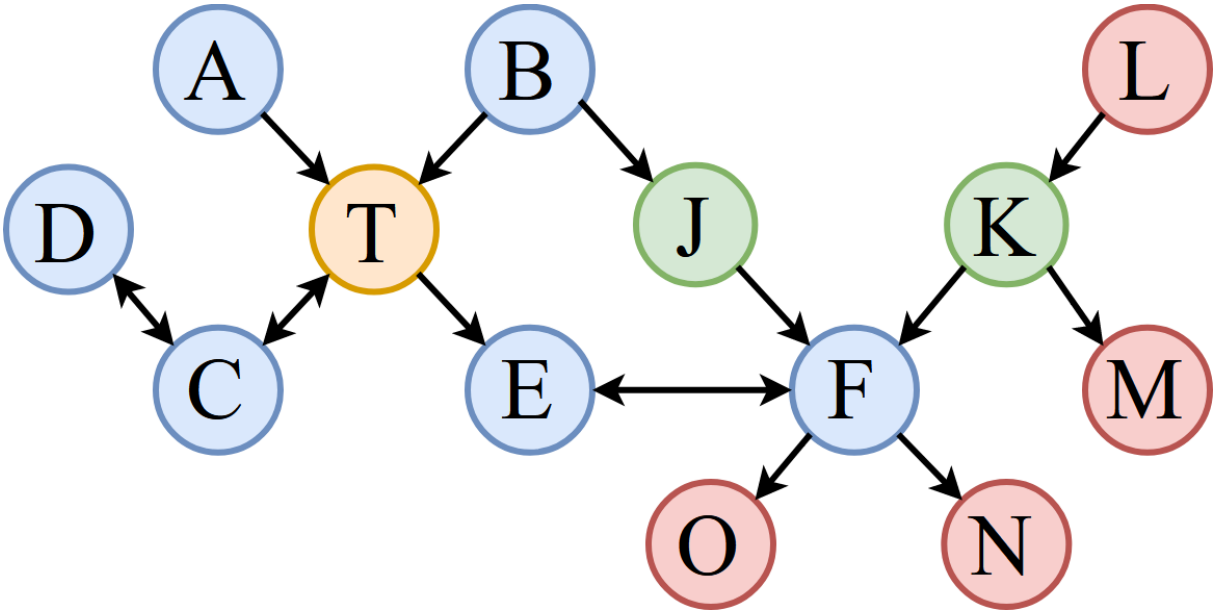


Figure 3 – Example of MB. The target is in orange. The nodes in blue are discovered by algorithms that assume causal sufficiency, while the nodes in green are only discovered by algorithms that do not assume causal sufficiency. The nodes in red are not part of the MB.

interventional experiments (SHEN et al., 2020; MEGANCK; LERAY; MANDERICK, 2007; SPIRTES et al., 2000).

A path μ is a sequence of adjacent nodes in a graph $[X_1, X_2, ..., X_n]$, or equivalently, a sequence of edges $[e_1, e_2, ..., e_{n-1}]$ connecting the nodes. A collider in a path is a variable X_i that has connections of the type $X_{i-1} \circ\!\!\rightarrow X_i \leftarrow\!\!\circ X_{i+1}$. It entails that X_{i-1} and X_{i+1} are only connected through this path when X_i is observed. FMMB and M3B search for members of the MB by building collider paths between the target and the features. A collider path is a path in which every node except the endpoints are a collider, e.g. in the path $X_1 \circ\!\!\rightarrow X_2 \leftrightarrow X_3 \leftarrow\!\!\circ X_4$ there are two colliders, X_2 and X_3, connecting the endpoints X_1 and X_4 (RICHARDSON; SPIRTES, et al., 2002). In Fig. 3, $[T, E, F, K, M]$ is a path between M and T where E and F are colliders. The path $[T, E, F, K, M]$ is not a collider path because M does not form a collider with its neighbour F, however the sub-path $[T, E, F, K]$ is a collider path, as well as the path $[T, C, D]$. A path μ *connects* X and Y given $\mathbf{Z}$ if every non-collider in μ is not in $\mathbf{Z}$, and every collider on μ either observed or is an ancestor of some node in $\mathbf{Z}$. If $\forall X \in \mathbf{X}$ and $\forall Y \in \mathbf{Y}$ all paths between X and Y are *not* connected given $\mathbf{Z}$, then $\mathbf{Z}$ *separates* $\mathbf{X}$ and $\mathbf{Y}$. If $\mathbf{Z}$ separates X and Y then $\mathbf{Z}$ is called a *sepset* or *separator* of X and Y. Denoted as $\sigma(X, Y) = \mathbf{Z}$ (RICHARDSON; SPIRTES, et al., 2002). In our example, that means that T and M are connected if:

1. Both E and F are observed and K is not observed;

2. E and a subset of $\{O, N\}$ are observed, and K is not observed. That is because F is

an ancestor of O and N, the collider is activate when either is observed.

Definition 1 (Depth of a node) *The depth of a node is equal to the length of the shortest collider path between X and T plus one. If a variable is adjacent to the target, we define its collider path to be the empty path $\mu = [\,]$.*

The structure defining a MB can be decomposed into the collider paths connecting each of the variables in the MB and the target (YU; LIU, et al., 2018). We define the depth of a node X in relation to the target T as of Def. 1. Algorithms that assume causal sufficiency are able to detect every member of the MB with depth ≤ 2, whereas algorithms that do not assume causal sufficiency can detect members at any depth (YU; LIU, et al., 2018). We denote the set of all nodes adjacent to a target T as $\mathbf{Adj}(T)$. In Fig. 3, when a subset of $\{B, J\}$ is observed, and E is not observed, F is separated from T. Since $B \in \mathbf{Adj}(T)$, a discovery algorithm will separate F from T, excluding it from the candidates for $\mathbf{Adj}(T)$. It will assign $\sigma(T, F) = \{B\}$. When discovering nodes at depth 2, it will find the node F through the path $[T, E, F]$, the only active path between F and the target when $\mathbf{Adj}(T)$ is known. The depth of F is therefore 2.

Formally, the graphical model that a discovery algorithm that do not assume causal sufficiency outputs "is a complete partial ancestral graph (CPAG), representing the Markov equivalence class of MAGs consistent with the data" – Meganck, Leray, and Manderick (2007). In a CPAG, we have to interpret the directed edges as ancestral relationships instead of immediate causal relationships. That means that if $X \circ\!\!\rightarrow Y$ then X is an ancestor of Y in the underlying DAG with all the causes observable, but X is not necessarily adjacent to Y in the DAG. To make causal inferences from these models, one first needs to convert the CPAG to a representation suitable for causal inference, where every directed edge does represent an immediate cause, and orient the necessary edges using interventional experiments. Meganck, Leray, and Manderick (2007) describe the procedure in detail.

2.2.5 MB discovery through the lens of Information Theory

Although the mutual information is not an addictive measurement, meaning that $I(X, \{Y, Z\}) \neq I(X, Y) + I(X, Z)$, a basic yet essential property is the chain rule of mutual information. It states that the MI between X and $\{Y, Z\}$ equals the information that Y has about X when solely observed plus the additional information that Z has about X given that we already know Y, formally expressed in Eq. (2.5).

$$I(X, \{Y, Z\}) = I(X, Y) + I(X, Z \mid Y) \tag{2.5}$$

The data processing inequality, that can be derived using the chain rule, says that if there is a mediator between two nodes, than the mutual information between the

endpoints is less than or equal to the mutual information between each endpoint and the mediator. That is, if $X \circ\!\!-\!\!\circ Y \circ\!\!-\!\!\circ Z$, then $I(X,Z) \leq \min\{I(X,Y), I(Y,Z)\}$ (COVER; THOMAS, 2006). A corollary of the data processing inequality is that the connection strength is monotonically non-increasing with the distance between the nodes. The greater the connection strength, the more likely is adjacency to the target. As the $\mathbf{Adj}(T)$ contains all the nodes that cannot be separated from the target, the connection strength between $X \notin \mathbf{Adj}(T)$ and T has to be smaller than the connection between the subset $\mathbf{S} \subseteq \mathbf{Adj}(T)$ that mediates the relationship between X and T. In a sense, a variable X that is part of the MB but is not adjacent to the target can be viewed as attached to a subset $\mathbf{S}$ of the adjacent variables, increasing the information that $\mathbf{S}$ have about the target. Moreover the increment of information between T and $\mathbf{S}$ that is gained when observing X is precisely the mutual information between X and T conditioned by the already known variables. This theorem (Thm. 1) is easily proved using the chain rule of mutual information (Proof 1).

Theorem 1 *If X is part of the MB of a target T but is not adjacent to T, then when observed it increases the mutual information between T and $\boldsymbol{S}$, where $S \subseteq \boldsymbol{Adj}(T)$ is the mediator between X and T.*

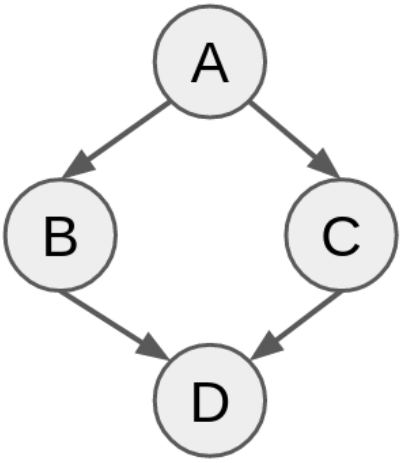
Proof 1 *If $X \notin \boldsymbol{Adj}(T)$, then it has a sepset $\sigma(X,T) \subseteq \boldsymbol{Adj}(T) \setminus \boldsymbol{S}$. Given that $\sigma(X,T)$ is observed, we can expand the MI between T and $\boldsymbol{S}$ in two different ways:*

1. $I(T, \boldsymbol{S} \mid \sigma(X,T)) = I(T, \boldsymbol{S} \mid \sigma(X,T)) + I(T, X \mid \boldsymbol{S}, \sigma(X,T))$
2. $I(T, \boldsymbol{S} \mid \sigma(X,T)) = \cancelto{0}{I(T, X \mid \sigma(X,T))} + I(T, \boldsymbol{S} \mid X, \sigma(X,T))$

The term $I(T, X \mid \sigma(X,T))$ is canceled by the definition of a separator. The simplified equation expresses that adding a node and the separator of the node to the observed variables has a net effect of zero information gain. From equating the expansions is follows that

$$I(T, \boldsymbol{S} \mid X, \sigma(X,T)) = I(T, \boldsymbol{S} \mid \sigma(X,T)) + I(T, X \mid \boldsymbol{S}, \sigma(X,T))$$
$$\Rightarrow I(T, \boldsymbol{S} \mid X, \sigma(X,T)) \geq I(T, \boldsymbol{S} \mid \sigma(X,T)) \blacksquare$$

We define the attached set of a target T, denoted $\mathbf{Att}(T)$, as the set of all members of the MB that are not adjancent to the target. Since every attached variable has to be connected to T through at least one collider path constituted of other MB members, we define the attached set of the MB as in Def. 2. Of course, an attached variable can be connected to T by many paths μ_i, being a member of every $\mathbf{Att}_T[\mu_i]$. And, from the



A	
T	0.25
F	0.75

	C	T			F		
	B	0	1	2	0	1	2
D	T	0.40	0.13	0.32	0.98	0.43	0.45
	F	0.60	0.87	0.68	0.02	0.57	0.55

Figure 4 – A Bayesian Network representing the factorization $P(\mathcal{U}) = P(A) \cdot P(B \mid A) \cdot P(C \mid A) \cdot P(D \mid B, C)$. Two of the four CPDs (with fictitious values) are displayed along with the structure.

definitions just made and fact the union of all collider paths form the MB, it is also true that $\bigcup_\mu \mathbf{Att}_T[\mu] = \mathbf{Att}_T$, and $\mathbf{MB}(T) = \mathbf{Adj}(T) \cup \mathbf{Att}_T$.

Definition 2 (Attached Set) *The attached set* $\boldsymbol{Att}_T[\mu] \subseteq \boldsymbol{Att}_T$ *is the set of all* $X \in \boldsymbol{Att}_T \setminus \mu$ *that are connected to* T *given* μ*, where* $[T] + \mu + [X]$ *is a collider path and the path* μ *is a nonempty subset of* $\boldsymbol{MB}(T)$*.*

2.3 Bayesian Networks

Formally, a BN is a factorization of a system's joint probability distribution that is represented as a pair $\langle \mathcal{G}, \Theta \rangle$ where $\mathcal{G}$ is a directed acyclic graph (DAG) that encodes the conditional independence relations between the system's variables $\mathcal{U}$, assuming the Markov Condition: every $X \in \mathcal{U}$ is independent of all non-descendants given it's parents. This DAG is called the structure of the BN, and Θ is the set of conditional probability distributions (CPDs) associated with each variable in $\mathcal{U}$. Fig. 4 shows an example of a BN.

BNs accommodate discrete and continuous variables, but most of the applications revolve around the discrete case. Usually, when dealing with Hybrid models, the continuous variables are discretized. However it requires managing the trade-off between the accuracy of the approximation and computational cost incurred by a finer-grained binning (FRIEDMAN; GOLDSZMIDT, et al., 1996). Moreover, discretization acts as a low-pass

filter, when correctly applied, reducing signal-to-noise ratio (PROAKIS; MANOLAKIS, 2004).

2.3.1 Parameter learning

The parameterization of a discrete BN can be expressed as $\Theta = \{\Theta_{X|\mathbf{f}}\}_{(X,\mathbf{f})}$ where $\Theta_{X|\mathbf{f}} = [\Theta_{X=x_1|\mathbf{f}}, ..., \Theta_{X=x_K|\mathbf{f}}]$ and $\mathbf{f}$ is some instantiation of the parents $\mathbf{F}$ of X. For example, in the BN of Fig. 4, $\Theta_{D|B=1,C=T} = [0.13, 0.87]$. Consider a structure $\mathcal{G}$ and a dataset $\mathcal{D}$ consisting of N samples of $P(\mathcal{U})$. Let $\mathbf{F}$ be the parents of $X \in \mathcal{U}$ and $N_{X=x_i|\mathbf{f}}$ denote the number of times $X = x_i$ and $\mathbf{F} = \mathbf{f}$ occurred simultaneously in $\mathcal{D}$. It is common practice to impose Dirichlet priors $\Theta_{X|\mathbf{f}} \sim Dir(m_1, ..., m_K)$ and assume $\mathcal{D}$ is Multinomial (KOLLER; FRIEDMAN, 2009). In this case, the posteriors are $Dir(m_1 + N_{X=x_1|\mathbf{f}}, ..., m_K + N_{X=x_K|\mathbf{f}})$ because the Dirichlet is the conjugate prior of the Multinomial distribution (AGRESTI, 2003). For example, let X have one parent Y, the prior for $\Theta_{X|Y=1}$ be $Dir(1, 1)$, suppose $(X = 0, Y = 1)$ occurred 10 times and $(X = 1, Y = 1)$ occurred 4 times. The resulting posterior is $Dir(11, 5)$.

2.3.2 Inference

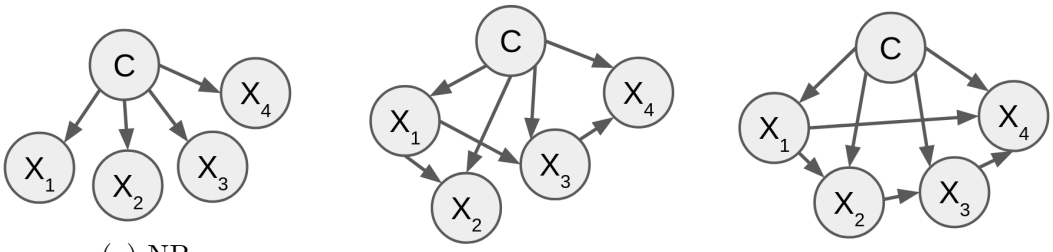
The simplest approach to answer probabilistic queries in BNs, is to use the Variable Elimination (VE) algorithm. The main idea of VE is marginalizing one variable at a time in the factorization, e.g. to answer the query $P(B \mid D = T)$ in the BN of Fig. 4, one might compute

$$
\begin{aligned}
P(B \mid D = T) &\propto \sum_{A,C} P(A, B, C, D = T) \\
&= \sum_{A,C} P(A)P(B \mid A)P(C \mid A)P(D = T \mid B, C) \\
&= \sum_{A} P(A)P(B \mid A) \sum_{C} P(C \mid A)P(D = T \mid B, C) \\
&= \sum_{A} P(A)P(B \mid A) \sum_{C} \phi(A, B, C) \\
&= \sum_{A} P(A)P(B \mid A)\tau(A, B) \\
&= \sum_{A} \phi(A, B) \\
&= \tau(B)
\end{aligned}
$$

where the τ's and ϕ's denote factors, generalizations of discrete CPDs that allow values greater then 1, being useful for postponing normalization until when necessary. After normalizing $\tau(B)$ one gets $P(B \mid D = T)$. Refer to Koller and Friedman (2009, p. 287) for a detailed explanation.

2.3.3 Bayesian Network Classifiers

There are many BNCs (BIELZA; LARRAÑAGA, 2014). The simplest among them is the Naive Bayes (NB) model (MARON; KUHNS, 1960), popular in text classification (EYHERAMENDY; LEWIS; MADIGAN, 2003; CHEN; HUANG, et al., 2009) and spam filtering (METSIS; ANDROUTSOPOULOS; PALIOURAS, 2006). It assumes the predictors $X_i \in \mathbf{X}$ to be conditionally independent given the class C. (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997) proposed relaxing this strong assumption about independence by extending NB with augmenting edges between predictors. If the structure of augmenting edges form a tree, it characterizes the Tree Augmented NB (TAN) classifier. Generally, the augmented structure can form any DAG, composing BN Augmented NB (BAN) classifiers. Fig. 5 exemplifies each variety.



