Casing the Joint: Steps Towards Evaluating the Joint Probability Distribution of a Bayesian Network of a Criminal Case

Ludi van Leeuwen[1,*], Leya Hampson[2]

[1]Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands
[2]Department of Transboundary Legal Studies, Faculty of Law, University of Groningen, The Netherlands

Abstract

Bayesian networks can be used to model evidential reasoning in criminal cases. A key challenge is that the probabilities in such networks are often subjective, as they concern past events for which no empirical data is available. Even when local probability estimates seem reasonable, the resulting joint distribution may have implications that are out of line with modelers' expectations. Identifying a misalignment between the modeler's intuition and the model output is the first step towards identifying whether this is due to probabilistic fallacies by the modeler, or a mistake when constructing the model. We apply three methods for evaluating such distributions to a Bayesian network of a real case: computing individual likelihood ratios, tracking cumulative posterior probabilities, and calculating posteriors for all relevant evidence combinations. The three methods operate at different levels of abstraction and each method identifies problems in the Bayesian network of the case. We argue these methods form a useful step in an iterative design cycle for constructing subjective legal Bayesian networks.

Keywords

Evidential reasoning, Bayesian networks, case study

1. Introduction

In order to come to a justified conclusion in a criminal court case and prevent miscarriages of justice, correct evidential reasoning is essential. Various approaches exist that allow for a formal representation of evidential reasoning [1]. One such approach that has emerged is scenario reasoning in Bayesian networks [2, 3, 4].

This approach combines two more fundamental approaches for reasoning about legal evidence, scenario reasoning [5, 6, 7, 8] and probabilistic reasoning [9, 10, 11]. In the scenario approach, multiple alternative narratives are put forward to explain the evidence that was found in a case. The scenarios are assessed based on their coherence, completeness and plausibility. By considering more than one scenario, fact-finders use one or more innocent scenarios to explain the evidence, rather than focusing only on how the evidence fits with the guilt scenario, thereby avoiding tunnel vision. However, the scenario approach does not prescribe *how* evidence should be weighted with respect to the question of guilt, especially when evidence is dependent or contradictory.

The probabilistic approach by itself cannot tell the modeler what relevant variables to consider and what evidence to collect in the real world. However, once the relevant variables are known, the probabilistic approach can tell a modeler how evidence should be weighted and combined, given that probability assessments have been made.

In a probabilistic framework, supporting evidence (e) makes a hypothesis ($\Pr(h \mid e)$) more probable than before the evidence was provided, and attacking evidence makes a hypothesis less probable than before, with reference to a contrasting hypothesis $\neg h$. This property of evidence can be expressed using the likelihood ratio (LR): $\frac{\Pr(e|h)}{\Pr(e|\neg h)}$. When $\frac{\Pr(e|h)}{\Pr(e|\neg h)} = 1$, the evidence e does not distinguish between h and

RELAF'26: Reasoning with Evidence in Law Enforcement and Forensics, Workshop at ICAIL 2026, Singapore
*Corresponding author.
l.s.van.leeuwen@rug.nl (L. v. Leeuwen); l.l.hampson@rug.nl (L. Hampson)
0000-0003-3165-4376 (L. v. Leeuwen); 0009-0008-0761-4802 (L. Hampson)

$\neg h$, when $\frac{\Pr(e|h)}{\Pr(e|\neg h)} > 1$, evidence e supports h over $\neg h$ and when $\frac{\Pr(e|h)}{\Pr(e|\neg h)} < 1$, e supports $\neg h$ over h, or, attacks h.

The likelihood ratio is a clear probabilistic interpretation of evidence strength. For individual pieces of forensic evidence, such as DNA analysis [10], the LR is well-understood and often reported. However, reasoning about the evidence for a scenario or multiple alternative scenarios, requires reasoning beyond a single source or activity level hypothesis: it requires taking known dependencies into account, with combinations of multiple (sub-)hypotheses.

Bayesian networks provide a natural framework for integrating scenario-based and probabilistic approaches to evidential reasoning. A Bayesian network is a compact graph representation of a joint probability distribution over a set of relevant variables [12, 13]. Bayesian networks are more appropriate for modeling scenarios than performing probability calculations by hand due to the number of elements in a scenario.

Bayesian networks have been used to model the weight of evidence in different criminal cases [14, 15, 3, 16, 17]. There is no standard method for modeling entire criminal cases in Bayesian networks, though various methods have been proposed [18, 19]. These methods mainly focus on selecting the relevant variables and constructing the graph of the Bayesian network. One open problem is that for these large networks, many probabilities need to be entered into the network. Of these probabilities, few can be obtained from empirical data[1] as they concern unique, hypothesized past events, collected into scenarios. Consequently, many probabilities in the network must be estimated directly by modelers rather than estimated empirically, before they can be entered into the conditional probability tables (CPTs). The resulting joint distribution therefore reflects the subjective judgments of the modeler.

However, it is a large step from the local conditional probabilities in conditional probability tables to the resulting joint probability distribution. Even when every local probability estimate seems reasonable to a modeler, the resulting distribution, which emerges from these individual conditional probabilities via the probability calculus, might have implications that are out of line with the expectations of the modeler. Evaluating the joint probability distribution of the network as a whole might expose modeling flaws. To our knowledge, no systematic evaluation approach exists for these types of subjective Bayesian networks in a legal context, specifically on evaluating the joint probability distributions, rather than on their graph structure.

