The Linear Conditional Probability Matrix Generator for IT Governance Performance Prediction

Mårten Simonsson, Robert Lagerström and Pontus Johnson

Department of Industrial Information and Control Systems, Osquldas väg 12, KTH, Royal Institute of Technology, 100 44 Stockholm, Sweden

{ms101, robertl, pj101}@ics.kth.se

Abstract. The goal of IT governance is not only to achieve internal efficiency in an IT organization, but also to support IT’s role as a business enabler. The latter is here denoted IT governance performance, and cannot be controlled by IT management directly. Their realm of control includes IT governance maturity, indicated by e.g. different IT activities, documents, metrics and roles. Current IT governance frameworks are suitable for describing IT governance, but lack the ability to predict how changes to the IT governance maturity indicators affect the IT governance performance. This paper presents a Bayesian network for IT governance performance prediction, learned with experience from 35 case studies. The network learns using the proposed Linear Conditional Probability Matrix Generator. The resulting Bayesian network for IT governance performance prediction can be used to support IT governance decision-making.

1 Introduction to IT Governance

The concept of IT governance emerged in the nineties. Henderson & Venkatraman [6] and Loh & Venkatraman [15] then used the term to describe the array of interfirm relationships involved in achieving strategic alignment of business and IT. Effective IT governance provides mechanisms that enable IS/IT management to develop integrated business and IT plans, allocate responsibilities, and prioritize IT initiatives [12],[23],[29]. It is important to ensure that the IT governance is not only designed to achieve internal efficiency in the IT organization, such as deploying good IT processes and making sure that the means and goals are documented. The final goal of good IT governance is to provide the business with the best support needed in order to conduct business in a good manner. The IT governance mechanisms should be chosen so that the impact on the business is maximized. There are many activities in the IT organization that can be changed, but clearly, not all changes affect the business in a positive way. From an IT manager’s point of view, it would be of great interest to know what impact each change made to the IT organization would have on the business, in order to choose the most beneficial way to govern IT.

There already exist several frameworks aiming to support IT governance. Weill & Ross have developed an IT governance framework based on just a few questions that can be used to assign responsibilities for high level IT decision making, but their work gives no further guidance on how the IT organization should actually transform
theory into practice [30]. The ISO/IEC 20000 and its predecessor IT Infrastructure Library (ITIL) might aid the creation of processes related to delivery and support [8],[19],[20]. ITIL also details establishment and maintenance of service level agreements (SLA). ITIL has traditionally provided little support for strategic IT concerns. However, this has been improved in recent ITIL v3 publications. Currently, the Control Objectives for Information and related Technology (COBIT) is most well-known framework for IT governance improvement, risk mitigation, IT value delivery and strategic alignment maturity assessments [2],[5],[7],[22],[28]. The COBIT framework was first issued by the IT Governance Institute, ITGI, in 1998 [9]. It describes the IT organization by means of 34 processes, within four domains: Plan & Organize, Acquire & Implement, Deliver & Support, and Monitor & Evaluate. A recent addition to COBIT is the Val IT framework, taking IT governance onto a higher level of abstraction by providing general directions on how to manage IT from a business point of view [10].

In this paper, the term IT governance performance is used to describe the goodness of an enterprise’s IT organization from a business point of view. The frameworks presented in the paragraph above are mainly of descriptive nature, i.e. they describe the state of an IT organization according to best practice on IT governance or IT management. None of them has however the ability to foresee how the IT governance performance is linked to the maturity of the IT organization in terms of its activities, level of documentation, etc. The purpose of this paper is to propose a method for prediction of IT governance performance within an enterprise. In particular, by using such method, it is possible to compare the current state with future scenarios. For instance, if the decision-making authority for acquisition of commodity software is moved from business unit level to IT operations level, how would that affect the IT governance performance? Making such predictions also enables prescription, i.e. not only evaluating different scenarios, but also to chose rationally between them. Figure 1 shows a conceptual view of the model for IT governance performance prediction proposed in this paper. On the left side, there is the actual, intrinsic, IT governance performance, as seen from the business point of view. Clearly, the aim of any organization would be to improve the IT governance performance to increase stakeholder satisfaction and make sure that business runs as smoothly as possible. The IT governance performance is not directly controllable by IT management, but IT processes for e.g. hardware acquisition, IT project management and IT strategy are in the realm of control.

