Transactions on Machine
Learning and Data Mining
Vol. 12, No. 2 (2019) 33 - 54

P-ISSN: 1865-6781, E-ISSN 2509-9337
ISBN: 978-3-942952-72-9

ibai Publishing
www.ibai-publishing.org

Applications of Bayesian Networks

Ron S. Kenett
KPA Ltd., Raanana and Samuel Neaman Institute, Technion, Israel
email: ron@kpa-group.com

Abstract. Modelling relationships between variables has been a major challenge for statisticians in a wide range of application areas. This capability is an essential component in an effort to generate information of high quality from a given data set. Bayesian Networks (BN) combine graphical analysis with Bayesian analysis to represent relations linking measured and target variables. Such graphical maps can be used for diagnostics and predictive analytics. The paper presents applications of Bayesian Networks to various domains such the evaluation of web site usability, the testing of web services, operational risks, biotechnology, customer satisfaction surveys, healthcare systems and an analysis of the impact of management style on statistical efficiency. These case studies provide complementary aspects of BN applications to emphasize the breath of potential applications and the various associated methodological challenges. Following the presentation of these case studies, a general section discusses various properties of Bayesian Networks, including the study of causality with BNs. Some references to software programs used to construct BNs are also provided. A concluding section summarises the paper main points and lists current research topics.

Keywords: Bayesian Networks, Directed Acyclic Graph, Conditional Probability Distribution, Information Quality.

1 Introduction to Bayesian Networks

Data analysis is about generating information from a given data set using applications of statistical methods. The quality of the information derived from data analysis is dependent on various dimensions, including the communication of results, the ability to translate results into actionable tasks and the capability to integrate various data sources [1]. This paper is about the application of Bayesian Networks to studies designed to derive high information quality from a given multivariate data set. We begin with an introduction to Bayesian Networks.

Bayesian Networks (BN) implement a graphical model structure known as a directed acyclic graph (DAG) that is popular in Statistics, Machine Learning and Artificial Intelligence. BN are both mathematically rigorous and intuitively understandable. They enable an effective representation and computation of the joint probability distribution over a set of random variables [2]. The structure of a DAG is defined by two sets: the set of nodes and the set of directed edges. The nodes represent random variables and are drawn as circles labelled by the variable-names. The edges represent links among the variables and are represented by arrows between nodes. In particular, an edge from node X_i to node X_j represents a relation between the corresponding variables. Thus, an arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i. This property is used to reduce, sometimes significantly, the number of parameters that are required to characterize the joint probability distribution (JPD) of the variables. This reduction provides an efficient way to compute the posterior probabilities given the evidence present in the data [3], [4], [5], [6], [7]. In addition to the DAG structure, which is often considered as the "qualitative" part of the model, one needs to specify the "quantitative" parameters of the model. These parameters are described by applying the Markov property, where the conditional probability distribution (CPD) at each node depends only on its parents. For discrete random variables, this conditional probability is often represented by a table, listing the local probability that a child node takes on each of the feasible values – for each combination of values of its parents. The joint distribution of a collection of variables can be determined uniquely by these local conditional probability tables (CPT). The case studies presented in this paper are focused on examples of discrete or discretized variables. This paper does not consider applications of BN to continuous data. In learning the network structure, one can include white lists of forced causality links imposed by expert opinion and blacklists of links that are not to be included in the network. Examples of Bayesian Networks are provided next. The case studies were chosen to demonstrate the breath of possible applications and motivate readers to apply BNs.

The case studies are described in detail in Section 2. Section 3 deals with the properties of the Bayesian networks. 4Software for Bayesian Network Applications are presented in Section 4. Summary and conclusions are given in Section 5.

2 The Case Studies

In applications of statistics one often needs to bridge between data analysis methodology and content expert`s knowledge and needs. Bayesian Networks offer a unique opportunity for conducting such a dialogue by combining graphical analysis and various statistical estimation methods. Two types of sensitivity and robustness studies will be mentioned. The first one relates to the BN DAG structure, the second one to the effect of upstream nodes on target variables. This section presents applications of BN to:

1. management efficiency [8],
2. web site usability [9],

3. operational risks [10],
4. biotechnology [11],
5. customer satisfaction surveys [12],
6. healthcare systems [13] and
7. testing of web services [14].

Section 3 presents methodological and theoretical aspects of Bayesian Networks.

2.1 Management: The Statistical Efficiency Conjecture

This case study is focused on the link between management maturity levels and the impact of statistical methods on process and product improvements and thereby on the competitive position of organizations. It shows benefits from process improvement and quality by design initiatives that can be implemented within and across organizations. The different approaches to the management of industrial organizations can be summarized and classified using a four steps Quality Ladder.
The four approaches are:

1. Fire Fighting,
2. Inspection,
3. Process Control and
4. Quality by Design and Strategic management [15].

To each management approach corresponds a particular set of statistical methods and the quality ladder maps management approach with the application of corresponding statistical methods. Managers involved in reactive fire-fighting need to be exposed to basic statistical thinking. Managers attempting to contain quality and inefficiency problems through inspection and 100% control can have their tasks alleviated by implementing sampling techniques. More proactive managers investing in process control and process improvement are well aware of the advantages of control chart and process control procedures. At the top of the quality ladder is the quality by design approach where up-front investments are secured in order to run experiments designed to impact product and process specifications. At that level, reliability engineering is performed routinely, and reliability estimates are compared to field returns data in order to monitor the actual performance of products and improve the organizations' predictive capability.

Efficient implementation of statistical methods requires a proper match between management approach and statistical tools. The case study in Kenett8, demonstrate the benefits achieved by organizations implementing process and quality improvements. The underlying theory behind the approach is that organizations that increase the maturity of their management system, moving from firefighting to quality by design, are experiencing increased benefits and significant improvements in their competitive position. A measure of practical statistical efficiency (PSE) is used to to assess the impact of problem-solving initiatives.

PSE involves 8 dimensions that can be assessed individually for specific projects. These are:

1. value of the data actually collected,
2. value of the statistical method employed (also called Pitman efficiency),
3. value of the problem to be solved,
4. probability that the problem actually gets solved,
5. value of the problem being solved,
6. probability the solution is actually implemented,
7. time the solution stays implemented and
8. expected number of replications.

As organizations move up the Quality Ladder, more useful data is collected, more significant projects get solved and solutions developed locally get replicated throughout the organization. Specifically, a BN analysis shows that increasing an organization's maturity by going up the Quality Ladder results in higher PSE and increased benefits are experienced. The data consists of 21 projects conducted in companies of various size and type of activity. Figure 1a presents a Bayesian Network of the collected data. From this figure we note that, overall in the sample, 11% experienced very high PSE and 54% very low and low PSE. Figure 1b presents the fitted data, conditioned on a company being located at the highest Quality by Design maturity level. In this group, 50% have very low or low PSE and 17% very high PSE. In the companies at the Inspection maturity level, only 8% experienced very high PSE. These are only initial indications of such possible relationships and more data, under better control, needs to be collected to validate such patterns. This finding has been labeled "The Statistical Efficiency Conjecture", i.e. companies higher up on the Quality Ladder experience higher efficiencies in problem solving with statistical and analytical methods.

The goal of web usability diagnostics is to identify design deficiencies that hamper a positive navigation experience on a specific web site. To understand the user experience, we need to compare the user activity to the user expectation. Both are not directly available from the server log file, but can be estimated by appropriate processing.

