To appear in IEE Proceedings Software Engineering, 1998

Assessing Dependability of Safety Critical Systems using Diverse Evidence

Norman Fenton
Bev Littlewood
Martin Neil
Lorenzo Strigini
Alistair Sutcliffe
David Wright

City University
Northampton Square
London EC1V OHB
Tel: 0171 477 8425
Fax: 0171 477 8585
n.fenton@csr.city.ac.uk

Version 2.1

12 May 1998

Abstract

A primary objective of the DATUM (**D**ependability **A**ssessment of safety critical systems **T**hrough the **U**nification of **M**easurable evidence) project was to improve the way dependability of software intensive safety-critical systems was assessed. Our hypothesis was that improvements were possible if we could incorporate multiple types of evidence. To achieve our objective we had to investigate how to get improved dependability predictions given certain specific information over and above failure data alone. We also had to provide a framework for modelling uncertainty and combining diverse evidence in such a way that it could be used to represent an entire argument about a system's dependability. We examined in depth the various methods and technologies for modelling uncertainty and selected a Bayesian approach as the most appropriate for our needs. To implement this approach for combining evidence we used Bayesian Belief Networks (BBNs). With the help of a BBN tool we provided a framework for dependability assessment that met our original objective and has subsequently proved to be practical and highly popular. A major benefit of this approach is that otherwise hidden assumptions used in an assessment become visible and auditable.

1. Introduction

With the increasing use of software components in safety-critical systems there has been a major concern about the dependability of such systems and how that dependability is assessed. The impact on overall system dependability of hardware is well understood, provided that it is free from design faults. However, some hardware failures and *all* software failures are due to design faults. Reliability in the presence of design faults and human operator errors is not well understood. To address the problem of assessing safety-critical systems in the presence of such errors the DATUM project was funded as part of the EPSRC/DTI

Safety Critical Systems Programme in 1993. DATUM (**D**ependability **A**ssessment of safety critical systems **T**hrough the **U**nification of **M**easurable evidence) was a three-year collaborative effort involving Lloyd's Register and three academic partners: Centre for Software Reliability and Centre for Human Computer Interface Design (both at City University), and Royal Holloway and Bedford New College.

DATUM's primary objective was to advance the state of the art in assessing and predicting the dependability of systems by combining diverse sources of relevant information. The traditional approach to assessing safety-critical systems has been highly subjective, relying heavily on the notion of 'engineering judgement'. Whilst this process is usually carried out responsibly, it is difficult for an outsider to analyse how the final judgement has been reached, and much has to be taken on trust. Moreover, there is some evidence of experts being unduly optimistic about their judgmental abilities. On the other hand, there are limitations to what can be claimed from purely objective evidence like failure data based solely on observing the system in test (or even operation). It was our belief that we could get improved assessments by taking account of, and combining, the many disparate types of evidence that might be available. This could include not only engineering judgement and failure data, but also evidence of the efficacy of the development methods used, experience in building similar systems in the past, competence of the development team, architectural details of the design (including especially the human computer interface), etc. The challenge for DATUM was to provide a rigorous measurement-based approach that could overcome the serious problems involved in combining disparate evidence in order to make a single evaluation of the overall dependability. A secondary objective for DATUM (which is beyond the scope of this paper) was to help developers determine how different development methods and system architectures contribute to the overall dependability argument.(see [Sutcliffe and Ryan 1996] for further details).

This paper summarises the work undertaken to meet the primary objective. Essentially, we had to tackle three problems:

1. Select an appropriate formalism for modelling uncertainty.
2. Provide a framework for combining diverse evidence in such a way that it could be used to represent an entire argument about a system's dependability. This framework also had to include guidelines on how to collect and record measurable evidence.
3. Quantify increased confidence in reliability predictions given certain *specific* information over and above the product failure data alone.

The way we tackled these three problems is described below in Sections 2-4 respectively. Our work on problem 2 is the main focus of this paper. Our selected approach, to use Bayesian Belief Nets, was novel for this application domain, but has since gained significant international interest. In Section 5 we summarise the extent of this interest by highlighting the impact of DATUM on subsequent work.