In this paper, we first present a Bayesian network graph that was constructed based on two independently modeled Bayesian networks of a specific case [17]. The probabilities for this new Bayesian network were estimated in collaboration by the two modelers.

We then evaluate the joint distribution of this Bayesian network in three different ways. First, we calculate the likelihood of evidence with respect to the hypothesis of guilt. Second, we calculate how evidence that was actually found in the case affects the posterior probability of guilt. Third, we calculate the posterior probability of guilt across all combinations of relevant evidence. For each of these calculations, we evaluate whether a misalignment indicates a modeling flaw or an error in a modeler's intuition. We conclude that these three methods together form a useful evaluation step in an iterative design cycle for legal Bayesian networks.

2. State of the art

Bayesian networks (BNs) are a way of representing how one should reason under uncertainty. A BN $\langle V, E, \Pr\rangle$ is a directed acyclic graph (DAG) $\langle V, E\rangle$ consisting of a set of random variables $V = \{V_1, ..., V_n\}$, the set E of directed edges $E \subset V \times V$ [12, 13]. These edges encode the independence relation between the variables V, and Pr is the joint probability distribution defined over those variables V. A variable is only conditioned on its parents in the graph, $\Pr(V_1 \mid \pi(V_1))$. This means that, given the independence relation is correctly represented, the probability distribution $\Pr(V)$ can be encoded compactly in the graph. The conditioning of the variable on its parents is done locally in its conditional probability tables (CPTs), where parameters are entered that represent the conditional probability distributions of the

[1] Even if based on empirical data, we would still need to overcome the problem of the reference class, [20].

child for each combination of parent values. This way, the joint distribution is uniquely defined as $\Pr(V_1, ..., V_n) = \Pi_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ and we can calculate prior distributions and condition on evidence to find posterior distributions.

Prior research has focused on how to construct larger networks from smaller building blocks called idioms [18]. These idioms are often quite abstract: a graph structure combined with a textual description, sometimes with constraints on the CPTs [21]. Idioms can be combined with each other, resulting in large-scale Bayesian networks. While we can often verify the behavior of simple idioms by putting in reasonable numbers in the CPTs and checking whether the posteriors predicted are aligned with either known frequencies or our intuitions, verifying and evaluating the correct behavior of larger-scale BNs that represent whole cases is an open problem. Methods exist for investigating parameter choices and evidential impact in Bayesian networks [22, 23]. However, these methods do not address whether the joint distribution that emerges from subjective CPT parameters is aligned with the expectations and intuitions of the modelers.

3. Method

In this section we first describe the modeled case and present the Bayesian network. Then we describe our three methods of evaluation.

3.1. Case study

The modeled case is inspired by a real criminal case. The whole verdict can be found online. [2] Essentially, the defendant is accused of robbing a supermarket. On the night of the robbery, the offender drove to the supermarket and parked his car, got out, ran to the supermarket with his face covered. In the supermarket, he threatened the cashiers and other witnesses with a gun, before forcing open the register, taking the cash, and running back to his car.

The prosecution offers several pieces of evidence that support the scenario of the defendant being the offender: his phone pinged near the crime scene, the car in his driveway is similar to the car in the CCTV footage, as determined by a forensic expert as well as a witness (W2), he matches the general description that the witnesses (P1, P2, W1) gave, and his gun showed some similarities to the gun used in the robbery. Additionally, his gait and movement were similar to the gait and movement of the offender, and a witness (W1) in the store recognized him post-hoc as her ex-partner.

As counterevidence, the defense provided that one of the witnesses claimed that the offender has a neck tattoo, which the defendant does not have. Additionally, the defendant's car is common, hence a match is not unexpected. Furthermore, all witnesses claim that the gun used in the robbery was silver, yet the defendant's gun is gold. Finally, there is the testimony from the defendant himself that states that he is innocent, and that night, he was at home and then went to work, as confirmed by his phone location.

The summary of the verdict states "A large number of facts and circumstances point towards the defendant as the offender. Individually, not all of these facts and circumstances carry significant probative value". [3] Combining the pieces of evidence coherently makes modeling this case a challenge. We model the evidence in a Bayesian network (Figure 1). The network has 51 nodes in total, with 67 edges. There are 23 evidence nodes in total, 13 nodes representing hypotheses and 15 reliability nodes. In this paper, rather than explaining the construction of the network, we focus on the evaluation of the joint distribution expressed by the network by looking at the following three calculations.