A system flagging possible usability design deficiencies requires a statistical model of server log data. How can we tell whether visitors encounter difficulty in exploring a particular page, and if so, what are the causes for this experience? We assume that the site visitors are task driven, but we do not know if the visitors' goals are related to a specific website. It may be that the visitors are simply exploring the site, or that they follow a procedure to accomplish a task. Yet, their behaviors reflect their perceptions of the site contents, and estimates of their effort in subsequent site investigation. Server logs provide time stamps for all hits, including those of page html text files, but also of image files and scripts used for the page display.

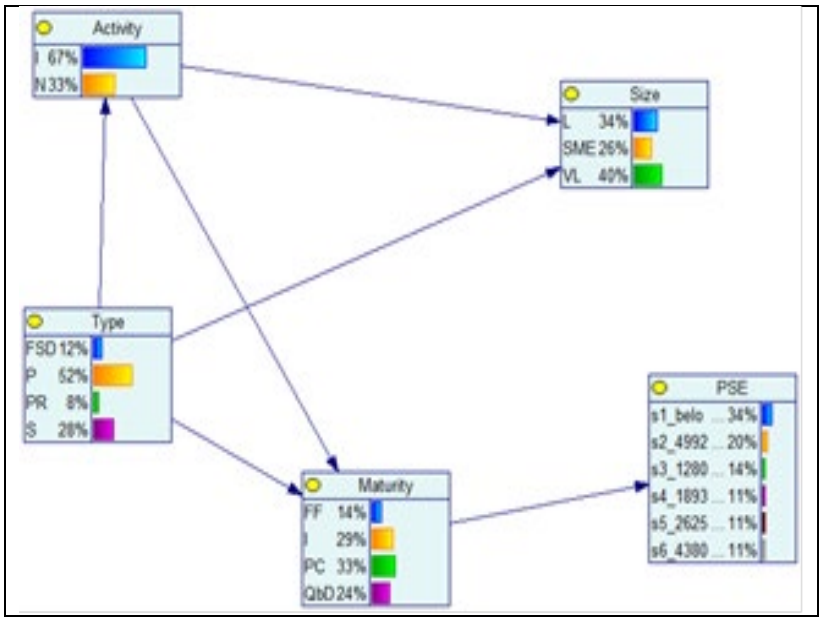

Fig. 1a. Bayesian Network of 21 companies in case study, Web Usability: Modeling Data Driven User Experience

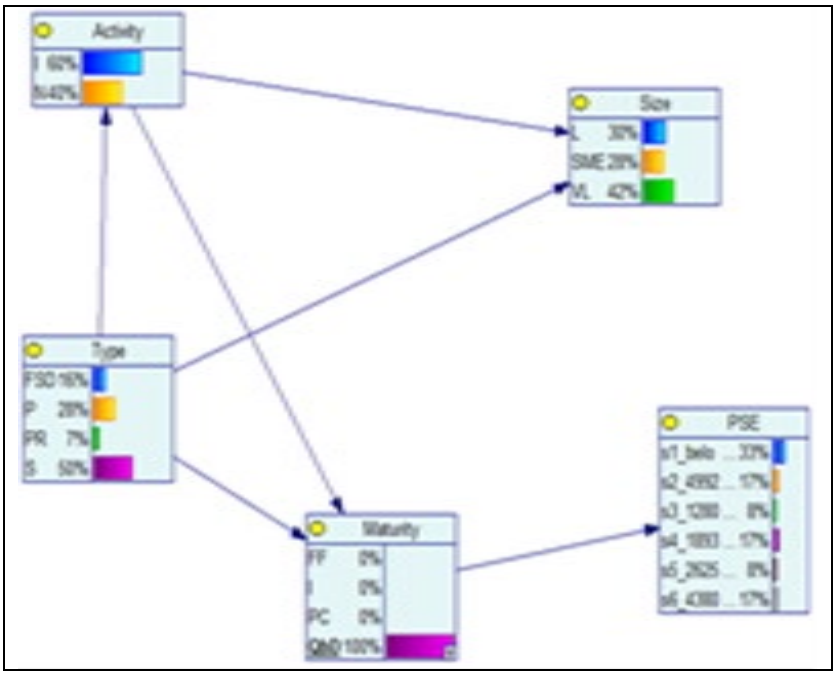

Fig. 1b. Bayesian Network conditioned on Quality by Design (QbD) maturity level.

The time stamps of the additional files enable us to estimate three important time intervals:

1. The time the visitors wait until the beginning of the file download, this is used as a measure of page responsiveness
2. The download time, this is used as a measure of page performance
3. The time from download completion to the visitor's request for the next page. This is a time interval where the visitor reads the page content, but also does other things, some of them unrelated to the page content.

The challenge is to decide, based on statistics of these time intervals, whether the visitors feel comfortable during the site navigation; when do they feel that they wait too long for the page download, and how do they feel about what they see on screen. To enable statistical analysis, we need to consider how the download time depends on the exit behavior. A Bayesian Network derived from analysis of web log analyzers is presented in Figure 2 which shows the links between measured variables. The network variables include page size, seeking time, download time and other statistics characterizing the web surfer's experience. The network itself has been learned from the data using algorithms presented in section 3. We can use the network for predicting behavior after conditioning on specific variables. For example, by conditioning the UsabilityScore to be at its lowest level we find that the percent of events with low AvgReadingTime dropped from 25% (without conditioning) to 13%. This drop on short reading times has a significant impact on usability score. For more details on this study see [l9].

2.2 ICT Operational Risks: Sensitivity Analysis of a Bayesian Network

In information and communication technology (ICT) operational risk analysis, different data sets are typically merged. Examples of such data sources include CRM call centre data, financial data from companies and log data utilized to monitor the provisioning of IT services [10]. Discrete variables BNs can easily merge ordinal qualitative data, such as severity impact of a system failure with nominal data, such as customer type and any discretised continuous data. the This case study is based on a firm marketing and operating telecommunication equipment in small, medium and large enterprises. The company records include, for each installed system, data such as "Customer type", "Number of extensions", "Number of smart phones", "Problem severity" and "Problem description". The data we analyse here consists of 4703 problems that occurred to a range of clients. The ordinal target variables is the "Severity" of loss due to the reported problem. Figure 3 presents the BN. On the basis of this BN, a sensitivity analysis was performed using statistically designed experiments (Cornalba16). Basically, we set up a variety of conditioning scenarios (i=1,…,120) using a full factorial experimental array. From a given BN with a set of variables set at prespecified levels, one can generate simulated outcomes (we used 1000 runs). Empirical goodness of fit (GoF) of a given BN model is computed using a distance measure between the simulated data and the real data. Among the various possible distance measures, we used a classification error defined as:

$\sum_{i=1}^{runs} I_{(s_i \neq \ddot{s}_i)}(x)$ where I is the indicator function of the subset of severities, s, of the set X and is defined as: $I = \begin{cases} 1 \text{ if } x \in \{s : s_i \neq \ddot{s}_i\} \\ 0 \text{ if } x \notin \{s : s_i \neq \ddot{s}_i\} \end{cases}$ wherее $\ddot{s}_i$ is the i-th simulated values of severity and s_i the i-th real data. As a response, we measure the GoF distribution, for a specific scenario.
The BN sensitivity analysis experiments involved three factors:

A: Lines - 4 levels;
B: Extensions - 5 levels;
C: Smart phones - 6 levels.
The full factorial experimental array consists of 120 experimental runs. For each run we get 1000 GoF responses which measure the fit of the model derived from comparing the conditioned BN simulation results with the original data. For each row, we compute the mean, trimmed mean, standard deviation, 5th and 95th percentile statistics of the empirical GoF distribution. Figure 4 presents main effects plot and contour plots for the GoF standard deviation. The higher the standard deviation, the less robust is the BN explicative capability.
The BN structure sensitivity presented in Figure 4 shows a highly dependency of the problem severity distribution on the number of smart phones and lines, and their interaction.