![Figure 1. The conceptual model for IT governance performance prediction. The IT governance performance, as seen from the business viewpoint, is not directly controllable. Within the realms of control for IT management are IT processes and IT governance maturity indicators.](image-url)
Such IT processes are difficult to measure directly, but they comprise numerous and measurable IT governance maturity indicators, including maturity of individual IT related activities, level of monitoring, level of documentation, level of role assignment, etc. It is reasonable to believe that some of the IT governance maturity indicators are correlated with the intrinsic IT governance performance, even though it might be hard to establish the strength of the correlation. A model for prediction of IT governance performance would need to take into account and define the impact of each one of the IT governance performance indicators.

Several prediction methods are used in the research community today, including Dempster-Shafer, Bayesian networks, neural networks, and multivariate analysis. Of the above presented methods, Bayesian networks fulfill most requirements as presented in [12]. Therefore, such networks are used for IT governance maturity prediction in this paper.

2 Bayesian Networks

Friedman describes a Bayesian network, \( B=(G, P) \), as a representation of a joint probability distribution, where \( G=(V, E) \) is a directed acyclic graph consisting of vertices, \( V \), and edges, \( E \) [3],[4]. The vertices denote a domain of random variables \( X_1, \ldots, X_n \), also denoted chance nodes. Each chance node, \( X_i \), may take on a value \( x_i \) from the finite domain \( \text{Val}(X_i) \). The edges denote causal dependencies between the nodes, i.e. how the nodes relate to each other. The second component, \( P \), of the network \( B \), describes a conditional probability distribution for each chance node, \( P(X_i) \), given its parents \( \text{Pa}(X_i) \) in \( G \). It is possible to write the joint probability distribution of the domain \( X_1, \ldots, X_n \) using the chain rule of probability, in the product form

\[
p(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{Pa}(X_i))
\]

In order to specify the joint distribution, the respective conditional probabilities that appear in the product form must be found. The second component \( P \) describes distributions \( P(x_i | \text{pa}(X_i)) \) for each possible value \( x_i \) of \( X_i \) and \( \text{pa}(X_i) \) of \( \text{Pa}(X_i) \), where \( \text{pa}(X_i) \) is the set of values of \( \text{Pa}(X_i) \). These conditional probabilities are represented in matrices, here on called conditional probability matrices (CPMs). Using a Bayesian network, it is possible to answer questions such as what is the probability of \( X = x_1 \) given that \( Y = y_2 \) and \( Z = z_1 \). An example of a Bayesian network and a CPM representing the chance nodes \( X, Y, \) and \( Z \) is shown in Figure 2. The CPM next to the network answers the question \( P(x_1 | y_2, z_1) \) stated above. More comprehensive treatment on Bayesian networks can be found in the literature [11], [18], [21], [24]. The generic process for constructing Bayesian networks consists of three steps to define the nodes, relations and conditional probability matrices. These are described in the context of IT governance performance prediction in the following sections.

![Figure 2. A Bayesian network and the conditional probability matrix for X given Y and Z.](image-url)
3 Defining Nodes

It was decided to base the predictive Bayesian network on COBIT, as the framework provides the most relevant and detailed support for IT governance. The motivation for this choice can be found in previous publications [25],[26].

The concept of IT governance as consisting of processes, activities, roles, documents and metrics was adopted from COBIT. The use of processes to describe an IT organization is commonly employed in many frameworks, and is also used in the herein proposed approach. Further, each process contains one or more activities, which represent the actual content of the work performed within the IT organization. The documents correspond to process inputs and outputs as stated in COBIT. Metrics are used to monitor the execution of each process, and a representation for metrics monitoring is also incorporated. The concept of Roles being responsible, accountable, consulted or informed on the execution of different activities is also incorporated. The role representation features the distinction between executives, business and IT as stated by Weill & Ross and IT Governance Institute, [10], [30] but also employs IT operations and audit roles taken from COBIT [9].