3.2. Calculating Individual Likelihood Ratios

We set the hypothesis node for *guilt* to *true*, and then, for each evidence node E, note the value of $E = true$. This is the conditional probability $\Pr(e \mid guilt)$. Then we do the same, but we set *guilt* to *false*,

[2] https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:GHARL:2025:3013

[3] Translated from https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:GHARL:2025:3013

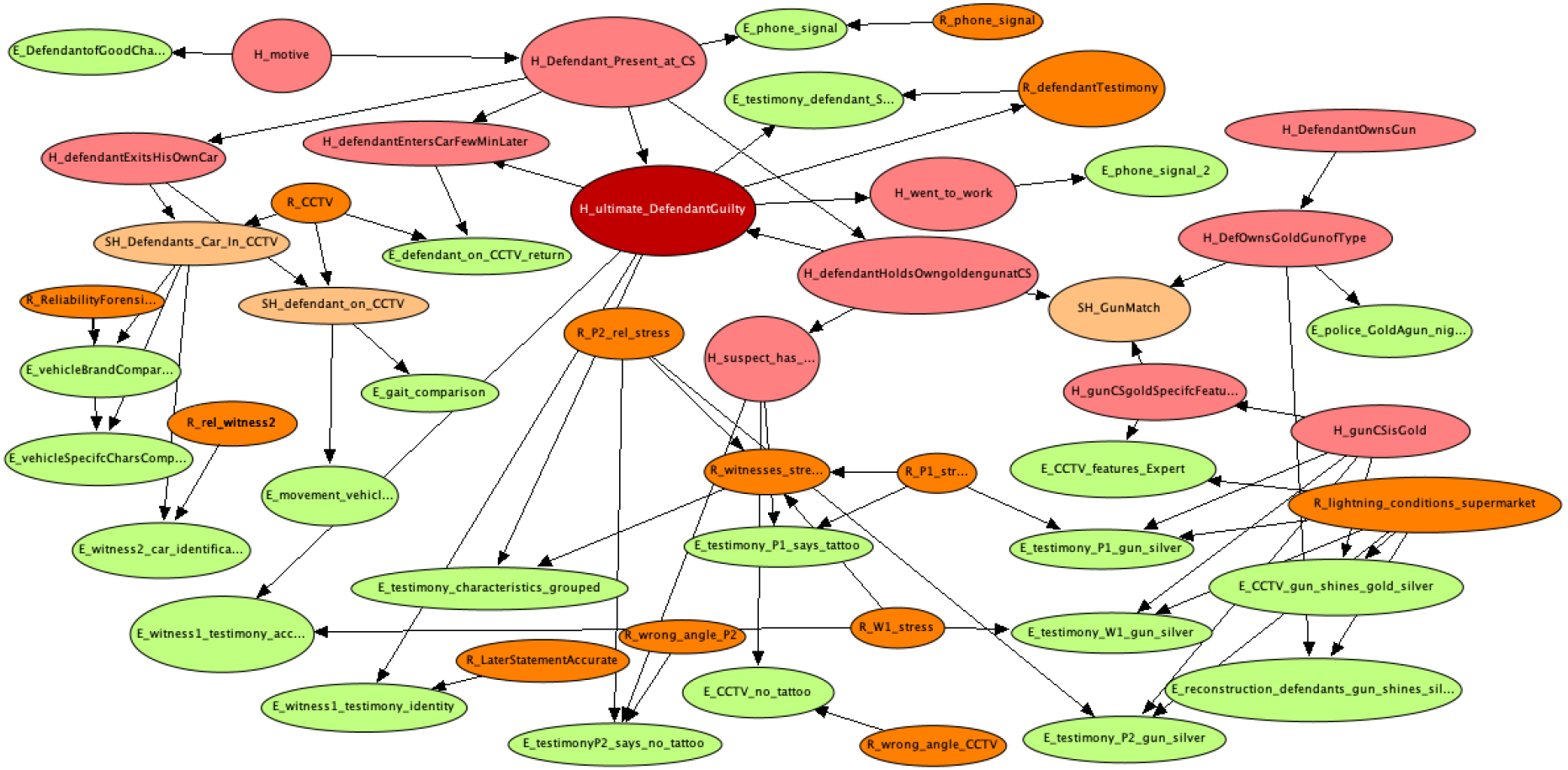

Figure 1: Bayesian network model of the case. The nodes are color-coded according to their function: Dark red is the ultimate hypothesis of guilt, light red are the scenario elements, pale orange are even narrower subhypotheses, green are evidence and orange are reliability nodes.

Evidence	LR	Favors	Verbal expression	Aligns?
phone signal work	0.23	Defendant	slightly more probable	Yes
defendant of good character	0.809	Defendant	slightly more probable	Yes
testimony defendant	0.90	Defendant	slightly more probable	Yes
testimony P1 tattoo	0.96	Defendant	slightly more probable	No
testimony P2 no tattoo	1		equally probable	No
video CCTV no tattoo	1		equally probable	No
gun (combined)	1		equally probable	No
testimony W1 accent	6	Prosecution	slightly more probable	Yes
vehicle brand	7	Prosecution	slightly more probable	Yes
movement	8	Prosecution	slightly more probable	Yes
testimony characteristics grouped	9	Prosecution	slightly more probable	Yes
video CCTV offender returning to car	10	Prosecution	slightly more probable	Yes
vehicle specific characteristics	12	Prosecution	more probable	Yes
testimony W2 car	14	Prosecution	more probable	Yes
phone signal crime scene	14	Prosecution	more probable	Yes
gait	90	Prosecution	more probable	Yes
testimony W1 identity	141	Prosecution	much more probable	No

Table 1
Ordered likelihoods for positive evidence (LR+) per evidence set, considered independently from each other, for 16 nodes, with one line in the table representing the combined strength of the gun evidence (6 nodes grouped together). Verbal expressions were translated from https://www.forensischinstituut.nl/vakbijlage/waarschijnlijkheidstermen

resulting in $\Pr(e \mid \neg guilt)$. We then calculate the likelihood ratio of each individual piece of evidence e: $\frac{\Pr(e|guilt)}{\Pr(e|\neg guilt)}$. This reflects the strength of each individual piece of evidence, which we can use to assess whether this strength aligns with our expectations as modelers.

