A Bayesian Networks Approach to Operational Risk

V. Aquaro[a,b], M. Bardoscia[*,c,d], R. Bellotti[c,d], A. Consiglio[a],
F. De Carlo[d], G. Ferri[e]

[a] *CARMA, Research Consortium for Risk Management Automation,
via Mitolo 17B, I-70124 Bari, Italy*

[b] *Formit Servizi S.p.A., via C.Conti Rossini 26, I-00147 Roma, Italy*

[c] *Dipartimento Interateneo di Fisica "M.Merlin", Università degli Studi di Bari e
Politecnico di Bari, via Amendola 173, I-70126 Bari, Italy*

[d] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari,
via Amendola 173, I-70126 Bari, Italy*

[e] *Dipartimento di Scienze Economiche e Metodi Matematici, Università degli Studi di
Bari, via C.Rosalba 53, I-70124, Italy*

Abstract

A system for Operational Risk management based on the computational paradigm of Bayesian Networks is presented. The algorithm allows the construction of a Bayesian Network targeted for each bank using only internal loss data, and takes into account in a simple and realistic way the correlations among different processes of the bank. The internal losses are averaged over a variable time horizon, so that the correlations at different times are removed, while the correlations at the same time are kept: the averaged losses are thus suitable to perform the learning of the network topology and parameters. The algorithm has been validated on synthetic time series. It should be stressed that the practical implementation of the proposed algorithm has a small impact on the organizational structure of a bank and requires an investment in human resources limited to the computational area.

Key words: Operational Risk, Complex Systems, Bayesian Networks,
Time Series, Value-at-Risk
PACS: 89.65.Gh, 05.45.Tp
2000 MSC: 91B30, 91B84, 37M10, 62M10

[*] Corresponding author mail: marco.bardoscia@ba.infn.it and tel: +390805442178

1. Introduction

In the last years a powerful set of tools to study complexity has been developed by physicists and applied to economic and social systems; among the several topics under investigation the quantitative estimation and management of several typologies of risks [1], like financial risk [2, 3, 4, 5, 6] and operational risk [7, 8] has recently emerged.

Operational Risk (OR) is defined as "the risk of [money] loss resulting from inadequate or failed internal processes, people and systems or from external events" [9], including legal risk, but excluding strategic and reputation linked risks. Since it depends on a family of heterogeneous causes, in the past only few banks dealt with OR management. Starting from 2005 the approval of *"The New Basel Capital Accord"* (Basel II) has substantially changed this picture: in fact OR is now considered a critical risk factor and banks are prescribed to cope with it setting aside a certain capital charge.

Basel II proposes three methods to determine this capital: i) the *Basic Indicator Approach* (BIA) sets it to 15% of the bank's gross income; ii) the *STandardized Approach* (STA or TSA) is a simple generalization of the BIA: the parcentage of the gross income is different for each Business Line (BL) and varies between 12% and 18%; iii) the *Advanced Measurement Approach* (AMA) allows each bank to use an internally developed procedure to estimate the impact of OR. Both the BIA and the STA seems overly simplistic, since in some way they suppose that the exposure of a bank to operational losses is proportional to its size. On the other side, an AMA not only helps a bank to set aside the correct capital charge, but may even allow the *OR management*, in the prospect of limiting the amount of future losses.

Each AMA has to take into account two types of historical operational losses: the internal ones, collected by the bank itself, and the external ones which may belong to a database shared among several banks. Nevertheless, due to the recent interest for OR, only small and not adequately accurate historical databases exist and this is why each AMA is required to use also assessment data produced by experts. In addition, Basel II provides a classification of operational losses in 8 BLs and 7 Loss Event Types (LETs) which has to be shared by all the AMAs. Finally, AMAs usually identify the capital charge with the 99.9% 1-year Value-at-Risk (VaR), i.e. the 99.9 percentile of the yearly loss distribution.