(a) NB (b) TAN (c) BAN

Figure 5 – Varieties of Naive Bayes

3 MATERIAL AND METHODS

In sec. 3.1, the contrast is made between direct and indirect discovery strategies by comparing FMMB with the state of the art, M3B, in MB discovery and as feature selection methods. We apply the algorithms to synthetic and real-world data. Then, in sec. 3.2 we discuss how to bootstrap FMMB and M3B using credit dat, and describe our proposed aggregation method. Finally, sec. 3.3 showcases how BNs can be used as a classifier for credit scoring, describing how to construct a BNC using CEM.

An information theoretical way of determining which discovered MB has the most information about a target T would be to directly measure the $I(T, \mathbf{MB}(T))$. Unfortunately, if the MB is too large, the number of samples required to accurately measure mutual information becomes prohibitive (KULLBACK, 1997). Since the Bayes error rate is bounded above and bellow by functions of mutual information (VERGARA; ESTÉVEZ, 2014), classification performance using the MB as predictors is a feasible proxy for comparing MB discovery algorithms (YU; GUO, et al., 2020). We use the area under the receiver operating characteristic curve (AUC) as the classification performance metric, as it is independent of the decision threshold and preferable for imbalanced datasets (LEE, 2019; FAWCETT, 2006). We use the G^2 CIT with a 0.01 significance level. Because this test is based on contingency tables, it is susceptible to loss of statistical power when the tables cannot be sufficiently populated, due to a large conditioning set or too few samples. In order to mitigate these issues, two heuristics presented in Aliferis et al. (2010a) were adopted. We limited the size of the conditioning set to a maximum of *max-k* variables, and assumed independence if the ratio of samples to cells in the contingency tables is smaller than a constant named *h-ps*. The only exception where the *max-k* heuristic is not applied is when FMMB searches for candidates. Experiments used the values $max\text{-}k = 5$ and $h\text{-}ps = 30$, based on the results of Aliferis et al. (2010a) and Aliferis et al. (2010b). Furthermore, the maximum depth was limited to 3 in both algorithms. Algorithms were implemented in Python 3, and experiments were conducted using an AWS EC2 c5.2xlarge instance running Amazon Linux. It is a high-performance computing machine, with 16GB of RAM and an Intel Xeon Platinum 8124M CPU @ 3.00GHz processor.

3.1 MB discovery without causal sufficiency

3.1.1 The FMMB algorithm

Here, we show that M3B is an extension of the IPCMB algorithm (FU; DESMARAIS, 2008), and that MB algorithms that do not assume causal sufficiency can be build by extending algorithms that assume sufficiency. At the same time, we define the Fast MAG Markov Blanket (FMMB) algorithm as an extension of EEMB (WANG; LING,

et al., 2020b).

We derived a generalized procedure that the algorithms use to discover nodes at depth d presented in Alg. 1, also illustrated it in Fig. 6. These algorithms discover, in depth first search (DFS) fashion, members of the MB at depths ≥ 2 starting from the structure already discovered by the base algorithm. Alg. 2 formalizes the employed DFS procedure. The differences between the algorithms lie in the inclusion and exclusion criteria they use to select and filter candidates. The criteria differentiating the algorithms are in Tab. 2. M3B uses a subroutine `Discover-Adj` that is the same procedure used by IPCMB to learn the nodes at depths 0 and 1. `Discover-Adj` selects as candidates every node that is connected to the target when no other nodes are observed. It than filters a candidate X if it can find a sepset $\sigma(X, T)$ among the other candidates (YU; LIU, et al., 2018; FU; DESMARAIS, 2008). If we follow M3B's algorithm as defined by Yu, Liu, et al. (2018) and stop the search at depth 2, M3B is equivalent to IPCMB. The exclusion criterion of FMMB and `Discover-Adj` in do not tell how the sepsets are found. The standard procedure in the literature and the way it as implemented in this work is as follows: Sort the candidates by connection strength with the target. Pick the candidates with the least connection strength first. Iterate through the powerset of all candidates minus the current candidate at consideration, in ascending subset size, and starting with the subsets that contain candidates with the most connection strength (YU; GUO, et al., 2020). This way, it is more likely that the false positives will be removed first, because of the data processing inequality, that guaranties that the connection strength of a node X that is not part of the MB is smaller than the connection strength between the target and the sepset of X. The EMMB algorithm as a base algorithm, and the unique extension procedure created by combining Alg. 1 and Alg. 2 with the criteria in Tab. 2, define the FMMB algorithm.

Now we give an example of how both algorithms would discover the nodes at depth 2 in Fig. 3, provided that a base algorithm already discovered the nodes at depths 0 and 1. FMMB searches the path $\mu_1 = [T, C, D]$ and does not discover any candidates. Then, it searches $\mu_2 = [T, E, F]$. It discovers the true positives J and K, and the false candidates L, M. Since they connect to T through the path $\mu_2 + [K]$. On the other hand, $\{N, O\}$ are never considered as candidates, because they are children of F, therefore are separated when F is observed. FMMB then filters L and M from the candidates, because they are separated from T when K is observed ($X \perp\!\!\!\perp T \mid \mu_2 \cup \sigma(X, T) \cup \{K\}$, where $\sigma(X, T) = \emptyset$, $\forall X \in \{L, M\}$). When searching μ_1, M3B calls `Discover-Adj(`D`)` and does not find any candidates as well. Then, it searches μ_2, calling `Discover-Adj(`F`)`. The nodes $\{J, K, N, O\}$ are considered as candidates, because they are adjacent to F. However, $\{L, \}$ were already eliminated by `Discover-Adj(`F`)` because they are independent of F given K. During the filter step, M3B asserts that J and K form a collider with F and E and eliminates $\{N, O\}$ because they fail to do so.

Algoritmo 1: `Discover-Att`

Inputs :
T: Target
μ: Collider Path
Output : $\mathbf{Att}_T[\mu]$: attached set of T given μ

`/* Initialization */`
$\mathcal{N} \leftarrow$ `Search-Space` (T, μ)
$\mathcal{C} \leftarrow \emptyset$; $\mathcal{A} \leftarrow \emptyset$; $\mathcal{E} \leftarrow \emptyset$
`/* Discover candidates */`
for $X \in \mathcal{N}$ **do**
  $\mathcal{N} \leftarrow \mathcal{N} \setminus \{X\}$
  if `Inclusion-Criteria`(X, T, μ) *is true* **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{X\}$
  else
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{X\}$
  end
end
`/* Filter candidates */`
for $X \in \mathcal{C}$ **do**
  $\mathcal{C} \leftarrow \mathcal{C} \setminus \{X\}$
  if `Exclusion-Criteria`(X, T, μ) *is true* **then**
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{X\}$
  else
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{X\}$
  end
end
return $\mathcal{A}$

Algoritmo 2: `DFS-Att-Discovery`

Inputs : T: Target, μ: Collider Path
$\mathbf{Att}_T[\mu] \leftarrow$ `Discover-Att`(T, μ)
for $X \in \boldsymbol{Att}_T[\mu]$ **do**
  `DFS-Att-Discovery`$(T, \mu + [X])$
end

3.1.2 Validation tests

To assess the ability of the proposed method to learn causal structures, two test cases (TCs) were elaborated using the Alarm Bayesian Network (BEINLICH et al., 1989). The experiment verifies that FMMB works as intended by applying it to data sampled from a distribution with a know underlying DAG. Fig. 7 displays Alarm's structure. TC 1 consists of discovering the MB of "VTUB" when "INT" and "PMB" are hidden. TC 2 is finding MB of "HR" where "HYP" and "STKV" are hidden, and the connection between "LVF" and "STKV" is removed (indicated by the dashed edge), making "LVF" only detectable at depth 2 through the path ["HR", "CO", "LVV", "LVF"]. Fig. 8 shows the

Table 2 – Criteria for selecting variables as part of the MB that define each discovery algorithm.

Algorithm	Inclusion Criteria	Exclusion Criteria	Search Space
Discover-Adj	$X \not\perp\!\!\!\perp T \mid \emptyset$	$\exists \mathcal{S} \subseteq \mathcal{A} \cup \mathcal{C} \setminus \{X\}, X \perp\!\!\!\perp T \mid \mathcal{S}$	$\mathcal{U} \setminus \{T\}$
M3B	$X \in$ `Discover-Adj`$(Y_m, [\,])^\dagger$	$X \perp\!\!\!\perp Y_{m-1} \mid Y_m \cup \sigma(X, Y_{m-1})^\dagger$	–
FMMB	$X \not\perp\!\!\!\perp T \mid \mu \cup \sigma(X, T)$	$\exists \mathcal{S} \subseteq \mathcal{A} \cup \mathcal{C} \setminus \{X\}, X \perp\!\!\!\perp T \mid \mu \cup \sigma(X, T) \cup \mathcal{S}$	$\mathcal{U} \setminus (\{T\} \cup \mathbf{Adj}(T) \cup \mu)$

$\dagger\ \mu = [Y_1, ..., Y_{m-1}, Y_m]$

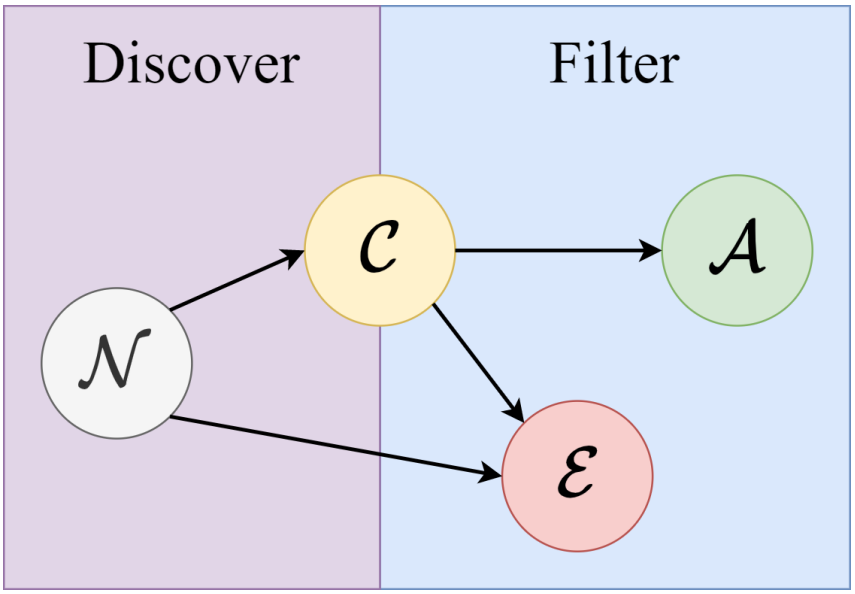


Figure 6 – Depiction of the states a node passed during the `Discover-Att` algorithm: not visited ($\mathcal{N}$), candidate ($\mathcal{C}$), admitted (($\mathcal{A}$), or excluded ($\mathcal{E}$).

correct structures of TC 1 and 2. The standard metrics were applied:

- Precision: the number of true positives in the output of the algorithm divided by the total size of the output of the algorithm. This measures the false positive rate in the output;
- Recall: the number of true positives in the output of the algorithm divided by the size of the true MB of the target. This reports the true positive rate in the output;
- F1: $f1 = 2 * precision * recall / (precision + recall)$ is the harmonic mean of precision and recall.
- CITs: total number of CITs conducted;
- Runtime: elapsed time of execution.

3.1.3 Feature selection experiment

As this class of algorithms were originally developed as tools for feature selection, a classification comparison after applying FMMB and M3B in ten real-world datasets is presented. Datasets were selected from the UCI Machine Learning Repository (DUA; GRAFF,

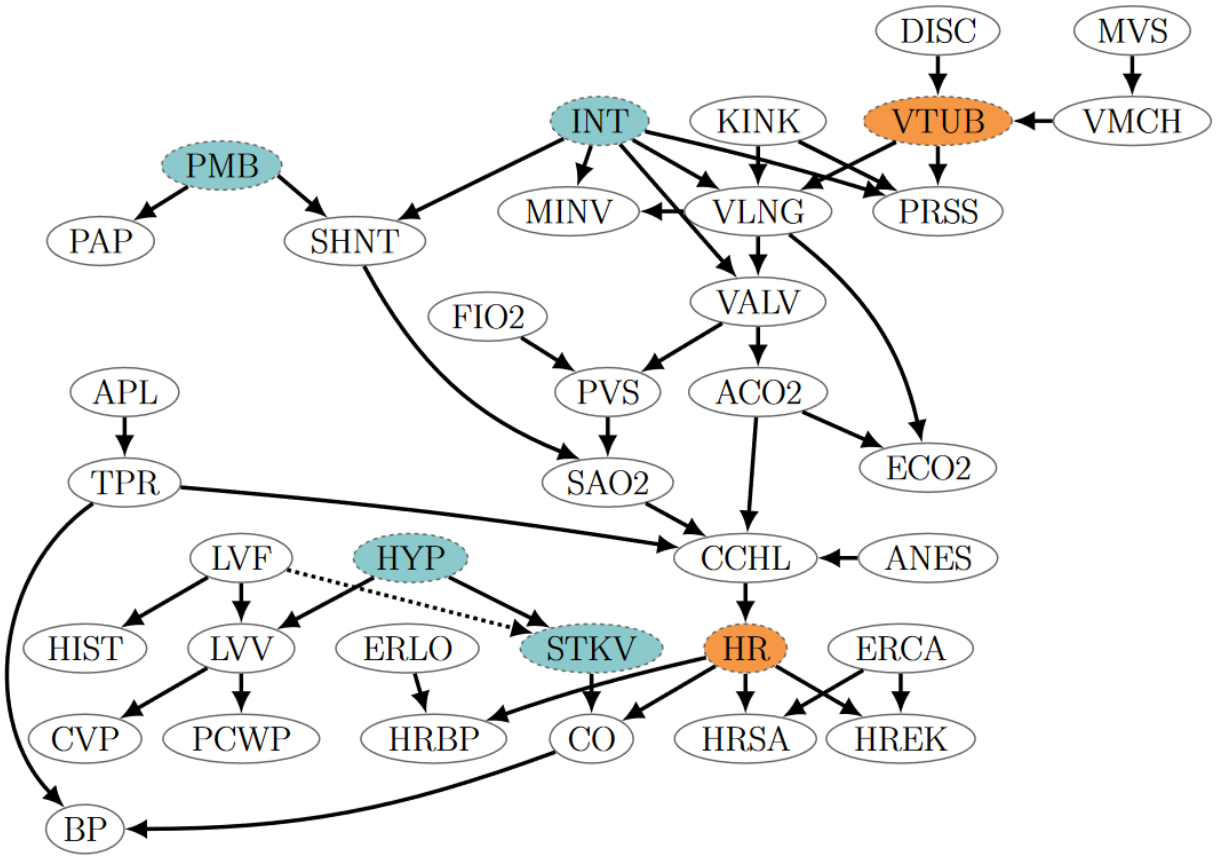


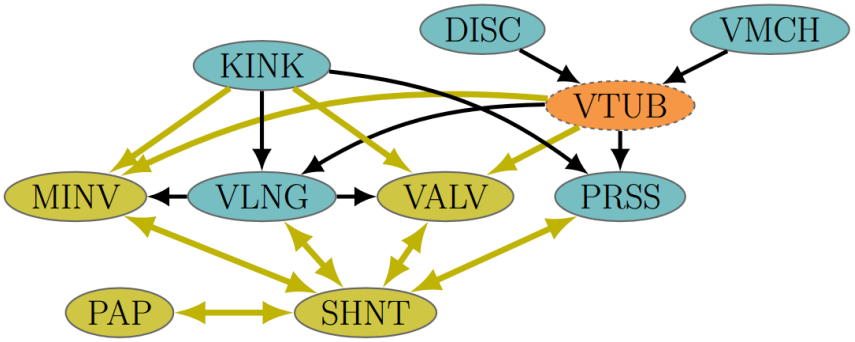
Figure 7 – Alarm, a well-known network designed for monitoring patients in intensive care unit, commonly used as a benchmark. From its structure, test cases (TCs) were elaborated for validating the proposed algorithm. For each test case, a node was selected as target and other nodes were hidden in order to violate causal sufficiency. The target nodes are shaded in orange and the hidden nodes are shaded in blue. The dashed arrow between "LVF" and "STKV" indicates that the connection is present in the original network, but was removed to create the test cases.

2017) and the KEEL dataset repository (ALCALÁ-FDEZ, 2011). Tab. 3 summarizes their characteristics. Naive Bayes (NB) and Gradient Boosted Decision Trees (GBDT) (KE et al., 2017) were chosen as classifiers due to NB's simplicity and GBDT's native capacity to handle mixed data types, as the selected datasets are comprised of categorical, integer, and continuous variables.

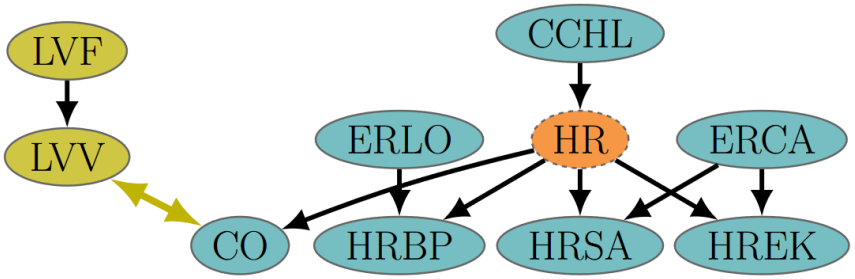
3.2 Aggregation of MB bootstraps and Application in Credit Data

3.2.1 Proposed Aggregation Method

The proposed method consists of first achieving a consensus in terms of nodes from the bootstrapped MBs, then constructing the aggregated structure from the union of the edges that form the most probable path $\mu^\star$ between the target T and each feature X, where $\mu^\star$ is defined in Eq. (3.1).