3.3. Calculating Cumulative Posterior Probabilities of Found Evidence

We cumulatively set the evidence in the network to the values that were found in the real case. Based on the temporal ordering in which the evidence was found in the real case, we calculate the posterior probability $\Pr(guilt \mid e)$ for every added piece of evidence. This allows us to follow how the posterior evolves as evidence accumulates, and to assess the combined effect of all evidence found in the case.

3.4. Calculating the Posterior Probability of Guilt Across All Combinations of Relevant Evidence

Each evidence node has two outcomes, either true or false. We enumerate all 2^n possible combinations of evidence node values, ranging from all nodes set to *true* to all nodes set to *false*. For each of these evidence combinations, we calculate the posterior probability of guilt. This reveals how the posterior probability responds to different evidence configurations, not only the evidence that was found in this specific case. Finally, we map the values of true and false in the outcomes to *favors prosecution* vs *favors defense*, depending on the meaning of the node. This altogether allows us to see the effect of the number of favorable pieces of evidence on the posterior. Note that since we have a binary mapping, 0 pieces of evidence that favor the prosecution means 16 pieces of evidence that favor the defense.

4. Results

In this section we present the results of our three evaluation methods.

4.1. Calculating Individual Likelihood Ratios

We present the likelihood ratios of each of the individual pieces of evidence in Table 1. Our analysis reveals evidence strength for individual pieces of evidence ranging from *slightly more probable* favoring the defendant to *much more probable* favoring the prosecution. The four pieces of evidence that favor the defendant are all relatively weak (0.23-0.96), while the 10 pieces of evidence that favor the prosecution range from being relatively weak (6) to stronger (LR=141). Three pieces of evidence have a likelihood of 1.

As stated in the summary of the verdict, individual pieces of evidence have limited evidential strength. This is reflected in the LR assessments of these pieces of evidence. However, there are several LRs that were not aligned with our expectations:

First, the gun evidence has an LR of 1. During the construction phase, both modelers expressed doubt about how to model this aspect of the case, where the defendant's gun is possibly matched with the gun at the crime scene, and this gun-match is in turn evidence for the defendant committing the crime. In the resulting BN, since the LR is 1, this evidence has no effect on the posterior. Intuitively, finding a matching gun and the contradicting witness testimonies about its color should affect the posterior.

The testimony of witness 1 carries disproportionate weight. This witness first testified that she recognized the accent of the offender. After she heard who the defendant was, she suddenly recognized her ex-partner. This evidence has the highest LR of all in this network, yet this recollection should intuitively carry less weight.

Finally, the tattoo evidence does not affect the posterior. If we believe witness P1, the defendant could not have been the offender, pointing strongly in favor of the defense. Yet, the LR is close to 1. The other two pieces of evidence related to the tattoo also have LRs of 1.

Hence, using this method we have identified 5 pieces of evidence whose strengths are misaligned with the intuition of the modelers.

4.2. Calculating Cumulative Posterior Probabilities of Found Evidence

We present the cumulative posterior probability of guilt in Figure 2. We add evidence in the chronological order that we find during the case: first the witness testimonies, then the CCTV footage of the movement and gait of the defendant, as well as the car, then the evidence offered by the defense. We end with the identification evidence of W1.

We start with a prior probability of 0.0001, which is based on a rough estimate of an opportunity prior [24] of gun-owning men in the area. Adding the initial evidence makes the posterior of guilt for the defendant larger, but not larger than 0.01. We see that the tattoo and gun evidence do not affect our posterior. Then, the gait evidence causes a large jump in the posterior of guilt, towards 0.32. Other evidence from the prosecution raises the posterior to a lesser degree, but altogether to 0.94. The evidence of the defense lowers the posterior probability again towards 0.78, and finally the evidence of the first witness causes the posterior probability of guilt to be 0.998.

The evaluation method of cumulative evidence tells us, like in the LR method, that in the next iteration of the Bayesian network we should reconsider the tattoo and the gun evidence, as these do not affect the posterior of guilt as we would expect. The rest of the evidence behaves in a way that aligns with our expectations: evidence weakly favoring the defendant lowers the posterior and vice versa. However, the ultimate hypothesis of guilt, which nears 1, might be overestimated.

In considering the cumulative evidence, we can see the impact of dependencies between evidence that was not visible in the LR approach. For example, finding the matching vehicle brand (LR=8) causes the posterior of guilt to jump from 0.73 to 0.87. However, when the other car evidence is added, the specific vehicle characteristics (LR=12) and the testimony of W2 about the car (LR=14), the posterior probability increases only to 0.89. When we change the order in which the evidence is added, the same effect occurs: the most change in posterior is caused by the first evidence that is added. Since these pieces of evidence are dependent in the network via the car evidence, this effect is not surprising. However, it does show the risk of only using the likelihood ratio to assess individual evidence strengths.