Among the AMA methods, the most widely used is the *Loss Distribution Approach* (LDA). In LDA the distribution of frequency and the distribution of impact (severity) modeling the operational losses are separately studied for each of the 56 pairs (BL, LET). LDA makes two crucial assumptions: i) frequency and severity distributions are independent for each pair; ii) the distributions of each pair are independent from the distributions of *all the other* pairs. In other words LDA neglects the correlations possibly existing between the frequency or the severity of the losses occurring in different pairs.

The idea of exploiting BNs to study OR has already been proposed in [10], and various approaches are possible. The main advantages offered by BNs are two:

- the possible correlations among different bank processes can be captured;
- the information contained into both assessments and historical loss data can be merged in a natural way.

One approach may be to design a completely different network for each bank process, trying to determine the relevant variables (in the context of each process) and the causal relationship among them; this kind of network has only one output node which typically represents the loss distribution for the process under investigation. This approach has several drawbacks: i) domain experts are needed for each process, in order to properly identify the variables and to define the topology of each network; ii) if the historical data needs to be used, a system monitoring all the included variables with an acceptable frequency and accuracy has to be built; since this kind of network can easily reach large sizes (tens of variables), managing such systems is quite challenging for a bank institution; iii) correlations across different processes are not taken into account.

Another approach [11] is to design a unique network composed by a node for each process which represents its loss distribution; all nodes are output nodes and the operational losses are sufficient to build a historical database, so that collecting the data and managing them is much more easier for a bank; in comparison with the previous approach even the experts' task becomes simpler since their assessment reduces to an estimate of the losses over a certain time horizon; obviously this kind of network is specifically designed for capturing the correlations among different processes. This approach resembles a way of reasoning typical of the field of the Complex Systems: all the "microscopic" details inherent to each process (that make the basis on which the first approach is built on) are not included in the model, assuming that they can be neglected to a certain extent.

Let us underline that, as regards the practical implementation inside a bank, the difference between the two approaches is huge: in the first approach tens of variables for each process need to be monitored, while in the second approach only one variable per process (the registered losses) has to; considering that an AMA-oriented bank has to track its own internal losses in any case, the cost of the proposed implementation is minimum.

Mixed approaches in which a subnetwork of the kind used in the first approach (but usually smaller) is nested into each node representing the loss distribution of a process are even possible.

2. Bayesian Networks

In order to define a Bayesian Network [12] two elements are necessary: a set of random variables $V = (X_1, X_2, \ldots, X_n)$ and a network of nodes corresponding to the random variables in V. In particular the network must be a *Directed Acyclic Graph* (DAG) and the joint *Probability Distribution Function* (PDF) $P(X_1, X_2, \ldots, X_n)$ must satisfy the Markov condition, i.e. each random variable X_i and the set of all its non-descendents must be conditionally independent, given the set of all its parents. It can be proved for discrete variables (which turns out to be our case) that the Markov condition easily allows to calculate the joint PDF as:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{N} P(X_i|\mathrm{Pa}_i), \tag{1}$$

where Pa_i is the set of random variables whose corresponding nodes are parents of the node associated with X_i.

Both the directed links appearing in the DAG and the values of the conditional probabilities $P(X_i|\mathrm{Pa}_i)$ can be learned from a dataset whose records hold the values assumed by each X_i in *independent* experiments. Even if we are not dealing here with the problem of a rigorous definition of what independent experiments are, we will be more formal about this point because it is the core of our implementation. Let us associate a random variable to each node, and to each experiment: $X_i^{(p)}$ is the random variable associated with the i-th node and with the p-th experiment. The p-th and the q-th experiments ($p \neq q$) are said to be independent if $X_i^{(p)}$ and $X_j^{(q)}$ are independent $\forall\, i$ and j.

3. Different-times correlations

One of the fundamental reasons to use BNs to estimate the OR is that if correlations do exist among different processes they can be captured through the network topology; however the correlation can extend arbitrarily over time: an example will help to clarify. Suppose that an employee violates the transaction control system with fraud purposes: he succeeds in his aim and a money loss is generated in some process of the bank. As a side effect a part of the IT infrastructure is damaged, but the failure is discovered and repaired only a week later: a loss is generated in the process of machinery servicing with a one week lag. At the same time the system remained partially unavailable and a certain amount of transactions failed, eventually generating losses delayed up to a week in many other processes.

