Bayesian Networks for the modeling and assessment of credit concentration risks

Tatjana Pavlenko* Oleksandr Chernyak† Annika Tillander‡

Abstract The main goal of this research is to demonstrate how probabilistic graphs may be used for modeling and assessment of credit concentration risk. The destructive power of credit concentrations essentially depends on the amount of correlation among borrowers. However, borrower companies correlation and concentration of credit risk exposures have been difficult for the banking industry to measure in an objective way as they are riddled with uncertainty. As a result, banks do not manage to make a quantitative link to the correlation driving risks and fail to prevent concentrations from accumulating. In this paper, we argue that Bayesian networks provide an attractive solution to the problems identified above and we show how to apply them in representing, quantifying and managing the uncertain knowledge in concentration of credits risk exposures. We first suggest the stepwise Bayesian network model building scheme and show how to incorporate expert-based prior beliefs regarding the risk exposure of a group of related borrowers and then update these beliefs through the whole model with the new information as it is learned. We then explore a specific graph structure, a tree-augmented Bayesian network, and show that this model provides better understanding of the concentration risk accumulating due to strong direct or indirect business links between borrowers. We also present two strategies of model assessment which exploit the measure of mutual information, and show that the constructed Bayesian network is a reliable model that can be implemented to identify and control threat from concentration of credit exposures. Finally, we demonstrate that suggested tree-augmented Bayesian network is also suitable for stress testing analysis, in particular, it can provide the estimates of the posterior risk of losses related to the unfavorable changes in the financial conditions of a group of related borrowers.

Keywords: Bayesian network, uncertainty, tree-augmented graph, mutual information, expert-based beliefs, related borrowers, posterior credit risk

1 Introduction

Concentration risks, especially concentrations in credit risk, have played a key role in the financial instability of the banking sector during the last years, see reports by Basel Committee on Banking Supervision (2006), Committee on the Global Financial System (2007), Bonti at al. (2006) and Das et al. (2007). In order to assess concentrations in credit risks, it is important to first accurately represent credit risk across key factors of exposure. The focus of our research is the risk of increased exposure to losses due to correlation among borrowers. Qualitative studies of the destructive power of this correlation are widely presented in credit risk assessment literature, see e.g. see York (2007), Kalapodas and Thomson (2006). However, until recently, borrower correlation and concentration risks have been difficult for the banking industry to measure in an objective way.

The traditional approaches for credit risk assessment apply credit value-at-risk (CVaR) and internal rating-based (IRB) risk-weight models. Both approaches are based on two key assumptions: a) no single exposure accounts for more than a vanishingly small share of the overall risk, and b) credit

*Department of Statistics, Stockholm University, Sweden
†Department Economic Cybernetics, Taras Shevchenko National University of Kyiv, Ukraine
‡Department of Statistics, Stockholm University, Sweden

assessments do not allow for a rich correlation structure between individual risks, which means that the corresponding credit model is additive. When these two assumptions hold, it is possible to show that the risk assessment of the entire portfolio can be conducted from the bottom up; see more details in e.g. Gordy and Lütkebohmert (2007). Statistical reasons behind the opting bottom-up models are relative simplicity of the fitting model parameters which is due to the additive model structure.

However when the two assumptions behind the CVaR and IRB are violated, there is no guarantee that the bottom-up methodology will be accurate. These assumptions are in particular unlikely to be exactly met by actual credit portfolios for those financial institutes that are smaller in size or relatively specialised. For such institutes, concentration risk can arise from a significant single exposure, from concentration on the specific business area, and from loss dependencies due to strong direct or indirect business links between borrowers. In such cases the marginal input to the overall risk by any single borrower will likely depend on the risk profile of the other related borrowers and can not be captured by an additive model. As a result, financial institutes do not manage to make a quantitative link to the correlation driving risks and fail to prevent concentrations from accumulating. Often the effect of correlations only become apparent when economic conditions turn sour, which results in "storing-up" losses and create a huge peak later on the trough of the economic circle, see Das et al. (2007), York (2007). Hence, the amount of correlation among the borrower counterparts seems to play an important part in the credit risk assessment. New modeling methods that allow to capture and control the concentration risk arising from distinct but strongly correlated exposures have therefore attracted a great deal of attention. Besides the goal of better understanding of concentration of credit risks, these models can support rational decision making in the tricky area of risk management.

The main challenge with the credit risk modeling and assessment is that they are riddled with uncertainty. Estimation of the probability of default (insolvency), modeling correlation structure for a group of connected borrowers and estimation of amount of correlation are the most important sources of uncertainty that can severely impair the quality of credit risk models.

In this research, we explore the Bayesian network (BN) methodology, a very promising modeling technique that provides a framework for representing, quantifying and managing the uncertain knowledge in concentration of credits risk exposures. At a general level, a BN model consists of two distinct, related components: one being a directed acyclic graph with *nodes* representing the random feature variables (risk related factors) and *edges* representing direct relationships (dependencies) among features (about which we may be uncertain); the other representing quantitative information about the strengths of the relationships, usually expressed in terms of a joint probability distribution. In conjunction, the graphical and probabilistic components of the model represent a unique multivariate probability distribution over the complete set of feature variables, thereby capturing and visually representing the uncertain statements involved. In addition, BN modeling technique provides the algorithms for the updating of probabilistic uncertainty in response to evidence and allows for rational decision making under uncertainty. BN models were extensively applied as classification and prediction tool in different domains; see e.g. Anderson et al. (2004), Cowell et al (2007), Sun and Shenoy (2007), Hosack et al. (2008), Neil et al. (2008).

The contribution of our paper is twofold. First, is advocates a BN modeling approach to the study of credit risk concentration in which a specific role for related parties is required. With this approach we can incorporate expert-based initial beliefs regarding the risk exposure of a group of related borrowers and then update these beliefs through the whole BN model with new information as it is learned. Second, it explores a specific graph structure, a tree augmented BN, which allows for better understanding of the concentration risks accumulating due to interaction between the borrowers and analysing the formal properties of the posterior credit risk in controlling and desicion-making processes. Thus, using the BN modeling approach we can understand the credit risk implications of excessive lending to a group of related borrowers in a concrete setting. Moreover, we use the BN graph representation to visual exploratory analyses of the effect of storing up risk of losses and suggest a method of concentration risk stress-testing by carrying out what-if scenarios.

The rest of this paper is set out as follows. In Section 2, we review the basic framework of BN graphical modeling and formalize our notion of prior and posterior credit concentration risk. Section 3 formalizes the concept of related parties and explores the stepwise BN model building using a concrete example of a private medium-sized bank in Ukraine and a number of credit risk related factors. In this section, we also examine two strategies for model assessment, both of which are based on the concept of mutual information. The first one adopts the threshold-based approach, where the reliability of an edge in the graph is estimated using the average mutual information with averaging over all possible graph edges. The second one is some what more complex and uses the approximate bootstrap-based distribution of the average mutual information. We show how this distribution in turn allows for specifying the edges of specified significance level for a BN graph. Further, the use of constructed BN model for concentration risk stress tests is demonstrated by Bayesian updating of the posterior risk. We conclude in Section 4.

