Analyzing the Simonshaven Case using Bayesian Networks

Norman Fenton*, School of Electronic Engineering and Computer Science, Queen Mary University of London

Martin Neil, School of Electronic Engineering and Computer Science, Queen Mary University of London

Barbaros Yet, Department of Industrial Engineering, Hacettepe Universitesi,Turkey

David Lagnado, Department of Experimental Psychology, University College London

5 November 2018

***Corresponding author. Address:** School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End, London E1 4NS, UK

n.fenton@qmul.ac.uk

Keywords: Bayesian networks, legal reasoning, probability, uncertainty, evidence, idioms

Word count: 7858

Abstract

This paper is one in a series of analyses of the Dutch Simonshaven murder case, each using a different modelling approach. We adopted a Bayesian Network (BN) based approach which requires us to determine the relevant hypotheses and evidence in the case and their relationships (captured as a directed acyclic graph) along with explicit prior conditional probabilities. This means that both the graph structure and probabilities had to be defined using subjective judgments about the causal, and other, connections between variables and the strength and nature of the evidence. Determining if a useful BN could be quickly constructed by a small group using the previously established idioms-based approach, which provides a generic method for translating legal cases into BNs, was a key aim. The model described was built by the authors during the course of a workshop dedicated to the case at the Isaac Newton Institute Cambridge in September 2016. The total effort involved was approximately 26 hours (i.e. an average of 6 hours per author). The paper describes a formal evaluation of the model, using sensitivity analysis, to determine how robust the model conclusions are to key subjective prior probabilities over a full range of what may be deemed ‘reasonable’ from both defence and prosecution perspectives. The results show that the model is reasonably robust (pointing to a high probability of guilt but generally below the 95% threshold). Given the constraints on building a complex model so quickly there are inevitably weaknesses, hence the paper describes these and how they might be addressed, including how to take account of supplementary case information not known at the time of the workshop.

1. Introduction

A Bayesian Network (BN) is a graphical model in which the nodes represent variables and the arcs represent causal, probabilistic, or influential relationships between variables. The strength of any relationship is specified by a probability distribution. Variables could represent unknown hypotheses such as Boolean variables like ‘suspect murdered the victim’ and ‘witness claims suspect was at scene of the crime’ or numeric variables like ‘number of people at scene of the crime’. Some variables, such as ‘suspect murdered the victim’, will generally never be observed and hence represent unknown hypotheses, while others, like ‘witness claims suspect was at scene of the crime’, may be observed (in this case with value yes or no) and so represent evidence. Hence, a BN is potentially a natural way to represent and communicate the relationships between different hypotheses and pieces of evidence in a complex legal argument. But, in addition to its powerful visual appeal, it has an underlying calculus (based on Bayes’ theorem) that determines the revised probability beliefs about all uncertain variables when any piece of new evidence is presented. However, despite its apparent obvious attractiveness, the take-up in practice of using BNs for legal arguments has been disappointing. One of the reasons for the lack of take up is the ad-hoc, and often complex and subjective approach to constructing an acceptable BN model in a legal case. A primary motivation for this paper is to demonstrate that it is possible to use a systematic idioms-based approach to easily develop a ‘consensus’ BN model for the Simonshaven case which represents a complex example legal case. The full details of the case are provided elsewhere in this journal issue, but in summary it concerns the violent murder of a woman who had been out walking with her husband in a quiet recreational area near the village of Simonshaven, close to Rotterdam, in 2011. The trial court of Rotterdam convicted the victim’s husband of murder by intentionally hitting and/or kicking her in the head and strangling her. For the appeal the defence provided new evidence about other ‘similar’ murders in the area committed by a person we shall refer to as the ‘man in the woods’.

The paper is structured as follows: In Section 2 we provide a full explanation of Bayes and BNs and their use in legal argumentation. In Section 3 we summarise the idioms-based approach to building BNs for legal cases and present the BN we developed for the Simonshaven case using this approach. In Section 4 we present the results of running the model using the evidence of the case. We also perform sensitivity analysis to test how robust the model conclusions are to adjustments made to the key subjective prior probabilities, given a full range of what might be considered ‘reasonable’ from both defence and prosecution perspectives. In Section 5 we describe how the model could be improved, taking account of the sensitivity analysis and providing some retrospective reflection in the absence of the time constraints imposed on the original model building, and further information about the case. In Section 6 we discuss the advantages and disadvantages of the BN approach from both a practical and legal perspective.

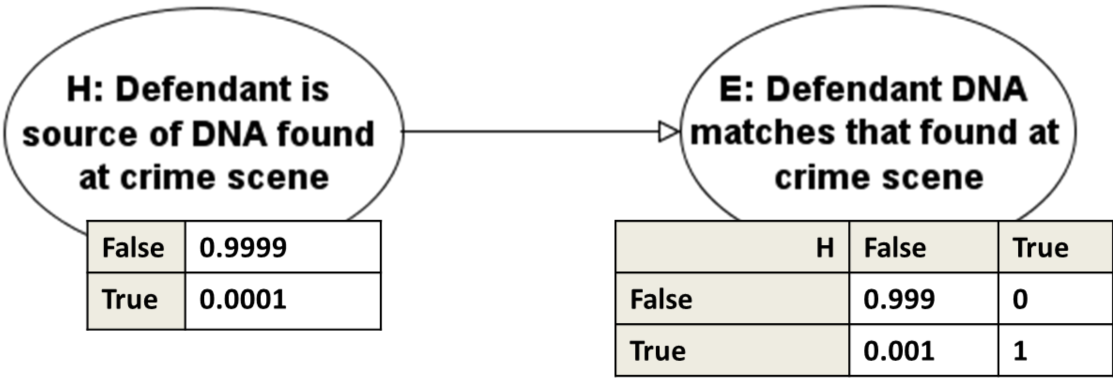
2. Bayes and Bayesian Networks in the legal context

To understand the context for the use of Bayes for legal argumentation we need some terminology and assumptions:

- A ***hypothesis*** is a statement (typically Boolean) whose truth value we seek to determine but is generally unknown - and which may never be known with certainty. Examples include:
    - ‘Defendant is guilty of the crime charged’ (this is an example of an ***offense level hypothesis*** also called the ***ultimate hypothesis***, since in many criminal cases it is ultimately the only hypothesis we are really interested in)
    - ‘Defendant was the source of DNA found at the crime scene’ (this is an example of what is often referred to as ***a source level hypothesis*** (Cook et al., 1998))
- A piece of ***evidence*** is a statement that, if true, lends support to one or more hypotheses.

The relationship between a hypothesis H and a piece of evidence E can be represented graphically as in the example in Figure 1 where we assume that:

- The evidence E is a DNA trace found at the scene of the crime (for simplicity we assume the crime was committed on an island with 10,000 people who therefore represent the entire set of possible suspects)
- The defendant was arrested and some of his DNA was sampled and analysed


H: Defendant is source of DNA found at crime scene
E: Defendant DNA matches that found at crime scene

False	0.9999
True	0.0001

H	False	True
False	0.999	0
True	0.001	1

Figure 1 Causal view of evidence, with prior probabilities shown in tables. This is a very simple example of a Bayesian Network

The direction of the causal structure makes sense here because H being true (resp. false) can 'cause' E to be true (resp. false), while E cannot 'cause' H[1]. However, inference can go in ***both*** directions. If we observe E to be true (resp. false) then our belief in H being true (resp. false) increases. It is this latter type of inference that people generally use (albeit informally) to revise their belief about an uncertain hypothesis after observing evidence and this is especially relevant to legal reasoning since, informally, lawyers and jurors are expected to:

- Start with some prior assumption about the ultimate hypothesis H (defendant is guilty) being true; for example, the assumption 'innocent until proven guilty' might equate to a belief that 'the defendant is no more likely to be guilty than any other able member of the population'.

[1] Note that throughout the paper when we speak of a causal relationship from *A* to *B* we are always using it in the sense of (Pearl et al., 2018) whereby *A* is an event that *can* cause *B* rather than an event that necessarily causes *B*.

- Update our belief about H once we observe evidence E based on the 'likelihood' of the evidence; specifically, the more unlikely we consider the evidence to have been if the defendant were not guilty (i.e *H* is false), the more our belief in the defendant being guilty (i.e. *H* is true) increases.