Indicators for IT governance maturity, as seen from IT’s viewpoint, can be obtained by gathering information on the above mentioned entities for each IT process. Then, the activity execution (A), metrics monitoring (M), documents in place (D), and the responsibility assignment (R) can be evaluated according to previous work by Simonsson [26]. These four are represented as chance nodes with maturity levels ml0-ml5 in the Bayesian network for IT governance performance prediction. The entire IT organization is represented by means of 136 different nodes, so called IT governance maturity indicators, that together form the 34 processes detailed in COBIT. As mentioned earlier the purpose of the Bayesian network is not solely to study the IT organization in terms of controllable maturity indicators but also to predict the uncontrollable business perception of IT governance performance by studying the controllable IT governance maturity indicators, cf. Figure 1. Weill & Ross have previously determined IT governance performance in 250 organizations by means of letting senior management judge their organization’s performance with respect to two objectives [30], cf. Table 1. The same objectives have been used in the research presented in this paper.

Table 1. Objectives employed in order to represent the IT governance performance node in the Bayesian network [30].

<table>
<thead>
<tr>
<th>O1. How important are the following outcomes of your IT governance, on a scale from 1 (not important) to 5 (very important)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost effective use of IT</td>
</tr>
<tr>
<td>Effective use of IT for growth</td>
</tr>
<tr>
<td>Effective use of IT for asset utilization</td>
</tr>
<tr>
<td>Effective use of IT for business flexibility</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>O2. What is the influence of IT governance in your business on the following measures of success, on a scale from 1 (not successful) to 5 (very successful)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost effective use of IT</td>
</tr>
<tr>
<td>Effective use of IT for growth</td>
</tr>
<tr>
<td>Effective use of IT for asset utilization</td>
</tr>
<tr>
<td>Effective use of IT for business flexibility</td>
</tr>
</tbody>
</table>
Weill & Ross’ objectives were aggregated and transformed into performance levels p0-p15 for the node ITG_Performance according to Formula (1).

\[
ITG\_\_Performance = 1.25 \left( \frac{\sum_{i=1}^{14} Q_i \cdot \alpha^2}{\sum_{i=1}^{14} Q_i} - 1 \right)
\]  

(1)

4 Defining Relations

In order to use a Bayesian network for predictions, not only knowledge about different nodes and their maturity levels or intrinsic performance levels is needed. It is also necessary to define how the nodes are related. The processes in COBIT well delimit the scope of the work performed by the IT organization [26]. The IT processes are controllable by IT management, and it is reasonable to believe that the maturity of an IT process is causally linked to the IT governance performance, but the IT processes are not measurable. However, each process consists of different activities, documents, metrics, roles and responsibilities. These are controllable by IT management and measurable in terms of IT governance maturity indicators. A causal relation exists between the IT governance maturity indicators, and the IT processes.

By determining only the maturity of one or more of the indicators as discussed in Section 3, the network can predict IT governance performance, which is the idea behind the Bayesian network, cf. Figure 3. The strength of the causal relations in the network is mathematically described as CPMs.

5 Defining Conditional Probability Matrices

The CPMs defining the chance nodes in the network must be learned, i.e. the parameters in the matrices need to be determined. The basic approach is to collect empirical data for the nodes by conducting case studies and then use Bayesian network learning algorithms to assign the parameters to the matrices.

5.1 Requirements on methods for learning Bayesian networks

There are several methods for learning Bayesian networks. This subsection presents a set of requirements that has been used in order to evaluate the four methods in focus of this paper. Performing case studies is a time consuming activity, which is also highly dependent on the number of accessible cases. The amount of data sets that the
network can learn from is often limited. Therefore, the learning method must be able
to obtain conditional probabilities based on a fairly small number of datasets. As
discussed previously, the structure of the network has already been determined and a
learning method should not change it. In other words, the user herself should be able
to choose network structure. The desired ability to learn conditional probabilities
without changing the structure of the network is denoted parameter learning. Finally,
it is deemed that the output of the method should be a conditional probability matrix.

5.2 Evaluation of methods for learning Bayesian networks

Four methods have been evaluated, including the Expectation Maximization (EM), B-
Course, Path Condition (PC), and the Necessary Path Condition (NPC) algorithms.
The evaluation is focused on the requirements presented in the previous subsection,
namely support for learning from a limited number of data sets, parameter learning,
user choice of structure, and method outcome.