2. Selecting an appropriate formalism for modelling uncertainty

Our first task was to examine in depth the various methods and technologies for modelling uncertainty, including Bayesian probability, Dempster-Shafer theory, fuzzy sets and possibility theory. In [Wright and Cai 1994] we provided a comprehensive overview and in-depth comparison of these approaches. We

identified the advantages and disadvantages of each method. For example, Dempster-Shafer's theory of belief functions was attractive to us since it can apparently deal with problems where the evidence is too weak to support a parametric model. However, this approach provides no capability to account for 'dependencies' between the two or more bodies of evidence which are combined, and there is a serious danger in the safety context of over optimistic assessment resulting from failure to recognise stochastic dependencies. Although no single formalism for uncertainty was perfect for our purposes the result of our study was a decision to adopt the most mature and well-developed (and arguably the most convincing) namely *Bayesian probability*. We provided specific justification for this in the context of systems dependability assessment in [Littlewood et al 1995]. Bayesian probability theory has been extended beyond merely a means of representing uncertainty, and in fact provides a decision theory: a coherent theory of how to *act* on the world in an optimal fashion under circumstances of uncertainty. Fuzzy logic and Dempster-Shafer theory do not as yet appear to offer anything comparable. Bayesian probability offers a language and calculus for reasoning about the beliefs that can be reasonably held, in the presence of uncertainty, about future events, on the basis of available evidence. *Prior* probabilities are thus updated, after new events are observed, to produce *posterior* probabilities. By repeating this process, the implications of multiple sources of evidence can be calculated in a consistent way.

3. A framework for combining diverse evidence: Bayesian Belief Nets (BBNs)

Having chosen Bayesian probability our next task was to 'put it into practice'. This meant that we had to provide a practical and usable method for assessors to use the formalism in order to take into account explicitly all the factors that affect the reliability of a software product: proficiency of developers, effectiveness of tools, effectiveness of inspections and debug testing, effects of specification and programming languages, specific difficulties of an application or a specific project, etc. The usual impediment to using this multiple evidence in Bayesian reasoning has been the overly complex computations that result. This has, in the past, severely restricted the scale of problems and evidence that could be tackled. However, a relatively new but rapidly emerging technology, *Bayesian Belief Networks* (BBNs), provided an elegant solution. Complete with appropriate computer tools, BBNs enable us to push back the boundary of the problems that could be attacked.

In this section we provide a very brief background and overview of BBNs (section 3.1), a summary of the benefits of using BBNs (Section 3.2) and the practical issues involved (Section 3.3).

3.1 Overview of BBNs

Bayesian Belief Networks (also known as Belief Networks, Causal Probabilistic Networks, Causal Nets, Graphical Probability Networks, Probabilistic Cause-Effect Models, and Probabilistic Influence Diagrams) have attracted much recent attention as a possible solution for the problems of decision support under uncertainty. Although the underlying theory (Bayesian probability) has been around for a long time, the possibility of building and executing realistic models has only been made possible because of recent algorithms and software tools that implement them (see [Jensen 1996] for a full description of the relevant

algorithms and tools). To date BBNs have proven useful in practical applications such as medical diagnosis and diagnosis of mechanical failures. Their most celebrated recent use has been by Microsoft who have, for example, used BBNs as the underlying technology for the help wizards in Microsoft Office. However, we believe that DATUM was the first project in which BBNs were applied extensively to the problems of assessing dependability of critical systems.

A BBN is a graphical network that represents probabilistic relationships among variables of interest. BBNs enable reasoning under uncertainty and combine the advantages of an intuitive visual representation with a sound mathematical basis in Bayesian probability. An informal judgement can be formalised into a BBN by specifying a series of links of the form the 'truth of statement A supports my belief in statement B'. Figure 1 is a simple (illustrative only) example of a BBN used for predicting software reliability that takes account of product and process information. Here, reliability is directly influenced by the number of (latent) faults and the amount of operational usage. Hence, we model this relationship by drawing arcs from the nodes 'number of latent faults and 'operational usage' to 'reliability'. The number of latent faults is influenced by the coders' performance which in turn is influenced by factors such as problem complexity, staff experience, and use of formal methods.