(a) Test Case 1: MB of "VTUB" when hiding "INT" and "PMB". There are also new bidirected connections between "PRSS" and every variable in {"VALV", "MINV", "VLNG"} that are not shown in the figure.



(b) Test Case 2: MB of "VTUB" when "HYP" and "STKV" are hidden, and "LVF" → "STKV" is removed.

Figure 8 – Correct structure of each test case created from the Alarm Network. The nodes in green only become members of the MB when the mentioned variables are hidden, while the nodes in blue are also part of the MB when all the variables are observable. Likewise, the arrows in green indicate connections that only appear in the MAG.

$$\mu^\star = \arg\max_{\mu} \prod_{e \in \mu} p(e) \tag{3.1}$$

Finding the most probable path is easily solvable using Dijkstra's algorithm by transforming the problem of searching for the path with the greatest product of probabilities into a shortest path problem (CORMEN et al., 2009). This is accomplished by applying the map $p \mapsto -\log(p)$ to the edge probabilities.

We consider two techniques to reach consensus. The first is modelling the probabilities of occurrence of features in the bootstrapped MBs as independent of each other and using a threshold τ to select features with frequency of occurrence $> \tau$. Scutari and Nagarajan (2013) have pointed out that most of the thresholds chosen in the literature are arbitrary, and proposed a statistically motivated threshold. Using the order statistics of the frequency of occurrence of edges, the method consists of solving for a threshold that minimizes the loss between the observed values and the theoretically optimal network. We

Table 3 – Summary of datasets used in feature selection experiment.

Dataset	Features	Samples
spect	22	267
spectf	44	267
splice	60	3190
chess	36	3196
mushroom	22	8124
polish	58	41340
connect-4	42	67557
kddcup99	39	100655
census	41	299285
covertype	54	581012

adapted Scutari's method by using the order statistics of the probability of occurrence of the nodes, instead of edges. The second is Hamming Clustering (HC) (WANG; PENG, 2014; GASIENIEC; JANSSON; LINGAS, 2004). It is realized by transforming the bootstrapped MBs to a set $\mathbf{S}$ of binary strings in which the i^{th} bit indicates the presence of the i^{th} feature in the MB, following an arbitrary but fixed ordering. The simplest HC technique is finding a consensus string that has the minimum Hamming distance (MHD) to every other string in $\mathbf{S}$. Formally, given a norm ν, the consensus string $s^\star$ is the answer to the minimization problem expressed in Eq. (3.2), where d_H denotes the Hamming distance (BULTEAU; SCHMID, 2020; CHEN; HERMELIN; SORGE, 2018).

$$s^\star = \arg\min_{s'} \left(\sum_{s \in \mathbf{S}} d_H^\nu(s, s') \right)^{1/\nu} \tag{3.2}$$

We compare aggregation by nodes with aggregation by edges, testing Scutari's and MHD thresholds for each aggregation strategy. It is easy to show that for $\nu = 1$ the problem is reduced to selecting the most frequent bit value at each position, which is equivalent to the first aggregation method with $\tau = 0.5$. We also tested $\nu = 2$ and $\nu = \infty$, but did not find any significant differences in the results. Thus, opted for the simplest approach, $\nu = 1$. Similarly, using MHD to aggregate by edges is reduced to selecting edges with frequency of occurrence $> 50\%$.

3.2.2 Experimental design

Fig. 9 is a flowchart of the whole process, from data preprocessing to result generation. During preprocessing, only completed loans were considered. Features with more than 85% of missing values were dropped, and then, remaining entries with missing values were dropped. The preprocessed datasets were split into 70% training and 30% testing portions. The training portion was used for MB learning and classifier fitting. The test portion was used to measure classification performance. We measured the performance of

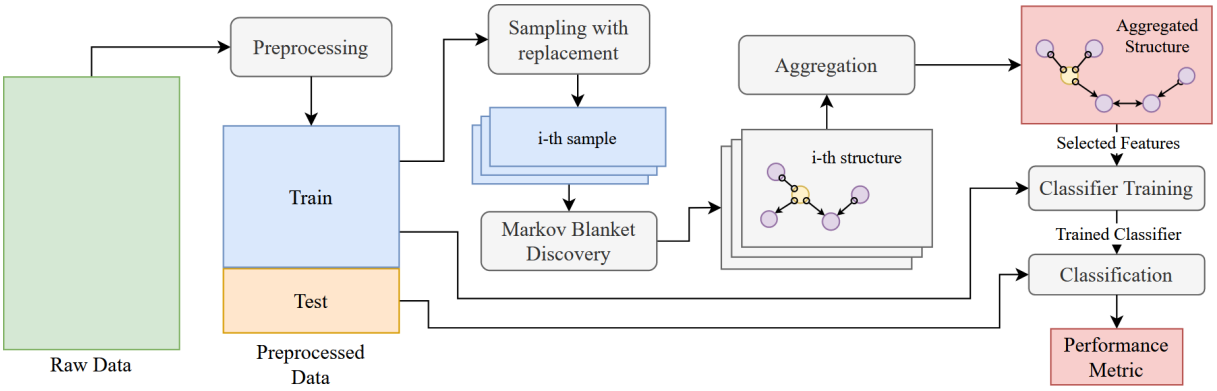


Figure 9 – Flowchart of the experimental design for studying aggregation methods.

Table 4 – Characteristics of credit datasets after preprocessing.

Dataset	Target	Features	Samples
Lending Club	Loan Default	29	1.74×10^6
Prosper	Loan Default	50	4.14×10^4
PAKDD 2009	Credit Card Default	20	4.00×10^4
Taiwan	Credit Card Default	23	3.00×10^4
Polish	Bankruptcy	58	4.13×10^4

the aggregated structure for each of the aggregation methods, and also of each bootstrapped structure, for FMMB and M3B. We tested Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Gradient Boosted Decision Trees (GBDT), and Random Forests (RF) as classifiers, but chose GBDT because it achieved the best classification performance overall, does not require much data processing, and is well suited for mixed datasets (KE et al., 2017; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). For each run, we generated 30 bootstraps of the data with 10% of the size of the original dataset. The hyper-parameters were chosen following best practices (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.2.3 Data description

Five publicly available credit related datasets were gathered. The *Polish* dataset, also present whose associated task is to predict the bankruptcy of Polish companies, is available at the UCI Machine Learning Repository. Also available at UCI is the *Taiwan* dataset, whose task is the prediction of the default of credit card clients in Taiwan. The *PAKDD 2009* dataset was provided during the PAKDD 2009 Data Mining Competition, with the goal of predicting the default of clients of a private label credit card of a major Brazilian retail chain. The *Prosper* data are provided by Udacity as part of their Data Analyst Nanodegree. Lastly, the *Lending Club* data, ranging from 2007 to 2020, are available at Kaggle [1]. Tab. 4 summarizes the characteristics of the datasets after preprocessing.

[1] https://www.kaggle.com/ethon0426/lending-club-20072020q1

3.3 Credit Scoring with Bayesian Network Classifier Ensembles

3.3.1 Learning BNCs using CEM

Aug	X_1	X_2	X_3	X_4	X_5
X_1		1	0	0	1
X_2			0	0	0
X_3				0	0
X_4					1
X_5					

Ch	X_1	X_2	X_3	X_4	X_5
	1	0	1	0	1

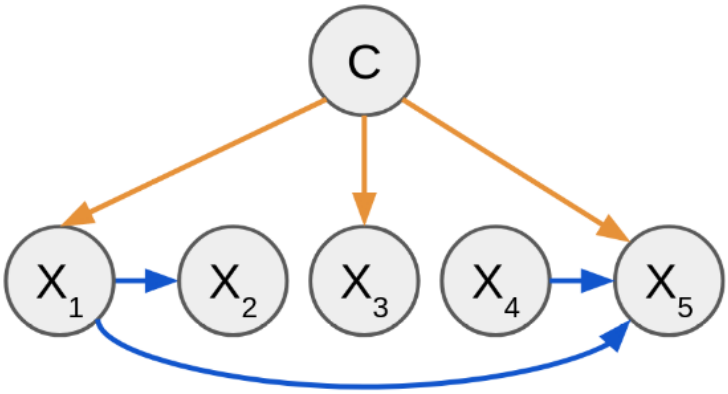


Figure 10 – Example of a mapping from a binary array to a BAN structure with five features. Assuming an ordering $\sigma = (X_1, X_2, X_3, X_4, X_5)$, children (ch) edges can be mapped to the initial positions of the array and augmenting (aug) edges to the remaining positions following an upper triangular matrix, e.g. $(1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1)$.

The method consists of a Bagging ensemble of BNCs, where each classifier is learned using CEM. In a nutshell, Bagging is the process of building and ensemble where, at each iteration, the original dataset is resampled with replacement (bootstrapping) and a classifier is trained using the bootstrapped data. The resulting ensemble uses majority voting to make predictions. CEM is an iteration based optimizer that samples from some distribution and selects the top performing members of the population according to a performance function. The cross-entropy minimization principle is then used to unite the selected members to update the distribution's parameters. In our case, the population members are BNCs with randomly generated structures, encoded by a binary array that CEM can optimize. We decided to limit the search space to BAN structures, encoding both augmenting edges, and edges between the class variable and the features (children edges), as follows:

- Let $\mathbf{Z} = (Z_1, ..., Z_m)$ be a array of Bernoulli random variables with probabilities of success $\mathbf{p} = (p_1, ...p_m)$, i.e. $\mathbf{Z} \sim \texttt{Ber}(\mathbf{p})$;
- Given a fixed ordering of variables, there are $m(m-1)/2$ possible augmenting edges in a DAG comprising m feature variables. We opted to randomly generate a

fixed ordering for each classifier in the ensemble, intending to increase search space exploration;

- Every possible augmenting and child edge is assigned to a random variable in $\mathbf{Z}$. The value 1/0 indicates the presence/absence of an edge.

Fig. 10 demonstrates the mapping between a realization of $\mathbf{Z}$ and the resulting structure. Next, consider the performance function $S(\mathbf{Z})$ that measures minus the log-loss (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) of a BNC with edges encoded by $\mathbf{Z}$ when predicting the training set. We have an optimization problem summarized in (3.3).

$$\mathbf{p}^{\star} = \underset{\mathbf{p}}{\operatorname{argmax}}\ S(\mathbf{Z}),\ \ \mathbf{Z} \sim \texttt{Ber}(\mathbf{p}) \tag{3.3}$$

The following variation of CEM, described in (DE BOER et al., 2005, p. 7), was used to solve (3.3). At each iteration:

1. Draw samples $\mathbf{Z}_1, ..., \mathbf{Z}_N$ from $\texttt{Ber}(\mathbf{p})$, generating a population of size N;
2. Let γ be the $(1-\rho)\%$ quantile according to the performance function $S(\mathbf{Z})$. Select top performing members of the population: $\{\mathbf{Z}_i \mid S(\mathbf{Z}_i) \geq \gamma\}$;
3. Update $\mathbf{p}$ using the cross-entropy minimization rule:

$$\mathbf{p} \leftarrow \frac{\sum_{i=1}^{N} I_{\{S(\mathbf{Z}_i)\geq\gamma\}} I_{\{Z_{ij}=1\}}}{\sum_{i=1}^{N} I_{\{S(\mathbf{Z}_i)\geq\gamma\}}} \tag{3.4}$$

4. Stop if γ has not changed for a fixed number K of iterations or $\mathbf{p}$ has converged to a binary array.

Following good practice, a smoothing during the update of the parameters was employed (RUBINSTEIN; KROESE, 2013). Let λ be the smoothing factor, and $\hat{\mathbf{p}}$ be the cross-entropy optimal parameter, calculated using Eq. (3.4). We update $\mathbf{p}$ using:

$$\mathbf{p}_k \leftarrow \lambda\hat{\mathbf{p}} + (1-\lambda)\mathbf{p}_{k-1} \tag{3.5}$$

3.3.2 Data preprocessing

Five publicly available credit related datasets were gathered. The *australian*, and *south-german* datasets are available at the UCI Machine Learning Repository (DUA; GRAFF, 2017). The *pakdd2009* dataset was provided during the PAKDD 2009 Data Mining Competition. Lastly, the Prosper data are provided by Udacity as part of their Data Analyst Nanodegree. After 2009, policy changes happened at Prosper intending to guarantee better credit grading and more security for lenders (YEN, 2019). Because

Table 5 – Summary of the preprocessed credit datasets used for comparing classifiers.

Dataset	Number of Instances	Number of Features	Bad Borrowers (%)
pakdd2009	39997	20	19.80
prosper-2009+	23028	28	23.14
prosper-2009	20008	26	35.42
south-german	1000	20	30.00
australian	690	14	55.51

of a possible heterogeneity, the data were split into two datasets: *prosper-2009* and *prosper-2009+*. Only completed loans were considered, and entries with missing values were dropped. Finally, the preprocessed datasets were split into 70% training and 30% testing data. Tab. 5 summarizes the five datasets. Notice that appart from *australian*, the datasets are imbalanced such that the bad borrowers represent the minority class, as expected.

3.3.3 Model training and evaluation

We compared the proposed method with Logistic Regression (LR) – a linear method, and Random Forests (RF) – an ensemble of decision trees (DTs). When making a critical or cost-sensitive decisions such as in credit scoring, well-calibrated probabilities become an issue (DAWID, 1982). While LR produces well-calibrated probabilities, DTs and BNCs might produce biased probabilities towards the edge of the probability space, that is, too close to 0 or 1 (ZHANG, 2005; NICULESCU-MIZIL; CARUANA, 2005; ZADROZNY; ELKAN, 2001; RISH et al., 2001; DOMINGOS; PAZZANI, 1997). For this reason, we applied Platt scaling to calibrate RF and the proposed BNC (PLATT et al., 1999). In the interest of increasing classification performance (MADDEN, 2008), BDeu-like priors with an equivalent sample size (η) of 10 were used to smoothen the CPDs. That is, the prior probability of every state is $\eta/(r_X q_X)$ where r_X is the cardinality of the variable X and q_X is the product of the cardinalities of the parents of X (CAMPOS, 2006). The models were trained using Bayesian hyper-parameter optimization (SCIKIT-OPTIMIZE, 2021) to decide the number of classifiers composing the ensemble in the case of RF and BNC, and in the case of LR, the L1 and L2 regularization penalties (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). The chosen values for the hyper-parameters of CEM were $N = 100$, $\rho = 0.1$, and $K = 5$, determined *ad hoc*.

Model comparison is made using these evaluation metrics:

- **Precision:** $tp/(tp + fp)$ where tp is the number of true positives, and fp is the number of false positives;
- **Recall:** $tp/(tp + fn)$ where fn is the number of false negatives;

- **F1:** the harmonic average of precision and recall;
- **AUC:** Area Under the (Receiver operating characteristic) Curve (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Since the financial cost of a false negative is greater than that of a false positive, AUC and recall should be preferred over other metrics. That is because wrongly classifying someone that will ultimately default their loan as a good borrower implies a principal loss, whereas classifying a potentially good borrower as bad only incurs an opportunity cost.