2 BN probabilistic graphs for modeling credit concentration risk

We start by reviewing the basic setting of BN graph modeling as a framework for capturing and representing uncertainty. For more thorough treatment of the topic see e.g ... Associated with a bank or a financial institute is a set of attributes, $X_1, \dots, X_d$ which in current study will be treated as discrete random variables representing highly informative credit risk accounting factors. In what follows we will use them to modeling relationships among the bank and corporate borrowers. We will use the notations $P_X(x) = P(X = x)$ and $P_{X|Y}(x|y) = P(X = x|Y = y)$ to describe the marginal and conditional probability distributions of the random variables X and $X|Y$, respectively. We will skip the index X when it does not make any confusion.

2.1 The formalism

A Bayesian network (BN) is an annotated directed acyclic graph that encodes a joint probability distribution over the set of random variables $\mathcal{X} = \{X_1 \dots, X_d\}$. Formally, a BN for $\mathcal{X}$ is a pair $\langle \mathcal{G}, P \rangle$. The first component, $\mathcal{G}$, is a directed acyclic graph whose nodes correspond to the random variables $X_1, \dots, X_d$, and whose edges encode the direct probabilistic influences between the variables. In this way, the graph $\mathcal{G}$ represents the independence assumption: each variable X_i is conditionally independent of its non-descendants given its parent nodes $\Pi_{[i]}$ in $\mathcal{G}$ for all $i = 1, \dots, d$. The second component of the pair, namely $P = \{P(x_1|\Pi_{[1]}), \dots, P(x_d|\Pi_{[d]})\}$ represents the set of d conditional probability distributions given the set of parent nodes $\Pi_{[i]}$ for each X_i, $i = 1, \dots, d$. For each node we assign a binary random variable X_i.

Informally speaking, a directed edge $X_i \to X_j$ between two nodes X_i and X_j in the BN is a statement (perhaps a belief) telling that there is an influence between the associated random variables; absence of an edge between X_i and X_j means that the corresponding random variables do not influencing each other directly. More formally, the graph structure can be given by the set of parents $\Pi = (\Pi_{[1]}, \dots, \Pi_{[d]})$ the associated set P defines the unique multivariate probability distribution of $X_1, \dots, X_d$ that factorizes over the graph structure as

$$P(x_1, \dots, x_d) = \Pi_{i=1}^{d} P(x_i|\Pi_{[i]}).$$

In general, the process of induction of a BN model can be divided into two main steps: *structural learning* and *parametric learning*. Structural learning usually involves a search procedure guided by a pre-specified score function which is defined over the space of all possible graph structures. The search procedures aims to optimize the score and finishes when the local optimum is found. Parametric learning consists of estimating parameters from the data. These parameters quantify the (in)dependence relations between the variables, represented by the BN graph.

As the graph structure increases in the size and complexity, the run-time complexity of probabilistic inference with BN becomes prohibitive. Complexity orders in terms of computational time and in terms of number of parameters estimated are extensively studied in statistical literature, see e.g.

Pérez et al. (2006), Ellis and Wong (2008) and references there in. These results clearly indicate the need to constrain the space of possible graph structures by using the domain knowledge about the context specific relationships among the feature variables.

In this study, we limit our attention to a class of the BN models whose graph $\mathcal{G}$ has a *tree* structure. This means that in the BN induced over the set of random variables $\mathcal{X} = \{Y, X_1, \ldots, X_d\}$, the nodes have different status. The variable Y in the tree-structured graph denotes the node which has no parents, that is $\Pi_{[Y]} = \{\emptyset\}$ and is referred to as a root of the tree. The remaining nodes, $X_i \in \mathcal{G}$ have as potential parents either the root variable Y only, or both Y and at least one other variable X_j, $i \neq j$.

Depending on the complexity of the underlying relationships between the domain variables, the BN tree-structure graph can be classified as a *naive* (N-BN), a *tree-augmented* (TAN), *k-dependence* (k-BN), etc. The N-BN represents the least complex graph structure requiring that the root variable Y be a parent of each variable X_i, i.e. $\Pi_{[i]} = \{Y\}$ for all $i = 1, \ldots, d$. The corresponding probabilistic relationship between the variables infers that all X_is are conditionally independent given the root variable Y. This in turn means that the joint probability distribution of $Y, X_1, \ldots, X_d$ can be factorized as

$$P_{Y,X_1,\ldots,X_d}(y, x_1, \ldots, x_d) = P(y) \cdot P(x_1, \ldots, x_d|y) = P(y) \cdot \Pi_{i=1}^d P(x_i|y), \tag{2.1}$$

which ensures that in this network structure, the probability $P(y|x_1, \ldots, x_d)$, the main term determining the converted conditional probability distribution of the root variable Y given observed $X_1, \ldots, X_d$ will be based on the marginal probability distribution, $P(x_i|y)$ thereby taking every variable X_i into account. The graph structure of N-BN is depicted in Figure 1a) and will be used in what follows as a base framework.

Even though the factorisation (2.1) induced by N-BN essentially simplifies probability calculations, the assumption about conditional dependence among the variables/attributes $\{X_1, \ldots, X_d\}$, is clearly not always realistic. Consider for example a network structure for modeling the risk in a loan application: is seems counter-intuitive to ignore the interactions between educational level, income and age. In order to capture possible interactions between the variables in the induced BN, we will use TAN model which imposes a tree-structure on the N-BN graph allowing additional edges among $\{X_1, \ldots, X_d\}$. In an augmented structure, en edge from X_i to X_j implies that the influence of X_i of the value of the root variable also depends on the evaluate of X_j. the tree-augmented graph structure can therefore be described by identifying the set of parents $\Pi_{[i]}$ of each node. The resulting BN converts the dependencies between $Y, X_1, \ldots, X_d$ to the approximation of joint probability distribution given by

$$P_{Y,X_1,\ldots,X_d}(y, x_1, \ldots, x_d) = P(y) \cdot P(x_i|y) \cdot \Pi_{j=1, j\neq i}^d P(x_j|x_i, y), \tag{2.2}$$

that is $\Pi_{[j]} = \{Y, X_i\}$ for each X_j, and X_i is usually referred to as a super parent. Figure 1b) illustrates one possible graph structure of TAN.