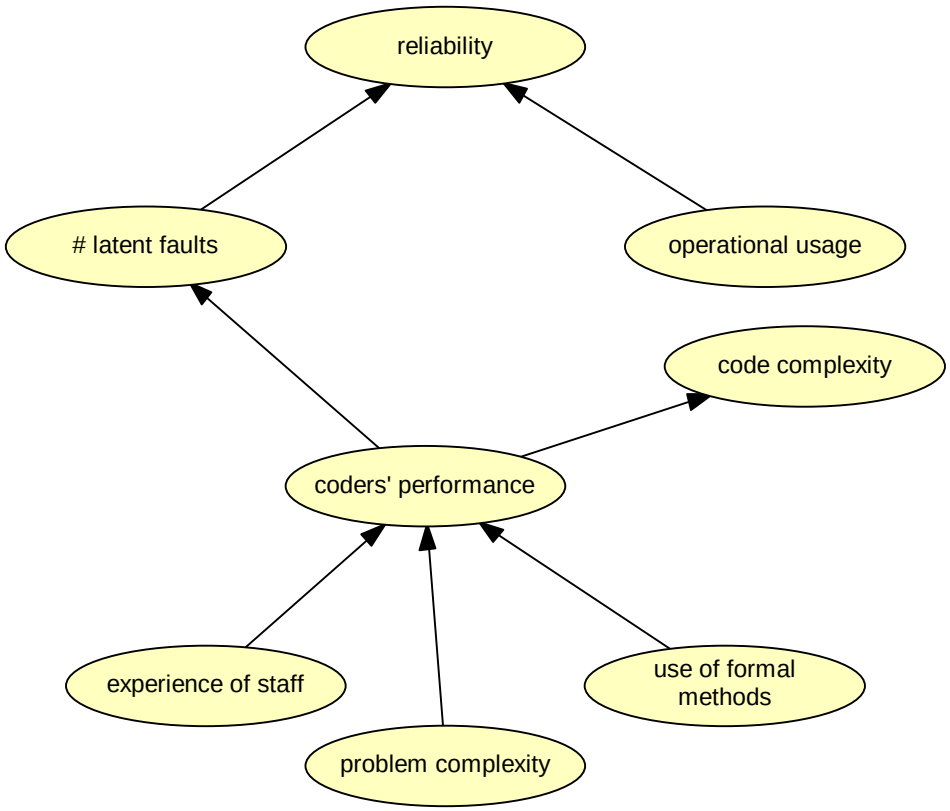

Figure 1: Simple BBN for software reliability taking account of process and product information

In our example, if we knew there were a high number of faults and high level of operational usage then the probability that the reliability is high would be less than if we knew there were few faults and low

operational usage. In the BBN we model this by filling in a node probability table (NPT). For example, Table 1 shows the NPT for the 'reliability' node.

	op usage	low			medium			high		
	# faults	low	medium	high	low	medium	high	low	medium	high
reliability	low	0.1	0.2	0.4	0.2	0.3	0.5	0.5	0.3	0.7
	medium	0.2	0.3	0.3	0.3	0.4	0.3	0.3	0.4	0.2
	high	0.7	0.5	0.3	0.5	0.3	0.2	0.2	0.3	0.1

Table 1: Node Probability Table (NPT) for the node 'reliability'

If the node has no incoming arcs (root node), the NPT lists the marginal probabilities of the possible values of the node; if it has incoming arcs, the table is a table of conditional probabilities (of the values of this node, conditional on the values of its parent nodes). Arcs and tables of conditional probabilities can thus be used to represent the fact that knowledge about one node is useful for predictions about another node. There may be several ways of determining the probabilities for the NPTs. One of the benefits of BBNs stems from the fact that we are able to accommodate both probabilities based on subjective judgements (elicited from domain experts) and probabilities based on objective data.

3.2 Benefits of using BBNs

The advantage of describing a probabilistic argument via a BBN, compared to describing it via mathematical formulas and prose, is that the BBN represents the structure of the argument in an intuitive, graphical format. The main use of BBNs is in situations that require statistical inference: in addition to statements about the probabilities of events, the user knows some evidence, that is, some events that have actually been observed, and wishes to infer the probabilities of other events, which have not as yet been observed. For example, suppose in Figure 1 we enter the evidence that formal methods are used, the experience of staff is high and the problem complexity is low. Using probability calculus and Bayes theorem it is then possible to update the values of all the other probabilities in the BBN (this is called propagation). Bayesian analysis can be used for both 'forward' and 'backward' inference. For example, we may know nothing about the coding process, but we may have evidence about the actual reliability. Suppose, for example, we know that the reliability is low. Then if we enter this fact and propagate the BBN it will update all of the other probabilities, giving for example a revised value for the probability that the staff were experienced.