4 RESULTS AND DISCUSSION

4.1 MB discovery without causal sufficiency

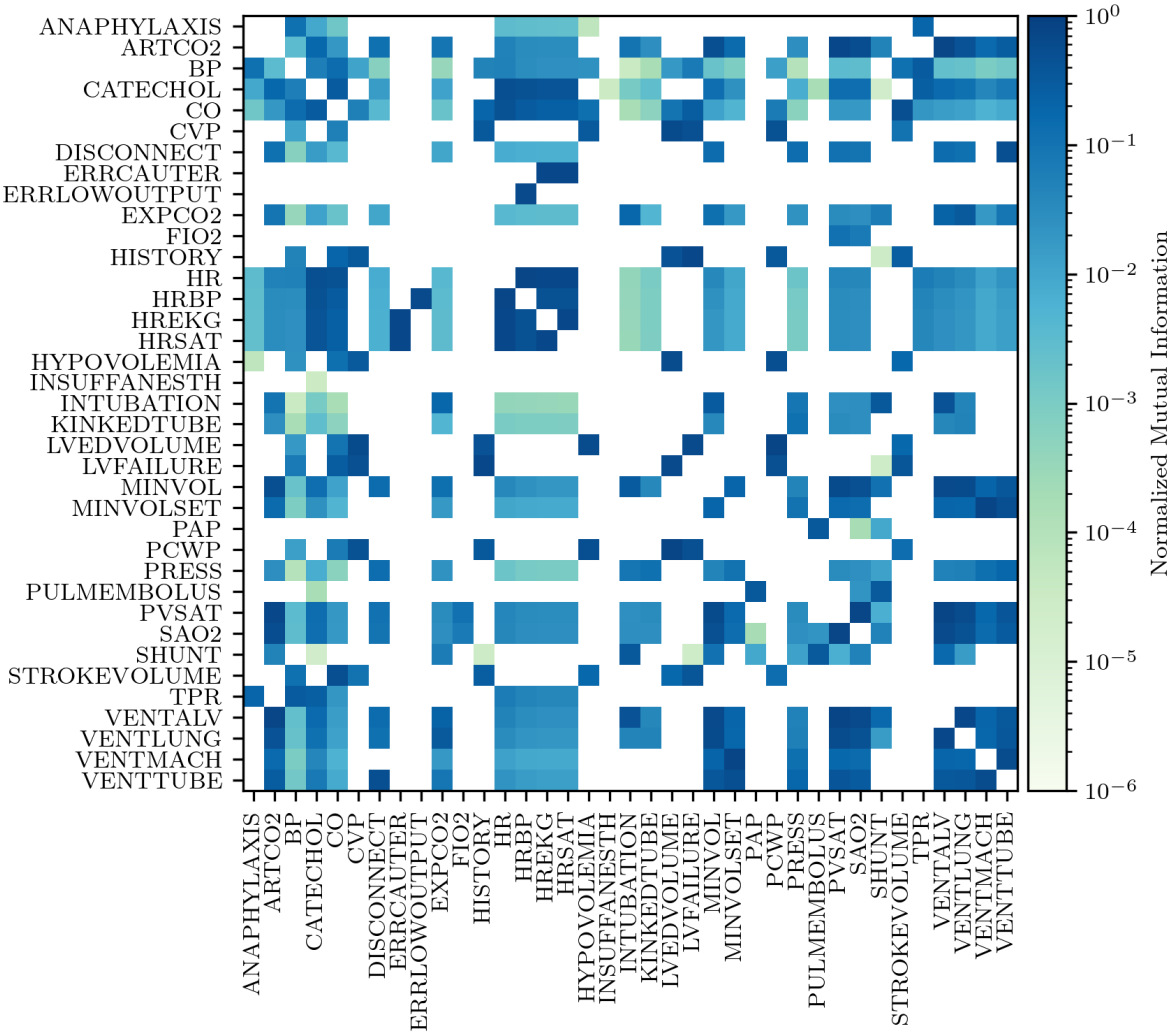


Figure 11 – Normalized Mutual Information between every pair of features in the Alarm Network.Entries in white indicate independence of the variables, given a G^2 test at a significance level of 0.01.

In Fig. 11 we can notice the high range of connection strength magnitudes present in the Alarm network. As the data was generated by sampling the network, it adheres perfectly to the structure. Naturally, the further the distance between each end, the lower the connection strength tends to be, as a consequence of the data processing inequality. It is also noticeable that the nodes that are adjacent to the targets of each test case are tightly coupled, with very high connection strengths.

Fig. 12 and 13 display the results of the structural learning experiments in the Alarm network. For small sample sizes, FMMB is unstable, producing poor precision and

recall values. Due to it's direct discovery strategy, as the conditioning sets increase in size, the power of the G^2 tests decrease. But given a large enough sample size, FMMB becomes capable of detecting the correct MB more often. Notice in Fig. 12 that the performance of M3B and FMMB decreases at 150000 samples, where a decay in precision is mainly attributed to spurious connections that are now detectable by the G^2 test that were not filtered. This can be attended with a higher *max-k* value. TC 2 was the most challenging for both algorithms, requiring a sample size of 20000 for M3B to achieve the correct result 50% of the time, whereas FMMB was not able to deliver a comparable performance even with a sample size of 150000. As the precision of FMMB also decreased at that sample size, a larger sample and increased *max-k* may be necessary. The main difference between the two test cases is that the connection strengths between the target and the deepest nodes are lower in TC 2, consequently being discovered less frequently. Nonetheless, when comparing the peak performance of both algorithms, there is a large gap in the number of CITs and execution time. For instance, in TC 2 at 50000 samples, FMMB conducted a median of 805.5 CITs in 5.60s *vs* 2182.50 CITs in 12.70s by M3B. Both methods were capable of discovering the correct structures in all the test cases. Overall, direct discovery proved to be faster, however, less consistent than the indirect method.

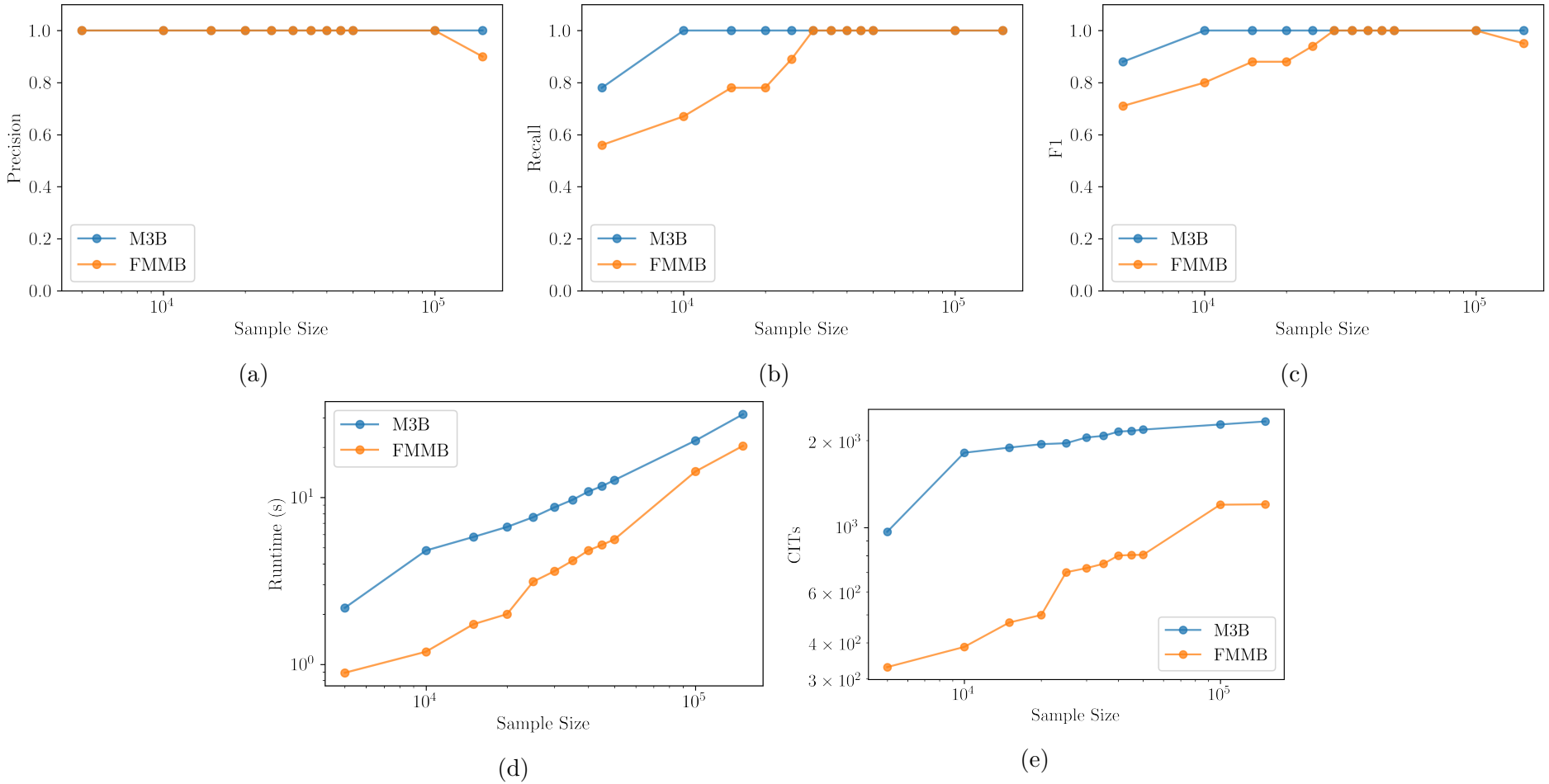


Figure 12 – TC1 – Estimation of the MB of "VTUB" when "INT" and "PMB" are hidden. Results are the median of 30 runs of the experiment.

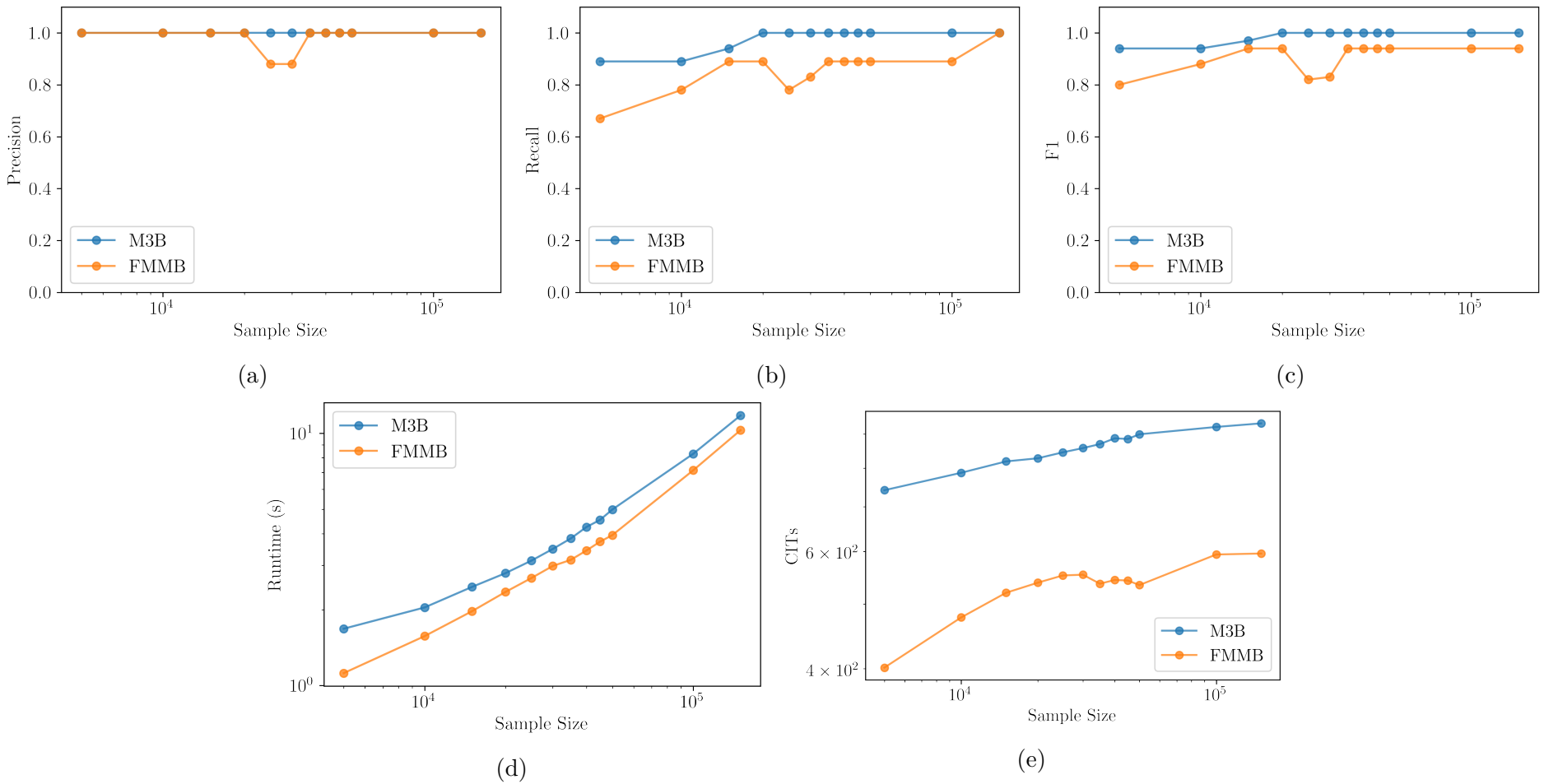


Figure 13 – TC2 – Estimation of the MB of "HR" when "HYP" and "STKV" are hidden, and the connection "LVF" → "STKV" is removed. Results are the median of 30 runs of the experiment.

Table 6 – Feature selection experiment results for M3B. Values in bold indicate a better performance than it's counterpart (FMMB), and values in italic indicate equality. 10-fold cross-validation was used in the five smaller datasets – the results are in the format *mean±std*. For the five larger datasets, a single 70%/30% train/test split was measured. Computational time for the execution of the feature selection algorithms was limited to 15000*s*. Elapsed time only encompasses feature selection, i.e. classifier training and prediction times were excluded from the measurements.

	M3B				
	NB Acc. (%)	GBDT Acc. (%)	MMB Size	Time (s)	CITs
spect	**78.97 ±4.10**	*79.36 ±3.68*	**2.30 ±0.46**	0.40±0.05	220.40±30.30
spectf	*80.46 ±8.76*	*68.49 ±9.18*	*2.00 ±0.00*	0.49±0.03	285.30±17.18
splice	*94.58 ±0.53*	*94.23 ±0.92*	*8.20 ±0.40*	4.77±0.29	2218.70±125.47
chess	89.05±1.98	98.40±0.99	**23.60 ±1.85**	20.25±3.17	9734.70±1491.28
mushroom	**98.82 ±0.40**	**99.41 ±0.24**	**2.00 ±0.00**	2.73±0.09	1073.20±8.95
polish	–	–	–	$\geq$ 15000	–
connect-4	**72.47**	**83.78**	32	2278.23	191249
kddcup99	*99.80*	*99.97*	*11*	183.47	14833
census	**93.73**	**94.12**	**3**	143.42	2313
covertype	*62.19*	*86.28*	*10*	614.81	5544

Table 7 – Feature selection experiment results for FMMB. Values in bold indicate a better performance than it's counterpart (M3B), and values in italic indicate equality. 10-fold cross-validation was used in the five smaller datasets – the results are in the format *mean±std*. For the five larger datasets, a single 70%/30% train/test split was measured. Elapsed time only encompasses feature selection, i.e. classifier training and prediction times were excluded from the measurements.

	FMMB				
	NB Acc. (%)	GBDT Acc. (%)	MMB Size	Time (s)	CITs
spect	76.38±2.58	*79.36 ±3.68*	2.60±0.49	**0.28 ±0.05**	**114.80 ±15.56**
spectf	*80.46 ±8.76*	*68.49 ±9.18*	*2.00 ±0.00*	**0.21 ±0.01**	**109.20 ±2.23**
splice	*94.58 ±0.53*	*94.23 ±0.92*	*8.20 ±0.40*	**1.92 ±0.16**	**804.40 ±59.44**
chess	**89.52 ±2.17**	**98.50 ±1.30**	25.10±1.97	**17.39 ±3.66**	**6548.50 ±1292.10**
mushroom	98.76±0.35	99.27±0.18	3.00±0.00	**0.39 ±0.02**	**137.40 ±5.39**
polish	**71.42**	**84.95**	**49**	**4859.74**	**464028**
connect-4	72.36	81.29	**22**	**343.29**	**25514**
kddcup99	*99.80*	*99.97*	*11*	**16.99**	**1391**
census	91.92	93.83	5	**24.36**	**401**
covertype	*62.19*	*86.28*	*10*	**61.49**	**944**

Classification accuracy results for each dataset obtained applying both classifiers, after feature selection using M3B and FMMB, are shown in Tab. 6 and 7. FMMB proved to be much fast, specially in the five larger datasets, while displaying equal or very close accuracy using both classifiers. Interestingly, with the exception of the *connect-4* dataset, the size of the MBs discovered by FMMB were also equal or slightly higher than those by

M3B. This may indicate a higher false positive rate, as expected from the worse precision observed in the structural learning experiment. The most discrepant gain in terms of execution time and number of CITs occurred in the *kddcup99* dataset, where FMMB was 10× faster and conducted 10× less CITs than M3B, while achieving identical accuracy in both classifiers. The results obtained in the *polish* dataset are worth highlighting. Although it's medium in samples, it has the most features among the selected datasets. While M3B could not terminate execution in a reasonable amount of time, FMMB was able to finish execution, showcasing the necessity for faster feature selection algorithms. However, large difference in the experimental design between sec. 3.1 and sec. 3.2 should be remembered. In this experiment, the algorithms are not bootstrapped, using the whole dataset as input, and producing a single output for the feature selection experiment. Whereas in sec. 3.2 the data is bootstrapped. Each bootstrap has a fraction of the size of the original data. This explains why the Polish dataset, present in both sections, did not produce any output for the M3B algorithm in a feasible amount of time, while it was possible to produce results for the bootstraps.

4.2 Aggregation of MB bootstraps and Application in Credit Data

From Fig. 14 and 15 we can infer that FMMB has an equivalent classification performance to M3B, which indicates that both methods discover MBs of equal quality. However, FMMB is significantly faster than M3B, with the largest speedup of two orders of magnitude in the Lending Club dataset, which is also where the speedup is most felt, since it is the dataset where the algorithms had the largest runtime. When comparing aggregation methods, the aggregation by nodes performed better than or equal to the aggregation by edges, and also produced MBs with sizes closer to the average size of the bootstrapped MBs. Moreover, the MHD threshold performed consistently better or equal to Scutari's. Specifically, the aggregation by edges using Scutari's threshold, when applied to the bootstrapped structures of the FMMB algorithm in the Taiwan dataset, resulted in an empty structure. Whereas the MHD threshold produced a result slight lower than the median, implying that Scutari's was too restrictive.

For each dataset, we measured the NMI between every pair of variables. The results are displayed in heat maps contained in Fig. 16, 18, 20, 22, 24, for Lending Club, Prosper, PAKDD2009, Taiwan, and Polish, respectively. Considering the MHD aggregation by nodes, as it is the best performing overall, figures were produced showing the MB discovered by FMMB and M3B for each of the datasets. The MB structures for Lending Club, Prosper, PAKDD2009, Taiwan, and Polish are, respectively, in Fig. 17, 19, 21, 23, 25.