In order to understand the importance of this point we need to look at the structure of a database of historical losses: each record holds the daily losses classified by the process in which they occur. The example should have made it obvious that the losses registered in different days cannot be considered originating from independent experiments (as defined in Section 2), so a database with this structure is in principle useless for learning purposes. To overcome this limitation we propose a new approach: the losses are averaged over a certain time interval T such that the correlations of the *averaged* losses vanish at different times, but are still present at the *same* time.

In such an approach the original database is replaced by a new database (which will be called the *extracted database*) of averaged losses whose number of records is $\frac{L}{T}$, being L the number of records into the original database. Suppose e.g. that $T = 90$ is one of the time intervals we are looking for and L equals to 1 year: this means that the *average* losses of a quarter of year are not correlated with the *average* losses of *another* quarter, but the *average* losses recorded by different processes in the *same* quarter are still correlated among themselves; different quarters may be considered independent experiments, thus the extracted database can be used for learning purposes.

4. Learning Bayesian Networks by aggregate losses

In Section 3 the idea of averaging the losses over a certain time interval is introduced. What we actually do is to *sum* all the losses belonging to the same process and the same time interval: the k-th record in the extracted database contains the aggregate loss of the records from $(k-1)\,T + 1$ to

$k\,T$, obviously retaining the process classification. Let us suppose again that $T = 90$: the first (second, …) record in the extracted database contains the aggregate loss of the records from 1 to 90 (91 to 180, …) in the original database. Summing is equivalent to averaging but, as we are going to see, makes much more sense in view of the VaR calculation.

After the new database has been extracted, we can start building the network defining the nodes and the allowed states of the associated variables: we set the number of states n to 5 for all the variables; the bins are equally spaced, being 0 the lower limit and the maximum *average* loss of each process the higher limit.

The extracted database is then used to learn the structure of the network and the conditional probabilities. As hinted in Section 1, another reason why BNs seems to be suitable for OR estimation is that they allow integrating of the information coming from the historical database with the information coming from experts' assessment. Topology constraints can be imposed before the structure learning is performed, while *a prior* knowledge can be embedded properly setting the marginal distributions of each variable before the conditional probability learning is performed. However, we are mainly interested in studying the correlations of the losses and thus we choose neither to impose topology constraints, nor to embed any prior knowledge about the marginal distributions of the variables.

The joint PDF can then be derived using (1) and the marginal PDF for each variable calculated. We recall here that the database entries are values assumed by the random variables associated with the nodes (see Section 2): if the database used for the learning procedure contains the cumulative losses of a quarter (classified by process) the marginal PDFs obtained as the output of the BN will be the loss distribution per quarter (classified by process). Let us note that, provided that $T = 90$ is such that the different-times correlations vanish, it is reasonable to consider the loss distributions relative to *different* quarters to be independent. Making the further assumption that the loss distributions per quarter are the same for each quarter it is possible to calculate the loss distributions over every time horizon, by numerically convoluting the loss distributions over the time horizon T an appropriate number of times. Indeed, in order to compare the results obtained for different values of T, we calculate the loss distributions and the VaR over a fixed time horizon: for this purpose L seems the most natural time horizon to fix.

Let P_i^T be the loss distribution of the i-th process over the time horizon T and $P_i^T(k)$ the value of P_i^T in the k-th bin; the convolution of P_i^T by itself is defined by:

$$\left(P_i^T * P_i^T\right)(k) = \sum_{m=\max(1,k+1-n)}^{\min(k,n)} P_i^T(m) P_i^T(k-m+1),$$

with $n = 5$ in our case. To obtain P_i^L (i.e. the loss distribution of the i-th process over the time horizon L) P_i^T has to be convoluted by itself a number of times equal to the closest integer to $\frac{L}{T}$:

$$P_i^L = P_i^T \underbrace{* \quad \ldots \quad *}_{\frac{L}{T} \text{ times}} P_i^T.$$

The VaR over the time horizon L for each process is the 99.9 percentile of the convoluted loss distribution and the total VaR is simply the sum of the VaRs of the single processes. The 99.9 percentile of the convoluted distribution (for each process) can be numerically determined in the following way: the convoluted distribution is sampled 10^3 times and the sample is arranged in increasing order: the second largest value is the 99.9 percentile of the convoluted distribution. Since this procedure involves sampling, it is repeated several times and the VaR is calculated as the mean of the obtained 99.9 percentiles.