The use of BBNs makes the decision process easier to describe, and thus to check, communicate and audit. In particular, BBNs can be used to check the consistency of individuals' beliefs and decisions. This was borne out in the results of a case study with Lloyd's Register in which assessors' arguments about a system's dependability and safety were described in terms of specific BBNs. The results were that assessors were able to better characterise and communicate their complex webs of inference; moreover, the very act of producing a BBN in conjunction with an expert in the formalism resulted in refinements and improvements to the safety case.

Because BBNs have a rigorous, mathematical meaning there are software tools that can interpret them and perform the complex calculations needed in their use. The specific tool used in DATUM was *Hugin Explorer* (from the Danish company Hugin A/S) (see www.hugin.dk) which provides a graphical front end for inputting the BBNs in addition to a computational engine for the Bayesian analysis. With the help of this tool we provided a framework for dependability assessment that met our original objectives and was also practical and highly popular with domain experts (it is described in more depth in [Neil, Littlewood, Fenton 1996], [Strigini and Fenton 1996], and [Fenton 1996]).

3.3 Practical issues in using BBNs

Building a BBN involves two difficult steps:

1. defining the BBN topology (that is, getting the 'right' collection of nodes and arcs)
2. defining the NPTs

DATUM investigated (and also applied) methods whereby NPTs could be extracted from untrained users by extrapolating from a small number of assumptions [Neil, Littlewood, Fenton 1996]. This work has since been developed much further as part of both the ESPRIT-funded SERENE project and the EPSRC-funded IMPRESS project (described below in Section 5). Both of these projects have also tackled the generic problem of building the 'correct' BBN topology for a given application domain.

DATUM also provided guidelines for populating NPTs with probabilities based on empirical data. On the one hand we described in [Fenton 1996] how relatively simple measurement programmes could provide the necessary data directly (a well managed software development project has a wealth of potentially important quantitative information to support safety assessment). On the other hand we produced a range of results that enable assessors to use quantitative data even if they have none available from their own projects. For example, in [Fenton, Neil, Ostrolenk 1995] we provided a wealth of data on fault densities after examining the published literature (we subsequently used such results to populate probability tables in our own BBNs). In [Fenton and Finney 1996] we provided data gathered from public sources on the likely impact on reliability from using formal methods during development. In [Littlewood and Wright 1995a] we provided evidence about the number of representative test cases required to obtain a high confidence (99%) that the probability of failure on demand is smaller than various specified limits. Our results included the situation where failures were observed during testing.

4. Gaining increased confidence in reliability predictions using specific additional information

We considered two especially important types of additional information common in safety critical systems:

The very common scenario whereby increased confidence in reliability is claimed as a result of evidence from previous experience of 'similar' products (in addition to evidence from testing the product itself). In [Littlewood and Wright 1995b] we examined this in depth. Importantly (but rather depressingly) we showed that this kind of additional evidence can only improve our confidence in the reliability of a product quite modestly. We used a Bayesian model in our analysis. Essentially, if you wished to claim that great

trust could be placed in a particular system following past experience of other systems, then you could only do this by starting with extremely strong a priori beliefs—essentially you need to bring to the problem the beliefs that you wish to claim.

The situation whereby improvements in reliability are expected as a result of using diversity in the product design. Recent models for the failure behaviour of systems involving redundancy and diversity have shown that common mode failures can be accounted for in terms of the variability of the failure probability of components over operational environments. Whenever such variability is present, we can expect that the overall system reliability will be *less* than we could have expected if the components could have been assumed to fail independently. In [Littlewood 1994] we generalised a well known model of hardware redundancy and showed that with forced diversity, this unwelcome result no longer applies: in fact it becomes theoretically possible to do *better* than would be the case under independence of failures. We also provided an example to show how the new model can be used to estimate redundant system reliability from component data.

5. Impact of DATUM and future research

The DATUM approach to dependability assessment using BBNs has achieved considerable appeal and impact. In 1996 the European Commission funded a major ESPRIT project SERENE (SERENE (SafEty and Risk Evaluation using Bayesian Nets) in which BBNs were being used as the major technology to support safety assessments. The project includes as partners key industrial assessors of critical systems across Europe. The ESPRIT basic research action DeVa (Design for VAlidation), which also began in 1996, is further investigating the use of BBNs for dependability assessment.