When analysing the results of the two P2P lending platforms, the heat maps paint similar pictures, with Prosper being slightly more sparse in connectivity than Lending

Club. In the structures it is noticeable that many of the most likely paths have edges with frequency $< 50\%$. In addition, the further the feature is from the target, the greater is the number of possible paths between it and the target. The features at depth 3, for example, were separated from the target when using aggregation by edges, because of the many possible paths that dilute the frequency of occurrence of the edges connecting the features to the rest of the MB. It is noticeable that the MB of the Prosper data is almost a subset of the MB of Lending Club. FMMB selected "Interest Rate", "Term" , the credit score rating – similar to the "FICO Rating" feature of Lending Club – and "Total Inquiries" – the closest feature in Lending Club is "Inquiries (6 mth)". The only difference being that "Employment Status" is present in Prosper whereas "Income Verified" is present in Lending Club, yet both are related to source of income. Results for M3B are also similar, containing "Interest Rate", the credit score rating, "Inquiries (6 mth)" – also present in Lending Club – and "Employment Status". The discovered MBs for Lending Club by M3B (Fig. 17a) and FMMB (Fig. 17b) agree upon which features are adjacent to the target, as well as on most of the connections. "Interest Rate" was selected in 100% of the bootstrapped structures in both algorithms as adjacent to the target, and is also the feature that shares the most amount of information with the target. (Fig. 16). From a lender's perspective, this shows that the models used for credit scoring at Lending Club and Prosper are reliable, because the interest rate is equal to a base rate, determined by the grade, adjusted for risk and volatility (CLUB, 2021), as was also concluded by other authors. However, there is still room for improvement, since there are many other features that compose the MB, thus providing additional information not completely captured in their models.

Both algorithms agree that "Annual Income" is at depth two, but disagree on the most likely collider. The most likely collider in M3B's structure is "Debt to Income Ratio". In contrast, FMMB's structure indicates a very low probability for this edge. Instead, it chooses "Installment", which has a high probability in both structures. Another difference is that while M3B selects "Revolving Utilization", the 12^{th} most informative when solely observed (Fig. 16), FMMB prefers "Total Credit Revolving Balance", which is the 5^{th} most informative and highly correlated with "Revolving Utilization". Both place similar features at depth 2, considering the most probable paths. Two nodes appear at depth 2 in the M3B structure that are not in FMMB's: "Total Credit Lines" and "Open Credit Lines", which are highly correlated yet both part of the MB. From the edge probabilities, we can see that both features were most likely discovered from different paths and rarely share an edge in the bootstrapped structures, which is confirmed by the following statistic: They appeared simultaneously in half of the structures but shared an edge only 20% of the time. This may be caused by inconsistencies in the indirect method employed by M3B, or is a sign of a partially unfaithful distribution, where there are two possible MBs, each exclusively containing one of the features. Both algorithms detected three nodes at depth

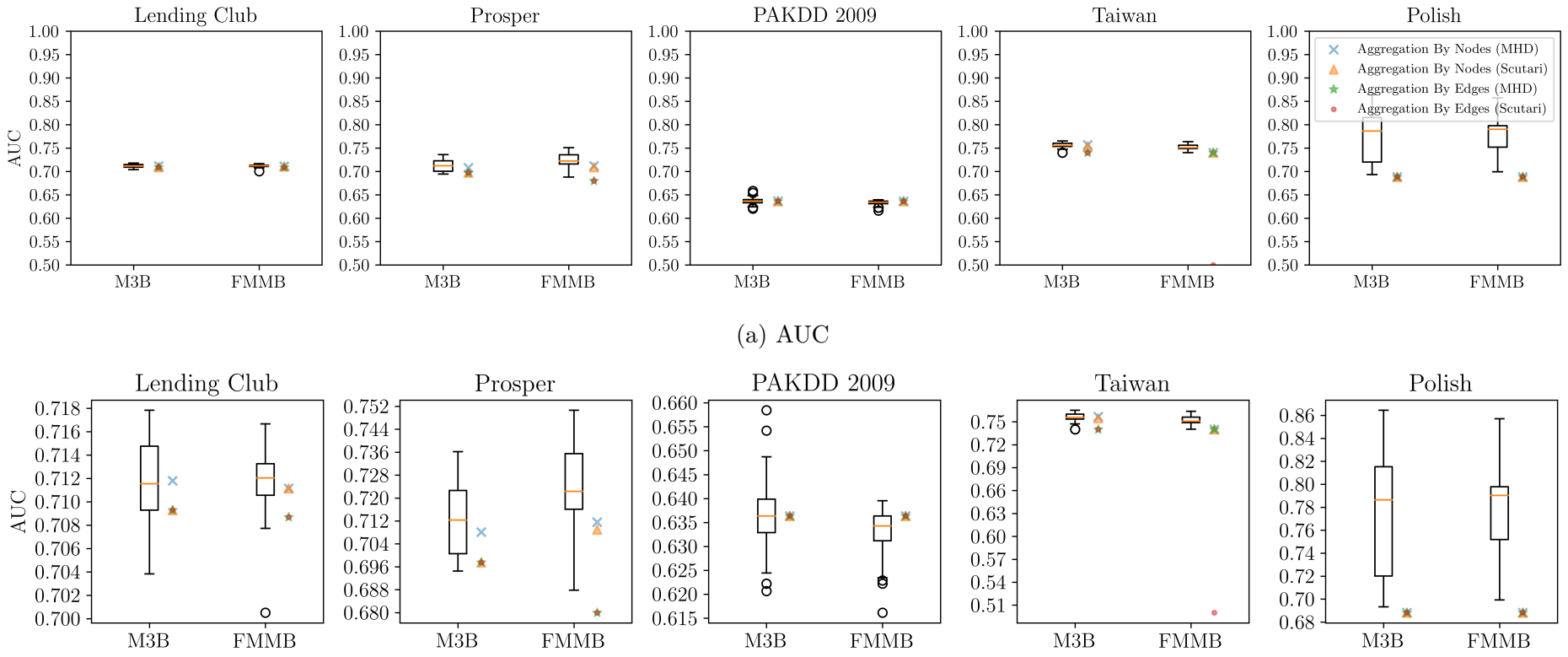
3, agreeing on “Verified Income” and “Earliest Credit Line” but only FMMB includes “Inquiries (6mth)”, and only M3B includes “Purpose”. While these are also concluded to be relevant factors when using linear methods, here, their relative importance is diminished because of their more distant relationship with the target.

Focusing on the FMMB structure, apart from “Installment”, which is the most likely collider between Revolving Balance and the target in the FMMB structure, the other members of $\mathbf{Adj}(T)$ are equally frequent as members of the sepset of the feature in the bootstrapped structures. Because the mutual information is so high, yet this feature appears at depth two, we can infer that there is at least one activate path between the feature and the target that is blocked by a subset $\mathbf{Adj}(T)$. Either this path is a subset of $\mathbf{Adj}(T)$ minus the collider (“Installment”), which is likely since the connection strength among the Revolving Balance and its most probable separators are high, as evidenced in Fig. 16. But, because these connections are not necessary to discover the $MB(T)$, they were not searched by the algorithm, therefore were omitted in the structure. Or, less likely but possible is the existence of a latent variable between the separators and the feature that creates an active path with the target when the separators are not observed. To further clarify the matter, the MB of the Revolving Balance would have to the discovered, which is beyond the scope of this work.

When comparing the two credit card related datasets, a major difference is the essence of the features. In Taiwan, most of the features are related to the payment of the last credit card bills. There are few social economical variables. Whereas in PAKDD2009 the situation is reversed. The features of this dataset bring almost no information about the costumer’s credit history, the only variable being the quantity of additional cards in the application. Most of the features are either social-economic or geographic. PAKDD2009 has the lowest average size of bootstrapped MBs. All the aggregation methods produced shallow structures with the same two nodes adjacent to the target: the age of the costumer and a Boolean variable that indicates whether the customer has a residential phone line. The presence of a phone in the residency is one of the social-economic variables considered in the census promoted by the Brazilian government and is an indicator of wealth. Age is also correlated with wealth. Since the consensus structure had median performance, we infer that these features carry most of the information about the target. The two outliers that performed best either were able to capture connections that require greater sample sizes to be detected consistently or are the product of noise during classification training. In Taiwan, we see that M3B performed marginally better than FMMB (Fig. 14b). In the discovered structures, while M3B discovered features up to depth 3, FMMB failed to discover more than the nodes adjacent to the target. While FMMB only discovered the Boolean variables indicating whether or not the payment due i months ago was payed, M3B recovered the bill amount and the amount actually paid as well. However, the connections of the M3B structure does not seam to have much meaning. From the heat map we can

see that the credit history features are extremely correlated with each other, and that the frequency of occurrence of the edges in the structure is directly proportional to the magnitude of the NMI between the variables. Nonetheless, a general observation can be inferred from the MB selected by both algorithms: The credit history is enough for determining the behaviour of the client, eliminating the influence social economic variables have on the target. It is an intuitive finding, since the social economic factors influence the credit history, but is the credit history itself what ultimately determines if a credit card has defaulted, since credit card default is when a client has not payed the bills for an extended period of time.

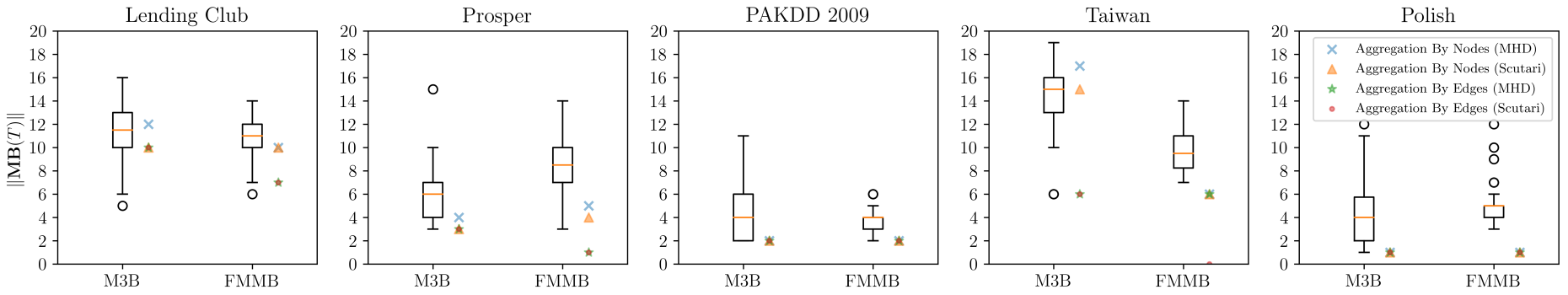
We calculated the average of the NMI measurements in the heat maps and found that Polish (0.052) and Taiwan (0.060) have averages more than two times higher than the other datasets (< 0.018). While Taiwan's results were similar to the other datasets, all aggregation methods performed poorly in the Polish data. The average sizes of the discovered MBs using M3B and FMMB in the Polish dataset were, respectively, 4.43 and 5.17. However, the aggregated structures of all methods had only 1 node, feature "X45" – the features of this dataset are anonymized. The high connectivity, increasing the possibility of existing many subsets of features that bring the same information about the target, and the fact that no other features were discovered consistently, denying the formation of a consensus structure of similar size, make us belief that the Polish data have an unfaithful distribution, explaining the poor results.



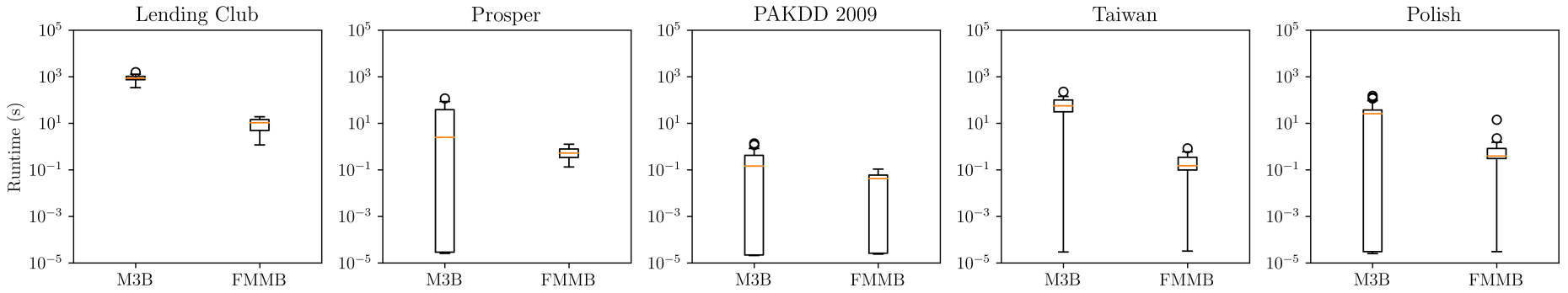
(a) AUC

(b) AUC (Amplified)

Figure 14 – Comparison of FMMB and M3B, as well as structure aggregation methods, in credit related datasets considering classification performance (AUC). Each boxplot summarizes the results of 30 bootstraps.



(a) Markov Blanket Size



(b) Runtime

Figure 15 – Comparison of FMMB and M3B, as well as structure aggregation methods, in credit related datasets considering MB size and Runtime. Each boxplot summarizes the results of 30 bootstraps. The runtime of the aggregations is in the order of milliseconds.

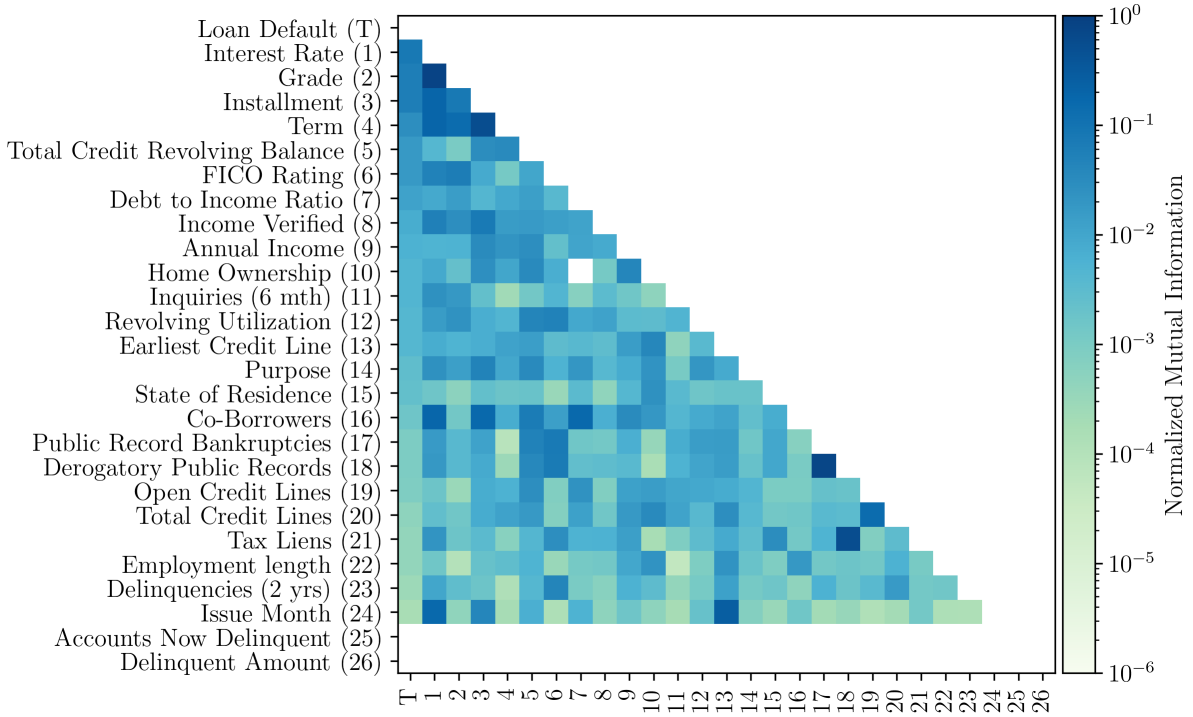
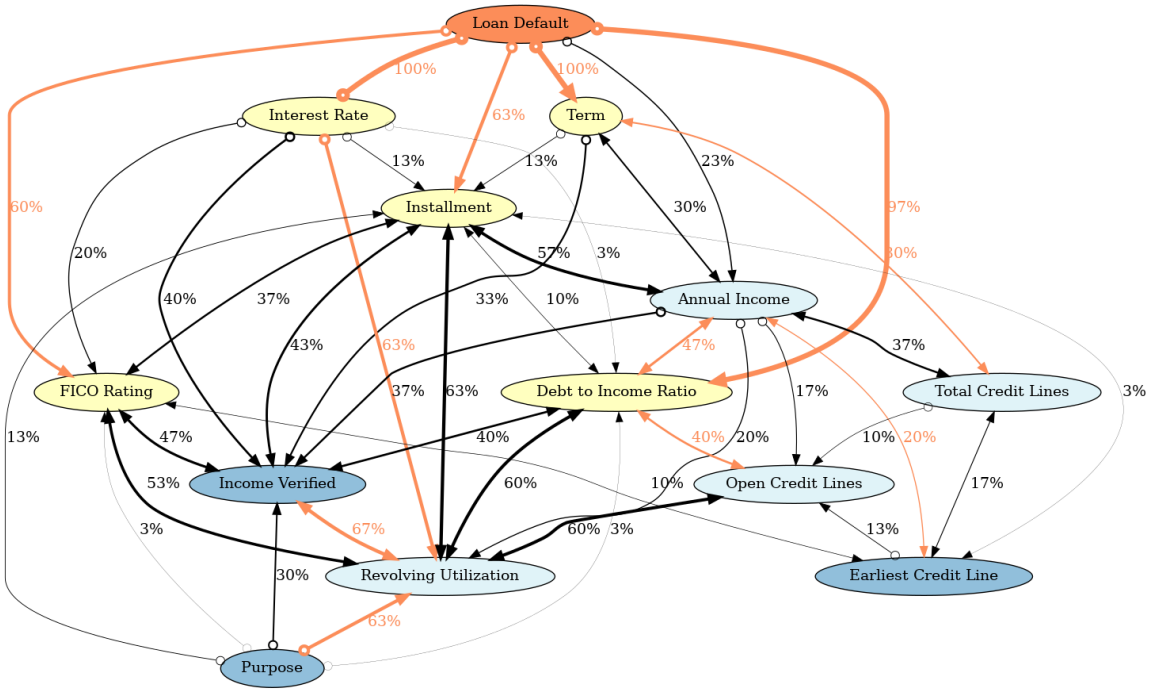
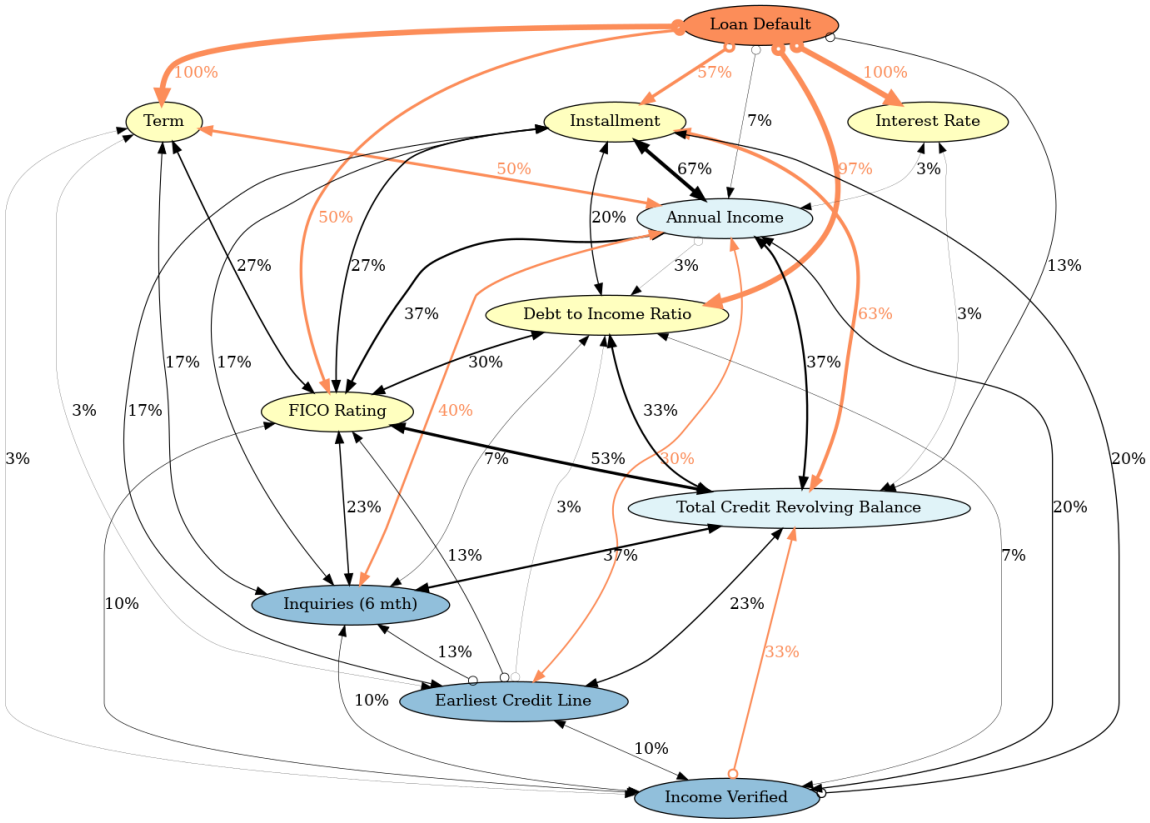