The most common method for learning Bayesian networks with statistical data is
called the EM algorithm \([3],[14]\). The main disadvantage with EM learning is that,
when using only a small number of datasets, the learning will result in conditional
probability matrices with zero entries. This means that if a set of values has not ap-
peared in any of the learning cases, the set cannot be used for prediction.

B-course is a web-based online data analysis method proposed by Myllymäki
\([17]\) that allows the user to analyze data for multivariate probabilistic dependencies.
The outcome of the method is a Bayesian network structure with learned conditional
probability matrices. The main drawback with this method is that it learns the struc-
ture from data. It is not possible to force an already set structure upon B-course and
only learn the parameters of the variables. B-course also requires large number of
datasets to provide useable conditional probability matrices.

The PC algorithm is a constraint-based learning algorithm. This means that the al-
gorithm uses statistical tests to derive a set of conditional independent and dependent
statements, and learns the structure of a Bayesian network. The NPC algorithm is an
enhancement of the PC algorithm which intends to bridge the latter’s deficiencies in
learning from small number of data sets. Both algorithms have the disadvantage that
their outcomes are structures and not CPMs \([16]\).

To summarize, the main requirement is that the method should be able to learn
parameters based on collected data, which excludes the PC and NPC algorithms.
Since data collection in the case of IT governance performance prediction is made
through case studies, a key requirement is the limited amount of data sets available.
Neither the EM algorithm nor B-course addresses this issue. The result of the evalu-
ation of the methods and the proposed approach is visualized in

Table 2. Linear regression is a commonly used method for prediction of the out-
come of one variable based on the information of other variables \([1],[27]\). It may thus
also be appropriate for learning conditional probability matrices in Bayesian net-
works. The main weakness of this approach is that the outcome of a linear regression
is not a conditional probability matrix, but rather an equation \(y=ax+b\). However, if the
outcome of linear regression could be translated into a CPM, the approach would be
appropriate for our purposes.
5.3 The Linear Conditional Probability Matrix Generator

Unfortunately, no linear learning approach with conditional probability matrices as outcome exists in the readily available tools for Bayesian statistics. Therefore, the Linear Conditional Probability Matrix Generator (LCPMG) was developed. In general, LCPMG takes into account gathered observation data, processes it, and returns a conditional probability matrix made with an assumption of linearity in the input data. The generator works according the following steps:

Observations on a quantitative scale of measurement are made and a structure is decided upon. For pedagogic purposes, assume 20 observations of the variable $X$ and 20 simultaneous observations of $Y$. The nodes $X$ and $Y$ are causally related to one another in the network, $X$ affecting $Y$. The observed values are on a continuous scale $x_i \in [0..5]$ and $y_i \in [0..5]$. The choice of scales of measure is due to the equally graded scales for IT governance maturity assessment used in the COBIT framework [9].

A linear regression on the observations is performed according to standard procedures described by e.g. Cohen and Walsh [1], [27]. The result is an equation $Y_{\text{estimate}} = aX + b$, where $a$ and $b$ are scalar constants. The residuals constitute the difference between the linear approximation that is fitted to the observations $(x_i, y_i)$, and the actual observations $R = Y - Y_{\text{estimate}}$. The standard deviation $S$ of the residuals, an approximation of the certainty with which the linear approximation is made, is calculated. The purpose of the LCPMG is to generate a discrete CPM from a continuous linear approximation $Y_{\text{estimate}}$. In order to do that, six different ranges for $y_i$ are created:

$-0.5 \leq y_1 < 0.5 \leq y_2 < 1.5 \leq y_3 < 2.5 \leq y_4 \leq 3.5 \leq y_5 < 4.5 \leq y_6 < 5.5$. 

Based on the linear approximation and the standard deviation $S$, the probability mass $P(y_i | x_i)$ in each cell of the CPM is calculated. The total probability $\sum_{y_i} P(y_i | x_i) = 1$ for each $x_i \in [0,1,2,3,4,5]$. As an example, if the linear approximation is $Y_{\text{estimate}} = 0.5x + 1$, and the standard deviation for the residual vectors equals 0.5, this corresponds to $P(y_i | x_i) = 0.6827 \%$ [1]. In summary, the LCPMG has now transformed two arrays with observations on a continuous scale, to a CPM describing the causal relation between two nodes in a Bayesian network. Returning to the requirements on methods for learning Bayesian networks and comparing the LCPMG to other already available methods one finds that the LCPMG fulfills all four requirements. Table 2 shows a final comparison of some common learning approaches for Bayesian networks and the here proposed LCPMG.