Finally we discuss k-dependence BN, that is a graph structure which extends TAN allowing each node X_i to have at most k parent variables plus the root variable Y for each feature variable, $\Pi_{[i]} = \{Y, X_{i_1}, \ldots, X_{i_k}\}$. This structure yields the following probability factorization

$$P_{Y,X_1,\ldots,X_d}(y, x_1, \ldots, x_d) = P(y) \cdot P(x_1|\Pi_{[1]}) \cdot \cdots \cdot P(x_{d-1}|\Pi_{[d-1]}) \cdot P(x_d|\Pi_{[d]}), \tag{2.3}$$

where $Y \in \Pi_{[i]}$ for all $i = 1, \ldots, d$. Figure 1c) illustrates one special case of k-BN with $k = 1$.

Observe that both TAN and k-BN induce tree-structure graph and TAN models are equivalent to k-BN models with $k = 1$. Observe also that according to the definition of k-BN the N-BN model is a 0-dependence BN.

All kinds of the structures $\mathcal{G}$, introduced above can be interpreted as simplification of the correspondent true multivariate probability distributions $P_{Y,X_1,\ldots,X_d}(y, x_1, \ldots, x_d)$. These simplifications are based on the relations of conditional dependency which are inferred from $\mathcal{G}$, and they are represented by means of factorizations (2.1)-(2.3) which in general require less parameters than fitting the joint probability distribution.

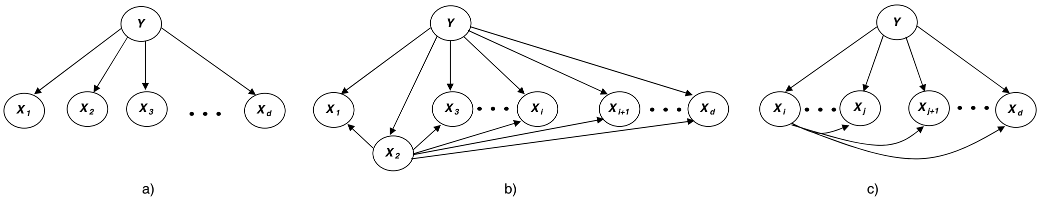

Figure 1: Three typical Bayesian Network structures: a) N-BN b) TAN c) k-BN

In what follows, we will explore the rooted and directed tree BN graphs and demonstrate that they provide an effective methodology for modeling and assessment the credit risk concentration.

2.2 Bayes rule and assessing posterior credit risk

Since the quantitative part of the BN model is represented by a set of conditional probabilities, we start from *Bayes' theorem* which is an essential part of probabilistic calculations. This theorem, for two subsets of random variables $\mathcal{X}$ and $\mathcal{Y}$ such that $P(\mathcal{X}) \neq 0$ determines the conditional probability distribution of $\mathcal{Y}$ given $\mathcal{X} = x$ as

$$P(\mathcal{Y} = y|\mathcal{X} = x) = P(y|x) = \frac{P(x|y)P(y)}{P(x)}. \tag{2.4}$$

One particular case of (2.4) is obtained when $\mathcal{Y}$ is a single variable, i.e. $\mathcal{Y} = \{Y\}$ and $\mathcal{X} = \{X_1, \ldots, X_d\}$ is a subset of variables. In this case,(2.4) becomes

$$P(y|x_1, \ldots, x_d) = \frac{P(y)P(x_1, \ldots, x_d|y)}{\sum_y P(y)P(x_1, \ldots, x_d|y)} \tag{2.5}$$

In the problem of credit risk assessment, Bayes' theorem can be exemplified as follows. Suppose the financial institute is interested to assess the credit risk arising from the uncertainty of a borrower companies ability to perform its obligations, i.e. to repay the loan. Assume for concreteness that the borrower solvency status, $\mathcal{Y}$, is represented by a single binary random variable Y that can take one of possible values, $\{\text{Solv}, \text{Ins}\}$, where $Y = \text{Solv}$ means the company is solvent, and $Y = \text{Ins}$ means company is insolvent. Suppose also for simplicity that the two characteristics, the stock market performance ($X_1 = \{\text{Good}, \text{Poor}\}$) and the guarantor's company financial condition ($X_2 = \{\text{Sound}, \text{Distressed}\}$), are thought to be associated with the borrower company credit strength. Suppose that the results of credit expert examination show that the borrower's stock market performance is poor and the guarantor is financially not in sound health, i.e. $\{x_1 = \text{Poor}, x_2 = \text{Distr}\}$. Now, given this cumulative evidence we wish to compute the probability that the company is insolvent, that is, $Y = \text{Ins}$. Then, using Bayes' theorem, we obtain

$$\begin{aligned} &P(Y = \text{Ins}|X_1 = \text{Poor}, X_2 = \text{Distr}) \\ &= \frac{1}{\mathcal{K}} \cdot P(Y = \text{Ins}) \cdot P(X_1 = \text{Poor}, X_2 = \text{Distr}|Y = \text{Ins}) \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} \mathcal{K} &= P(X_1 = \text{Poor}, X_2 = \text{Distr}|Y = \text{Ins}) \cdot P(Y = \text{Ins}) \\ &+ P(X_1 = \text{Poor}, X_2 = \text{Distr}|Y = \text{Solv}) \cdot P(Y = \text{Solv}). \end{aligned}$$

Interpretation: The probability $P(Y = \text{Ins})$ is the *prior* probability of the borrower companies insolvency, because it can be obtained *before* knowing the companies characteristics X_i. A numerical value of this probability will be referred to as the *prior credit risk* that can be assigned by expert-based initial belief regarding the risk exposure, or estimated as for example, an average percentage of companies of the same financial level in the whole population that have recently been declared insolvent.

The probability $P(Y = \text{Ins}|X_1 = \text{Poor}, X_2 = \text{Distr})$ is the *posterior* probability of the insolvency, $Y = \text{Ins}$, because it is computed *after* knowing the outcomes $X_1 = \text{Poor}, X_2 = \text{Distr}$. This posterior probability represents the *revised* experts' belief for the credit risk given that X_1 and X_2 have been observed and will be referred to as the *posterior credit risk*.

The probability $P(X_1 = \text{Poor}, X_2 = \text{Distr}|Y = \text{Ins})$ is referred to as the *likelihood* that an insolvent borrower company $Y = \text{Ins}$, will show the outcomes states $X_1 = \text{Poor}, X_2 = \text{Distr}$. $\mathcal{K}$ plays the role of normalizing constant in (2.6).

Thus, the credit risk assessment in the example above is based on updating the posterior probability distribution of Y using both the prior and the likelihood. All component probability terms in the expressions (2.6) can be evaluated from relatively few observed cases and easily updated with new data/information as it is learned.