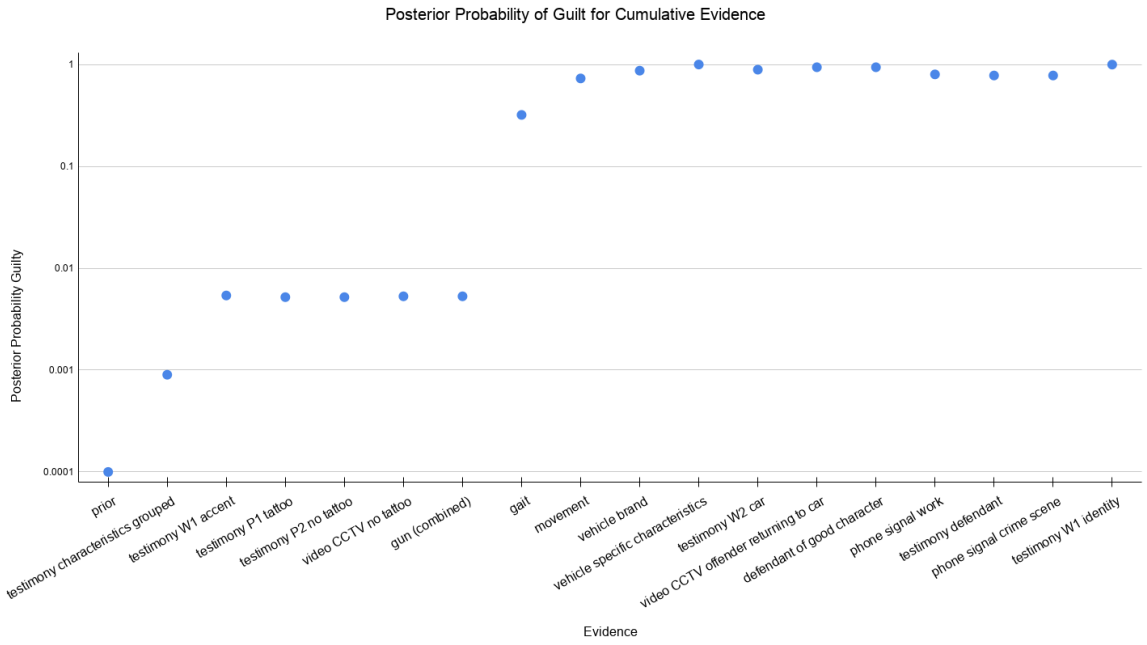

Figure 2: Posterior probability of guilt given cumulative evidence. Note the log-scale on the y-axis, in order to show the value of the prior.

In the context of the network, the combined evidential strength might be lower than when the pieces are considered individually.

4.3. Calculating the Posterior Probability of Guilt Across All Combinations of Relevant Evidence

We exclude the gun evidence from these calculations since its LR is 1. This reduces the computational complexity, resulting in 16 pieces of evidence in total.

We plot, for each number of evidence nodes set in favor of the prosecution, the distribution of posterior probabilities that are found. For 0 or 16 favorable evidence pieces, there is only one possible combination, yielding a single datapoint. For 1 or 15 favorable evidence pieces, there are 16 datapoints, and so on.

In Figure 3, we see that the Bayesian network represents a range of different posterior probabilities of guilt (from 0 to 1) for given evidence sets.

When there are fewer than 6 pieces of evidence in favor of the prosecution, the posterior probability is always near 0. When there are 6 favorable pieces of evidence, the maximum posterior probability that can be found is around 0.5. When we have between 7 and 12 favorable pieces of evidence, the posteriors take on a range: here it depends, for some combinations of evidence the posterior of guilt is near 1, while for others the posterior is near 0, and everywhere in between. We see a slight bimodal distribution in the plot at this range. When 13 pieces of evidence are favorable to the prosecution, the posteriors cluster around 1 but it is still possible that the probability is as low as 0.2. From 14 pieces of evidence onward, all posteriors are above 0.8 at least.

From this plot we see that the network is able to represent both innocent and guilty posteriors for the defendant, depending on the evidence that is found. We can see that we cannot deem the defendant guilty when there are too few pieces of evidence that support guilt: so even one very strong piece of evidence would not be enough to convict if the other pieces of evidence point towards innocence. From this plot we see that at least 7 pieces of prosecution-favoring evidence are needed to get to a high posterior probability of guilt. However, not any arbitrary 7 evidence pieces. In this way, the network