Table 2. A comparison of different learning approaches for Bayesian networks.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>EM Algorithm</th>
<th>B-Course</th>
<th>PC Algorithm</th>
<th>NPC Algorithm</th>
<th>LCPMG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for learning from a limited number of data sets</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Parameter learning</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>User choice of structure</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Method outcome</td>
<td>CPM</td>
<td>CPM &amp; Structure</td>
<td>Structure</td>
<td>Structure</td>
<td>CPM</td>
</tr>
</tbody>
</table>
5.4 Using LCPMG for creation of a Bayesian network for IT governance performance prediction

The LCPMG is suitable for generating the CPMs of nodes that are linearly related to one another. In the case of IT governance maturity prediction, the Bayesian network has three hierarchical levels. The first level contains the measureable, yet not controllable IT governance performance node. The second level contains 34 IT process nodes that are controllable, but not measureable. On the third level, 136 measureable and controllable IT governance maturity indicator nodes reside, cf. Figure 1. The CPMs of all nodes at all levels must be defined, and the LCPMG can be applied stepwise in order for the network to learn the CPMs.

Calculate the regressions for all IT governance maturity indicator nodes and the IT governance performance node. Use the regressions to assign normalized weights $w_i$ to each of the four node types; activities, metrics, documents and responsibilities.

The maturity for an IT process, $m_p$, can be calculated as $m_p = w_1 \cdot ml_1 \cdot m_j + w_2 \cdot ml_2 \cdot m_j + w_3 \cdot ml_3 \cdot m_j + w_4 \cdot ml_4 \cdot m_j$. Calculate the $m_p$ for each of the N*34 IT processes, where N represents the number of different observations made. Use LCPMG to determine the CPMs for each of the 34 IT process nodes, based on the $m_p$:s and the ITG_Performance node. Use LCPMG to determine the CPMs for each of the 136 maturity indicator nodes, based on maturity levels for the maturity indicators, and the $m_p$:s. Finally, the prior of the ITG_Performance node is set by analyzing the occurrence of each one of the possible levels p0-p15.

![Figure 4. Calculations of IT process PO4’s maturity and observations of ITG_Performance](image)

S denotes the standard deviation of the residuals [1],[27]. A small S indicates a good fit to the linear model. If only a limited amount of datasets have been used in order for the Bayesian network to learn, all levels of ITG_Performance have perhaps not been observed. This can be corrected for by using Laplace’s estimation, i.e. add 1 to the number of observations assigned to each state [10]. In this way, no zeros will be present in the resulting CPM and it is thus resulting in a better and more smoothly predicting Bayesian network. Figure 4 shows observations for $Y = PO4$ (Define the IT processes, Organization and Relationships) and $X = ITG_Performance$, the linear approximation and a graphic representation of the probability mass for each cell in
the CPM. The darker red the color of a bubble, the higher the probability mass 
\( P(y|x) \) of the corresponding cell in the CPM.

### 6 Discussion & Conclusions

As of March 2008, about 160 interviews have been conducted within 35 different organizations. The collected data spans a variety of industries, including banks, the public sector, telecommunications, electric utilities and manufacturing. The LCPMG has been applied upon the collected data in order for the Bayesian network for IT Governance Performance to learn. In spite of the variety of empirical data, correlations between IT governance performance and IT governance maturity indicators are clearly visible, and the Bayesian network is already usable for making predictions. Based on the current sets of data it seems that the maturity indicators that most strongly correlate with IT governance performance do not differ among industries. In summary, this paper has been discussing the use of Bayesian networks for prediction of IT governance performance. The Linear Conditional Probability Matrix Generator, LCPMG, is proposed as a way for Bayesian networks to learn from small datasets. The resulting network can be employed to make well-informed decisions regarding IT governance performance. Finally, the authors would like to thank Professor Stefan Arnborg for his valuable input on Bayesian statistics.

### References