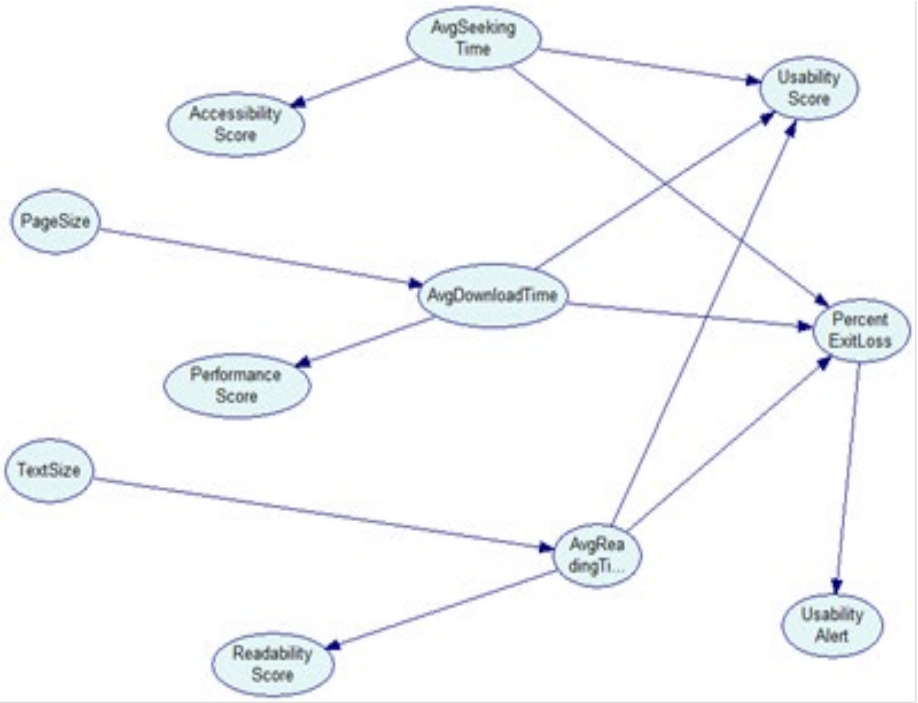

Fig. 2. Bayesian Network of weblog data

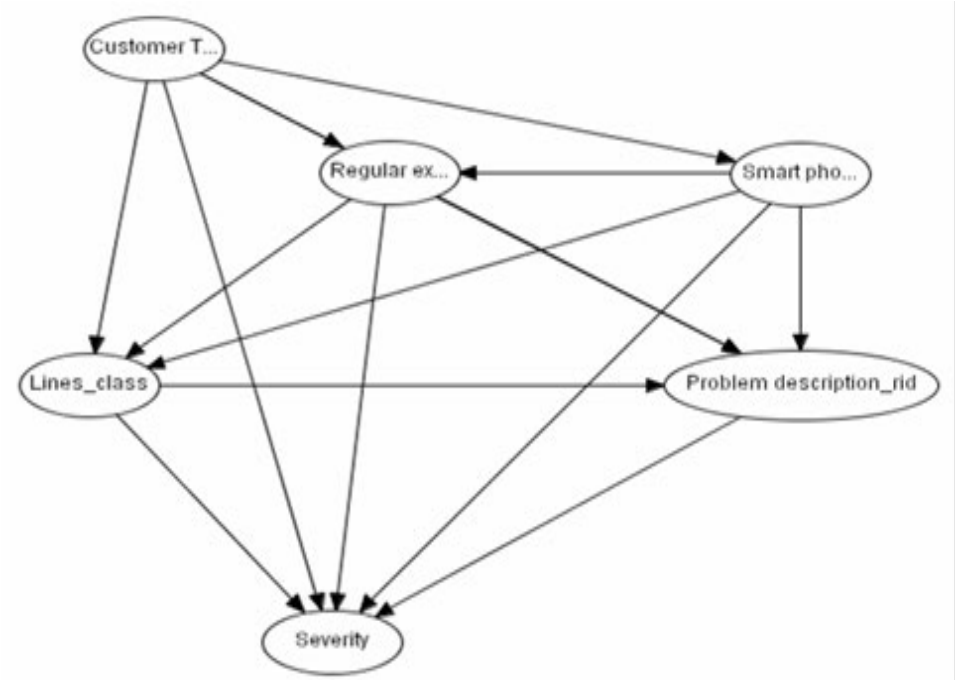

Fig. 3. Bayesian Network of ICT case study

2.3 Biotechnology: An Example of Tracking Performance over Time

In this case study, derived from [11], we consider 4 bioreactors operating in parallel for 4 weeks. Several amino acids in the medium composition are tracked periodically. These include: Taurine, Aspartic acid, Hydroxyproline, Threonine, Serine, Asparagine, Glutamic acid, Glutamine Proline, Glycine, Alanine, Valine, Cystine, Methionine, Isoleucine, Leucine, Tyrosine, Phenylalanine, Ornithine, Lysine, Histidine and Arginine. The control parameters include: IGF and levels of two control factors, A and B. The target variables consist of: Volumetric productivity, Ps, Titer, Max Cell and Diamid%. This example is an application of BN to analyse data collected over time. The goal is to generate insights on the behavior of the bioreactors for improved operation and monitoring. We aim at better understanding how control factors affect the target response variables so that we can optimize the process and generate early warning signals during production for mid-course corrections.

Figure 5 presents a BN conditioned on the first (5a, left) and last stage of operation (5b, right). Each node represents a discretized variable. Some variables are naturally discrete such as the bioreactor number or the week of operation labeled “Stage”. One can see that the variable “Stage” is affecting the composition of many of the amino acids. As an example, on the left panel of Figure 5 we can see that, according to the BN model, at Stage I the highest compositions of Isoleucine, Alanine and Arginine

correspond to 63%, 12% and 62% respectively. As we move to Stage IV (right panel), these numbers become, respectively, 13%, 25%, 12% with a dramatic drop in high values of Isoleucine and Arginine and an increase of 100% in the high values of Alanine. This example shows how a Bayesian Network can be used to predict an outcome when a process is set at a certain level of control parameters.