The BN model for credit risk in the example above can be represented by the graph $\mathcal{G}$ consisting of the set of random variables (graph nodes) $\mathcal{X} = \{X_1, X_2, Y\}$ and associated dependencies (edges) between them. The calculations in (2.6) are performed assuming the complete graph structure. However, depending on the problem at hand the relationships can be restricted to a specific graph structure and the factorisation of the correspondent probability distribution can be used. For example, by restricting the relationships between X_1 and X_2 to conditional independence given Y, a special case of N-BN model (see Figure 1a)) can be induced. In this case, $\Pi_{[i]} = Y$ for each X_i and, since the graph has no edges between X_1 and X_2,the joint likelihood decomposes according to the N-BN graph structure (see (2.1)) into a product of two marginal conditional probabilities,

$$P(X_1 = \text{Poor}, X_2 = \text{Distr}|Y = \text{Ins}) =$$
$$= P(X_1 = \text{Poor}|Y = \text{Ins}) \cdot P(X_2 = \text{Distr}|Y = \text{Ins})$$

which essentially simplifies calculations.

An important aspect of the BN framework is that the constructed model can be used for a classification task where the goal is to accurately predict the value of the class variable Y given a set of attributes $X_1, \ldots, X_d$. By using Bayesian minimum-error classification rule see e.g. Friedman et al. (1997), we then assign an observed vector X to the class with the highest posterior probability, $P(y|x_1, \ldots, x_d)$, specified by (2.5). This entails the use the winner-takes-all rule, see e.g Ekdahl and Koski (2006), Pavlenko and Fridén (2007). When represented as a BN classifier, the model can be used for customer credit evaluation. As in our example above with given binary class variable, Y, one can pre-specify two classes of borrowers, one is highly reliable (sure to repay in time) and the other one is not reliable (default certainly), and then assign a new borrower $x_1, \ldots, x_d$ to one of the classes by maximizing the posterior probability, $P(y|x_1, \ldots, x_d)$. Classification accuracy and computational efficiency of the BN classifiers were studied in statistical data analysis and classification literature, see e.g. Ekdahl and Koski (2006), Pavlenko and Fridén (2007), Corrander et al (2009).

3 Description of the Empirical Framework

The problem of statistical learning of a BN model consists of finding the graph structure and estimating correspondent probability parameters. However, in the real problem it is very difficult to find good data, especially when the studied problem is complex. What we present in the current study is the stepwise BN construction which combines expert knowledge with available data.

3.1 Model construction

We focus on a private mid sized bank in Ukraine whose aim was to design a credit risk model in which a particular role of the related borrowers exposure can be analysed as a risk aggregating factor.

To identify significant risk sources and capture the threat from concentration of credit risk, the credit experts select five financial institutes (FIs) and five private enterprises and single borrowers whose entire credit grant accounts for a substantial proportion of the bank's total capital funds.

To capture potential correlations among these borrowers, a questionnaire with a number of compulsory fields was stored and analysed, which is a part of the bank credit evaluation and approval process, (a different structure of the questionnaire was used for FIs and private borrowers). The questionnaire was developed according to IAS 24, Related Party Disclosures requirements, see IAS 24 (2008), and comprises the following compulsory characteristics

Role	Characteristic
1	Founder of bank or business partner
2	Director of a business partner
3	Depositor/Guarantor
4	Connected persons (e.g family members)

Table 1. Decoding codes for the related party disclosures.

The screening process yields a group of connected borrowers together with the directed relationships which are illustrated in Figure 2. The directed links in Figure 2 can be interpreted as follows. The

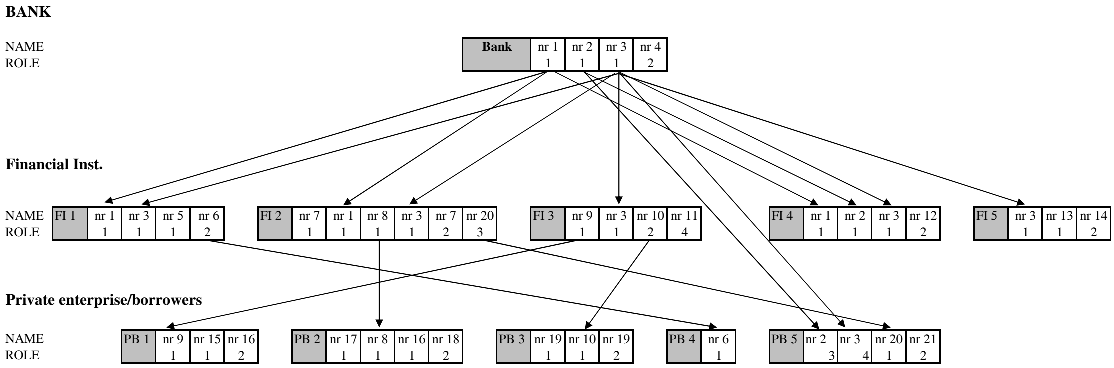

Figure 2: Related party disclosure. Decoding code names are given in Table 1.

three links between the bank and the fourth FI, for example, indicate that they have joint founders. The two links from the third FI to the first and third private enterprises indicate that they have joint founders or business partners.

To formalize the two previous stages and induce a BN model, we introduce a set of random variables, Y, S_i and T_j where Y denotes the bankruptcy status of the bank, S_i denotes the solvency status of the ith FI, $i = 1, \dots, 5$ and T_j denotes the solvency status of jth private enterprise, $j = 1, \dots, 5$. Then, the resulting BN graph, $\mathcal{G}$, represented in Figure 3 consists of the following nodes, $\mathcal{X} = \{Y, S_1, \dots, S_5, T_1, \dots, T_5\}$, and edges representing direct dependencies between the nodes in accordance with credit experts evaluation represented in Figure 2.

All the nodes in $\mathcal{G}$ are treated as binary random variables that assume the values $\{b, nb\}$, i.e. bankrupt or non-bankrupt for Y, and $\{s, ns\}$, i.e solvent or insolvent for both S_i and T_j. Observe that the induced graph has a tree augmented structure, more precisely this is a special case of TAN graph model introduced in Section 2. To show this we investigate a set of parents of one of T_is. The node T_2, for example, has the set of parent nodes, $\Pi^T_{[2]} = \Pi^T_{[5]} = \{Y, S_2\}$, where the root node, Y, represents the bank status and the super parent node, S_2, represents one of FIs. The structure of $\mathcal{G}$ also can be interpreted as a particular case of k-BN with $k = 1$.

Recall that the BN is a pait $\langle \mathcal{G}, P \rangle$ that encodes a joint probability distribution over a set of random variables $\mathcal{X}$. Hence, in order to fully specify the network, the induced graph $\mathcal{G}$ will need to be populated with the conditional probabilities $P(x_i|\Pi_{[i]})$ for each node in $\mathcal{X}$. Observe that by the tree structure of $\mathcal{G}$, the S_is are conditionally independent given Y, and T_j are also conditionally independent given their direct predecessors, S_i and Y. The absence of an edge indicates thelack of direct association.