This informal reasoning is a perfect match for Bayesian inference where the prior assumption about H and the likelihood of the evidence E are captured formally by the probability tables shown in Figure 1. Specifically, these are the tables for the ***prior probability*** about H**,** written $P(H)$, and the conditional probability of E given H, which we write as $P(E|H)$. Bayes' theorem provides the formula for updating our prior belief about H in the light of observing *E* to arrive at a ***posterior probability*** about $H,$ which we write as $P(H|E)$. In other words, Bayes calculates $P(H|E)$ in terms of $P(H)$ and $P(E|H)$. Specifically:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|not\ H)P(not\ H)}$$

The first table (the probability table for *H*) captures our knowledge that the defendant is one of 10,000 people who could have been the source of the DNA. The second table (the probability table for $P(E|H)$ captures the assumptions that:

- The probability of correctly matching a DNA trace is one (so there is no chance of a false negative DNA match). This probability $P(E|H)$ is called the ***prosecution likelihood*** for the evidence $E.$
- The probability of a match in a person who did not leave their DNA at the scene (the 'random DNA match probability') is 1 in 1,000. This probability $P(E|not\ H)$ is called the ***defence likelihood*** for the evidence $E.$

With these assumptions, it follows from Bayes' theorem that, in our example, the posterior belief in H after observing the evidence E being true is about 9%, i.e. our belief in the defendant being the source of the DNA at the crime scene moves from a prior of 1 in a

10,000 to a posterior of 9%. Alternatively, our belief in the defendant not being the source of the DNA moves from a prior of 99.99% to a posterior of 91%.

One of the reasons legal professionals are reluctant to endorse the use of Bayes is because it requires us to assign prior probabilities that, in many situations, are necessarily subjective. However, an equivalent formulation of Bayes (called the 'odds' version of Bayes) enables us to interpret the value of evidence E without ever having to consider the prior probability of H. Specifically, this version of Bayes' tells us:

the posterior odds of H are the prior odds of H times the likelihood ratio

where the Likelihood Ratio (LR) is simply the prosecution likelihood of E divided by the defence likelihood of E:

$$LR = \frac{P(E|H)}{P(E|not\ H)}$$

In the example in Figure 1 the prosecution likelihood for the DNA match evidence is 1, while the defence likelihood is 1/1,000. So, the LR is 1,000. This means that, whatever the prior odds were in favour of the prosecution hypothesis, the posterior odds must increase by a factor of 1,000 as a result of seeing the evidence. In general, if the LR is bigger than 1 then the evidence results in an increased posterior probability of H (with higher values leading to the posterior probability getting closer to 1), while if it is less than 1 it results in a decreased posterior probability of H (and the closer it gets to zero the closer the posterior probability gets to zero). If the LR is equal to 1 then E offers no value since it leaves the posterior probability is unchanged.

The LR is therefore an important and meaningful measure of the probative value of evidence. However, while forensic scientists and lawyers find it attractive because it avoids having to consider the prior probability of *H*, no conclusions about the posterior probability of *H* can be drawn from the LR without explicitly considering the prior. In our example the fact that the DNA match evidence had a LR of 1000 meant the evidence was highly probative in favour of the prosecution. But as impressive as that sounds, whether or not it is sufficient to convince you of which hypothesis is true still depends entirely on the prior $P(H)$. If $P(H)$ is, say 0.5 (so the prior odds are evens 1:1), then a LR of 1000 results in posterior odds of 1000 to 1 in favour of H. That may be sufficient to convince a jury that H is true. But if $P(H)$ is very low - as in our example (9999 to 1 against) - then the same LR of 1000 results in posterior odds that still strongly favour the defence hypothesis by 10 to 1.

There are also severe problems with using the LR when we move beyond the case of a single hypothesis H and a single piece of evidence E (Fenton, Berger, et al., 2013). In practice, real legal arguments normally involve multiple hypotheses and pieces of evidence with complex causal dependencies, as is certainly the case in Simonshaven. Even the simplest instance of one piece of DNA evidence strictly speaking involves three unknown hypotheses and two pieces of evidence with the causal links shown in Figure 2 (Dawid et al., 1998; Fenton et al., 2014) once we take account of the possibility of different types of DNA collection and testing errors (Koehler, 1993; Thompson et al., 2003).

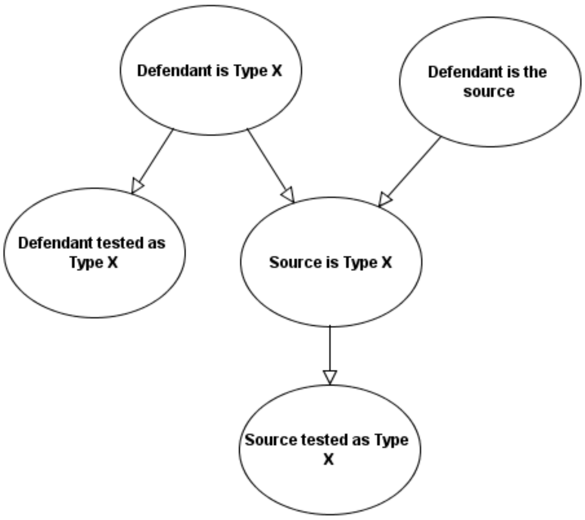


Figure 2 Bayesian network for DNA match evidence. Each node has states *true* or *false*

Moreover, there are further crucial hypotheses not shown in Figure 2 such as: ‘Defendant was at the scene of the crime’ and the ultimate hypothesis ‘Defendant committed the crime’. Figure 2 is an example of a Bayesian Network (BN). By convention (which is applied in the modelling that follows) an arc linking two nodes represents a direct dependency between the nodes. By definition nodes have to be directed and there must be no cycles. It is normal to draw the direction from cause to effect, where *A* and *B* are causally linked in the sense describe above. If *A* and *B* are directly – but not causally - related (for example, if *A* and *B* represented respectively a person’s *height* and *weight*) then the arc direction is usually chosen on the basis of which one is more convenient for specifying the conditional probability relationship. Crucially, the absence of an arc between *A* and *B* means that *A* and *B* are conditionally independent. A BN – like any model – is an approximation of reality and there is a necessary trade-off between reality and efficiency when selecting whether there should be an arc between two nodes. The more arcs, the more difficult it is to construct and run the model.

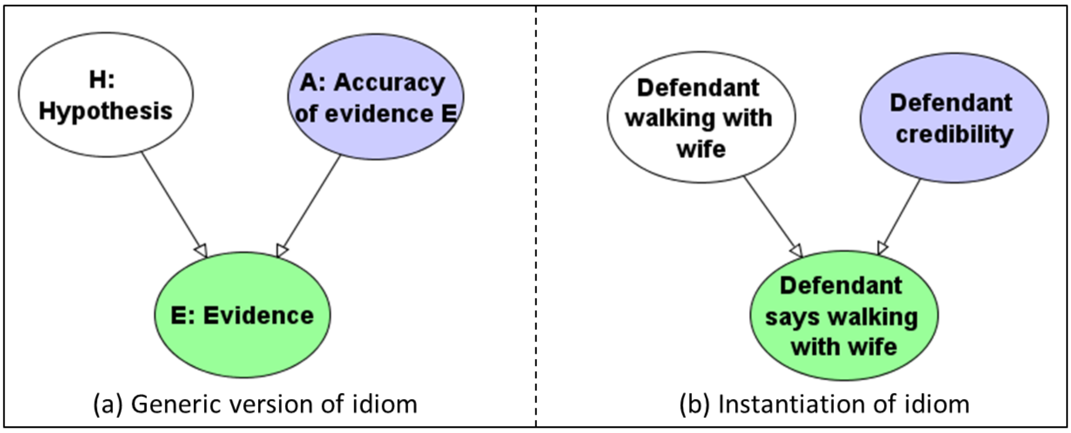
To perform the correct Bayesian inference once we observe evidence we need to know the prior probabilities of the nodes without parents and the conditional prior probabilities of the nodes with parents. If it is possible to obtain suitable estimates of these prior probabilities, the bad news is that, even with a small number of nodes, the calculations necessary for performing correct probabilistic inference are far too complex to be done manually. Moreover, until the late 1980’s there were no known efficient computer algorithms for doing the calculations. This is the reason why, until relatively recently, only rather trivial BNs could be built and used. However, algorithmic breakthroughs in the late 1980s (Pearl, 1988) made it possible to perform correct probabilistic inference efficiently for a wide class of BNs and these algorithms have subsequently been incorporated into widely available graphical toolsets that enable users without any statistical knowledge to build and run BN models (Fenton et al., 2012). This now includes accurate inference with BNs containing numeric variables without the need to manually discretize such variables -the first generation of BN software did not have this capability and as such this was considered something of an ‘Achilles heel’ (Neil et al., 2007).