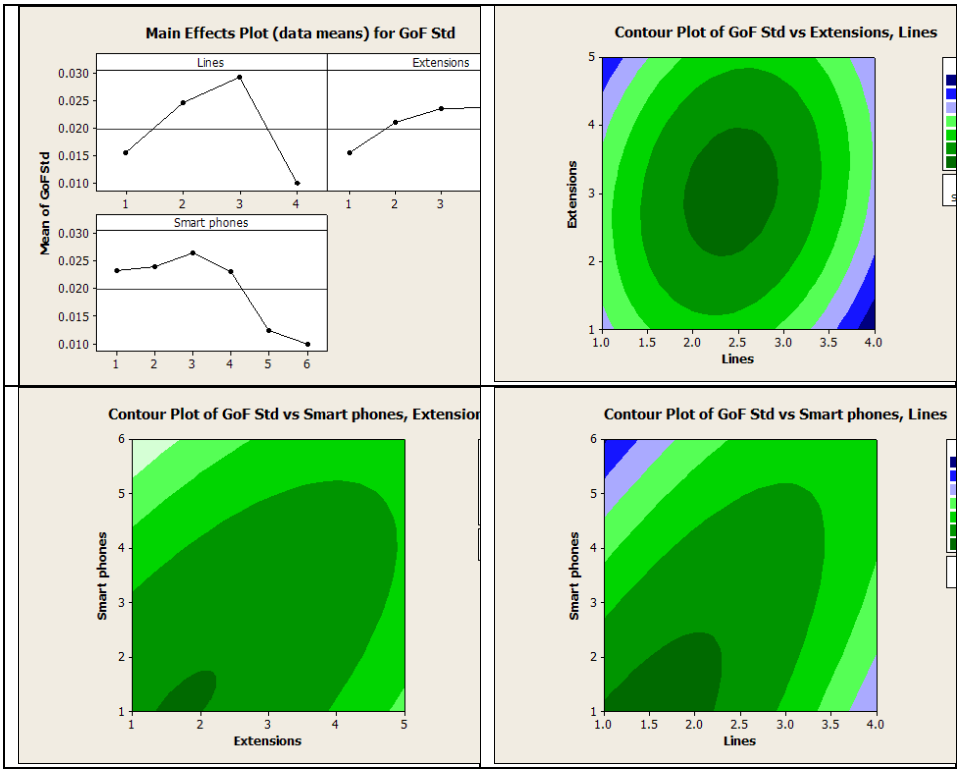

Fig. 4. Main effect plot and contour plots of mean GoF

2.4 Customer Surveys: Analysis of Ordinal Data

Self-declared or interview-based surveys are a prime research tool in social science research, customer management and marketing. In such surveys, target individuals are requested to fill in questionnaires which can have five or over one hundred questions [12]. Take for example an Annual Customer Satisfaction Survey directed at customers of an electronic product distributed world-wide. The survey is assessing satisfaction levels of customers from different features of the product and related services. The questionnaire is composed of 81 questions including demographics and overall satisfaction from the company. An important output of the survey is to determine which aspects of the product and services influence customer overall satisfaction, recommendation level and repurchasing intentions.