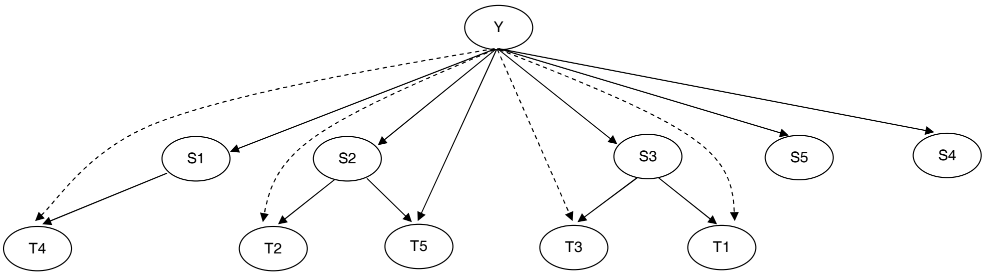

Figure 3: TAN representation of the group of related borrowers.

Both prior marginal and conditional probability distributions for the root node Y and the set of nodes S_i and T_j respectively, were assest by credit experts and summarized in the following table

$P(Y=b)$	$P(Y=nb)$
0.5	0.5

	$y=nb$	$y=b$
$P_{S_1\|Y}(s\|y)$	0.7	0.2
$P_{S_1\|Y}(ns\|y)$	0.3	0.8
$P_{S_2\|Y}(s\|y)$	0.7	0.2
$P_{S_2\|Y}(ns\|y)$	0.3	0.8
$P_{S_3\|Y}(s\|y)$	0.6	0.3
$P_{S_3\|Y}(ns\|y)$	0.4	0.7
$P_{S_4\|Y}(s\|y)$	0.8	0.1
$P_{S_4\|Y}(ns\|y)$	0.2	0.9
$P_{S_5\|Y}(s\|y)$	0.6	0.3
$P_{S_5\|Y}(ns\|y)$	0.4	0.7

	$x=s$	$x=ns$
$P_{T_1\|S_3}(s\|x)$	0.6	0.3
$P_{T_1\|S_3}(ns\|x)$	0.4	0.7
$P_{T_2\|S_2}(s\|x)$	0.6	0.3
$P_{T_2\|S_2}(ns\|x)$	0.4	0.7
$P_{T_3\|S_3}(s\|x)$	0.55	0.25
$P_{T_3\|S_3}(ns\|x)$	0.45	0.75
$P_{T_4\|S_1}(s\|x)$	0.65	0.3
$P_{T_4\|S_1}(ns\|x)$	0.35	0.7
$P_{T_5\|S_2}(s\|x)$	0.52	0.45
$P_{T_5\|S_2}(ns\|x)$	0.48	0.55

	$y=nb$	$y=b$
$P_{T_1\|Y}(s\|y)$	0.5	0.3
$P_{T_1\|Y}(ns\|y)$	0.5	0.7
$P_{T_2\|Y}(s\|y)$	0.5	0.4
$P_{T_2\|Y}(ns\|y)$	0.5	0.6
$P_{T_3\|Y}(s\|y)$	0.5	0.35
$P_{T_3\|Y}(ns\|y)$	0.5	0.65
$P_{T_4\|Y}(s\|y)$	0.6	0.5
$P_{T_4\|Y}(ns\|y)$	0.4	0.5
$P_{T_5\|Y}(s\|y)$	0.6	0.3
$P_{T_5\|Y}(ns\|y)$	0.4	0.7

Table 2. Prior marginal and conditional probability distributions for the induced TAN.

Summarizing from above we see that the graph $\mathcal{G}$ together with the table of conditional probabilities constitutes the full TAN model. The tree dependent probability distribution of $\{Y, S_i, T_j, \}$ can now be factorized along $\mathcal{G}$ as it is given in (2.2) for general TAN models. We will use the probability facctorizaton for model assessment and updating algorithms.

3.2 Model Assessment/Validation

With the previous developmental stages, a completely specified TAN is obtained. However before the TAN can be used in practice, its accuracy and consistency for modeling the credit risk have to be established. Assessment of the sensitivity to the apriori assumptions must be also evaluated.

The problem of assessing the reliability of the constructed BN model consists of two parts : the first one is how to measure the degree of association between each variable X_i and its parents $\Pi_{[i]}$ in the TAN, and the second one is how to determine whether an edge connecting X_i to X_j provides an appreciable amount of information, i.e. is significant.

The main idea underlying the approach we shall specialize here is based on the *mutual information* measure that captures the mutual dependency across the set of variables X_i and their parents $\Pi_{[i]}$, given the factorization of probability distribution of $\{X_1, \ldots, X_d\}$ along the graph tree $\mathcal{G}$, and is defined as

$$MI_{X_i,\Pi_{[i]}} = \sum_{x,y} P_{X_i,\Pi_{[i]}}(x,y) \log\Big(\frac{P_{X_i,\Pi_{[i]}}(x,y)}{P_{X_i}(x)\cdot P_{\Pi_{[i]}}(y)}\Big),$$

where $P_{X_i,\Pi_{[i]}}(x,y)$ represents the joint probability distribution of X_i and its parent variables $\Pi_{[i]}$, and $P_{X_i}(x)$ and $P_{\Pi_{[i]}}(y)$ are marginal distributions of X_i and $\Pi_{[i]}$. Various properties of $MI_{X_i,\Pi_{[i]}}$ are studied in in Friedman et al (1997) and de Campos (2006) in the context of BN model learning. In order to simplify calculations of the joint distributions of X_i and $\Pi_{[i]}$ we exploit the recursive property of the mutual information (Friedman et al (1997): $MI_{X,Y,Z} = MI_{X,Z} + MI_{X,Y|Z}$ where the second term,

$$MI_{X,Y|Z} = \sum_{x,y,z} P(x,y,z) \log \frac{P(x,y|z)}{P(x|z)P(y|z)} \tag{3.1}$$

is called *conditional mutual information* measuring the information that Y provides on X when the value of $Z = z$ is known. Let's assume now that $\Pi_{[i]} = \{X_{i_1}, \dots, X_{i_q}\}$, i.e. the variable X_i has q parents, and then apply the recursive property to $MI_{X_i,\Pi_{[i]}}$. This gives

$$\begin{aligned} MI_{X_i,\Pi_{[i]}} &= MI_{X_i,X_{i_1},\dots,X_{i_{q-1}}} + MI_{X_i,X_{i_q}|X_{i_1},\dots,X_{i_{q-1}}} = \\ = MI_{X_i,X_{i_1},\dots,X_{i_{q-2}}} &+ MI_{X_i,X_{i_{q-1}}|X_{i_1},\dots,X_{i_{q-2}}} + MI_{X_i,X_{i_q}|X_{i_1},\dots,X_{i_{q-1}}} = \dots \\ &= MI_{X_i,X_{i_1}} + \sum_{k=2}^{q} MI_{X_i,X_{i_k}|X_{i_1},\dots,X_{i_{k-1}}}. \end{aligned} \tag{3.2}$$

The first term in the final expression of the decomposition above can be calculated using the estimates of the two dimensional probability distributions $P(x_i, x_{i_1})$ as well as the marginal distributions $P(x_i)$ and $P(x_{i_1})$. It will be interpreted as the degree of interaction between the variables X_i and X_{i_1}, i.e. by inserting the edge $X_{i_1} \to X_i$. If we insert the edge $X_{i_2} \to X_i$ given that X_{i_1} is already a parent of X_i, the interaction degree between X_{i_2} and X_i is measured the conditional mutual information $MI_{X_i,X_{i_2}|X_{i_1}}$.