As hinted before, the VaR may be calculated over every desired time horizon tuning the number of convolutions; in particular the time horizon can be set to 1 year, as required by Basel II, performing $\frac{365}{T}$ convolutions.

5. Synthetic Data

In order to investigate our approach, we developed a reliable and tunable database of synthetic internal losses: in this way we are able to control the correlations between the different processes and some inherent features of each process.

We consider the historical losses of each process as a time series and, inspired by [13], generalize a stochastic algorithm for generating multiple time series. We point out that this procedure allows to impose, at least in principle, arbitrary cross-correlation functions between each pair of generated time series, as well as the auto-correlation function and distribution for each generated time series.

The steps of the algorithm are the followings: i) for each process, L values are drawn from an arbitrary distribution; the order in which the values are extracted is considered to be a temporal order, so let us call the extracte values $l_i(s)$, where the subscript $i = 1, \dots, N$ indexes the process and the argument $s = 1, \dots, L$ defines the temporal ordering. ii) The following quantity is calculated:

$$\sum_{i,j=1}^{N} \sum_{t=1}^{L-1} \left[c_{ij}(t) - C_{ij}(t)\right]^2, \tag{2}$$

where N is the number of processes, C_{ij} are the imposed cross-correlation (or auto-correlation) functions, while c_{ij} are the cross-correlation (or auto-correlation if $i = j$) functions calculated from the generated data:

$$c_{ij}(t) = \frac{1}{\mathrm{cov}(l_i, l_j)} \left[\frac{1}{L-t} \sum_{s \leq L-t} l_i(s) l_j(s+t) - \langle l_i \rangle \langle l_j \rangle \right], \tag{3}$$

with $\langle l_i \rangle = \frac{1}{L} \sum_{s \leq L} l_i(s)$ and $\mathrm{cov}(l_i, l_j) = \big\langle \big(l_i - \langle l_i \rangle\big)\big(l_i - \langle l_i \rangle\big) \big\rangle$. From (3) it follows that $c_{ij}(0) = 1$: in other words, because of its normalization, c_{ij} carries no information about the same-time correlations; in order to make the whole procedure consistent $C_{ij}(0)$ must also be equal to 1: this explains why the summation over t in (2) starts from 1 and not from 0. iii) Two values belonging to a randomly selected series are randomly chosen and exchanged, and the quantity (2) is recalculated. iv) If (2) has decreased the exchange between the two values performed in the step (iii) is accepted, otherwise it is rejected. As c_{ij} are not limited, (2) cannot be normalized and thus a threshold below which the algorithm is halted cannot be set. We rather choose to iterate the algorithm until (2) reaches a plateaux.

Since we are interested in the change of the correlation between different processes with respect to the time interval T over which the losses are averaged, we imposed auto-correlation and cross-correlation functions of the form:

$$C_{ij}(t) = e^{-\frac{t}{\tau_{ij}}}, \tag{4}$$

in fact making such a choice implies that the different-times correlation between the processes i and j should be significantly reduced averaging over a time interval $T \simeq \tau_{ij}$.

Even though the algorithm allows to impose both distributions and C_{ij}, in practice a certain degree of compatibility may exist between them: this means that, even if (2) reaches a plateaux, still c_{ij} and C_{ij} are significantly different. In order to overcome this limitation the algorithm is slightly modified in the following way: we generate series which are indeed longer than L so that a larger basin of values that may fulfill the imposed constraints is available; e.g. suppose that the values of the series are drawn from a uniform distribution and that the imposed C_{ij} have an higher degree of compatibility with another distribution: a subset of values belonging to this distribution will be selected by the algorithm. The modified algorithm obviously alters the imposed distributions; however we see no reasons to impose strict constraints on the distributions and, on the other hand, as we are interested in studying the correlations between the processes, need a high accuracy in reproducing the C_{ij}.