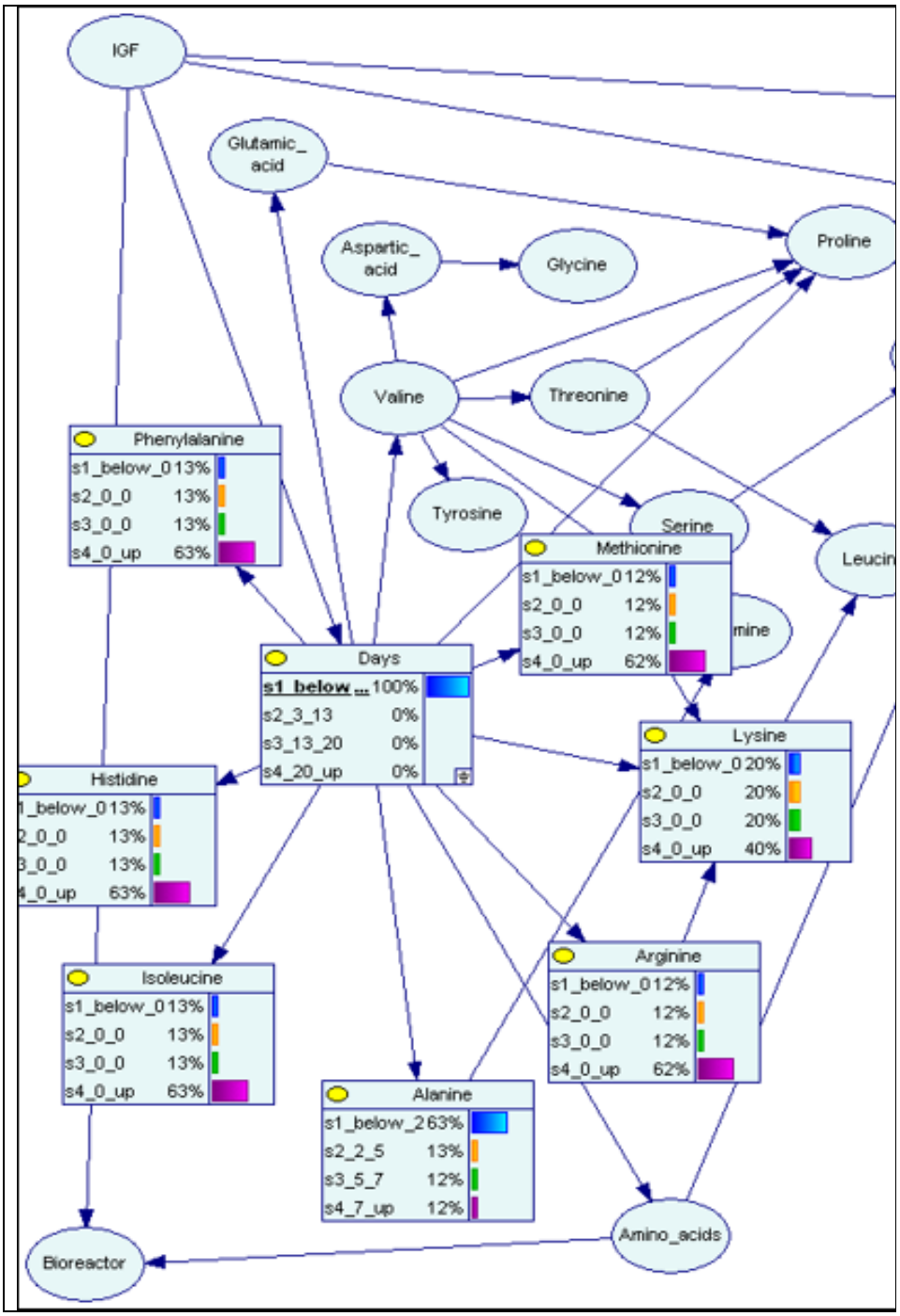

Fig. 5a. Bayesian Network of bioreactor data, conditioned on Stage I

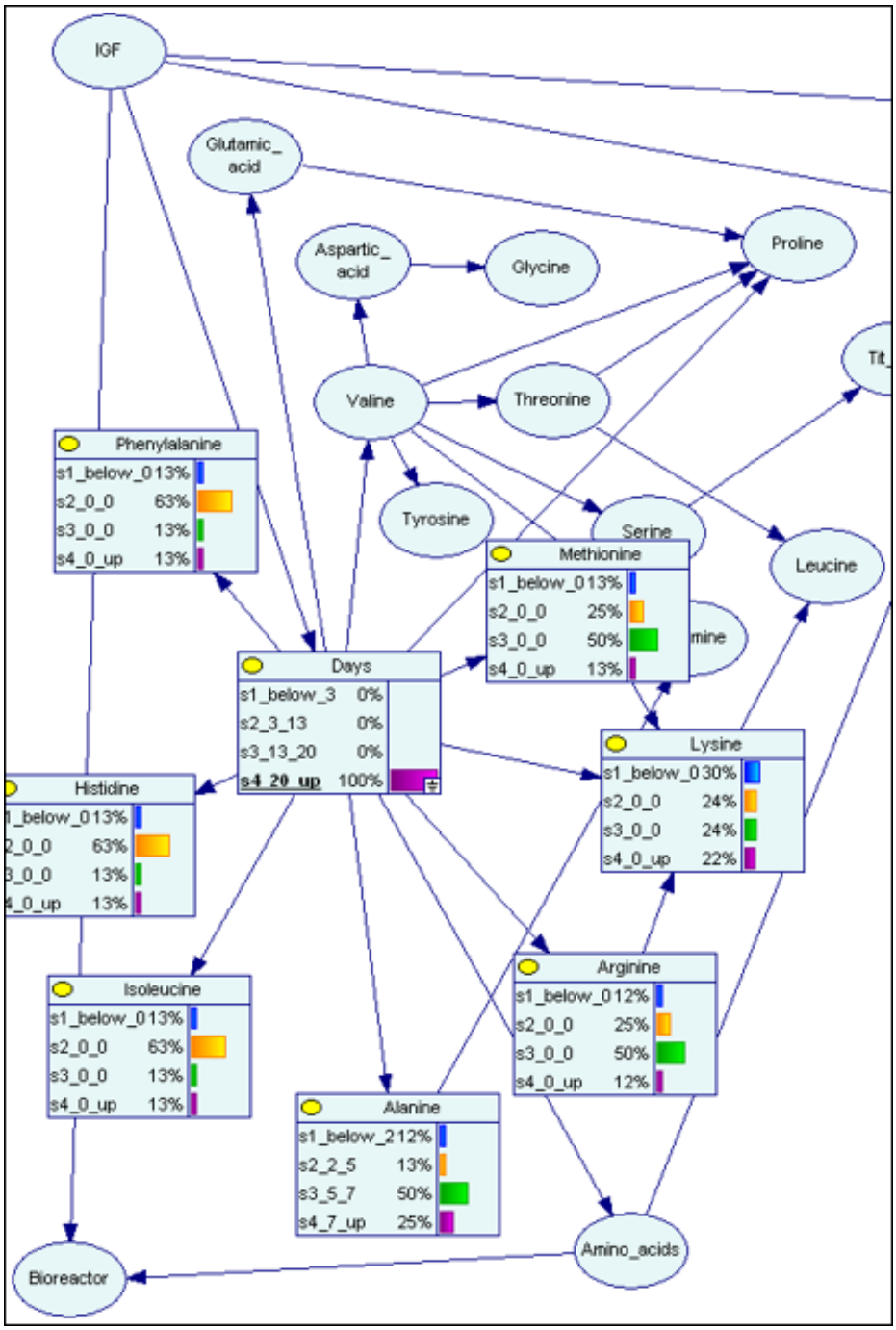

Fig. 5b. Bayesian Network conditioned on Stage IV

The topics covered by the survey include: Equipment, Sales Support, Technical Support, Training, Customer Portal, Administrative Support, Terms and Conditions and Site Planning and Installation. Demographic variables that can help profile customer responses include Country, Industry type and Age of equipment. A BN has been applied to data collected from 266 customers participating in an Annual Customer Satisfaction Survey (ACSS). As described above, the data refers to responses to a questionnaire composed of 81 questions. The BN derived from this data is presented in Figure 6. On the basis of the network, we can perform various diagnostic checks. For example, we can compute distribution of responses to various questions for customers who indicated that they are very likely to recommend the product to others. Such an analysis allows to profile loyal customers and design early warning indicators that predict customer dissatisfaction. More on this analysis in [12].

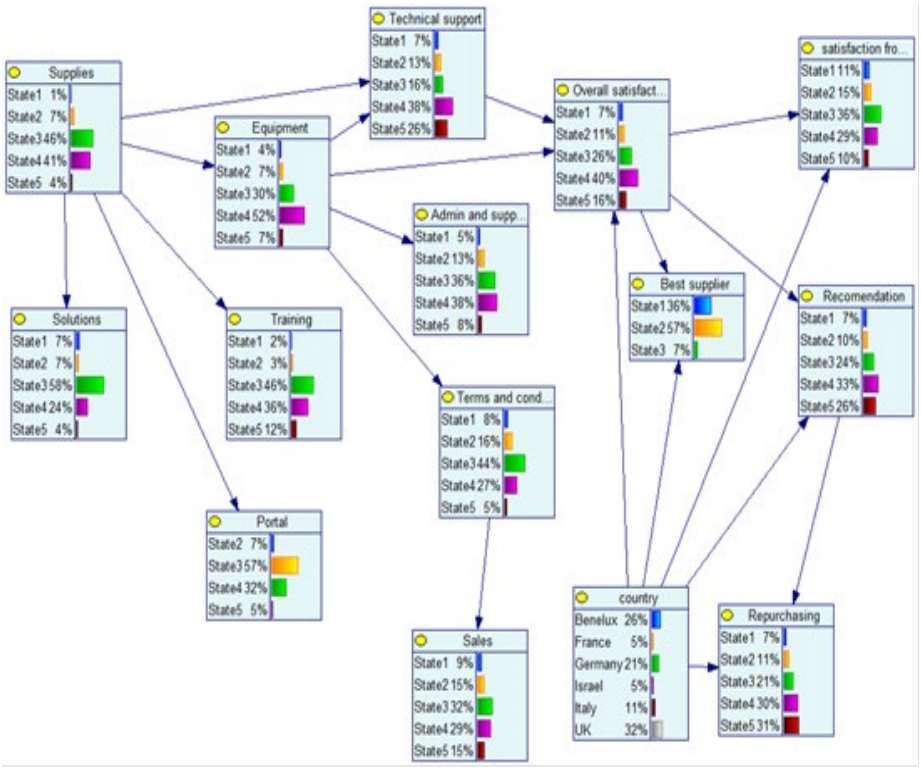

Fig. 6. Bayesian Network of responses to satisfaction questions from various topics, overall satisfaction, repurchasing intentions, recommendation-level and country of respondent.

2.5 Healthcare Systems: A Decision Support System Case Study

Health care organizations apply risk management as a key element in the improvement of service delivery and patient outcomes. In this context, clinical and operational risks are quantified and managed.

A risk assessment involves two elements that are:

1. a probability or frequency with which an event might take place and
2. an assessment of impact severity or consequences from such events.

Physicians usually evaluate and forecast adverse events that may provoke morbidity, mortality or a longer hospital stay for a patient. A patient risk profile is an assessment of a patient's medical parameters using probability distributions, given the patient's status and prior medical knowledge. Physicians typically summarize risk probability distributions through a percentile and on this basis decide acceptability of risks. Unacceptable risks are addressed by a detailed risk mitigation plan. Economic losses and costs due to adverse events are typically evaluated, mainly to choose convenient forms of insurance; furthermore, for better governance, it is useful to understand risk levels and how each risk contributes to economic losses.

When data is scarce, the experience of physicians often offers a good source of information. Bayesian methodology can be used for the estimation of operational and clinical risk profiles. The approach described in this case study includes an example involving health care management of End Stage Renal Disease (ESRD). The main goal is to support nephrologist and risk managers who have to manage operational and clinical risk in health care [13]. Decisions models have to be considered in order to realize a fully integrated risk management process. The integration between risk estimations and decision making can be achieved with Bayesian Networks.

Given medical parameters of a patient, $X_1, \ldots, Xn,$, a physician wants to estimate both mortality and hospitalization risk and the failure risks of a device. In the ESRD case, more than one target variable is analyzed under the hypothesis that X1, . . . , Xj ($j \leq n$) are positive dependent, with targets, and the combinations (Xi, Xk), where X1, . . . , Xq $q \leq j \leq n$, i k, are either positive dependent or independent. Dependencies and independence between variables are typically determined in medicine through scientific studies and clinical research. Different sources of knowledge such as subjective information (e.g. expert opinions of nephrologists, knowledge from literature) and data can be integrated with a BN. In our case, a BN offers several advantages: 1) the method allows to easily combine prior probability distributions, 2) the complexity of the ESRD domain and the relationships among medical parameters can be intuitively represented with graphs and, 3) utility or loss functions can be included in the model. The complex domain of ESRD is represented in [13] by a BN with 34 variables used to describe dialysis. Each variable being classified into one group of causes, such as Dialysis Quality Indexes = {Dialysis adequacy (Kt/V), PTH pg/ml, Serum albumin g/dl} and HD Department Performances = {Serum phosphorus PO4 mg/dl, Potassium mEq/l, Serum calcium mg/dl}.