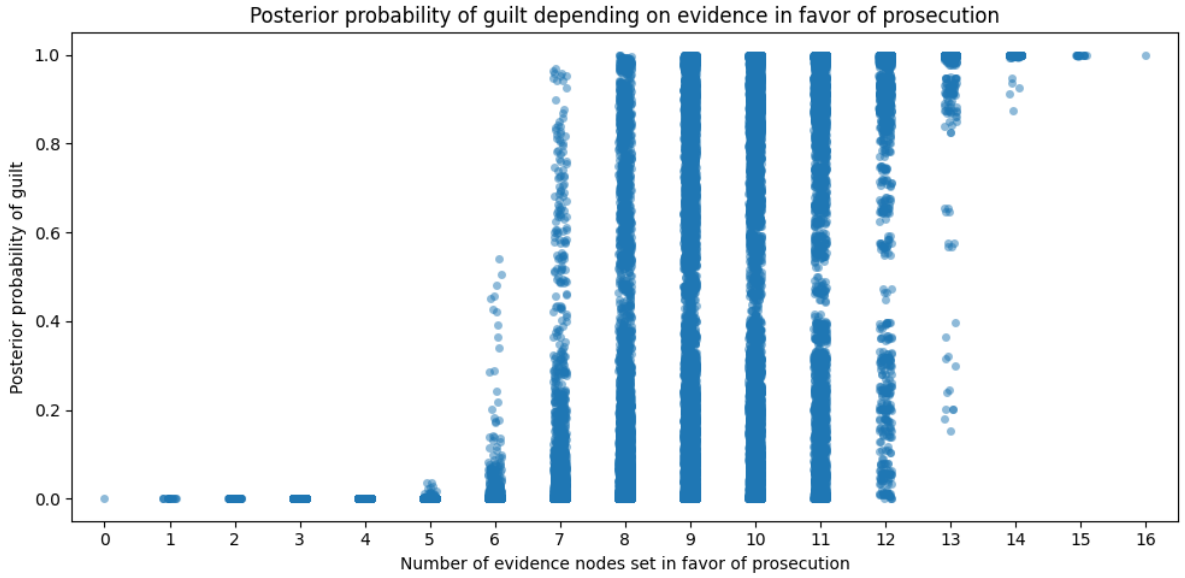

Figure 3: The effect of the number of evidence nodes set in favor of the prosecution on the posterior probability

does represent the burden of proof.

We further investigate which evidence combinations are sufficient for the network to predict a high posterior probability of guilt. We find that there are two sufficient evidence sets of 6 pieces of evidence for which, when they are true, the probability of guilt is at least 0.5. These are: (*gait, testimony characteristics grouped, movement, testimony W1 identity, phone signal crime scene, testimony W1 accent*) and (*gait, testimony characteristics grouped, movement, testimony W1 identity, phone signal crime scene,* ¬ *phone signal work*). [4]

Interestingly, these are not the 6 most prosecution-favored pieces of evidence according to the LRs in Table 1. For instance, no car matching evidence is included (LRs 14, 12), even though *movement* (LR=8) is. This suggests that the independence relation in the Bayesian network causes these pieces of evidence to be less important than their LRs alone would suggest.

According to the network, if these combinations of evidence are set, there is no other evidence that we can set in the network such that the posterior probability of guilt drops below 0.5. However, both these sets of evidence seem too weak for such a high posterior probability of guilt. If it turns out that the defendant's car does not match the car at the crime scene, and that one witness saw a tattoo on the offender, with no conflicting evidence, the posterior probability of guilt would still be at least 0.5, yet this conflicts with our intuition. Therefore, in the next iteration of the network, we need to make these pieces of evidence weigh more strongly towards an innocent defendant.

5. Discussion

We present a Bayesian network model of a criminal case and evaluate the joint probability distribution that it implies. We evaluate it from three different perspectives: 1) the likelihood ratios of individual pieces of evidence, independent of each other, 2) a comparison of how the posterior probability changes cumulatively when evidence is set and 3) the posterior probability of guilt under different evidence valuations.

When using the likelihood ratio method of evaluating the joint probability distribution of the BN, we find that the gun evidence is underestimated, the testimony of witness 1 is overestimated, and the tattoo evidence is both under and overestimated, depending on the specific piece. This method of evaluation

[4] Analysis code and Bayesian network will be made available.

makes visible where the LRs in the BN are misaligned with our estimations of the evidence and hence is a useful method for evaluating the joint distribution of the BN.

The cumulative evidence showed us how the posterior changes under the evidence that we found in this specific case. The total collection of evidence found in the actual case is enough to get to a high posterior probability of guilt, while taking dependencies between pieces of evidence into account in a way that the LR method neglects.

Calculating the posterior probability of guilt across all combinations of relevant evidence gives us a holistic perspective on the distribution presented by the Bayesian network. No individual piece of evidence is strong enough to reach a high probability of guilt. We also found that it is not enough to count the number of favorable or unfavorable evidence pieces: finding 13 pieces of favorable evidence can result in a posterior of 0.2, under some evidence combinations. However, if one finds the right 6 evidence pieces, the probability of guilt is always higher than 0.5.

The three methods of evaluation complement each other by operating at different levels of abstraction. The LR method gives a direct, interpretable assessment of individual evidence strength, but is blind to dependencies between evidence nodes. The cumulative evidence method makes those dependencies visible, but only for the specific evidence that was actually found in this case. The effect of evaluating all evidence combinations captures the full range of what the network can express, but loses the narrative thread of the case. In our evaluation, all three methods converged on the same problematic evidence: the gun, tattoo and W1 testimony. This strengthens our confidence that these are genuine modeling flaws rather than artifacts of any single method. The method of calculating guilt across all combinations of relevant evidence revealed that the car match evidence was weaker than we would expect. This was not revealed by the other methods.