Figure 16 – Normalized Mutual Information between every pair of features in the Lending Club dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01. Notice that "Accounts Now Delinquent" and "Delinquent Amount" are not associated with any other feature. An explanation is that these features have a very low percentage of nonzero samples, making it unfeasible to detect a connection if any.



(a) M3B



(b) FMMB

Figure 17 – Discovered structure of Lending Club using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue.

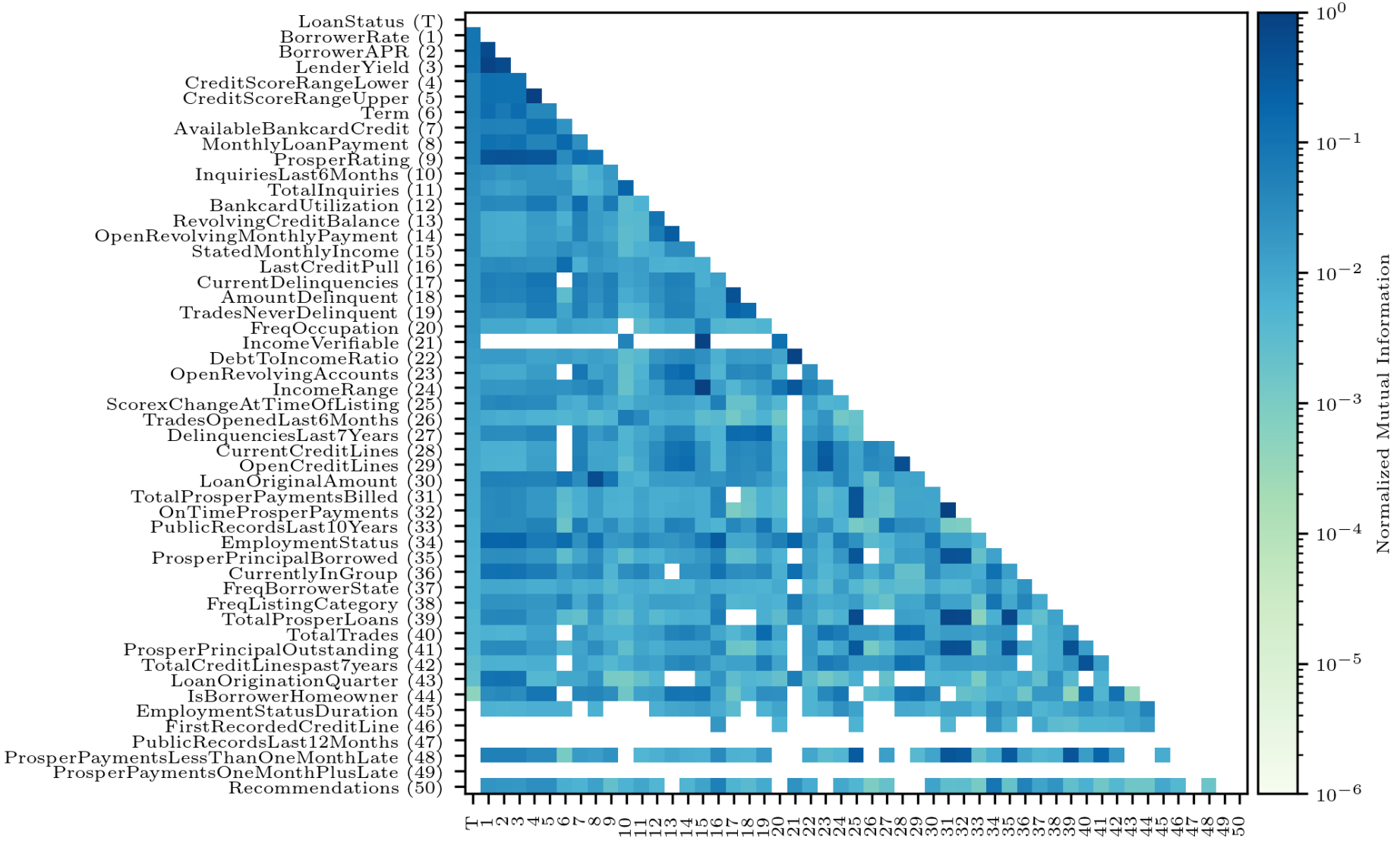
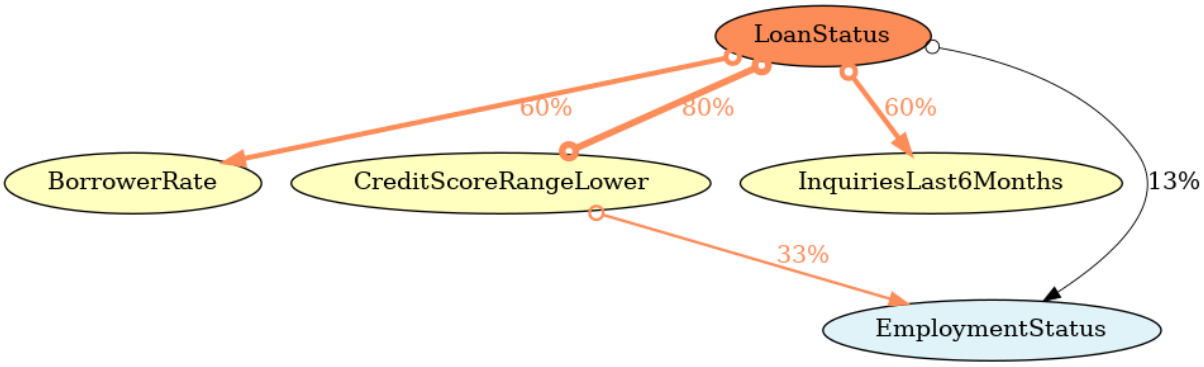
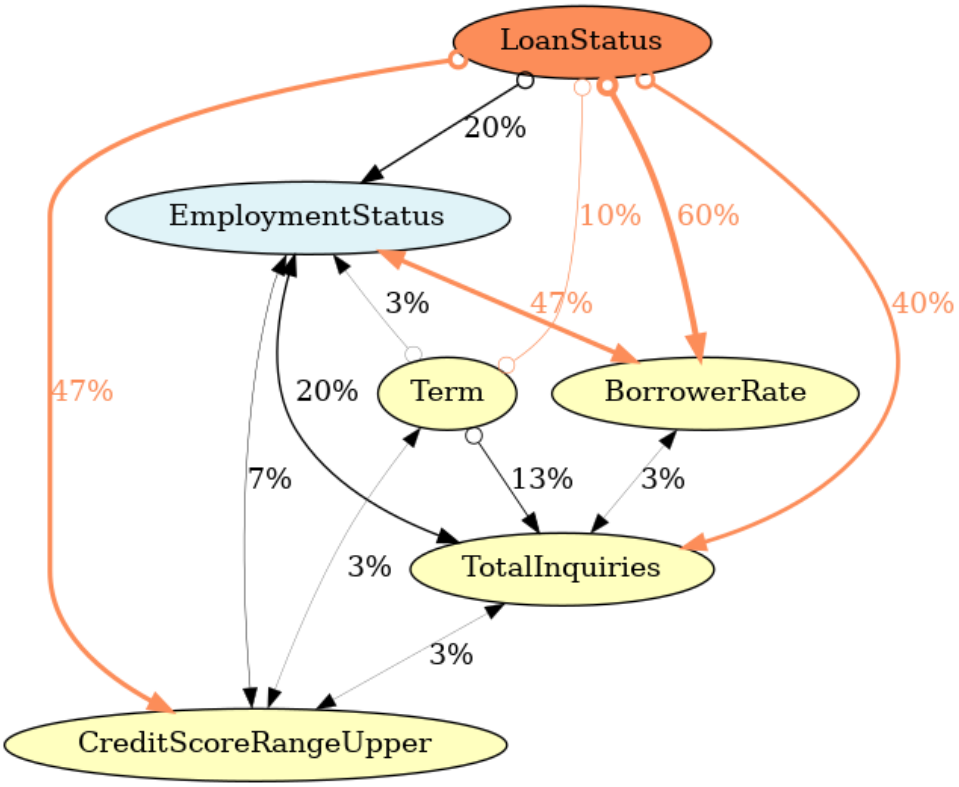


Figure 18 – Normalized Mutual Information between every pair of features in the Prosper dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01.



(a) M3B



(b) FMMB

Figure 19 – Discovered structure of Prosper using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue.

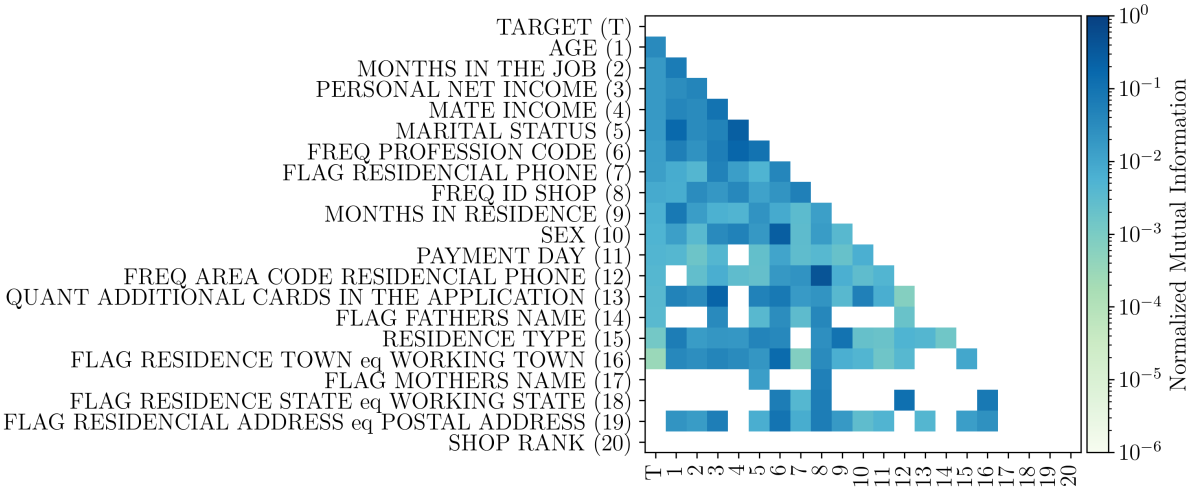
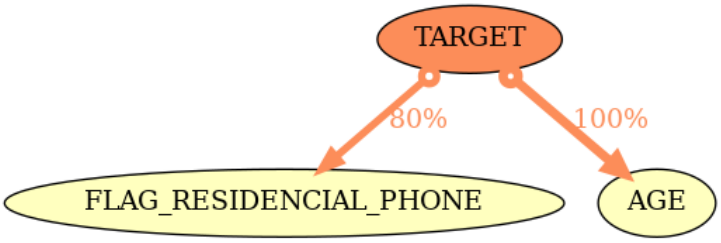
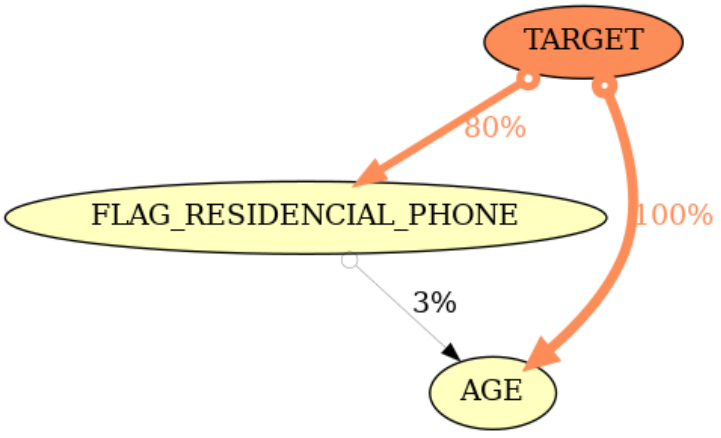


Figure 20 – Normalized Mutual Information between every pair of features in the PAKDD2009 dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01.



(a) M3B



(b) FMMB

Figure 21 – Discovered structure of PAKDD2009 using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue.