The idea of using BNs for legal arguments is by no means new. Many, e.g., see (Aitken et al., 1995; Dawid et al., 1997; Huygen, 2002; Jowett, 2001; Kadane et al., 1996; Taroni et al., 2014) have explicitly used BNs to model legal arguments probabilistically. Indeed, (Edwards, 1991) provided an outstanding argument for the use of BNs in which he said of this technology: “I assert that we now have a technology that is ready for use, not just by the scholars of evidence, but by trial lawyers.” He predicted such use would become routine within “two to three years”. Unfortunately, he was grossly optimistic for reasons that are explained in (Fenton et al., 2011) and (Fenton, Neil, & Berger, 2016). One of the reasons for the lack of take up is the ad-hoc, and often extremely complex and subjective, approach to constructing an acceptable BN model in a legal case. This lack of a systematic, repeatable method for modelling legal arguments as BNs has been addressed in (Fenton, Lagnado, et al., 2013; Hepler et al., 2007; Lagnado et al., 2013) with the use of a small set of common patterns or ‘idioms’. This approach has been used to present full case studies of actual legal cases: the Dutch Anjum murders case involving complex evidence (Vlek et al., 2014); and a murder case involving DNA evidence leading to acquittal (Vlek et al., 2016). However, more examples are required to demonstrate its practicality, and the Simonshaven case provides an ideal example to demonstrate that the idioms-based approach can be used effectively to build a ‘consensus’ BN model.

3. Description of method and Simonshaven BN model

While the following description is intended to be self-contained, further details of the idioms and notation used are found in (Fenton, Lagnado, et al., 2013).

The simplest type of BN relevant to a legal case is the one whose structure we already saw in Figure 1, namely a two node model with an unknown hypothesis node H and a child node E representing observable evidence. While this can be considered as a simple idiom (the ‘evidence idiom’), in almost all cases it is also necessary to consider the accuracy of the evidence. For example, one hypothesis in the case is ‘defendant was walking with his wife’ and the evidence to support this is ‘defendant says he was walking with his wife’. Clearly the extent to which the evidence supports the unknown hypothesis depends on the accuracy/credibility[2] of the witness. Hence, one of the key and most commonly used idioms is the ***evidence accuracy idiom*** shown in Figure 3.


H: Hypothesis
A: Accuracy of evidence E
E: Evidence
Defendant walking with wife
Defendant credibility
Defendant says walking with wife

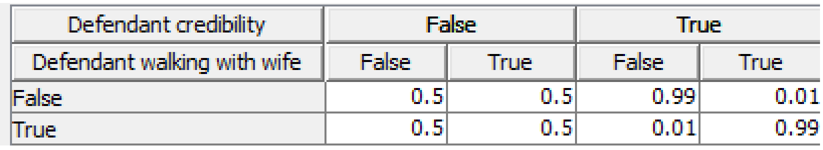
(a) Generic version of idiom

(b) Instantiation of idiom

Figure 3 Evidence accuracy idiom generic and instantiated versions)

In the simplest case the accuracy/credibility node is Boolean, in which case a reasonable Node Probability Table (NPT) of the evidence node is the one show in Table 1:

Table 1 Node Probability Table (NPT) for evidence node



Defendant credibility	False		True	
Defendant walking with wife	False	True	False	True
False	0.5	0.5	0.99	0.01
True	0.5	0.5	0.01	0.99

This essentially says that if the defendant is credible then we can be fairly certain that if he was/wasn’t walking with his wife then he will say he was/wasn’t (in each case the LR of the

[2] The notion of accuracy/credibility for witnesses is actually composed of several distinct attributes including *veracity*, *objectivity*, and *competence* (Schum, 1989). Ideally these should be explicitly represented as separate nodes as described in (Fenton, Lagnado, et al., 2013). However, for simplicity they are combined into one here.

evidence is 99), whereas if he is not credible we learn nothing from whatever he says (in each case the LR of the evidence is 1).

Different instances of idioms, such as the evidence accuracy idiom, are joined together by the ***cause-consequence idiom*** – when one hypothesis is a cause of another. For example, in Simonshaven we have three causally linked hypotheses:

> 'Defendant owned gun' → 'Defendant gun used in killing' →'Weapon discarded at pump station'.

When put in the context of their respective evidence accuracy idioms we obtain the model fragment shown in Figure 4.

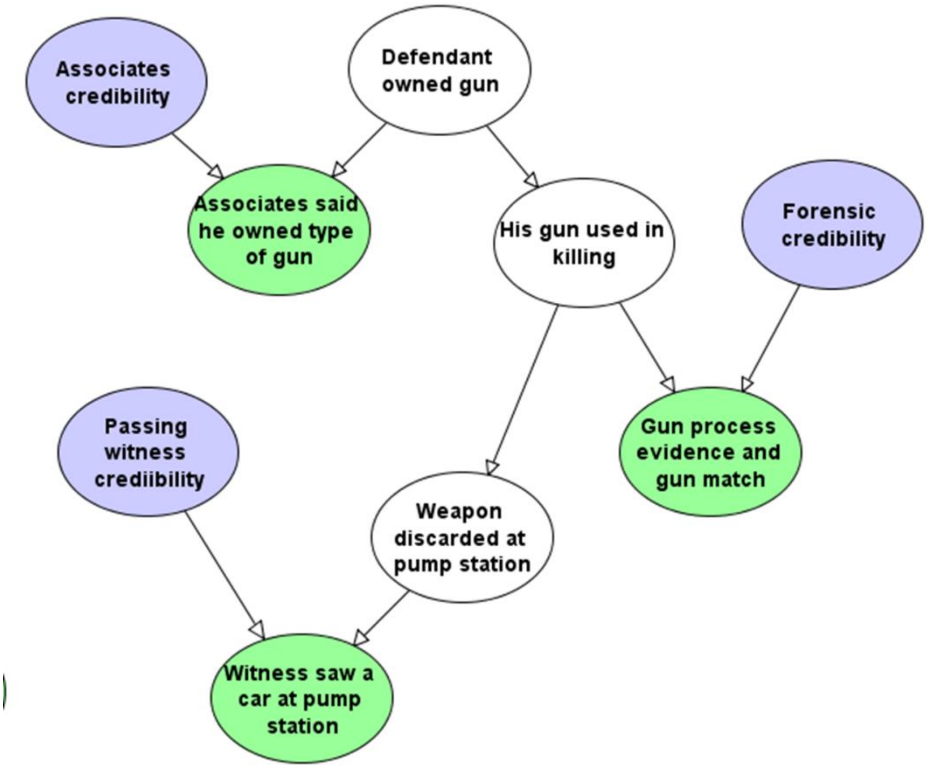

Associates credibility
Defendant owned gun
Associates said he owned type of gun
His gun used in killing
Forensic credibility
Passing witness crediibility
Gun process evidence and gun match
Weapon discarded at pump station
Witness saw a car at pump station

Figure 4 Three instances of evidence accuracy idiom linked by cause consequence idiom

While the evidence accuracy idiom should normally be used for each piece of evidence, there are two idioms – the ***opportunity*** idiom and the ***motive*** idiom - that normally only appear once, and are usually combined as in Figure 5.

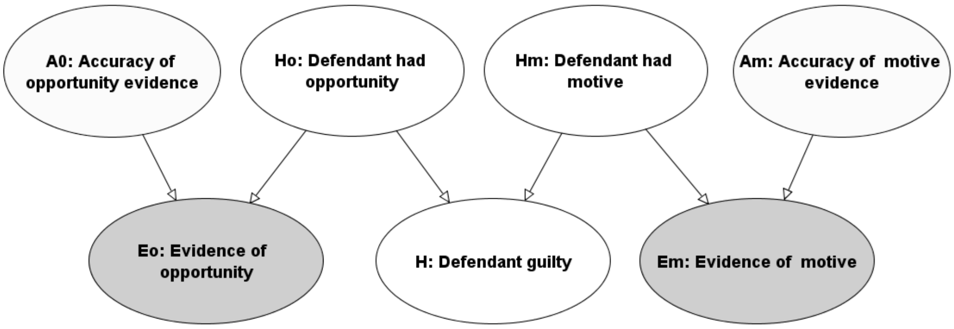

A0: Accuracy of opportunity evidence
Ho: Defendant had opportunity
Hm: Defendant had motive
Am: Accuracy of motive evidence
Eo: Evidence of opportunity
H: Defendant guilty
Em: Evidence of motive

Figure 5 Generic version of combined opportunity and motive idiom

The idiom is based around three hypotheses: the ultimate hypothesis of the case (normally ‘defendant is guilty’, which in Simonshaven corresponds to ‘Defendant killed her’) and two hypotheses that are assumed to be necessary causes, namely ‘opportunity’ and ‘motive’.