6. Results

We investigate a sort of *toy model* whose number of processes is limited to $N = 3$; this choice is the result of a trade-off between our need to considering a system complex enough to have a reasonable number of correlated processes and the convenience of using series longer enough to be able to carry out the average over time and still have a sufficient number of data to perform the learning of the network. With $L = 5000$ it is possible to average over 240 steps and still have 20 patterns left for the learning.

The negative exponential distribution has shown to be compatible with (4) if the decay matrix is homogeneous, i.e. $\tau_{ij} = \tau, \quad \forall\ i$ and j and if $\frac{\tau}{L}$ is not too large. In the top panel of Fig.1 both c_{ij} and C_{ij} are shown for $\tau = 25$. In order to simulate different kinds of processes their means have been set respectively at 100, 50 and 10. Using a larger basin of values as described in Section 5 both the mean and the variance of the distributions do not significantly change, but a heavier tail appears.

As it is shown in the bottom panel of Fig.1, averaging over a time interval T leaves the form (4) unchanged with a new decay time equal to $\frac{\tau}{T}$. This actually means that, at the cost of reducing the length of the time series, averaging effectively removes the different-times correlations: in particular when $\frac{\tau}{T} = 2.4$ ($T \simeq 60$) all the different-times correlations are reduced to 0.1 and for $T \geq 80$ they can be considered effectively extinguished.

Since C_{ij} carry no information about the same-time correlations (see Section 5), in order to study them we look at the learned structure of the networks: in Tab.1 it is shown that the number of links decreases as T increases.

This is somewhat expected since, as T increases, the size of the extracted database reduces and it becomes more and more difficult to learn from it. However for $T \geq 80$ the algorithm of structure learning still detects the presence of some links: since the different-times correlations are extinguished for such a large T, they must be due to the survived same-time correlations.

To evaluate the consistency of the whole procedure we require that, for values of T such that the different-times correlations can be neglected, the value of VaR does not depend on T.

In Fig.2 the values of VaR with respect to T are represented; each point is the mean over 30 realizations of the procedure described in Section 4 and the standard deviations are also shown. Indeed from Fig.2 it can be seen that for $T \geq 60$ the values of VaR are compatible among themselves. On the other hand, for $T < 60$ the different-times correlations are still present and the records belonging to the extracted database cannot be considered independent; nevertheless the learning algorithm for BNs considers them to be independent (see Section 2) and returns unreliable loss distributions: the corresponding VaR values are consequently also unreliable.

7. Conclusion

A novel approach, based on Bayesian Networks, has been proposed for the quantitative management of Operational Risk in the framework of The New Basel Capital Accord. The principal features of the proposed approach are the following: 1) the whole topology of the network is derived from data of operational losses; each node in the network corresponds to a bank process and the links between the nodes, which are drawn learning from data, model the causal relationships between the processes; this scheme seems more flexible than the classification in 56 pairs (BL, LET) prescribed by Basel II and has the advantage of representing both the units that generate operational losses and the relationships between them. 2) For the first time a Bayesian Network is used to represent the influence between correlated operational losses that take place in different days exploiting a dataset whose records represent losses occurred over T days: using such a dataset the nodes in the network represent the aggregate loss over T and the VaR over a time horizon T can be computed. The extension to the VaR over the time horizon L requires an additional assumption (see Section 4) and is performed by convoluting the probability density functions $\frac{L}{T}$ times and extracting the 99.9 percentile of the convoluted distribution.

Acknowledgements

One of us (M. Bardoscia) would like to thank M.V. Carlucci for the countless suggestions and useful disscussions.