From the BN one can determine that the most important adverse event is due to an incorrect dose of erythropoietin. Moreover, it is possible to also explore marginal posterior probability distributions of the target variables. For example, during the first update both of therapeutic protocol and data collection, the mortality risk of a patient increases. To restore the correct risk profile, the nephrologist can add a dose of erytropoietin. To complete the risk management process, physicians have to make deci-

sion either on patient's treatment or device's substitution. The set of decision d and the corresponding set of actions A for each decision are the following:

d1. "Time": keep or add 30 minutes to dialysis session.
d2. "Ca-based therapy": treat hypercalcemia; continue current therapy; decrease vitamin D dose to achieve ideal Ca; decrease Ca-based phosphate binders; decrease or discontinue vitamin D dose/Ca-based phosphate binders; decrease Ca dialysate if still needed; assess trend in serum PTH as there may be low turnover.
d3. "Phosphate binder": assess nutrition, discontinue phosphate binder if being used; being dietary counseling and restrict dietary phosphate; start or increase phosphate binder therapy; being short-term Al-based phosphate binder use, then increase non-Al based phosphate binder; begin dietary counseling and restrict dietary phosphate increase dialysis frequency.
d4. "Diet?": apply an "hypo" diet or keep his/her diet.
d5. "QB": increase, keep or decrease QB.
d6. "Erythropoietin": keep, decrease or increase (1 EPO) the current dose.
d7. "Iron management": keep the treatment; iron prescription.

Analyzing the most important causes and the consequence of each action, it is possible to assess each scenario and prioritize actions that should be taken for a given patient. With the approach presented in this section the physician can recommend the best treatment (for more on this topic see Kenett6 and references therein).

2.6 System Testing: Risk Based Group Testing

Testing is necessary to ensure the quality of web services that are loosely coupled, dynamic bound and integrated through standard protocols. Exhaustive testing of web services is usually impossible due to the unavailable source code, diversified user requirements and the large number of service combinations delivered by the open platform. This case study outlines a risk-based approach for selecting and prioritizing test cases to test service-based systems. We address here the problem in the context of semantic web services and analyze the service structure from various perspectives such as dependency, usage and service workflow. This is used to identify the factors that contribute to the risks of the services. Risks are assessed from two aspects: failure probability and importance. These are measured and predicted using BN analysis techniques. With this approach, test cases are associated to the service features and scheduled based on the risks of their target features. As a statistical testing technique, the proposed approach aims to detect, as early as possible, the problems with highest impact on the users. For more details on this big data application see [14].

3 Properties of Bayesian Networks

3.1 Parameter Learning

In order to fully specify a BN and thus fully represent the joint probability distribution it represents it is necessary to specify for each node X the probability distribution for X conditional upon X's parents. The distribution of X, conditional upon its parents may have any form. Sometimes only constraints on a distribution are known. One can then use the principle of maximum entropy to determine a single distribution, i.e. the one with the greatest entropy given the constraints [17].

Often these conditional distributions include parameters which are unknown and must be estimated from data, for example using the maximum likelihood approach. Direct maximization of the likelihood (or of the posterior probability) is often complex when there are unobserved variables. A classical approach to this problem is the expectation-maximization (E-M) algorithm which alternates computing expected values of the unobserved variables conditional on observed data, with maximizing the complete likelihood assuming that previously computed expected values are correct. Under mild regularity conditions this process converges on maximum likelihood (or maximum posterior) values for parameters [18[.

A more fully Bayesian approach to parameters is to treat parameters as additional unobserved variables and to compute a full posterior distribution over all nodes conditional upon observed data, then to integrate out the parameters. This approach can be expensive and lead to large dimension models, so in practice classical parameter-setting approaches are more common [19].

3.2 Structure Learning

BNs can be specified by expert knowledge (white lists and blacklists) with network structure learned from data. The parameters of the local distributions are learned from data, priors elicited from experts, or both. Learning the graph structure of a BN requires a scoring function and a search strategy. Common scoring functions include the posterior probability of the structure given the training data, the Bayesian information criteria (BIC) or Akaike information criteria (AIC). When fitting models, adding parameters increases the likelihood, which may result in over-fitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model with the penalty term being larger in BIC than in AIC. The time requirement of an exhaustive search returning back a structure that maximizes the score is super-exponential in the number of variables. A local search strategy makes incremental changes aimed at improving the score of the structure. A global search algorithm like Markov chain Monte Carlo can avoid getting trapped in local minima. For more on BN structure learning see [17].

3.3 Causality and Bayesian Networks

Causality analysis has been carried out from two main different points of view, the "probabilistic" view and the "mechanistic" view. Under the probabilistic view, the causal effect of an intervention is judged by comparing the evolution of the system when the intervention is and when it is not present. The other point of view focuses on understanding the mechanisms determining how specific effects come about. The interventionist and mechanistic viewpoints are not mutually exclusive. For examples, when studying biological systems scientists carry out experiments where they intervene on the system, for instance by adding a substance or by knocking out a gene. However, the effect of a medication introduced into the human body cannot be decided only in the laboratory. A mechanistic understanding based on pharmacometrics models is needed in order to determine if a certain medication ought to work and studied in order to elucidate biological mechanisms used to intervene and either prevent or cure disease.

This section provides a general introduction to causality analysis and its links to BN. The concept of potential outcomes is present in the work on randomized experiments by Fisher and Neyman in the 1920s and was then extended by Rubin in the 1970s to non-randomized studies and different modes of inference. In their work, causal effects are viewed as comparisons of potential outcomes, each corresponding to a level of the treatment and each observable, had the treatment taken on the corresponding level with at most one outcome actually observed, the one corresponding to the treatment level realized. In addition, the assignment mechanism needs to be explicitly defined as a probability model for how units receive the different treatment levels. With this perspective, a causal inference problem is viewed as a problem of missing data, where the assignment mechanism is explicitly modeled as a process for revealing the observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects [20].

In a recent paper, Imai et al [21] study how to design randomized experiments to identify causal mechanisms, i.e. a causal process through which the effect of a treatment on an outcome comes about. They study designs that are useful in situations where researchers can directly manipulate the intermediate variable that lies on the causal path from the treatment to the outcome. Such a variable is often referred to as a 'mediator'. Under the parallel design, each subject is randomly assigned to one of two experiments; in one experiment only the treatment variable is randomized whereas in the other both the treatment and the mediator are randomized. Under the crossover design, each experimental unit is sequentially assigned to two experiments where the first assignment is conducted randomly, and the subsequent assignment is determined without randomization on the basis of the treatment and mediator values in the previous experiment. They propose designs that permit the use of indirect and subtle manipulation. Under the parallel encouragement design, experimental subjects who are assigned to the second experiment are randomly encouraged to take (rather than assigned to) certain values of the mediator after the treatment has been randomized. Similarly, the crossover encouragement design employs randomized encouragement rather than the direct manipulation in the second experiment. These two designs gen-

eralize the parallel and crossover designs, allowing for imperfect manipulation, thus providing informative inferences about causal mechanisms by focusing on a subset of the population. The diagrams for causal mechanisms are similar to BNs.