Note that this combined version of the idiom also incorporates two instances of the evidence accuracy idiom.

Typically, ‘defendant has opportunity’ – is synonymous with whether or not the defendant was present at the crime scene. However, in this case there is no doubt the defendant was present and so what is relevant in determining the effect of ’opportunity’ is the number of other people who were also present. Hence in the subsequent model ‘defendant had opportunity’ is replaced with ‘number of people in the wood’. Also, in addition to motive and opportunity, in Simonshaven we add ‘capability’ as a necessary causal parent of the ultimate hypothesis since there is uncertainty about whether he was physically capable of killing her as claimed. This results in the model component shown in Figure 6.

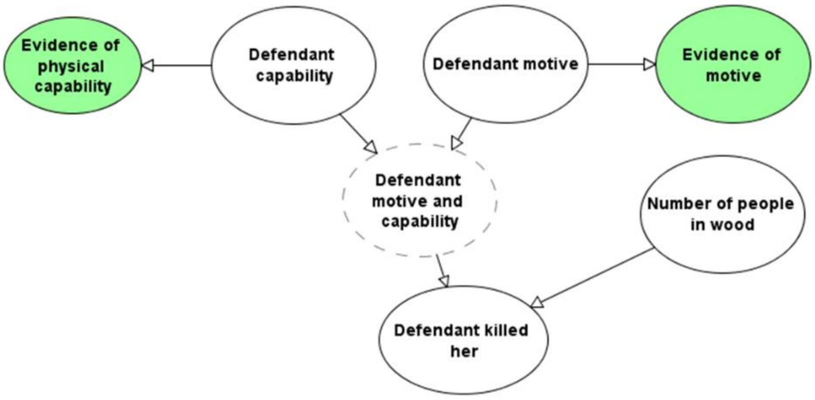


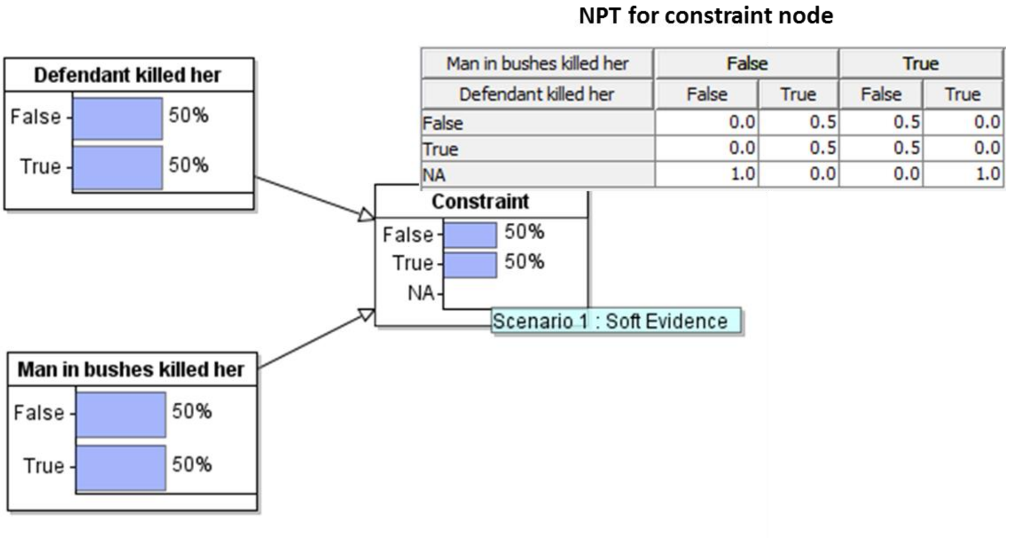
Figure 6 Motive, opportunity and capability idioms applied in Simonshaven[3]

Although the defence has no obligation to provide an 'alternative' narrative, in the Simonshaven case there is such a narrative, namely the 'man in the bushes' as the alternative potential murderer. The problem of integrating alternative narratives into a single BN model is non-trivial, as discussed in (Fenton, Neil, Lagnado, et al., 2016; Verheij et al., 2016; Vlek et al., 2014, 2016), because it requires us to ensure that nodes of the 'alternative' narrative part of the model are 'mutually exclusive' from those of the normal narrative. In particular, we would like to create a separate alternative hypothesis node (in this case 'man in the bushes killed her') that should be true if and only if the 'defendant killed her' node is false[4]. This is because the alternative hypothesis node will generally have its own parents (e.g. motive and opportunity nodes) and children (evidence) nodes that need to be kept 'separate'. However, the BN semantics do not enforce such mutual exclusivity (Fenton, Neil, Lagnado, et al., 2016). In (Fenton, Lagnado, et al., 2013) we proposed the mutual exclusivity idiom whereby a constraint node is introduced to ensure

[3] Some nodes are introduced only for simplification (such as the node 'Defendant motive and capability,' which is simply the AND function of its parent nodes). These nodes have a dashed line and are subsequently hidden in subsequent displays of the model. Where such nodes appear on a path the arcs will be dashed rather than solid.

[4] In this case we ignore any other alternative because none was proposed by the Defence, although one attendee at the Cambridge workshop suggested that there may have been a possible 'mafia type hit'.

that two separate Boolean nodes are mutually exclusive, provided soft evidence[5] is set in the constraint node as shown in Figure 7.



NPT for constraint node

Man in bushes killed her	False		True	
Defendant killed her	False	True	False	True
False	0.0	0.5	0.5	0.0
True	0.0	0.5	0.5	0.0
NA	1.0	0.0	0.0	1.0

Figure 7 Idiom to ensure mutually exclusivity between two Boolean nodes. Constraint node has an NA state as defined in the NPT. Soft evidence is set to ensure NA has zero probability