References

[1] A.J. McNeil, R. Frey, P. Embrechts, *Quantitative Risk Management* (Princeton University Press, Princeton, 2005)

[2] R.N. Mantegna, H.E. Stanley, *An Introduction to Econophysics: Correlation and Complexity in Finance* (Cambridge University Press, Cambridge, 2000)

[3] J-P. Bouchard, M. Potters, *Theory of Financial Risk and Derivative Pricing, From Statistical Physics to Risk Management* (Cambridge University Press, Cambridge, 2003)

[4] J.L. McCauley, *Dynamics of Markets* (Cambridge University Press, Cambridge, 2004)

[5] B.K. Chakrabarti, A. Chakraborti, A. Chatterjee, *Econophysics and Sociophysics: Trends and Prespectives* (Wiley-VHC, London, 2006)

[6] N. Basalto, R. Bellotti, F. De Carlo, P. Facchi, E. Pantaleo, S. Pascazio, *Physical Review E*, **78**, 046112 (2008)

[7] J. Voit, *Physica A*, **321**, 286 (2003)

[8] R. Khun, K. Anand, *Physical Review E*, **75**, 016111 (2007)

[9] Basel Committee on Banking Supervision, *International convergence of capital measurement and capital standards*, Bank for International Settlements Press & Communications, 2005

[10] M.G. Cruz, *Modeling, Measuring and Hedging Operational Risk* (Wiley, London, 2002)

[11] C. Cornalba, P. Giudici, *Physica A*, **338**, 166 (2004)

[12] R.E. Neapolitan, *Learning Bayesian Networks* (Prentice Hall, New York, 2003)

[13] I.W. Hunter, R.E. Kearney, *Biological Cybernetics* **47**-2, 141 (1983)

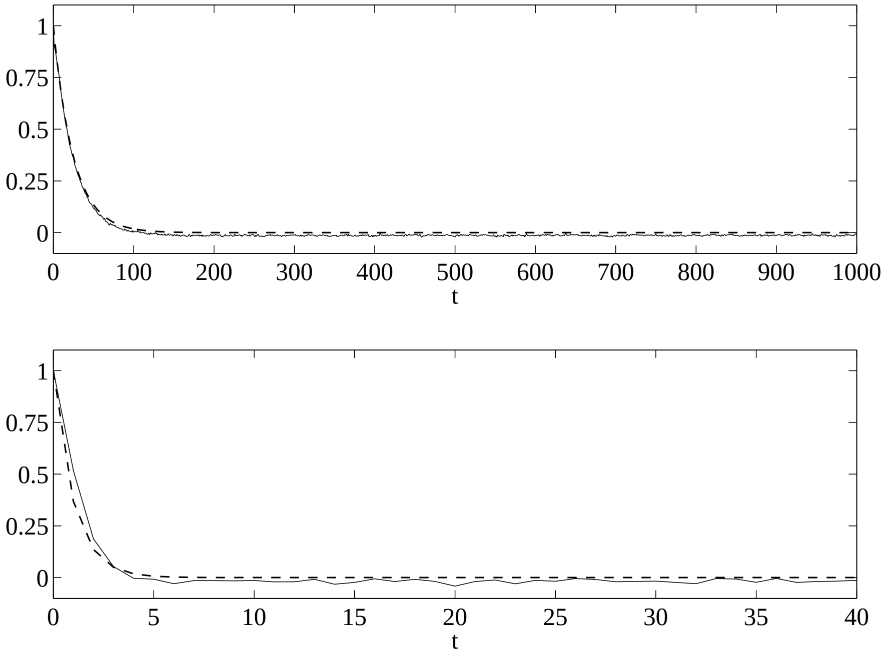


Figure 1: Top panel: imposed cross-correlation function C_{12} (dashed line) and obtained cross-correlation function c_{12} (solid line), with no average on time; for the sake of readability only the first 1000 values are shown. Bottom panel: imposed cross-correlation function C_{12} with scaled decay time $\frac{\tau}{25}$ (dashed line) and obtained cross-correlation function c_{12} (solid line) averaged over a time interval $T =$ 25; for the sake of readability only the first 40 values are shown.