The term 'causal inference' denotes different ways to approach causal aspects of statistical analysis. Causal Bayesian Networks are BNs where the effect of any intervention can be defined by a 'do' operator that separates intervention from conditioning. The basic idea is that intervention breaks the influence of a confounder so that one can make a true causal assessment. The established counterfactual definitions of direct and indirect effects depend on an ability to manipulate mediators. A BN like graphical representations, based on local independence graphs and dynamic path analysis, can be used to provide an overview of dynamic relations (Aalen22). On the other hand, the econometric approach develops explicit models of outcomes, where the causes of effects are investigated and the mechanisms governing the choice of treatment are analysed. In such investigations, counterfactuals are studied (Counterfactuals are possible outcomes in different hypothetical states of the world). The study of causality in studies of economic policies has involved:

(a) defining counterfactuals,
(b) identifying causal models from idealized data of population distributions and
(c) identifying causal models from actual data, where sampling variability is an issue [23].

Judea Pear, the 2011 Turing Medallist, developed BNs as the method of choice for uncertain reasoning in artificial intelligence and expert systems replacing earlier ad hoc rule based systems. His extensive work covered topics such as: causal calculus, counterfactuals, Do calculus, transportability, missingness graphs, causal mediation, graph mutilation and external validity [24]. In a heated head to head debate between probabilistic and mechanistic view, Pearl has taken strong standings against the probabilistic view, See for example [25] and [26]. The work in [22] and [21] show how these approaches can be used in complementary ways. The examples in this paper present BNs as descriptive tools that enhance the quality of information derived from a specific data set. Causal BNs involve additional analysis and data collection plans that were not considered here.

4 Software for Bayesian Network Applications

Packages for generating and analyzing Bayesian networks are:

- ***GeNIe*** (Graphical Network Interface) is the graphical interface to SMILE (Structural Modelling, Inference, and Learning Engine), a fully portable Bayesian inference engine developed by the Decision Systems Laboratory of the University of Pittsburgh and thoroughly field tested since 1998. GeNIe can be freely downloaded by academic institutions from https://www.bayesfusion.com/ with a user guide and related documentation.

- ***Hugin*** (http://www.hugin.com) is a commercial software which provides a variety of products for both research and non-academic use. The HUGIN Decision Engine (HDE) implements state-of-the-art algorithms for Bayesian networks and influence diagrams such as object-oriented modeling, learning from data with both continuous and discrete variables, value of information analysis, sensitivity analysis and data conflict analysis.
- ***IBM SPSS Modelle***r (http://www-01.ibm.com/software/analytics/spss/) is a general application for analytics that has incorporated the Hugin tool for running BNs (http://www.ibm.com/developerworks/library/wa-bayes1). IBM SPSS is not free software.
- The ***R bnlearn*** package is powerful and free. Compared with other available BN software programs, it is able to perform both constrained-based and score-based methods. It implements five constraint-based learning algorithms (Grow-Shrink, Incremental Association, Fast Incremental Association, Interleaved Incremental association, Max-min Parents and Children), two scored based learning algorithms (Hill-Climbing, TABU) and two hybrid algorithms (MMHC, Phase Restricted Maximization).
- ***Bayesia*** (http://www.bayesia.com/) developed proprietary technology for Bayesian network analysis. In collaboration with research labs and big research projects the company develops innovative technology solutions. Its products include 1) BayesiaLab, a Bayesian network publishing and automatic learning program which represents expert knowledge and allows one to find it among a mass of data, 2) Bayesia Market Simulator, a market simulation software package which can be used to compare the influence of a set of competing offers in relation to a defined population, 3) Bayesia Engines, a library of software components through which can integrate modelling and the use of Bayesian networks and 4) Bayesia Graph Layout Engine, a library of software components used to integrate the automatic position of graphs in specific application.
- ***SamIam*** (http://reasoning.cs.ucla.edu/samiam/) is a comprehensive tool for modeling and reasoning with Bayesian networks, developed in Java by the Automated Reasoning Group at UCLA. Samiam includes two main components: a graphical user interface and a reasoning engine. The graphical interface lets users develop Bayesian network models and save them in a variety of formats. The reasoning engine supports many tasks including: classical inference; parameter estimation; time-space tradeoffs; sensitivity analysis; and explanation-generation based on MAP and MPE.
- ***BNT*** (https://code.google.com/p/bnt/) supports many types of conditional probability distributions (nodes), decision and utility nodes, static and dynamic BNs and many different inference algorithms and methods for parameter learning. The source code is extensively documented, object-oriented, and free, making it an excellent tool for teaching, research and rapid prototyping.
- ***Agenarisk*** (http://www.agenarisk.com/) handles continuous nodes without the need for static discretization. It enables decision-makers to measure and compare different risks in a way that is repeatable and auditable and is ideal for risk scenario planning.

5 Summary and Conclusions

This paper presents examples of Bayesian Networks designed to illustrate their wide range of application. The examples complement each other and range from exploring a conjecture in management science, using 21 case studies, to modelling data driven user experience in an analysis of usability of web sites using web logs. In another example, the effect of system variables on a target variable representing severity of operational risk events are evaluated using statistical designed factorial experiments applied to a BN. This non-standard approach to sensitivity analysis can be incorporated in any application of BN. That case study also showed how various data sources and data types can be combined with BNs. Another case study deals with tracking a very large number of variables over time and demonstrates how BNs can be used to set up a control strategy by producing expected performance characteristics. Control is established by comparing actual performance to expected performance and acting on observed discrepancies. We also show how BNs can be used to analyse customer survey data, a natural application given the natural discretization of responses to a survey questionnaire. Other applications covered include a decision support system to help manage patients undergoing dialysis and a risk-based approach to test web services. The seven examples in section 2 demonstrate the application of BNs in:

1. Analysis of small data sets with few variables
2. Data driven modelling of user experience
3. Sensitivity analysis for determining robustness of a BN model and data integration
4. Analysis of large data sets with many variables
5. Application of BN to ordinal data such as customer satisfaction surveys
6. Use of BN in decision support systems, with a healthcare application
7. Testing of web services using a dynamic adaptable approach.

In all these domains, using BNs has helped enhance the quality of information derived from an analysis of the available data sets.

In section 3 we describe various technical aspects of BNs, including estimation of distributions and algorithms for learning the BN structure. In learning the network structure, one can include white lists of forced causality links imposed by expert opinion and blacklists of links that are not to be included in the network, again using inputs from content experts. This essential feature permits an effective dialogue with content experts who can impact the model used for data analysis. In addition, as presented in example 2.3, BNs offer a unique capability of combining data from various sources. For a related analysis of this problem see Dalla Valle [27]. We also briefly discuss statistical inference of causality links, a very active area of research. In general, BNs provide a very effective descriptive causality analysis, with a natural graphical display.

The examples presented above shop how a BN can be used as a decision support tool for determining which predictor variables are important on the basis of their effect on target variables. In such an analysis, choosing an adequate BN structure is a critical task. In practice, there are several algorithms available for determining the BN structure, each on with its specific characteristics. For example, the R package

bnlearn includes eleven algorithms: two scored based learning algorithms (Hill-Climbing with score functions BIC and AIC and TABU with score functions BIC and AIC), five constraint based learning algorithms (Grow-Shrink, Incremental Association, Fast Incremental Association, Interleaved Incremental association, Max-min Parents and Children), and two hybrid algorithms (MMHC with score functions BIC and AIC, Phase Restricted Maximization). In this section we present an approach for performing a sensitivity analysis of BN, across various structure learning algorithms, in order to assess the robustness of the specific BN one plans to use. Following the application of different learning algorithms to set up a BN structure, some arcs in the network are recurrently present and some are not. As a basis for designing a robust BN, one can compute how often an arc is present, across various algorithms, with respect to the total number of networks examined. A method to select a robust network and perform what-if sensitivity scenario analysis is presented in [28]. These scenarios are computer experiments on a BN performed by conditioning on specific variable combinations and predicting the target variables using empirically estimated network.