5.1. Limitations and Future Research

In the original network, many reliability nodes are present. These nodes are useful in considering different perspectives on the same evidence. However, at scale they make a network difficult to evaluate due to the increased complexity. Therefore, in this case we never set evidence on the reliability nodes and they are summed out in the calculations. However, instantiating reliability nodes could affect the dependency relations in the graph, depending on its d-separation properties. Including the reliability nodes in the evaluation would be valuable as it would allow us to reason about different possible standpoints: for example, the weight of W1's testimony is very different depending on her reliability. This will be left to future work.

Additionally, we evaluated the methods only on a single Bayesian network on a single case. If these methods are to be used, they should be evaluated and validated on a range of different Bayesian networks: from simple idioms to larger scale networks.

In the case of the individual LR method, we now consider each piece of evidence separately. This does not allow us to look at how probabilistic dependencies between variables have been modeled.

In the cumulative evidence method, we add evidence in a way that is aligned with the time at which the evidence was discovered. However, this is somewhat misleading, because at the time of the first testimony, no defendant was identified yet. Therefore, we could not meaningfully talk about the probability of guilt for this defendant. Additionally, in a timeline of an investigation, evidence can be discovered based on evidence that you already have. In a Bayesian network, no new evidence can be uncovered. Instead, this is a post-hoc representation of how a posterior changes over time.

In the combination of relevant evidence method, the posterior of guilt seems to be divided in a bimodal way for evidence combinations between 8 and 12 prosecution-favorable pieces. That is, there seem to be combinations for which the posterior approaches 1, and combinations for which the posterior approaches 0. In future work we would like to analyze this in a more fine-grained way. This is also the least scalable method with regards to complexity.

These evaluation tools can only evaluate relative to our modeler's intuition. For example, even though the LRs for some pieces of evidence correspond to the modelers' intuition, we do not know if this correlates with the judges' estimation of the strength of this evidence in the actual case, or with

objective frequencies, insofar as they exist. Therefore, these evaluation methods should be seen as a diagnostic tool for improving a subjective Bayesian network, rather than a method for validating a Bayesian network against an empirical joint probability distribution in the real world, that we do not have access to. Integrating other perspectives, for example, with independent auditors, forensic advisors, or other experts, is necessary for justifiable legal reasoning.

There is the possibility to combine the different evaluation tools: one could combine the LR approach with the cumulative evidence approach, to discover the LR of a single piece of evidence conditioned on all evidence that was found before. This will be explored in further research.

In these evaluation tools, we contrasted the model's output with the intuition of the modelers. When these diverged, we took the side of the modeler's global intuition and deemed that the model was wrong in some way. However, the model always draws probabilistically correct conclusions from the local probabilities. From this, we assumed that the modeler made a mistake in filling out the CPTs. However, it might instead be the case that the modeler's intuition of the evidence strength at the global level is wrong. It is known that people have cognitive biases when reasoning probabilistically [25, 26, 27]. The extent to which modeler's intuition can diverge from model output and whether this is caused by a cognitive fallacy or a modeling mistake is to be explored in future research.

5.2. Conclusion

The three evaluation methods identify specific pieces of evidence that should be reconsidered in order to create a Bayesian network that is more aligned with the intuitions of the modelers. We showed that three evaluation steps could be part of an iterative design cycle towards constructing subjective Bayesian networks for modeling evidential reasoning.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Sonnet 4.6) in order to assist with formatting, paraphrasing and rewording, improving writing style and checking for grammar and spelling mistakes.

References

[1] M. Di Bello, B. Verheij, Evidential reasoning, in: G. Bongiovanni, G. Postema, A. Rotolo, G. Sartor, C. Valentini, D. N. Walton (Eds.), Handbook of Legal Reasoning and Argumentation, Springer, Dordrecht, 2018, pp. 447–493.

[2] C. S. Vlek, H. Prakken, S. Renooij, B. Verheij, Building Bayesian networks for legal evidence with narratives: a case study evaluation, Artificial Intelligence and Law 22 (2014) 375–421. URL: http://link.springer.com/10.1007/s10506-014-9161-7. doi:10.1007/s10506-014-9161-7.

[3] N. Fenton, M. Neil, B. Yet, D. Lagnado, Analyzing the Simonshaven Case Using Bayesian Networks, Topics in Cognitive Science 12 (2020) 1092–1114. URL: https://onlinelibrary.wiley.com/doi/10.1111/tops.12417. doi:10.1111/tops.12417.

[4] L. van Leeuwen, B. Verheij, R. Verbrugge, S. Renooij, Building a stronger case: Combining evidence and law in scenario-based bayesian networks, in: F. Lorig, J. Tucker, A. D. Lindstrom, F. Dignum, P. Murukannaiah, A. Theodorou, P. Yolum (Eds.), Hybrid Human AI Systems for the Social Good - Proceedings of the 3rd International Conference on Hybrid Human-Artificial Intelligence, IOS Press, Malmö, 2024, pp. 291–299.

[5] N. Pennington, R. Hastie, Inside the Juror: The Psychology of Juror Decision Making, Cambridge Series on Judgment and Decision Making, Cambridge University Press, Cambridge, 1993.