In the UK the Defence Evaluation Research Agency has recently begun funding a collaborative project with CSR at City University to use BBN technology to improve its predictions of reliability of complex vehicles and equipment. Some of the key technical issues identified in DATUM associated with building realistic BBNs are being addressed in the EPSRC project IMPRESS (IMproving the software PRocESS using bayesian nets) that began in 1997. Specifically, IMPRESS is developing methods for building large NPTs; this project is also extending the use of BBNs for system quality evaluation in the commercial software domain.

Full details of the projects described here may be found on <www.city.csr.ac.uk>

6. Conclusions

The DATUM project's primary objective was to improve the state-of-the-art of assessing the dependability of safety-critical systems. We believed it was possible to provide a rigorous, empirically-based and logically consistent approach that was able to take account of diverse sources of information that were often uncertain and subjective. We examined in depth the various methods and technologies for modelling uncertainty and selected a Bayesian approach as the most appropriate for our needs. To implement the Bayesian approach we used the technology of BBNs. With the help of a BBN tool we provided a framework for dependability assessment that met our original objectives. Using BBNs developers and

assessors are able to model the many (normally hidden) assumptions that contribute to an argument about a system's dependability and safety. The technology takes the strain of combining subjective criteria with quantitative evidence about the process and product to obtain predictions of attributes such as reliability. Although the predictions still cannot be validated without operational testing, the major benefit of this approach is that the entire set of assumptions used to assess a system's dependability become visible and auditable. The very act of modelling the argument with a BBN leads assessors to formalise their assumptions and arrive at a safety argument which they feel is improved and more accurate.

DATUM's approach to safety argument quantification using BBNs is likely to have a significant impact in the coming years as a result of its adoption in both the prestigious ESPRIT project SERENE and also in recent MoD assessment activities. The results of DATUM may also ensure that issues such as the limits to quantified safety and dependability targets become a prominent feature in future standards.

The results of DATUM may eventually be adopted by a wide community beyond those concerned only with safety-critical software systems. This includes a large number of practitioners faced with the problem of assessing complex systems in the face of diverse and uncertain evidence.

Acknowledgements

The work described here was originally funded under the UK Safety Critical Systems Programme (Project IED/1/9314) and specifically by the EPSRC (grant number GR/H89944). The authors are indebted to the support of EPSRC and the encouragement of the Monitoring Officer Dr Dominic Semple. The authors also acknowledge the contributions of their colleagues at Lloyd's Register and Royal Holloway, and the helpful comments of the anonymous referees.

7. References

Fenton NE, The role of measurement in software safety assessment, in 'Safety and Reliability of Software Based Systems' (Ed Shaw, R), Springer Verlag, 217-248, 1996.

Finney K and Fenton NE, Evaluating the effectiveness of using Z: the claims made about CICS and where we go from here, J Systems Software, Nov, 1996.

Jensen FV, An Introduction to Bayesian Networks, UCL Press, 1996.

Littlewood B and Wright DR, On a Stopping Rule for the Operational Testing of Safety-Critical Software, Digest of IEEE 1995 FTCS, 25th Annual International Symposium Fault-Tolerant Computing, (Pasadena), IEEE Computer Society Silver Spring, Md., pp 444-451, 1995a.

Littlewood B and Wright DR, A Bayesian model that combines disparate evidence for the quantitative assessment of system dependability, Proc 14th International Conference on Computer Safety (SafeComp'95), pp 173-188, Springer, 1995b.

Littlewood B, Neil M and Ostrolenk G, The Role of Models in Managing Uncertainty of Software-Intensive Systems, Reliability Engineering and System Safety, 46, 87-95, 1995.

Neil M, Littlewood B, Fenton NE, Applying Bayesian belief networks to systems dependability assessment, in Proceedings of 4th Safety Critical Systems Symposium, Springer Verlag, 1996.

Strigini L and Fenton NE, Rigorously assessing software reliability and safety, Proc Product Assurance Symposium and Software Product Assurance Workshop, 19-21 March 1996, ESA SP-377, May, 1996.

Sutcliffe A.G. and Ryan M, Evaluating safety critical systems design and user-system interfaces, Centre for HCI Design Report, School of Informatics, City University, 1996.

Wright D and Cai K-Y, Representing uncertainty for safety critical systems, DATUM/CITY/02, City University, London EC1V OHB, 1994.