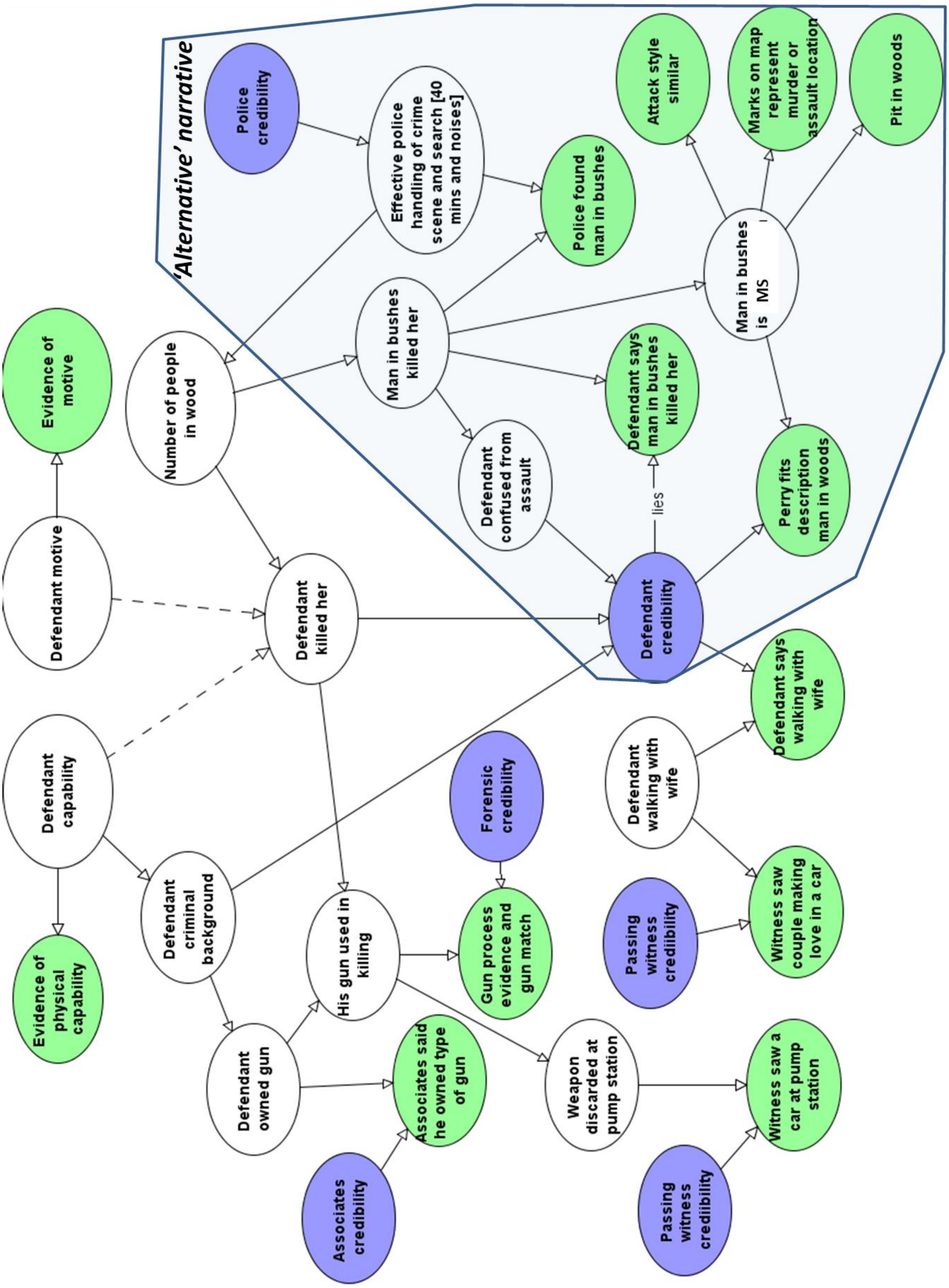
The full Simonshaven BN model that was developed at the Cambridge workshop is the one shown in Figure 8, where we have also highlighted the separation of the alternative narrative. Two nodes (namely the constraint node and the synthetic node that combines capability and motive as shown previously in Figure 6) are specified as 'hidden' nodes because they are used for model convenience and, since we do not need to change their observations, are not shown. The model is an aggregation of the idiom components described above. The production of this BN model was the result of an iterative process starting the day before the September workshop in Cambridge when members of the group worked individually (ranging from 30 minutes to 3 hours each) to create their own first draft of the model using the BN software AgenaRisk (Agena Ltd, 2018). Because each member of the group applied the idioms-based framework their models were reasonably consistent both in terms of variables used and causal structure. The main differences were in the level

[5] The standard type of evidence entered into a BN is 'hard evidence' whereby we specify that a node is in a particular state (e.g. "True"). However, sometimes we wish to enter 'soft evidence' such as here where we are specifying that the state is "either False or True but definitely not NA".

of granularity. Those who spent more time incorporated more of the evidence of the case as additional nodes. The group then worked together for 2 hours on the first day of the workshop on a consolidated version, agreeing on a level of granularity which was a compromise. We decided to omit evidence nodes that we felt were least probative to keep the model sufficiently comprehensible for presentation at the workshop. A further two iterations on the second day resulted in the model in Figure 8. The total effort involved was approximately 24 hours (with the highest individual effort being 10 hours and lowest 3).

In summary, the prosecution narrative relies on establishing motive and opportunity and witness evidence about the suspect's movements and location at and around the crime scene at the time of the murder. Inconsistencies between the suspect's statements and those provided by independent witnesses would suggest the defendant's evidence is not credible and he is not telling the truth.

Critical to the alternative (defence) narrative is the existence of the man in the bushes whom the defendant claims committed the murder. However, given that the police failed to find a man in the bushes after the murder, it could logically be argued that 'absence of evidence' is not equivalent to 'evidence of absence' simply because the police failed to effectively 'lock down' the crime scene within a reasonable time, thus allowing the possibility of escape. Also, the defence narrative suggests evidence that the police failed to investigate noises, heard by police officers on the scene, emanating from the woods. Subsequent evidence became available of an investigation into similar crimes, where the main the suspect, who we shall refer to as "MS", might be claimed to have been the 'man in the bushes'.



'Alternative' narrative
Police credibility
Effective police handling of crime scene and search [40 mins and noises]
Police found man in bushes
Attack style similar
Marks on map represent murder or assault location
Pit in woods
Man in bushes is MS
Man in bushes killed her
Defendant says man in bushes killed her
Perry fits description man in woods
Defendant confused from assault
lies
Defendant credibility
Number of people in wood
Evidence of motive
Defendant motive
Defendant killed her
Defendant says walking with wife
Defendant walking with wife
Defendant capability
Evidence of physical capability
Defendant criminal background
Forensic credibility
His gun used in killing
Gun process evidence and gun match
Witness saw couple making love in a car
Passing witness credibility
Defendant owned gun
Associates said he owned type of gun
Weapon discarded at pump station
Witness saw a car at pump station
Associates credibility
Passing witness crediibility

Figure 8 Full Simonshaven model, subdivided into the prosecution and alternative narratives

It is worth noting that the node ‘defendant credibility’ plays a centralising role in the BN model, given that it is affected by three different hypothetical explanations for the defendant’s behaviour: he has a criminal background (and might therefore lie); he killed her (and would also have a self-interest in lying); or he was genuinely assaulted and was confused (giving rise to unreliable testimony and suggesting innocence). Given the professed ability of BNs to ‘explain away’ competing causal explanations for events this provides a difficult test case for our analysis.

The probability assignments for the credibility nodes in the model are given in Table 2.

Table 2 Probability assignments for credibility nodes

Credibility node	***Probability credible (%)***
Police credibility	90
Forensic credibility	90
Defendant credibility (in absence of any evidence, except existence of crime). *Note that the figure here is determined automatically by the priors for this node’s parent nodes.*	53
Associates credibility (perhaps criminal?)	30
Passing witness credibility (pump station)	90
Passing witness credibility (car)	90

We are assuming that the independent witnesses are generally credible (90%, meaning that there is a 90% chance that what they say is reliable), whereas the associate credibility is not very high (30%).

4. Model results and sensitivity analysis

When we run the model with only the prior assumptions, that is those before any evidence is entered into the model, the probability that the defendant killed the victim is just over 1% as shown in the fragment in Figure 9.

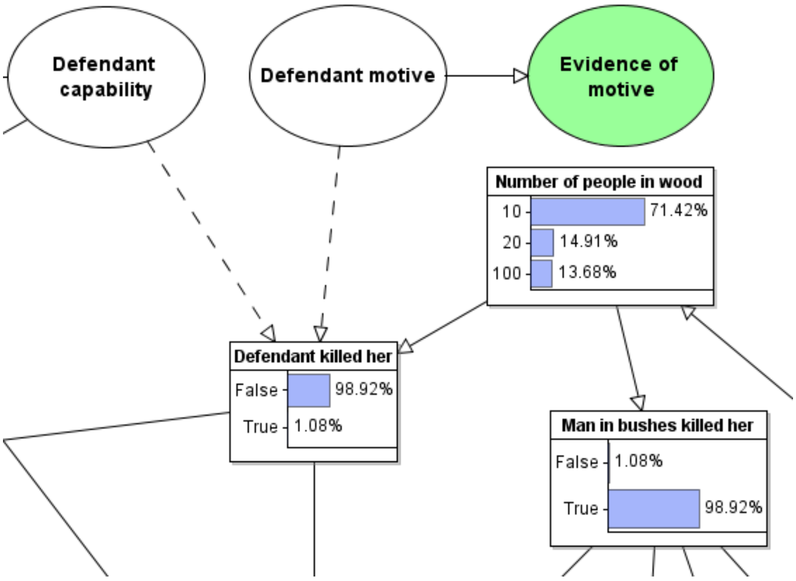


Figure 9 Initial state of model before evidence observed (

As we add the evidence (where (P) denotes prosecution evidence and (D) defence evidence) the probability changes as shown in Table 3. Note, however, that at no point have we added any observation about the credibility of the evidence directly – all credibility nodes are inferred, based on our prior probability assignments and as posterior updates resulting from the evidence entered into the model. This can have a major impact, and this is especially true of the forensic evidence: if, for example, we determine that the forensic evidence has no credibility (i.e. we set this to false) then the probability of guilt will drop significantly.