[6] W. A. Wagenaar, P. J. Van Koppen, H. F. Crombag, Anchored Narratives: The Psychology of Criminal Evidence, St Martin's Press, New York, NY, 1993.

[7] F. J. Bex, Evidence for a Good Story: A Hybrid Theory of Arguments, Stories and Criminal Evidence, Dissertation University of Groningen, Groningen, 2009.
[8] A. Mackor, H. Jellema, P. J. van Koppen, Explanation-based approaches to reasoning about evidence and proof in criminal trials, in: Law and mind: A survey of law and the cognitive sciences, Cambridge University Press, Cambridge, 2021, pp. 431–470.
[9] C. Aitken, A. Gammerman, Probabilistic reasoning in evidential assessment, Journal of the Forensic Science Society 29 (1989) 303–316. URL: https://linkinghub.elsevier.com/retrieve/pii/S0015736889732709. doi:10.1016/S0015-7368(89)73270-9.
[10] A. Biedermann, F. Taroni, Bayesian networks for evaluating forensic DNA profiling evidence: A review and guide to literature, Forensic Science International: Genetics 6 (2012) 147–157. URL: https://linkinghub.elsevier.com/retrieve/pii/S1872497311001359. doi:10.1016/j.fsigen.2011.06.009.
[11] M. Vink, J. de Koeijer, M. Sjerps, A template Bayesian network for combining forensic evidence on an item with an uncertain relation to the disputed activities, Forensic Science International: Synergy 9 (2024) 100546. URL: https://www.sciencedirect.com/science/article/pii/S2589871X24000937. doi:https://doi.org/10.1016/j.fsisyn.2024.100546.
[12] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
[13] F. V. Jensen, T. D. Nielsen, Bayesian Networks and Decision Graphs, Information science and statistics, 2nd ed ed., Springer, New York, 2007.
[14] J. B. Kadane, D. A. Schum, A Probabilistic Analysis of the Sacco and Vanzetti Evidence, Wiley, Chichester, 1996.
[15] C. S. Vlek, When Stories and Numbers Meet in Court: Constructing and Explaining Bayesian Networks for Criminal Cases with Scenarios, Ph.D. thesis, Rijksuniversiteit Groningen, Groningen, 2016.
[16] L. van Leeuwen, B. Verheij, A comparison of two hybrid methods for analyzing evidential reasoning, in: Legal Knowledge and Information Systems, IOS Press, Amsterdam, 2019, pp. 53–62.
[17] L. Hampson, L. van Leeuwen, Investigating the value of qualitative bayesian net-works of complete cases as "double-check" tools on traditional judicial reasoning: An exploratory study., in: Proceedings of the AI4EVIR Workshop on AI for Evidential Reasoning, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4157/paper5.pdf.
[18] D. A. Lagnado, N. E. Fenton, M. D. Neil, Legal idioms: a framework for evidential reasoning, Argument and Computation 4 (2012) 1–18.
[19] C. Vlek, H. Prakken, S. Renooij, B. Verheij, A method for explaining Bayesian networks for legal evidence with scenarios, Artificial Intelligence and Law 24 (2016) 285–324. doi:10.1007/s10506-016-9183-4.
[20] M. Colyvan, H. M. Regan, S. Ferson, Is it a Crime to Belong to a Reference Class, Journal of Political Philosophy 9 (2001) 168–181. URL: https://onlinelibrary.wiley.com/doi/10.1111/1467-9760.00123. doi:10.1111/1467-9760.00123.
[21] A. Onnes, M. Dastani, R. Dobbe, S. Renooij, Extending idioms for Bayesian network construction with qualitative constraints, in: Information Processing and Management of Uncertainty in Knowledge-Based Systems - 20th International Conference, IPMU 2024, Proceedings, Lecture Notes in Networks and Systems, Springer, United States, 2024, pp. 415–426. doi:10.1007/978-3-031-74003-9_33.
[22] L. C. van der Gaag, S. Renooij, V. M. Coupé, Sensitivity Analysis of Probabilistic Networks, in: Advances in Probabilistic Graphical Models, Springer, Berlin, Heidelberg., 2007, pp. 103–124.
[23] S. Renooij, Relevance for robust Bayesian network MAP-explanations, in: A. Salmerón, R. Rumí (Eds.), Proceedings of The 11th International Conference on Probabilistic Graphical Models, volume 186 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 13–24. URL: https://proceedings.mlr.press/v186/renooij22a.html.
[24] N. Fenton, D. Lagnado, C. Dahlman, M. Neil, The opportunity prior: a proof-based prior for criminal cases, Law, Probability and Risk 18 (2019). doi:10.1093/lpr/mgz007.
[25] W. C. Thompson, E. L. Shumann, Interpretation of statistical evidence in criminal trials: The

prosecutor's fallacy and the defense attorney's fallacy, Law and Human Behaviour 11 (1987) 167–187.

[26] C. G. G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd Edition, Wiley, Chichester, 2004.

[27] C. Dahlman, A systematic account of probabilistic fallacies in legal fact-finding, The International Journal of Evidence & Proof 28 (2023) 45–64. URL: https://journals.sagepub.com/doi/10.1177/13657127231209019. doi:10.1177/13657127231209019.