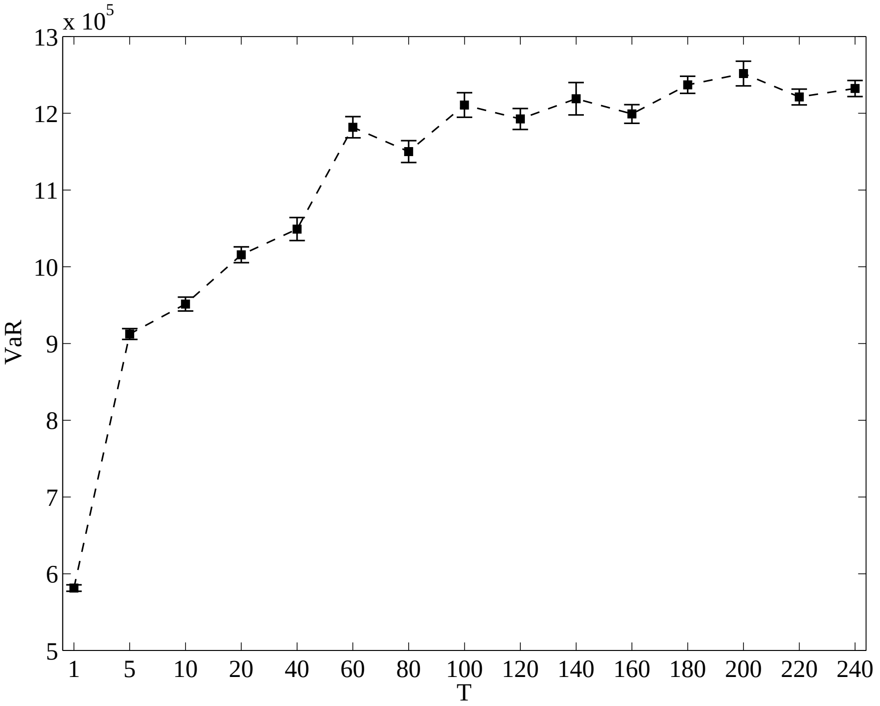


Figure 2: VaR with respect to the time interval T over which the average of the losses is performed; each point is the mean over 30 realizations of the procedure described in Section 4 and the error bars span over one standard deviation. For $T \geq 60$ the values of VaR are compatible among themselves. For $T < 60$ the values are not reliable because the records in the extracted database cannot be considered independent.

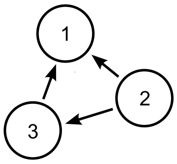
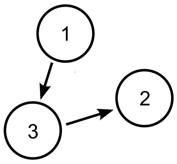
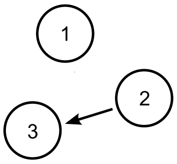
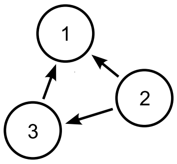
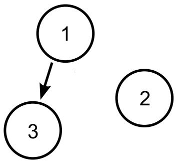
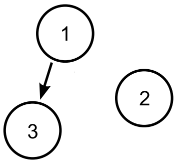
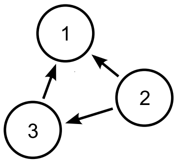
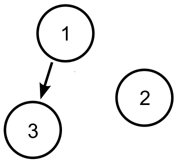
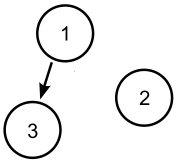
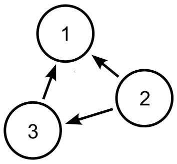
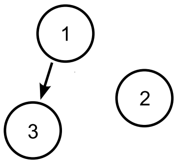
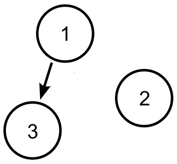
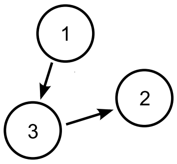
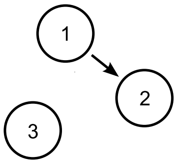
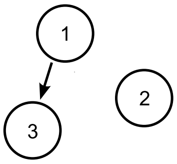
T		T		T	
1		60		160	
5		80		180	
10		100		200	
20		120		220	
40		140		240	

Table 1: The topology of BNs as the time interval T over which the losses are averaged varies. The links are more difficult to detect as T increases because the size of the extracted database used for the structure learning reduces. For $T \simeq 60$ the different-times correlations are reduced to 0.1, while for $T \geq 80$ they are extinguished and only the same-time correlations remain.