We can see from Table 3 that, as we add prosecution evidence, the probability of guilt increases to 96% and alongside this the credibility of the defendant falls to less than 1%. Once we consider the evidence under the alterative narrative, firstly relating to the police handling of the crime scene, the probability of guilt falls to 80%. Next, as we add the supporting evidence relating to the identity of an alternative suspect the probability of guilt falls to 46% and the defendant's credibility rises slightly to 6%. However, the fact that the defendant's description of the man in the woods fails to match the alternative MS suspect means that the defendant's credibility falls again, to 4%, and the probability of guilt jumps to 74%.

Table 3 Changes to probability of guilt, and defendant credibility, as evidence is entered in model (P refers to prosecution evidence and D to defence evidence)

Evidence (cumulative)	***Probability defendant guilty (%)*** *[rounded down]*	***Probability defendant credible*** *[rounded down]*
None	1	55
Evidence physical capability and Evidence of motive (P)	21	41
Associates said he owned type of gun + witness saw car at pump station (P)	53	25
Gun process evidence and gun match (P)	93	5
Witness saw couple making love on car (P) but defendant says walking with wife at time (D)	96	< 1
Police failed to find man in bushes and poor handling of crime scene (D)	80	2
Various bits of MS evidence {attack style, marks on map, pit in woods} and fact that defendant says man in bushes killed her (D)	46	6
MS does not fit suspect's description of the man in the woods (P)	74	4

We can use AgenaRisk to perform a sensitivity analysis of selected unobserved variables on the target node 'Defendant killed her'. When we do this for the credibility nodes we get the Tornado graph[6] results shown in Figure 10. This identifies which sources of evidence the conclusions of the case might rely most heavily on. Clearly the defendant's credibility is crucial – if he is a perfectly honest oracle then he cannot possibly be guilty, hence the 0% result for probability of guilt. Next most important is the forensic evidence; if this was to be false then the probability of guilt would plummet to approximately 18%. The passing witness evidence relating to the petrol pump and car are clearly less important, as is the

[6] In a Tornado graph the length of the bars corresponding to each node represent a measure of the impact of that node on the target node.

associate’s credibility. Finally note that the credibility of the police remains unchanged in the sensitivity analysis because its effect on the rest of the BN is blocked by the evidence entered in the model about the handling of the crime scene.

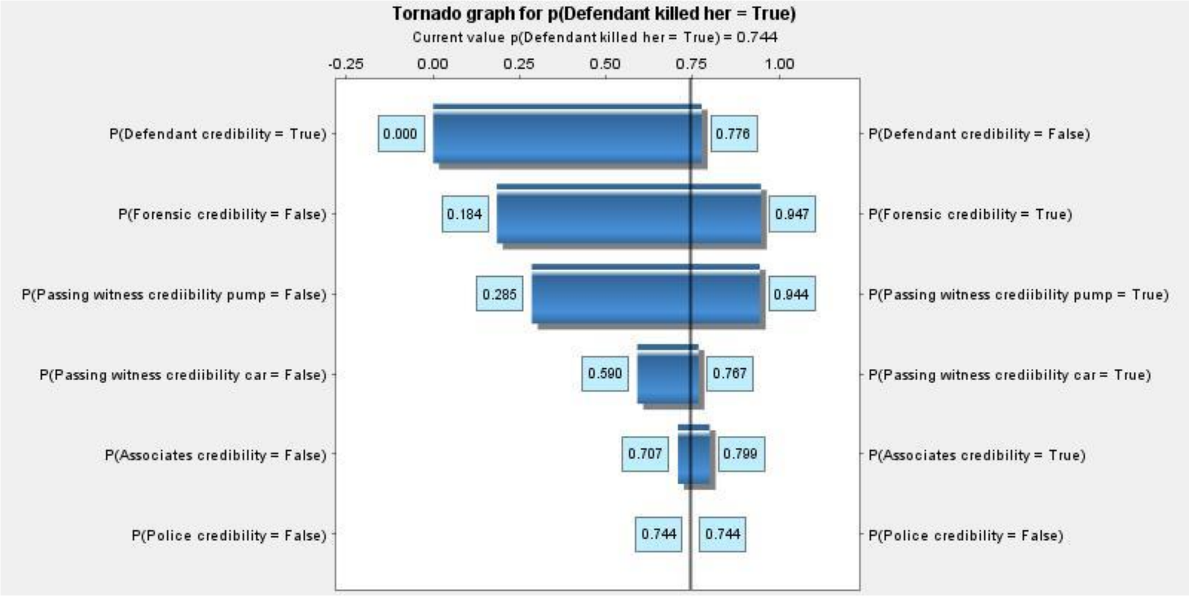


Figure 10 Sensitivity analysis (witness cedibility nodes on defendant guilty, given all cumulated evidence in Table 3)

One of the main objections to the use of BNs for legal arguments has been the necessity and difficulty of assigning probability values to NPTs based on subjective judgements. However, research in other domains has shown that BNs are often robust to changes in their parameters if the conditional dependence relations are correctly modelled in the BN structure (Druzdzel et al., 2000; Oniśko et al., 2013). Several approaches have been developed to assess the sensitivity of an individual BN model to changes in its parameters (Kjærulff et al., 2000; Laskey, 1995).

In the Simonshaven case, we investigated the parameter changes that would change the result of the BN model to ‘beyond reasonable doubt’. As shown above, with our initial prior probability assignments, the posterior probability of guilt is 0.74 after the available evidence is entered to the BN model. We changed each prior probability value individually and examined which prior probability values would make the posterior probability of guilt greater than 0.95 and 0.99 with the same evidence. Table 4 shows the current probability values, and the probability values that would make the posterior of guilt greater than 0.95 and 0.99. For example, the prior probability of ‘Forensic Credibility = False’ is 0.1 in our model. If we decrease this probability to zero, then the posterior probability of guilt would be 0.95 with the same evidence. Since the probability of this parameter cannot be further decreased, changing this parameter alone cannot lead to 0.99 posterior probability of guilt.

Table 4: Sensitivity Analysis of Hypothesis Probability Values

Parameter	*Type**	*Current Value*	*Value For 0.95 Guilt Posterior*	*Value For 0.99 Guilt Posterior*
Pr(Forensic credibility = False)	C	0.100	0	-
Pr(Police found man in bushes = False | Effective police handling of crime scene = False , Man in bushes killed her = True)	E	0.500	0.077	0.015
Pr(Gun process evidence and gun match = False | Forensic credibility = False , His gun used in killing = False)	E	0.500	1	-
Pr(Witness saw a car at pump station = False | Passing wittness credibility = False , Weapon discarded at pump station = False)	E	0.500	1	-
Pr(Man in bushes killed her = False | No. of People in Woods =100)	H	0.010	0.290	0.808
Pr(Man in bushes killed her = False | No. of People in Woods =20)	H	0.050	0.472	-
Pr(No. of People in Woods = 10 | Effective police handling of crime scene = False)	H	0.063	0.787	-

**C: Credibility, E: Evidence, H: Hypothesis*

Parameter	Type*	Current Value	Value For 0.95 Guilt Posterior	Value For 0.99 Guilt Posterior
Pr(Forensic credibility = False)	C	0.100	0	-
Pr(Police found man in bushes = False | Effective police handling of crime scene = False , Man in bushes killed her = True)	E	0.500	0.077	0.015
Pr(Gun process evidence and gun match = False | Forensic credibility = False , His gun used in killing = False)	E	0.500	1	-
Pr(Witness saw a car at pump station = False | Passing wittness credibility = False , Weapon discarded at pump station = False)	E	0.500	1	-
Pr(Man in bushes killed her = False | No. of People in Woods =100)	H	0.010	0.290	0.808
Pr(Man in bushes killed her = False | No. of People in Woods =20)	H	0.050	0.472	-

Pr(No. of People in Woods = 10 | Effective police handling of crime scene = False)	H	0.063	0.787	-

**C: Credibility, E: Evidence, H: Hypothesis*

Our model is robust to the changes in its parameters in the following sense: The BN model contains 99 independent parameters (prior probability values), yet changes to only 7 of these values (on their own) can make the posterior of guilt more than 0.95, and only 2 can make it more than 0.99 when the same evidence is entered. Figures 11 and 12 illustrates the parameter changes required to make the posterior of guilt 0.95 and 0.99 respectively. Note that, the parameters about forensic credibility, accuracy of gun evidence and witness observation need to be made certain to have a 0.95 posterior of guilt from the model. Those would not be reasonable assumptions considering their inherent uncertainty. The other parameter changes in Figures 11 and 12 include assuming that a few people were in the woods even when the police handling of crime scene were poor, lower probability of guilt for other suspects, and that other suspects could be found more easily even when the crime scene was poorly handled.