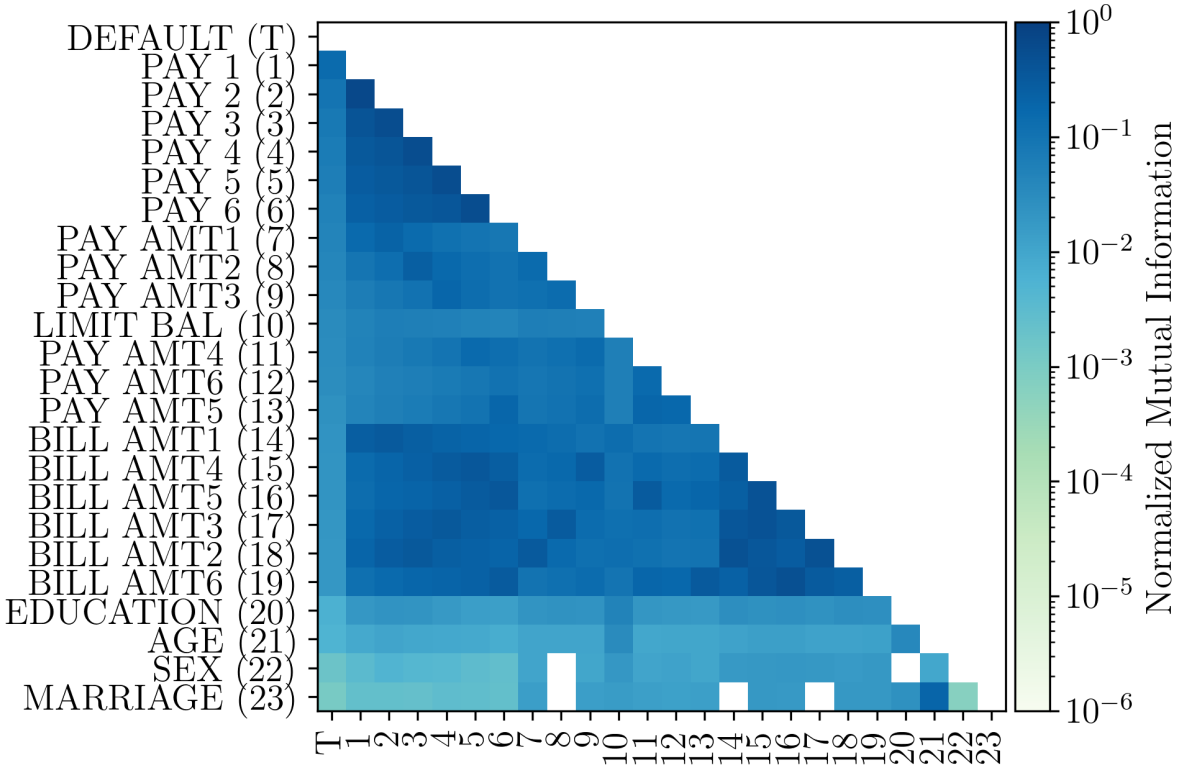
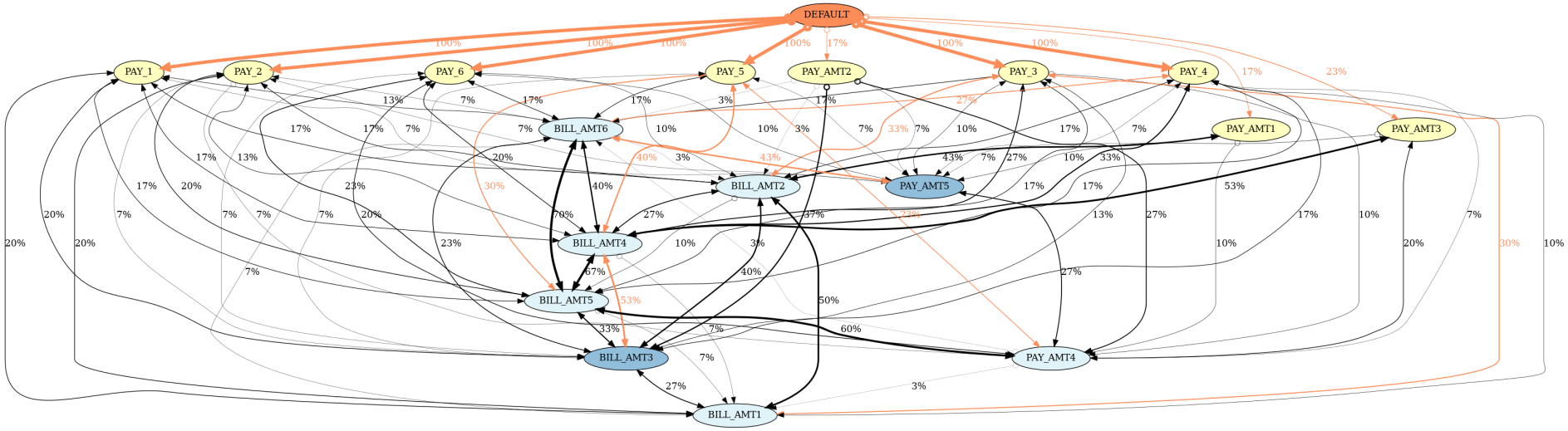
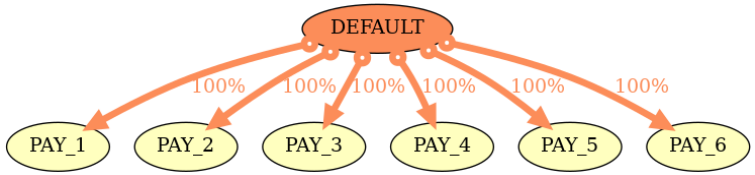


Figure 22 – Normalized Mutual Information between every pair of features in the Taiwan dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01.



(a) M3B



(b) FMMB

Figure 23 – Discovered structure of Taiwan using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue.

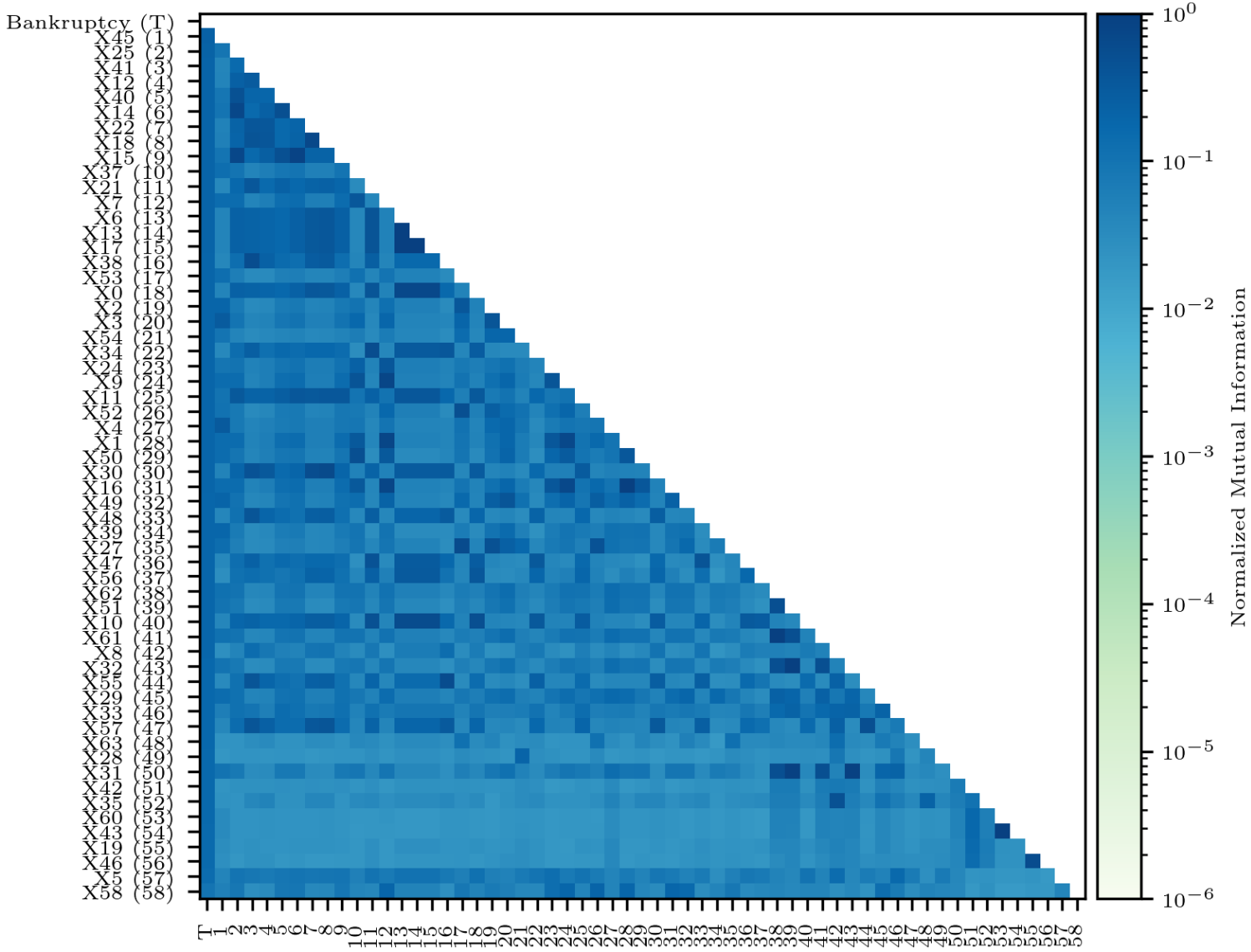
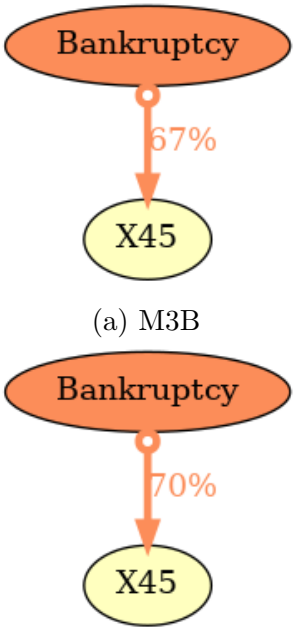


Figure 24 – Normalized Mutual Information between every pair of features in the Polish dataset. Features are ordered by decreasing connection strength with the target. Entries in white below the main diagonal indicate independence of the variables, given a G^2 test at a significance level of 0.01.



(a) M3B

(b) FMMB

Figure 25 – Discovered structure of Polish using aggregation by nodes and the MHD threshold. The percentages represent the frequency of occurrence of each edge in the bootstraps. The target, and the most probable path between each feature in the MB and the target, is highlighted in orange. Nodes adjacent to the target are colored in yellow. Nodes at depth 2 are colored in light blue, and nodes at depth 3 in dark blue.

4.3 Credit Scoring with Bayesian Network Classifier Ensembles

Tab. 8 shows the experimental results. In four of the five datasets, the proposed method achieved superior AUC and F1 scores, in spite of a worse precision than RF in three of the five. The high precision value and negligible recall of RF in *pakd2009* and *prosper-2009+* is an indication that RF was heavily biased towards the majority class, explaining it's higher precision over the BNC. The recall obtained in *pakdd2009* was very low for all models, and the AUC indicates that only the proposed method was capable of performing better than assuming every borrower to be good for this dataset. As most of the datasets are unbalanced, cost-sensitive training, that is, penalizing with heavier weight classification mistakes of the minority class, might be necessary to improve the performance of RF, whereas the proposed method displayed good performance even under class imbalance. The proposed method was the most effective in the *australian* dataset and least effective in the *south-german* dataset, however the better performance of LR indicates that most of the relationships present in this particular dataset are linear. To summarize, the proposed method proved to be a viable interpretable alternative to linear models, being robust to class imbalance and delivering good recall and AUC overall.

The 100 iterations were never reached, taking approximately 20 iterations for the

algorithm to converge. Since the core computations of the BNC were implemented in Python, whereas the others were optimized in C, the average conversion time was of about 1 hour, contrasted with a few minutes necessary to train the other classifiers. A fair comparison would necessitate the optimization of the implementation the proposed method. Another perspective is comparing the number of intensive computations required by each method. Training RF requires hundreds of interactions, an order of magnitude more than the proposed BNC. On the other hand, LR is very cheap to train, and achieved similar performance to RF. Which algorithm to choose, BNC or LR, is a trade-off between better classification performance and available computational resources and time.

Table 8 – Experimental results. Values in **bold** indicate a clear winner when considering each metric individually, whereas values in *italic* indicate a draw.

	pakdd2009			prosper-2009+			prosper-2009			south-german			australian		
	BNC	RF	LR	BNC	RF	LR	BNC	RF	LR	BNC	RF	LR	BNC	RF	LR
Precision	0.429	**0.778**	0.459	0.426	**0.635**	0.58	0.573	**0.635**	0.616	0.781	0.74	**0.787**	**0.818**	0.752	0.759
Recall	**0.119**	0.003	0.007	**0.422**	0.075	0.115	**0.509**	0.401	0.401	0.933	**0.962**	0.914	0.88	*0.924*	*0.924*
F1	**0.186**	0.006	0.014	**0.424**	0.134	0.192	**0.539**	0.491	0.486	**0.85**	0.836	0.846	**0.848**	0.829	0.833
AUC	**0.54**	0.501	0.503	**0.625**	0.531	0.545	**0.651**	0.637	0.632	0.661	0.587	**0.668**	**0.862**	0.84	0.845

5 CONCLUSION

We revisited the problem of MB discovery without assuming causal sufficiency by introducing the Fast MAG Markov Blanket (FMMB) algorithm, presented at CBA2020 (JERONYMO; MACIEL, 2020), with the aim to mitigate the curse of dimensionality by using a faster method of discovery. Its key idea is the strategy of direct non-adjacent member discovery, improving the state-of-the-art in terms of time efficiency by reducing the number of CITs needed to find the MB of a target. In structural learning, the experiments showed that the main drawback of FMMB is the requirement of a large enough sample size for getting accurate outputs. The strategy used by M3B is data efficient, yielding precise tests, but given a large enough sample size, FMMB's strategy leads to accurate outputs in significantly less time, however at the cost of consistency. In Feature Selection, FMMB proved to be a viable alternative to M3B, achieving comparable classification accuracy in ten real-world datasets, and again consuming much less execution time. In particular, the results achieved in the Lending Club dataset were to two orders of magnitude faster. These results inspire the application of MB feature selection in domains where it was previously unfeasible. We conclude that the equivalence of classification performance with lower runtime of FMMB, when compared to M3B, makes direct discovery preferable to indirect discovery.

We also explored the aggregation of bootstrapped MB structures, as it is a crucial step in learning a reliable structure. We showed that aggregating by nodes is better suited for discovering MB structures because it produces MBs with sizes closer to the average size of the bootstrapped MBs than aggregating by edges, consequently having a smaller loss of information. Altogether, the methods that were analysed are relatively simple. They do not demand much computation, but still fail to achieve median performance in some datasets. In particular, they straggled when applied to the P2P lending platforms' data. However, they share an important property, that is the justification of the chosen threshold. However, more elaborate, and computationally intensive methods, could be considered. For example, using Bayesian techniques.

We also contributed to the open problem of determining if a dataset is adequate for graphical modelling by using the discrepancy between the average size of the bootstrapped structures and the size of the aggregated structure as an indicator of unfaithfulness. A critical path for progress is exploring the relaxation of other strong theoretical assumptions, often violated in practice, such as the faithfulness assumption. Another important next step towards better MB discovery is to develop new methods for testing conditional independence that are both fast and resilient to the loss of statistical power when the size of the conditioning set increases.

In addition, we proved the necessity of handling latent factors by detecting features at depths > 2, justifying the development of algorithms that do not assume causal sufficiency. In the discussion, we highlighted the points of agreement and disagreement between the results generated by both algorithms. We belief that such detailed comparisons are fundamental for further development, yet lacking in the literature, so we encourage further research. This work has implications in local-to-global structural learning methods that rely on conditional independence based MB discovery algorithms. With faster structural learning capabilities, FMMB inspires the development of new local-to-global structural learning algorithms. While the proposed aggregation method challenges the literature, where aggregation by edges predominates.

In credit scoring, we demonstrated the usefulness of graphical models and MB discovery in studying the causes that lead to default, since many of the findings agree with the literature, and others challenge it, such as the diminished importance of some features, such as the purpose of the loan, in relation to the results from linear modelling. We concluded that the models used at the P2P lending platforms analysed are effective in grading bad borrowers because the interest rate feature, provided by the platforms, is the most informative. However, they still have room for improvement, given that many other features appear in the MB.

Furthermore, we introduced a novel Monte-Carlo based method for BNC learning using CEM and Bootstrap Aggregation, as an interpretable alternative for RFs and LR. From the results, we conclude that the proposed algorithm is promising when applied to unbalanced datasets, and as a model for credit scoring and prediction of default. The main drawback of the final model is its training time. While other classifiers take at most few minutes to train, the proposed classifier can take hours depending on the hyper-parameters chosen. A fair point is that the other classifiers were implemented efficiently in C, and merely wrapped in Python, whereas the implementation of the proposed method is purely in Python, and can be hugely optimized. The results bring interesting ideas of structural learning using importance sampling in the space of DAGs. A natural extension of the method is adapting it to learn general purpose BNs, where the focus is joint probability estimation. The performance function can be easily switched to a traditional score for BNs such as BDeu. Another possibility is merging the learned ensemble into a single, more robust, structure with the goal of gaining structural knowledge. We hope to inspire further research in the applications of CEM or related probabilistic methods of structural learning. Future steps are incorporating cost-sensitive learning and decision making. For further assessment, we intend to apply this algorithm to more datasets such as Lending Club and Taiwan, and compare it to more classifiers, and other meta-heuristically trained models, such as Genetic Algorithms, as well.

In summary, we fulfilled the main objective of developing a faster method of MB

discovery, motivated by the limitations imposed by the super-exponential search space, and the need of practical results in a viable amount of time. We leave with open theoretical and practical questions. This super-exponential nature still currently limits the applicability of current MB discovery methods to the order of < 100 variables. Parallel processing could be applied to some extent, but the dependencies in the depth first search based algorithms limit the achievable parallelism. How to mitigate these dependencies, and ideally create embarrassingly parallel algorithms? Domains where variables are highly correlated are also problematic, as we saw in the Polish data. A possibility is using dimensionality reduction techniques such as PCA. However, the warping of the variables caused by the transformations of the data in euclidean space could jeopardize interpretability. Another possibility is using auto-encoders, yet, they require very large datasets to train properly, and are computationally expensive. A suitable method should be computationally fast, cheap to train, and preserve interpretability. How can MB discovery algorithms handle highly correlated data better? Frequentist modelling is at the essence of the MB discovery methods. Are there improvements to be made by replacing the CITs with Bayesian alternatives? Can we extend these MB discovery methods for time series data? What results and insights can be generated by applying these methods in other domains?