Using the interpretation of the mutual information given above one can see that $MI_{X_i,\Pi_{[i]}}$ is null when the two sets of variables are independent and maximum when they are functionally dependent. We use the decompositions (3.1) and (3.2) to assess the reliability of the edges inserted according to Table 2 for the group of related borrowers.

In order to assess the strength of dependence between the nodes in the TBN graph we need to compute MI_{X_i,X_j} for $d\cdot(d-1)$ different pairs of indices. We begin by analysing the edges in $\mathcal{G}$ relating the bank, Y to financial institutes S_i, $i = 1, \dots, 5$ and to private enterprises T_j, $j = 1, \dots, 5$, and estimate the strength of these interactions assuming only one parent variable for both S_i and T_j, i.e. $\Pi_{[i]} = \{Y\}$ and $\Pi_{[j]} = \{Y\}$ for S_i and T_j, respectively. Using the first term in the right-hand side of the decomposition (3.2), we compute the numbers

$$MI_{S_i,Y} = \sum_x \sum_y P_{S_i,Y}(x,y) \log \frac{P_{S_i,Y}(x,y)}{P_{S_i}(x)P_Y(y)}, \tag{3.3}$$

for each S_i, where the marginal distribution of Y is specified in Table 2. The marginal distribution of S_i is calculated using the total probability theorem as

$$P_{S_i}(x) = \sum_{y\in\{b,nb\}} P_Y(y) \cdot P_{S_i|Y}(x|y), \quad x \in \{s, ns\}, \quad i = 1, \dots, 5. \tag{3.4}$$

By analogy to (3.3) and (3.4) we can also compute $MI_{T_i,Y}$ marginal distribution of T_i. Both marginal distributions of S_i and T_j are presented in Table 4.

$P_{S_i}(x)$	$x = s$	$x = ns$
S_1	0.45	0.55
S_2	0.45	0.55
S_3	0.45	0.55
S_4	0.45	0.55
S_5	0.45	0.55

$P_{T_i}(x)$	$x = s$	$x = ns$
T_1	0.4	0.6
T_2	0.45	0.55
T_3	0.425	0.575
T_4	0.55	0.45
T_5	0.45	0.55

Table 3. Marginal probability distributions for S_i and T_j.

The joint distribution of S_i, Y is calculated using the table of conditional distributions of $S_i|Y$ and marginal distribution of Y as

$$P_{S_i,Y}(x,y) = P_{S_i|Y}(x|y) \cdot P_Y(y). \tag{3.5}$$

The same technique is used for calculating $P_{T_i,Y}(x,y)$. The resulting values of $MI_{S_i,Y}$ and $MI_{T_i,Y}$ are given in Table 4, indicating much weaker direct interactions among Y and T_is than among Y and S_is.

i	1	2	3	4	5
$MI_{S_i,Y}$	0.1325	0.1325	0.0462	0.2753	0.0462
$MI_{T_i,Y}$	0.0211	0.0051	0.0116	0.0051	0.0463

Table 4. Mutual information $MI_{S_i,Y}$ and $MI_{T_i,Y}$.

In order to capture the relationship between S_i and T_i given the root variable Y, we focus on the conditional mutual information between X and its parent variables Π_X given a set of variables Z and explore the second term of the decomposition (see the decomposition of $MI_{X,Y|Z}$, given by (3.2)). Observe that for the TAN graph structure $MI_{X_i,\Pi_{[i]}|Z}$ can be calculated by

$$MI_{X_i,\Pi_{[i]}|Z} = \sum_z \Big(P(z) \sum_{x,x_{\Pi_{[i]}}} P(x, x_{\Pi_{[i]}}|z) \cdot \log \Big(\frac{P(x, x_{\Pi_{[i]}}|z)}{P(x|z)P(x_{\Pi_{[i]}}|z)} \Big) \Big). \tag{3.6}$$

see Friedman et al. (1997). Given a root variable Y, the conditional mutual information between T_i and S_j for all pairs of indices (i,j) is computed using (3.6). The resulting values of $MI_{T_i,S_j|Y}$ are summarised in the following table

$MI_{S_i,T_j|Y}$	T_1	T_2	T_3	T_4	T_5
S_1	0	0	0	0.0670	0
S_2	0	0.0344	0	0	0.0419
S_3	0.0515	0	0.0484	0	0
S_4	0	0	0	0	0
S_5	0	0	0	0	0

Table 5. Conditional mutual information between T_i and S_j given Y.

Now the joint mutual information MI for the whole TAN is obtained by emerging Tables 4 and 5 are recorded in Table 6.

T_i	$\Pi[i]$	$MI_{T_i,\Pi[i]}$
T_1	(S_3, Y)	0.0726
	$(S_i, Y), \quad i \neq 3$	0.021
T_2	(S_2, Y)	0.0395
	$(S_i, Y), \quad i \neq 2$	0.0051
T_3	(S_3, Y)	0.060
	$(S_i, Y) \quad i \neq 3$	0.0116
T_4	(S_1, Y)	0.0721
	$(S_i, Y), \quad i \neq 1$	0.0051
T_5	(S_2, Y)	0.0882
	$(S_i, Y), \quad i \neq 2$	0.0463

Table 6. Joint mutual information $MI_{T_i,\Pi[i]}$ for the whole TAN.

In order to filter out the non-reliable edges from $\mathcal{G}$ different strategies can be considered. The simplest one is to adopt a *threshold based* approach that uses the average mutual information, $\widehat{MI}_{S,T|Y}$ as a cut-off value, where the averaging is performed over all possible edges between T and S. The values of $MI_{S_i,T_j|Y}$ that are higher than the cut-off are considered to be reliable. The data from Table 5) gives $\widehat{MI}_{S,T|Y} = 0.0097$, which indicates that all the edges that were specified in the TAN by expert knowledge are reliable. The reliability of the edges connecting Y, T_i and S_j, can be estimated by the joint mutual information whose average value yields $\widehat{MI}_{S,T,Y} = 0.0276$. All the triple edges connecting Y, S_i, T_j in the Figure 3 have the value of joint mutual information higher than $\widehat{MI}_{S,T,Y}$ and are therefore reliable. Observe that some of evaluated edges appear to be reliable even though they have not been induced by the experts. For example, the connection between Y, S_1 and T_5 appear to be stronger than the average $\widehat{MI}_{S,T,Y}$ which is due to the strong direct connection between Y and T_5 and no conditional connection between S_1 and T_5, i.e.