Current research topics in this area include, among others, the development of solutions to BNs using continuous data, handling the effect of outliers in the context of BN structure and BN estimators, algorithmic solutions to permit big data analysis and leverage RESTful and cloud hosting technologies. Additional applications of Bayesian Networks are presented in [29], [30] and [31]. In conclusion, this paper aims to show that Bayesian Networks offer unique opportunities for statisticians to work collaboratively with content experts in a wide range of application domains and in addressing challenging methodological and theoretical issues. The global objective is to generate information quality [32].

References

1. Kenett RS. and Shmueli, G. On Information Quality, Journal of the Royal Statistical Society (Series A),2014; 177, Part 1, pp. 3–38.
2. Pearl J. Causality: Models, Reasoning, and Inference, 2nd ed., Cambridge University Press, UK, 2009.
3. Jensen FV. Bayesian Networks and Decision Graphs, Springer, 2001.
4. Ben Gal I. Bayesian Networks, in Encyclopedia of Statistics in Quality and Reliability, Ruggeri, F., Kenett RS. and Faltin, F. (editors in chief), John Wiley and Sons, Chichester: UK, 2007.
5. Koski T. and Noble J. Bayesian Networks – An Introduction, John Wiley and Sons, Chichester: UK, 2009.
6. Pourret O, Naïm P and Marcot B. Bayesian Networks: A Practical Guide to Applications, John Wiley and Sons, Chichester: UK, 2008.
7. Fenton N and Neil M. Risk Assessment and Decision Analysis with Bayesian Networks, CRC Press, 2012.
8. Kenett RS, De Frenne A, Tort-Martorell X and McCollin C. The Statistical Efficiency Conjecture, in Applying Statistical Methods in Business and Industry – the state of the art, Greenfield, T., Coleman and Montgomery, R. (editors), John Wiley and Sons, Chichester: UK, 2008.

9. Harel A, Kenett RS. and Ruggeri F. Modeling Web Usability Diagnostics on the basis of Usage Statistics in Statistical Methods in eCommerce Research, W. Jank and G. Shmueli (editors), John Wiley and Sons, Chichester: UK, 2008.
10. Kenett RS. and Raanan Y. Operational Risk Management: a practical approach to intelligent data analysis, John Wiley and Sons, Chichester: UK, 2010.
11. Peterson J. and Kenett RS. Modelling Opportunities for Statisticians Supporting Quality by Design Efforts for Pharmaceutical Development and Manufacturing. Biopharmaceutical Report, ASA Publications 2011; 18 (2): 6-16.
12. Kenett RS and Salini S. Modern Analysis of Customer Satisfaction Surveys: with applications using R, John Wiley and Sons, Chichester: UK, 2011.
13. Kenett RS. Risk Analysis in Drug Manufacturing and Healthcare, in Statistical Methods in Healthcare, Faltin F, Kenett RS and Ruggeri F. (editors in chief), John Wiley and Sons. Chichester: UK, 2012.
14. Bai X, Kenett RS. and Yu W. Risk Assessment and Adaptive Group Testing of Semantic Web Services. International Journal of Software Engineering and Knowledge Engineering 2012; 22(5):565-620.
15. Kenett RS and Zacks S. Modern Industrial Statistics with applications in R, MINITAB and JMP, 2nd edition, John Wiley and Sons, Chichester: UK, 2014.
16. Cornalba C, Kenett RS and Giudici P. Sensitivity Analysis of Bayesian Networks with Stochastic Emulators. 2007 ENBIS-DEINDE proceedings, University of Torino, Turin, Italy.
17. Gruber A. and Ben Gal I. Efficient Bayesian Network Learning for System Optimization in Reliability Engineering," Quality Technology & Quantitative Management, 2012; 9 (1), 97-114.
18. Heckerman D. A tutorial on learning with Bayesian networks. Microsoft Research tech. report MSR-TR-95-06. 1995, http://research.microsoft.com.
19. Neapolitan ER. Learning Bayesian Networks. Prentice Hall, 2003.
20. Frosini B. Causality and causal models: a conceptual perspective. International Statistical Review 2006; 74: 305-334.
21. Imai K., Tingley D. and Yamamoto T. Experimental designs for identifying causal mechanisms. Journal of the Royal Statistical Society (Series A) 2013; 176(1): 5-51.
22. Aalen O., Røysland K. and Gran JM. Causality, mediation and time: a dynamic viewpoint. Journal of the Royal Statistical Society (Series A) 2012; 175(4): 831-861.
23. Heckman J. Econometric Causality. International Statistical Review 2008; 76: 1-27.
24. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988.
25. Baker S. Causal inference, probability theory, and graphical insights. Statistics in Medicine 2003; doi: 10.1002/sim.5828
26. Pearl J. Comment on Causal inference, probability theory, and graphical insights (by Stuart G. Baker). UCLA Cognitive Systems Laboratory, Technical Report (R-412), June 2013, Statistics in Medicine. ftp.cs.ucla.edu/pub/stat_ser/r412.pdf
27. Dalla Valle, L. and Kenett, R.S. Social Media Big Data Integration: A New Approach Based on Calibration, Expert Systems with Applications, 2018: 111, pp. 76–90..
28. Cugnata, F., Kenett, R.S., and Salini, S. Bayesian networks in survey data: Robustness and sensitivity issues. Journal of Quality Technology, 2016, 48, pp. 253-264.
29. Kenett, R.S. On Generating High InfoQ with Bayesian Networks, Quality Technology and Quantitative Management, 2016: 13(3), pp 309-332.

30. Kenett, R.S. Bayesian networks: Theory, applications and sensitivity issues, Encyclopedia with Semantic Computing and Robotic Intelligence, World Scientific Press, 1(1), pp 1-13., 2017.
31. Kenett, R.S. Cause and Effect Diagrams in Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, 2019; DOI: 10.1002/9781118445112.stat03928.pub2.
32. Kenett RS. and Shmueli, G. Information Quality: The Potential of Data and Analytics to Generate Knowledge, John Wiley and Sons, Chichester: UK, 2016.

Professor Ron Kenett is Chairman of the KPA Group, Israel, Senior Research Fellow at the Neaman Institute, Technion, Haifa and Professor at University of Torino, Torino, Italy. He is an applied statistician combining expertise in academic, consulting and business domains. Ron is Past President of the Israel Statistical Association (ISA) and of the European Network for Business and Industrial Statistics (ENBIS). He authored and co-authored over 250 papers and 14 books on topics such as biostatistics, healthcare, industrial statistics, customer surveys, multivariate quality control, risk management and information quality. The KPA Group he founded in 1994, is a leading Israeli firm focused on generating insights through analytics. He is editor in chief of Wiley's StatsRef, serves on the editorial board of several international journals and was awarded the 2013 Greenfield Medal by the Royal Statistical Society and, in 2018, the Box Medal by the European Network for Business and Industrial Statistics for outstanding contributions to applied statistics. He founded the point and click translator company, Babylon.com and is member of the board of several startup companies. Ron holds a BSc in Mathematics (with first class honors) from Imperial College, London University and a PhD in Mathematics from the Weizmann Institute of Science, Rehovot, Israel. See also www.amazon.com/author/rkenett