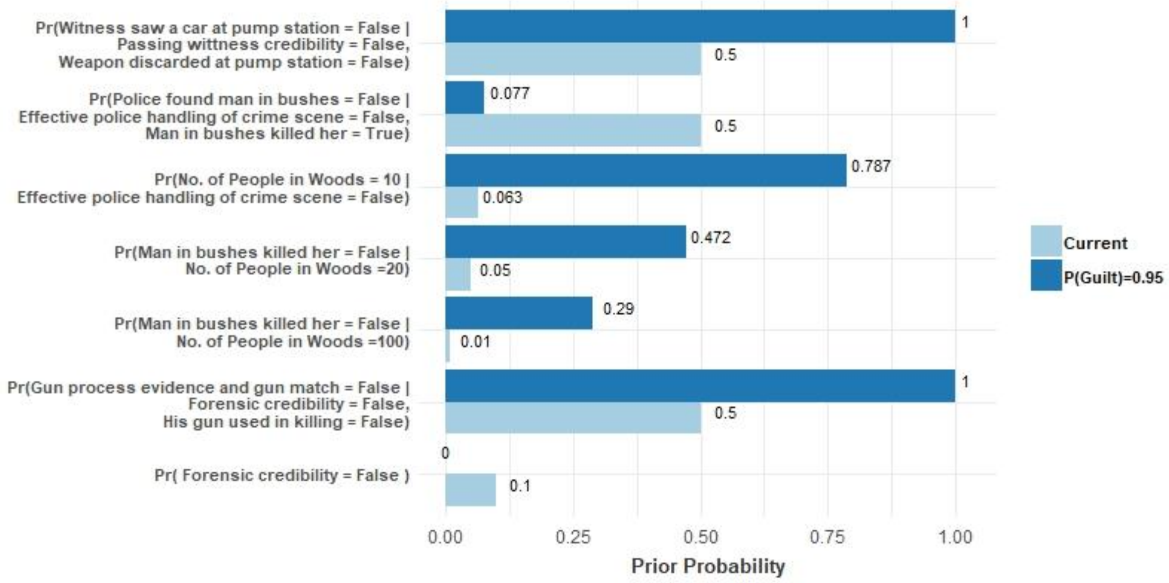


Figure 11 Sensitivity Analysis of Parameters for P(Guilt)=0.95

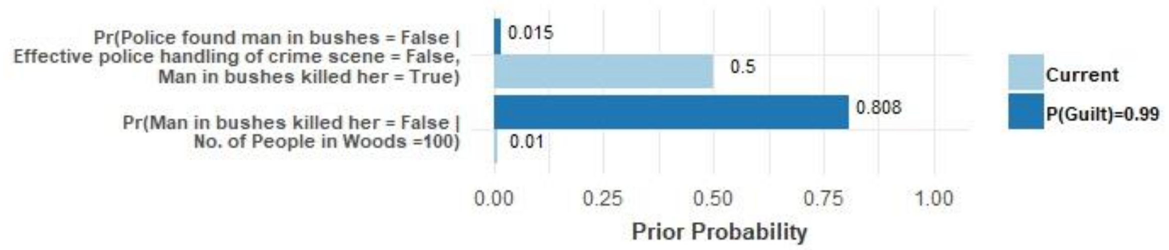


Figure 12 Sensitivity Analysis of Parameters for P(Guilt)=0.99

5. Weakness of the model and potential improvements.

There was much discussion both at the workshop and in the period that followed about the crucial role of the 'opportunity' evidence. In cases like this the number of people who were physically present at the crime scene (which was a very small defined location) should clearly play a major role in the prior probability of guilt. These discussions led to a new approach to modelling and calculating the opportunity prior that is described in (Fenton et al., 2017). In the current model the prior marginal for the node 'Number of people in wood' is at least 10 (as shown in Figure 9). However, using the approach described in (Fenton et al., 2017) would likely have led to a lower number. Indeed, if we knew there were only 10 people then, entering this value as an observation in the current model along with all the other evidence in Table 3, results in a posterior probability of guilt of just over 95% (going below 10 has little impact). Likewise, how we have modelled the police credibility, the effectiveness of the police handling of the crime scene, and how it relates to the possible number of people in the wood, was subsequently discussed and may require reappraisal.

Moreover, with reflection on both the model and additional evidence in the case, we would propose a number of improvements to the model. Because of the (self-imposed) time constraint for modelling this case it was not possible to include all of the evidence. For example, during the trial witnesses were heard that suggested the relation between the suspect and the victim had become friendlier over the last couple of weeks and/or the mismatch between the number and severity of wounds between the victim and the suspect has not yet been addressed. Furthermore, the problem that no weapon has been found was presented as a pivotal part of the defence argument. The prosecution suggests that the suspect could have thrown the weapon away at a nearby pump station. This hypothesis can be evaluated in accordance with the presented time frame of events. We could expand this list endlessly, but also recognize that for both such a probabilistic model and in legal practice, it becomes necessary to combine many findings in one node/conclusion due to time constraints. When constructing and presenting these types of models, the precise definition of a node needs to be specified in order to come up with reasonable probability assignments. For example, in this network, the node 'defendant capability' is a simplification of multiple things. The evidence of physical capability considers eyewitness accounts of the defendant carrying some heavy car parts (during the appeal the defence objected to this evidence claiming that these parts were not heavy enough and provided further evidence opposing this claim), but the link between defendant capability and defendant criminal background should be associated with being mentally capable. Defining what this node represents exactly is challenging and the model would most likely benefit when this part is expanded further. Extending the time frame (for the probabilistic event of how the defendant got rid of the weapon, given that he is the offender) should also be pursued for events before the assault. In the current model, a distinction between the defendant walking with his wife and him making love in the car is made. Apart from the fact that it is possible that both happened at different times, the link between the defendant being credible and him claiming that he walked with his wife is up for discussion.

Currently we do not distinguish between being credible and being accurate. For example, as part of the appeal process the crime scene was visited to establish if it was possible to see whether there were people in a car at the car park from the place where the witness claimed he observed them making love. It was recognized that it would be difficult to see people, let alone observe what they were doing. Hence, although it is likely that this witness is credible, accuracy is another question. In other words, credible means different things when talking about a suspect and an innocent witness. Currently, 'credible' summarizes whether someone is telling the truth and whether someone remembers what happened accurately.

6. Discussion and Conclusions

We focus our discussion on the four questions posed to the authors of each of the different approaches presented in this special issue:

To what extent is the analysis objective and to what extent is it based on subjective beliefs, assumptions and choices?

The main criticisms of the BN approach relate to over-reliance on subjectivity. Specifically:

1. Constructed BNs tend to be completely ad-hoc in the sense that different modellers end up with different models
2. The task of building a BN is complex and error prone
3. It is necessary – and extremely difficult – to provide the prior probability table values and the inevitable lack of data means that most of these values are based on subjective judgments.

However, by using the idioms-based approach in a tool like AgenaRisk we have demonstrated that it is possible to overcome objections 1 and 2 in a realistic case. A team of four people completed the model with total effort of approximately 24 hours. Using the completed model in AgenaRisk it is possible to run multiple scenarios that compute the updated probabilities of all unobserved variables (notably the 'defendant killed her' node).

With respect to objection 3 we recognise that in this case the prior probabilities in the BN model were primarily subjective (although because of the idioms-based approach and the AgenaRisk tool they were not difficult to construct). However, the causal structure of the model places many strong constraints on the feasible range of probabilities that could be reasonably assumed. Our sensitivity analysis considered the full range of priors for all of the nodes in the model. The fact that the results were reasonably robust across the full range demonstrates that the ‘subjectivity’ of the probabilities may be an exaggerated impediment. In any case fact finders would have to arrive at their own personal judgements for all nodes so at the very least we are simply replacing tacit implicit subjectivity with a numerical form that is openly subject to test and challenge.

How natural is the analysis from a cognitive and legal point of view?