$$MI_{S_1,T_5,Y} = MI_{T_5,Y} + MI_{S_1,T_5|Y} = 0.0463 + 0 = 0.0463.$$

The more subtle strategy to assess the edges reliability is based on so-called permutation test, see e.g. []. We explore the empirical distribution of estimates of conditional and joint mutual information for permuted data to judge whether the calculated values of $MI_{S,T|Y}$ and $MI_{S,T,Y}$ are higher than we would expect when connecting the nodes in G by chance only. We perform 1000 bootstrap replicates for the data given in Tables 5 and 6 and apply the kernel smoothing density technique, (see []) to approximate the empirical distributions of the conditional and joint information. The results given in Figure 4 (see the legend to Figure 4 for detailed of the density approximation) and are very satisfactory. Values in the range of $MI_{S,T|Y}$ and $MI_{S,T,Y}$ which we obtain from the original TAN probability distribution were never achieved with any of the permuted values of MIs. This corresponds to an empirical p-value of zero in the permutation test for our entire measure of edge reliability based on mutual information. We thus can surely reject the hypothesis that the values of $MI_{S,T|Y}$ and $MI_{S,T,Y}$ found on the original probability distribution tables are irrelevant and just a noise artifact. Moreover, we observe that the spread of permuted values of both conditional and joint mutual information is much higher, clearly exceeding the typical values of $MI_{S,T|Y}$ and $MI_{S,T,Y}$, see Tables 5 and 6.

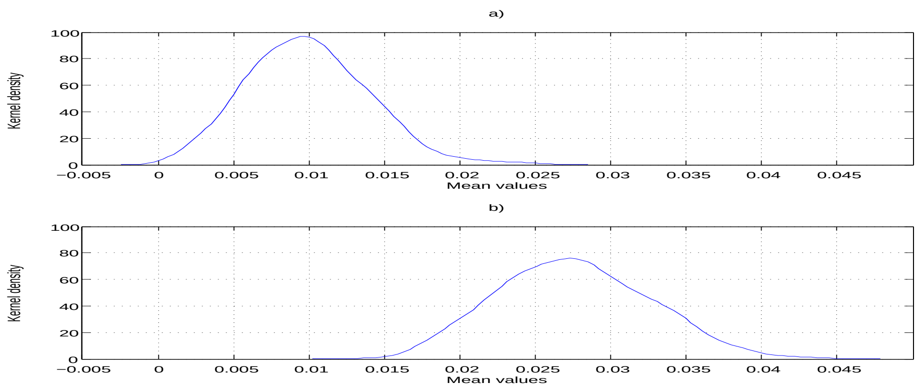

Figure 4: Kernel density approximation of the average values of $MI_{S,T|Y}$ and $MI_{S,T,Y}$ computed by a normal kernel smoother based on 1000 bootstrap replicates from Table 5 (top panel) and from Table 6 (bottom panel), respectively. the kernel-smoothing window. The bandwidth of the kernel-smoothing window was chosen to be optimal for estimating normal densities. The number of equally spaced points in both bootstrap samples were equal to 100.

In summary, we conclude that our assessment algorithm assigns relatively high values of mutual information to the edges in $\mathcal{G}$ representing a group of related borrowers, so that in this sense the induced TAN is a reliable model that can be very accessible to identify and control threat from concentration of credit exposures.

3.3 Updating Algorithms and Stress Testing

In order to evaluate the TAN output under extreme unfavorable changes in the economical conditions of the group of related borrowers and the implications to the bank, we focus on the stress testing technique. Here we run a number of hypothetical "what if?" scenarios on the constructed model under various assumptions about the prior probability of insolvency for a specific set nodes in $\mathcal{G}$.

Firstly we look at the case when only one of FI's, for example S_2 declared insolvensy, i.e. the credit risk analyst observed $S_2 = ns$. To update the prior distribution of Y we use Bayes theorem given by (2.4) and get the the posterior probability of $Y|S_2$ as

$$P_{Y|S_2}(b|ns) = \frac{P_{S_2|Y}(ns|b) \cdot P_Y(b)}{P_{S_2}(b)} = \frac{0.8 \cdot 0.5}{0.55} = 0.73.$$

Hence, changing in the solvency status of just one FI, S_2 seriously increase the bankruptcy risk from the prior value of 0.5 to posterior estimator of 0.73. The next scenario demonstrates how to capture the effect of concentration of the credit risk exposures due to insolvency of two FIs. Given that the second and forth FIs are classified by the risk analyst as insolvent, the posterior credit risk can be updated using TAN graph structure, i.e. using the conditional independence of S_2 and S_4 nodes, given the root Y. This yields

$$P_{Y|S_2,S_4}(b|ns,ns) = \frac{P_Y(b)P_{S_2|Y}(ns|b)P_{S_4|Y}(ns|b)}{\sum_{y\in\{b,nb\}} P_Y(y)P_{S_2|Y}(ns|y)P_{S_4|Y}(ns|y)} = 0.923,$$

which implies that the increase of the bank posterior bankruptcy risk from 0.5 to 0.923 in the case of insolvency of these two FIs is much more pronounced.

The effect of insolvency of the private enterprises and single borrowers on the risk of bankruptcy can also be investigated. Let's assume that T_2 and T_5 declared insolvency. Observe that due to the TAN structure, both T_2 and T_5 have the same set of parents $\Pi^T_{[2]} = \Pi^T_{[5]} = \{Y, S_2\}$. Using (2.5) and the probability factorization for TAN models from the Section 2.1 we get

$$P_{Y|T_2,T_5}(b|ns,ns) = \frac{\sum_{x\in\{s,ns\}} P_{T_2|S_2}(ns|x)P_{T_5|S_2}(ns|x)P_{S_2|Y}(x|b)P_Y(b)}{\sum_{y\in\{b,nb\}}\sum_{x\in\{s,ns\}} P_{T_2|S_2}(ns|x)P_{T_5|S_2}(ns|x)P_{S_2|Y}(x|y)P_Y(y)} = 0.581.$$

Notice that T_2 and T_3 are related to Y through their super parents, S_is, the increase of the credit risk from its prior value of 0.5 to the posterior estimation, 0.581, is not that serious.

In the group of related borrowers, the private enterprises and single borrowers could have different parents among the FI's, like for example T_2 and T_3 whose set of parents are $\Pi^T_{[2]} = \{Y, S_2\}$ and $\Pi^T_{[3]} = \{Y, S_3\}$, respectively, that might be interesting to run a scenario on the TAN model under assumptions that $T_2 = ns, T_3 = ns$, i.e. two private borrowers declared unsolvency. The posterior risk for this case will be

$$P_{Y|T_2,T_3}(b|ns,ns) = \frac{P_{T_2|Y}(ns|b)P_{T_3|Y}(ns|b)P_Y(b)}{\sum_{y\in\{b,nb\}} P_{T_2|Y}(ns|y)P_{T_3|Y}(ns|y)P_Y(y)} = 0.6094,$$

and the increase is observed from 0.5 to the posterior value of 0.6094.