REFERENCES

AGRESTI, Alan. **Categorical data analysis**. [S.l.]: John Wiley & Sons, 2003. v. 482.

ALCALÁ-FDEZ, J. et al. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. **Journal of Multiple-Valued Logic and Soft Computing**, 17: 2-3, p. 255–287, 2011. Available from: <https://sci2s.ugr.es/keel/index.php>.

ALIFERIS, Constantin F. et al. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. **Journal of Machine Learning Research**, v. 11, Jan, p. 171–234, 2010.

_____. Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions. **Journal of Machine Learning Research**, v. 11, Jan, p. 235–284, 2010.

ANDERSON, Billie. Using Bayesian networks to perform reject inference. **Expert Systems with Applications**, Elsevier, v. 137, p. 349–356, 2019.

BAESENS, Bart; ROESCH, Daniel; SCHEULE, Harald. **Credit risk analytics: Measurement techniques, applications, and examples in SAS**. [S.l.]: John Wiley & Sons, 2016.

BEINLICH, Ingo A et al. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In: AIME 89. [S.l.]: Springer, 1989. P. 247–256.

BELGHAZI, Mohamed Ishmael et al. Mutual information neural estimation. In: PMLR. INTERNATIONAL Conference on Machine Learning. [S.l.: s.n.], 2018. P. 531–540.

BIELZA, Concha; LARRAÑAGA, Pedro. Discrete Bayesian network classifiers: a survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 47, n. 1, p. 1–43, 2014.

BULTEAU, Laurent; SCHMID, Markus L. Consensus strings with small maximum distance and small distance sum. **Algorithmica**, Springer, v. 82, n. 5, p. 1378–1409, 2020.

CAMPOS, Luis M. de. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. **Journal of Machine Learning Research**, v. 7, Oct, p. 2149–2187, 2006.

CHANG, KC et al. Bayesian networks applied to credit scoring. **IMA Journal of Management Mathematics**, Oxford University Press, v. 11, n. 1, p. 1–18, 2000.

CHEN, Jiehua; HERMELIN, Danny; SORGE, Manuel. On Computing Centroids According to the p-Norms of Hamming Distance Vectors. **arXiv preprint arXiv:1807.06469**, 2018.

CHEN, Jingnian; HUANG, Houkuan, et al. Feature selection for text classification with Naïve Bayes. **Expert Systems with Applications**, Elsevier, v. 36, n. 3, p. 5432–5435, 2009.

CHEN, Serena H; POLLINO, Carmel A. Good practice in Bayesian network modelling. **Environmental Modelling & Software**, Elsevier, v. 37, p. 134–145, 2012.

CLUB, Lending. **Rate Information**. [S.l.: s.n.], 2021. Available from: <https://www.lendingclub.com/foliofn/rateDetail.action>.

COLOMBO, Diego et al. Learning high-dimensional directed acyclic graphs with latent and selection variables. **The Annals of Statistics**, JSTOR, p. 294–321, 2012.

COOPER, Gregory F; HERSKOVITS, Edward. A Bayesian method for the induction of probabilistic networks from data. **Machine learning**, Springer, v. 9, n. 4, p. 309–347, 1992.

CORMEN, Thomas H et al. **Introduction to algorithms**. [S.l.]: MIT press, 2009.

COVER, Thomas M.; THOMAS, Joy A. **Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)**. [S.l.]: Wiley-Interscience, 2006.

CROUX, Christophe et al. Important factors determining Fintech loan default: Evidence from a lendingclub consumer platform. **Journal of Economic Behavior & Organization**, v. 173, p. 270–296, 2020. ISSN 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2020.03.016.

DAWID, A Philip. The well-calibrated Bayesian. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 77, n. 379, p. 605–610, 1982.

DE BOER, Pieter-Tjerk et al. A tutorial on the cross-entropy method. **Annals of operations research**, Springer, v. 134, n. 1, p. 19–67, 2005.

DOMINGOS, Pedro; PAZZANI, Michael. On the optimality of the simple Bayesian classifier under zero-one loss. **Machine learning**, Springer, v. 29, n. 2-3, p. 103–130, 1997.

DUA, Dheeru; GRAFF, Casey. **UCI Machine Learning Repository**. [S.l.: s.n.], 2017. Available from: <http://archive.ics.uci.edu/ml>.

EXPERIAN. **Scorecards**. [S.l.: s.n.], 2021. Available from: <https://www.experian.nl/en/business/analytics-and-decisioning/decision-analytics/scorecards>.

EYHERAMENDY, Susana; LEWIS, David D; MADIGAN, David. On the naive bayes model for text categorization. Citeseer, 2003.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FRIEDMAN, Nir; GEIGER, Dan; GOLDSZMIDT, Moises. Bayesian network classifiers. **Machine learning**, Springer, v. 29, n. 2-3, p. 131–163, 1997.

FRIEDMAN, Nir; GOLDSZMIDT, Moises, et al. Discretizing continuous attributes while learning Bayesian networks. In: ICML. [S.l.: s.n.], 1996. P. 157–165.

FRIEDMAN, Nir; GOLDSZMIDT, Moises; WYNER, Abraham. Data analysis with Bayesian networks: A bootstrap approach. **arXiv preprint arXiv:1301.6695**, 2013.

FU, Shunkai; DESMARAIS, Michel C. Fast Markov blanket discovery algorithm via local learning within single pass. In: SPRINGER. CONFERENCE of the Canadian Society for Computational Studies of Intelligence. [S.l.: s.n.], 2008. P. 96–107.

GASIENIEC, Leszek; JANSSON, Jesper; LINGAS, Andrzej. Approximation algorithms for Hamming clustering problems. **Journal of Discrete Algorithms**, Elsevier, v. 2, n. 2, p. 289–301, 2004.

GLYMOUR, Clark; ZHANG, Kun; SPIRTES, Peter. Review of Causal Discovery Methods Based on Graphical Models. **Frontiers in Genetics**, v. 10, p. 524, 2019. ISSN 1664-8021. DOI: 10.3389/fgene.2019.00524.

GROSS, Tadeu Junior et al. An analytical threshold for combining Bayesian Networks. **Knowledge-Based Systems**, Elsevier, v. 175, p. 36–49, 2019.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer Science & Business Media, 2009.

HOU, Wen-hui et al. A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. **Knowledge-Based Systems**, Elsevier, v. 208, p. 106462, 2020.

JERONYMO, Pedro VB; MACIEL, Carlos D. Fast Markov Blanket Discovery Without Causal Sufficiency. In: 1. CONGRESSO Brasileiro de Automática - CBA. [S.l.: s.n.], 2020. v. 2.

KE, Guolin et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, p. 3146–3154, 2017.

KLEITER, Gernot D. The posterior probability of Bayes nets with strong dependencies. **Soft Computing**, Springer, v. 3, n. 3, p. 162–173, 1999.

KOLLER, Daphne; FRIEDMAN, Nir. **Probabilistic graphical models: principles and techniques**. [S.l.]: MIT press, 2009.

KOLLER, Daphne; SAHAMI, Mehran. **Toward optimal feature selection**. [S.l.], 1996.

KULLBACK, Solomon. **Information theory and statistics**. [S.l.]: Courier Corporation, 1997. P. 155–176.

LEE, Jong-Seok. AUC4.5: AUC-Based C4.5 Decision Tree Algorithm for Imbalanced Data Classification. **IEEE Access**, v. 7, p. 106034–106042, 2019. DOI: 10.1109/ACCESS.2019.2931865.

LEONG, Chee Kian. Credit risk scoring with bayesian network models. **Computational Economics**, Springer, v. 47, n. 3, p. 423–446, 2016.

LESSMANN, Stefan et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. **European Journal of Operational Research**, v. 247, 1 2015. ISSN 03772217. DOI: 10.1016/j.ejor.2015.05.030.

LI, Zhiyong et al. Reject inference in credit scoring using Semi-supervised Support Vector Machines. **Expert Systems with Applications**, v. 74, 2017. ISSN 09574174. DOI: 10.1016/j.eswa.2017.01.011.

LIPTON, Zachary C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. **Queue**, ACM New York, NY, USA, v. 16, n. 3, p. 31–57, 2018.

MADDEN, Michael G. On the classification performance of TAN and general Bayesian networks. In: SPRINGER. INTERNATIONAL Conference on Innovative Techniques and Applications of Artificial Intelligence. [S.l.: s.n.], 2008. P. 3–16.

MANCISIDOR, Rogelio A. et al. Deep generative models for reject inference in credit scoring. **Knowledge-Based Systems**, v. 196, 2020. ISSN 09507051. DOI: 10.1016/j.knosys.2020.105758.

MARGARITIS, Dimitris; THRUN, Sebastian. Bayesian network induction via local neighborhoods. In: ADVANCES in neural information processing systems. [S.l.: s.n.], 2000. P. 505–511.

MARON, Melvin Earl; KUHNS, John Larry. On relevance, probabilistic indexing and information retrieval. **Journal of the ACM (JACM)**, ACM New York, NY, USA, v. 7, n. 3, p. 216–244, 1960.

MASMOUDI, Khalil; ABID, Lobna; MASMOUDI, Afif. Credit risk modeling using Bayesian network with a latent variable. **Expert Systems with Applications**, v. 127, 2019. ISSN 09574174. DOI: 10.1016/j.eswa.2019.03.014.

MCDONALD, John H. **Handbook of biological statistics**. [S.l.]: sparky house publishing Baltimore, MD, 2009. v. 2.

MCNALLY, Richard J; HEEREN, Alexandre; ROBINAUGH, Donald J. A Bayesian network analysis of posttraumatic stress disorder symptoms in adults reporting childhood sexual abuse. **European journal of psychotraumatology**, Taylor & Francis, v. 8, sup3, p. 1341276, 2017.

MEGANCK, Stijn; LERAY, Philippe; MANDERICK, Bernard. Causal graphical models with latent variables: Learning and inference. In: SPRINGER. EUROPEAN Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. [S.l.: s.n.], 2007. P. 5–16.

METSIS, Vangelis; ANDROUTSOPOULOS, Ion; PALIOURAS, Georgios. Spam filtering with naive bayes-which naive bayes? In: MOUNTAIN VIEW, CA. CEAS. [S.l.: s.n.], 2006. v. 17, p. 28–69.

MOSCATO, Vincenzo; PICARIELLO, Antonio; SPERLÍ, Giancarlo. A benchmark of machine learning approaches for credit score prediction. **Expert Systems with Applications**, v. 165, 2021. ISSN 09574174. DOI: 10.1016/j.eswa.2020.113986.

NICULESCU-MIZIL, Alexandru; CARUANA, Rich. Predicting good probabilities with supervised learning. In: PROCEEDINGS of the 22nd international conference on Machine learning. [S.l.: s.n.], 2005. P. 625–632.

PEARL, J. **Probabilistic reasoning in intelligent systems: networks of plausible inference**. [S.l.]: San Francisco: Morgan Kaufmann, 1988.

PENA, Jose M et al. Towards scalable and data efficient learning of Markov boundaries. **International Journal of Approximate Reasoning**, Elsevier, v. 45, n. 2, p. 211–232, 2007.

PLATT, John et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. **Advances in large margin classifiers**, Cambridge, MA, v. 10, n. 3, p. 61–74, 1999.

POLENA, Michal; REGNER, Tobias. Determinants of borrowers' default in P2P lending under consideration of the loan risk class. **Games**, Multidisciplinary Digital Publishing Institute, v. 9, n. 4, p. 82, 2018.

PROAKIS, John G; MANOLAKIS, Dimitris G. Digital signal processing. **PHI Publication: New Delhi, India**, 2004.

RICHARDSON, Thomas; SPIRTES, Peter, et al. Ancestral graph Markov models. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 30, n. 4, p. 962–1030, 2002.

RÍO, Sergio del; VILLANUEVA, Edwin. A Novel Method to Estimate Parents and Children for Local Bayesian Network Learning. In: SPRINGER. PROCEEDINGS of SAI Intelligent Systems Conference. [S.l.: s.n.], 2021. P. 468–485.

RISH, Irina et al. An empirical study of the naive Bayes classifier. In: 22. IJCAI 2001 workshop on empirical methods in artificial intelligence. [S.l.: s.n.], 2001. v. 3, p. 41–46.

RODGERS, Rachel F et al. Structural differences in eating disorder psychopathology after history of childhood abuse: Insights from a Bayesian network analysis. **Journal of abnormal psychology**, American Psychological Association, v. 128, n. 8, p. 795, 2019.

RUBINSTEIN, Reuven Y; KROESE, Dirk P. **The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning**. [S.l.]: Springer Science & Business Media, 2013.

SADEGHI, Kayvan. Faithfulness of probability distributions and graphs. **Journal of Machine Learning Research**, Microtome Publishing, v. 18, n. 148, p. 1–29, 2017.

SCIKIT-OPTIMIZE. **scikit-optimize: Sequential model-based optimization in Python - scikit-optimize 0.7.3 documentation**. [S.l.: s.n.], 2021. Available from: <https://scikit-optimize.github.io/>.

SCUTARI, Marco; NAGARAJAN, Radhakrishnan. Identifying significant edges in graphical models of molecular networks. **Artificial intelligence in medicine**, Elsevier, v. 57, n. 3, p. 207–217, 2013.

SERRANO-CINCA, Carlos; GUTIÉRREZ-NIETO, Begoña; LÓPEZ-PALACIOS, Luz. Determinants of default in P2P lending. **PloS one**, Public Library of Science San Francisco, CA USA, v. 10, n. 10, e0139427, 2015.

SHEN, Xinpeng et al. Challenges and opportunities with causal discovery algorithms: application to Alzheimer's pathophysiology. **Scientific reports**, Nature Publishing Group, v. 10, n. 1, p. 1–12, 2020.

SPIRTES, Peter et al. **Causation, prediction, and search**. [S.l.]: MIT press, 2000.

TSAMARDINOS, Ioannis; ALIFERIS, Constantin F; STATNIKOV, Alexander. Time and sample efficient discovery of Markov blankets and direct causal relations. In: PROCEEDINGS of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.: s.n.], 2003. P. 673–678.

TSAMARDINOS, Ioannis; ALIFERIS, Constantin F. Towards principled feature selection: relevancy, filters and wrappers. In: AISTATS. [S.l.: s.n.], 2003.

TSAMARDINOS, Ioannis; ALIFERIS, Constantin F.; STATNIKOV, Alexander R; STATNIKOV, Er. Algorithms for large scale Markov blanket discovery. In: FLAIRS conference. [S.l.: s.n.], 2003. v. 2, p. 376–380.

VERGARA, Jorge R; ESTÉVEZ, Pablo A. A review of feature selection methods based on mutual information. **Neural computing and applications**, Springer, v. 24, n. 1, p. 175–186, 2014.

WANG, Hao; LING, Zhaolong, et al. Towards efficient and effective discovery of Markov blankets for feature selection. **Information Sciences**, Elsevier, v. 509, p. 227–242, 2020.

_____._____. **Information Sciences**, v. 509, p. 227–242, 2020. ISSN 0020-0255. DOI: https://doi.org/10.1016/j.ins.2019.09.010.

WANG, Ru; PENG, Jie. Learning directed acyclic graphs via bootstrap aggregating. **arXiv preprint arXiv:1406.2098**, 2014.

WEINBERGER, Naftali. Faithfulness, coordination and causal coincidences. **Erkenntnis**, Springer, v. 83, n. 2, p. 113–133, 2018.

XIAO, Jin et al. Cost-sensitive semi-supervised selective ensemble model for customer credit scoring. **Knowledge-Based Systems**, Elsevier, v. 189, p. 105118, 2020.

YANG, Xianglin et al. Three-Fast-Inter Incremental Association Markov Blanket learning algorithm. **Pattern Recognition Letters**, Elsevier, v. 122, p. 73–78, 2019.

YARAMAKALA, Sandeep; MARGARITIS, Dimitris. Speculative Markov blanket discovery for optimal feature selection. In: IEEE. FIFTH IEEE International Conference on Data Mining (ICDM'05). [S.l.: s.n.], 2005. 4–pp.

YEN, Lorna. P2P Lending Platform Data Analysis: Exploratory Data Analysis in R - Part 1. **Medium**, Towards Data Science, Jan. 2019.

YU, Kui; GUO, Xianjie, et al. Causality-based feature selection: Methods and evaluations. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 53, n. 5, p. 1–36, 2020.

YU, Kui; LIU, Lin, et al. Mining markov blankets without causal sufficiency. **IEEE transactions on neural networks and learning systems**, IEEE, v. 29, n. 12, p. 6333–6347, 2018.

YU, Kui; WU, Xindong, et al. Markov blanket feature selection using representative sets. **IEEE transactions on neural networks and learning systems**, IEEE, v. 28, n. 11, p. 2775–2788, 2016.

ZADROZNY, Bianca; ELKAN, Charles. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: IN Proceedings of the Eighteenth International Conference on Machine Learning. [S.l.]: Morgan Kaufmann, 2001. P. 609–616.

ZHANG, Harry. Exploring conditions for the optimality of naive Bayes. **International Journal of Pattern Recognition and Artificial Intelligence**, World Scientific, v. 19, n. 02, p. 183–198, 2005.

ZHANG, Tao et al. Multiple instance learning for credit risk assessment with transaction data. **Knowledge-Based Systems**, v. 161, 2018. ISSN 09507051. DOI: 10.1016/j.knosys.2018.07.030.

ZHANG, Wenyu et al. A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring. **Expert Systems with Applications**, v. 165, 2021. ISSN 09574174. DOI: 10.1016/j.eswa.2020.113872.