Although a BN model is not as easily accessible for legal professionals as narrative/scenario based approaches, we believe it is a natural representation of a legal case and its evidence. This is because the Bayesian approach is a formalisation of the standard approach to legal reasoning which is to continually revise any prior beliefs about guilt/innocence as new evidence is presented. Above all, and in contrast to alternative non-Bayesian modelling approaches (Bex, 2011; Prakken, 2010; Verheij et al., 2016) considered at the workshop and presented in this special issue, a BN model can provide an explicit probability of guilt, given all prior assumptions and evidence observed. This is the ultimate requirement for a judge or jury member in reaching a verdict.

Did the analysis identify errors or biases in the reasoning of the judge, prosecutor or defence?

It was beyond the scope of this particular study to read through any statements made by the judges or lawyers. However, as demonstrated comprehensively in (Fenton et al., 2011), one of the major benefits of formulating a BN model is that it is also able to identify common errors and biases in legal reasoning.

Does the analysis respect the legal constraints, such as the burden and standard of proof and the right to remain silent?

The analysis respects the most important legal constraints in ways that are very explicit. In particular:

- The ‘innocent until proven guilty’ requirement is explicitly modelled by the prior probability associated with the number of people in the wood - the assumption is that the defendant was no more likely to have committed the crime than any other person who was there.
- The burden and standard of proof are explicitly captured by the target probability for guilt.
- The right to remain silent would be modelled by simply presenting no defence evidence at all in the model. If the model – when run only with the prosecution evidence – results in a posterior probability of guilt that lies below an agreed threshold, the defendant would be considered not guilty.

Acknowledgements

This work was supported by EPSRC grant no EP/K032208/1, the Simons Foundation, and the ERC (ERC-2013-AdG339182-BAYES_KNOWLEDGE). The participants thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the FOS programme, where much of this work was undertaken. We also thank the three reviewers for their detailed comments.

References

Agena Ltd. (2018). AgenaRisk. http://www.agenarisk.com. Retrieved from http://www.agenarisk.com

Aitken, C. G. G., Connolly, T., Gammerman, A., Zhang, G., Bailey, D., Gordon, R., & Oldfield, R. (1995). Bayesian belief networks with an application in specific case analysis. In A. Gammerman (Ed.), *Computational Learning and Probabilistic Reasoning*. John Wiley and Sons Ltd.

Bex, F. J. (2011). Arguments, stories and criminal evidence: a formal hybrid theory. *Artificial Intelligence and Law*, *18*(2), 123–152.

Cook, R., Evett, I. W., Jackson, G., Jones, P., & Lambert, J. A. (1998). A hierarchy of propositions: deciding which level to address in casework. *Science and Justice*, *38*, 231–239.

Dawid, A. P., & Evett, I. W. (1997). Using a graphical model to assist the evaluation of. complicated patterns of evidence. *Journal of Forensic Sciences*, *42*, 226–231.

Dawid, A. P., & Mortera, J. (1998). Forensic identification with imperfect evidence. *Biometrika*, *85*(4), 835–849. https://doi.org/10.1093/biomet/85.4.835

Druzdzel, M. K., & van der Gaag, L. C. (2000). Building probabilistic networks: where do the numbers come from? *IEEE Transactions on Knowledge and Data Engineering*, *12*(4), 481–486.

Edwards, W. (1991). Influence diagrams, Bayesian imperialism, and the Collins case: an appeal to reason. *Cardozo Law Review*, *13*, 1025–1079.

Fenton, N. E., Berger, D., Lagnado, D. A., Neil, M., & Hsu, A. (2013). When 'neutral' evidence still has probative value (with implications from the Barry George Case). *Science and Justice*. https://doi.org/http://dx.doi.org/10.1016/j.scijus.2013.07.002

Fenton, N. E., Lagnado, D. A., Dahlman, C., & Neil, M. (2017). The opportunity prior: a simple and practical solution to the prior probability problem for legal cases. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law,* (pp. 69–7). New York: ACM Press.

Fenton, N. E., Lagnado, D. A., & Neil, M. (2013). A general structure for legal arguments using Bayesian networks. *Cognitive Science*, *37*(61–102). https://doi.org/http://dx.doi.org/10.1111/cogs.12004.

Fenton, N. E., & Neil, M. (2011). Avoiding legal fallacies in practice using Bayesian networks. *Australian Journal of Legal Philosophy*, *36*, 114–150.

Fenton, N. E., & Neil, M. (2012). *Risk Assessment and Decision Analysis with Bayesian Networks*. Boca Raton: CRC Press. Retrieved from www.bayesianrisk.com

Fenton, N. E., Neil, M., & Berger, D. (2016). Bayes and the law. *Annual Review of Statistics and Its Application*, *3*(1), 51–77. https://doi.org/10.1146/annurev-statistics-041715-033428

Fenton, N. E., Neil, M., & Hsu, A. (2014). Calculating and understanding the value of any type of match evidence when there are potential testing errors. *Artificial Intelligence and Law*, *22*, 1–28. https://doi.org/http://dx.doi.org/10.1007/s10506-013-9147-x

Fenton, N. E., Neil, M., Lagnado, D. A., Marsh, W., Yet, B., & Constantinou, A. (2016). How to model mutually exclusive events based on independent causal pathways in Bayesian network models. *Knowledge-Based Systems*, *113*, 39–50. https://doi.org/10.1016/j.knosys.2016.09.012

Hepler, A. B., Dawid, A. P., & Leucari, V. (2007). Object-oriented graphical representations of complex patterns of evidence. *Law, Probability and Risk*, *6*(1–4), 275–293. https://doi.org/doi:10.1093/lpr/mgm005

Huygen, P. E. M. (2002). Use of Bayesian nelief networks in legal reasoning. *17th BILETA Annual Conference*. Free University, Amsterdam. Retrieved from http://www.bileta.ac.uk/02papers/huygen.html

Jowett, C. (2001). Lies, damned lies, and DNA statistics: DNA match testing. Bayes' theorem, and the criminal courts. *Med Sci. Law*, *41*(3), 194–205.

Kadane, J. B., & Schum, D. A. (1996). *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. John Wiley & Sons.

Kjærulff, U., & van der Gaag, L. C. (2000). Making sensitivity analysis computationally efficient. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence* (pp. 317–325). Morgan Kaufmann Publishers Inc.

Koehler, J. J. (1993). Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics*, *34*, 21–39.

Lagnado, D. A., Fenton, N. E., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument and Computation*, *4*(1), 46–63. https://doi.org/dx.doi.org/10.1080/19462166.2012.682656

Laskey, K. B. (1995). Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, *26*(6), 901–909.

Neil, M., Tailor, M., & Marquez, D. (2007). Inference in hybrid Bayesian networks using dynamic discretization. *Statistics and Computing*, *17*(3), 219–233. Retrieved from http://dx.doi.org/10.1007/s11222-007-9018-y

Oniśko, A., & Druzdzel, M. J. (2013). Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2013.01.004

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Palo Alto, CA: Morgan Kaufmann.

Pearl, J., & Mackenzie, D. (2018). *The book of why : the new science of cause and effect*. New York: Basic Books.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument & Computation*, *1*(2), 93–124. https://doi.org/10.1080/19462160903564592

Schum, D. A. (1989). Knowledge, probability, and credibility. *Journal of Behavioral Decision Making*, *2*(1), 39–62. https://doi.org/10.1002/bdm.3960020104

Taroni, F., Aitken, C. G. G., Garbolino, P., & Biedermann, A. (2014). *Bayesian Networks and Probabilistic Inference in Forensic Science* (2nd ed.). Chichester, UK: John Wiley & Sons.

Thompson, W. C., Taroni, F., & Aitken, C. G. G. (2003). How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Sciences*, *48*(1), 47–54.

Verheij, B., Bex, F., Timmer, S. T., Vlek, C. S., Meyer, J.-J. C., Renooij, S., … M., W. (2016). Arguments, scenarios and probabilities: connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk*, *15*(1), 35–70. https://doi.org/10.1093/lpr/mgv013

Vlek, C. S., Prakken, H., Renooij, S., & Verheij, B. (2014). Building Bayesian networks for legal evidence with narratives: a case study evaluation. *Artificial Intelligence and Law*, *22*(4), 375–421. https://doi.org/10.1007/s10506-014-9161-7

Vlek, C. S., Prakken, H., Renooij, S., & Verheij, B. (2016). A method for explaining Bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, *24*(3), 285–324.

https://doi.org/10.1007/s10506-016-9183-4