To update the local risk concentration, i.e. posterior distribution of S_i, we consider the scenario where T_2 and T_5 become insolvent. The posterior insolvency risk for the parent FI, S_2, is calculated by

$$P_{S_2|T_2,T_5}(ns|ns,ns) = \frac{P_{T_2|S_2}(ns|ns)P_{T_5|S_2}(ns|ns)P_{S_2}(ns)}{\sum_{x\in\{s,ns\}} P_{T_2|S_2}(ns|x)P_{T_5|S_2}(ns|x)P_{S_2}(x)} = 0.7233.$$

The increase of the posterior risk of insolvency of S_2 from 0.55 (see Table 4) to 0.7233. Observe that the marginal distribution of S_i is not a prior distribution in the same sense as that for Y. This distribution was calculated using the prior distribution of Y, but not specified a priori by experts.

Clearly, these scenarios are by the problem statement relatively simple. However,in practice, by using a tool such as the induced TAN model, our approach captures "storing-up" of the bank posterior credit risk which arises if the borrowers in one portfolio have a particularly strong correlations with one another.

4 Conclusions and scope for future

The primary contribution of this study has been to demonstrate that the BN methodology has a number of advantages for adapting it to modeling of credit concentration risk. The main advantage of BNs is their ability to integrate uncertain expert knowledge (e.g. expert estimates of the risk exposure of a group of related borrowers) with data . This potentially provides economists with the ability to update prior knowledge with new information as it is learned, and to built a solution of increasing scope and complexity. Another advantage of the BN modeling approach is that both the graph structure and probability parameter estimate can easily be updated on a periodic basis, thereby taking into account changes in the borrowers credit strength and economic conditions.

Two types of BN models, TAN and a special case of k-BN were explored and shown to be the most suitable graphs for capturing and visually representing aggregating of credit risk for a group of related borrowers. The established models can be applied to forming posterior credit risks based on the evidence observed, update marginal and conditional probability distributions of each risk related node of the BN graph, and simulate scenarios for stress testing. These models are now successfully implemented in the Ukrainian bank as a part of the global program on improvement of monitoring and management of concentration risk in the banking industry of Ukraine.

The measure of mutual information suggested in this study for model assessment is also applicable to learning graph structure directly from data, i.e. without any expert information on a tree dependent probability distribution, see e.g. Pérez et al(2006), Ekdahl and Koski (2006). Given the set of nodes and the relevant sample frequencies, the procedure of structure learning can be started from computing the mutual information between all pairs of nodes, and then the best tree structure is selected as the one that gives the maximum overall mutual information.

Current models were based on the binary graphs and were structurally relatively simple networks that can be extended a multinomial model which is more precise instrument for quantifying the risk exposure. Furthermore, introducing nodes with continuous probability distributions and increasing the graph complexity allows for applying the BN methodology for the modeling of more sophisticated multiple correlations across exposures in management of concentration risk.

Acknowledgements.
The authors would like to thank Prof. Daniel Thorburn, Stockholm University for useful discussions related to this work, and to Ludmila Kucheruk for providing the credit risk questionnaires analyses and generating the group of related borrowers in the mid-sized Ukrainian bank. The first author was supported in part by the Swedish Research Council under VR Grant 421-2008-1966. The second author is supported in part by the Swedish Institute under Grant 00962/2007-2010.

References

Anderson RD, Mackoy RD, Thompson VB, Harrell G. A Bayesian network estimation of the service-profit chain for transport service satisfaction. *Decision sciences* 2004, 35(4):665-690.

Basel Committee on Banking Supervision. *International convergence of capital measurements and capital standards*, Basel, 2006.

Bonti G, Kalkbrener M, Lotz C, Stahl G. Credit risk concentrations under stress. *Journal of Credit Risk* 2006, 2(3):115-136.

de Campos LM, A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *The Journal of Machine Learning Research* 2006, 7:2149-2187.

Commitee on the Global Financial System. Stress testing at major financial institutions: survey results and practice, Basel, 2007.

Corrander J, Gyllenberg M, Koski T. Bayesian unsupervised classification framework based on stochastic partitions of data and parallel search strategy. *Advances in Data Analysis Classification* 2009, 3:3-24.

Cowell RG, Verrall RJ, Yoon YK. Modeling operational risk with Bayesian networks. *The Journal of Risk and Insurance* 2007, 74(4):795-827.

Das SR, Duffie D, Kapadia N, Saita L. Common failings: How corporate defaults are correlated? *Journal of Finance* 2007, 12(1):93-117.

Ekdahl M, Koski T. Bounds for the loss in probability of correct classification under model base approximation. *Journal of Machine Learning Research* 2006, 7:2449-2480.

Ellis B, Wong WH. Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association* 2008, 103(482):778-789.

Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Machine Learning* 1997, 29(2):1-36.

Gordy M, Lütkebohmert E. Granuality adjustment for Basel II. *Banking and Financial Studies* 2007, 1:1-23.

Hosack RH, Hayes KR, Dambacher MJ. Assessing model structure uncertainty though an analysis of system feedback and Bayesian networks. *Ecological Applications* 2008, 18(4):1070-1082.

IAS 24 Related Party Disclosures. *International Accounting Standards Board*, *http://www.iasplus.com/agenda/relatedparty.htm*, 2008.

Jing Y, Pavlović V, Rehg MJ. Boosted Bayesian network classifiers. *Machine Learning* 2008, 73:155-184.

Kalapodas E, Thomson M. Credit risk assessment: a challenge for financial institutions. *Journal of Management Mathematics* 2006, 17:25-46.

Neil M, Marquez D, Fenton N. Using Bayesian networks to model the operational risk to information technology infrastructure in financial institutions. *Journal of Financial Transformation* 2007, 2:131-138.

Pavlenko, T, Fridén, H. Scoring feature subsets for separation power in supervised Bayesian classification. *Soft Methodology and Integrated Uncertainty Modeling* 2007, 37:383-391.

Pavlenko, T, Fridén, H. Learning class discriminative Bayesian networks in high-dimensional framework. To be published in Theory of Stochastic Processes: Applications to Financial Mathematics and Risk Theory, *Springer*, 2010.

Péretz A, Larrañaga P, Inza I. Supervised classification with conditional Gaussian networks: Increasing the structure from naive Bayes. *International Journal of Approximate Reasoning* 2006, 43:1-25.

Sun L, Shenoy PP. Using Bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research* 2007, 180: 738-753.

York J. Bank concentration risk. *Risk Management Analysis Journal* 2007, 09:52-57.