Follow the trace

Evaluating physical and digital forensic findings given activity level propositions using Bayesian networks

Vink, M.

Publication date
2025

Chapter 1

INTRODUCTION

"Doubt is not a pleasant condition, but certainty is absurd."

Voltaire (1770)

Figure 1.1: The CSI-team finds a toenail fragment in the shoe and Grissom tries to match it with the toenail clippings of Mr. S (artwork created using ChatGPT-4o (OpenAI, 2024)).

1.1 "Whodunnit?"

"Whodunnit?" has been a popular formula in TV series for years. Consider one of the cases[1] from the pilot episode of *CSI: Crime Scene Investigation* CSI in 2000, where a man is found dead with a gunshot wound in a house. The homeowner, Mr. S, admits shooting the man but claims it was self-defense because the man had forcibly entered his home and attacked his family. A crime scene investigator finds a broken toenail fragment inside the victim's shoe. Back in the lab, Gil Grissom, a forensic scientists, investigates the striation pattern of the toenail fragment along with Mr. S's toenail clipping and finds a matching striation pattern between the two. Grissom calls his colleague: "It is Mr. S." Grissom concludes that Mr. S murdered the man and that he must have worn the victim's shoes after death to stage a home invasion. Here, Grissom finds that the toenail fragment and Mr. S's clippings lead directly to Mr. S as the perpetrator—as if the toenail fragment itself describes the crime.

[1] The case scenario has slightly been altered for the sake of introducing the readers to the key topics of this thesis.

In reality, this represents a clear example of flawed reasoning: while we may infer someone's identity from a trace ("who"), without additional information we cannot infer "what" that person did—let alone establish what offense (a murder) occurred[2].

We can refer to the "hierarchy of propositions" (Cook *et al.*, 1998) to explain why inferring an "activity" from only "source" evidence is often problematic. The hierarchy has three main levels: source level, activity level, and offense level. The levels represent the nature of the issue[3] that forensic scientists address, as well as how closely that issue relates to the ultimate question in court (such as "Did Mr. S murder the victim?").

In general, the higher the level, the more comprehensive our information must be in order to address the issue level. Therefore, jumping directly from source evidence (a matching striation pattern) via activity level claims (Mr. S wore the victim's shoe after the shooting) to offense level conclusions (that Mr. S committed murder) without additional supporting information is problematic[4]. Activity level issues can be addressed by forensic scientists, but require information such as shedding, transfer, persistence, prevalence, and recovery of the evidential traces—information that the scientists potentially have. Offense level issues, however, typically fall outside the domain of forensic

[2]For readers who are familiar with forensic science practices, note that we present Grissom in evaluative mode rather than investigative mode. He is addressing issues centered around the defendant Mr. S to introduce key concepts: the hierarchy of propositions, probabilistic reasoning, and Bayesian networks. In investigative mode, it would be appropriate for Grissom to make educated guesses about what might have occurred during the incident based on his findings. See Jackson *et al.* (2006) for an elaboration on the difference between these modes and conclusions drawn in these modes.

[3]I am intentionally staying away from the topic of propositions for now but will address them shortly when discussing the Likelihood Ratio framework. For now I call it "addressing activity level issues".

[4]The attentive reader immediately sees that, apart from the fact that we cannot simply jump between levels, this is not a form of balanced and logically sound reasoning. I will delve deeper into this fallacy in the next section, when I introduce the Likelihood Ratio framework.

scientists since they require information that scientists usually cannot access[5].

While Cook *et al.* (1998) first described the hierarchy of propositions, Evett (1983) had earlier emphasized that asking "What is the probability that this blood came from that person?" is typically not the most meaningful question for forensic scientists to answer. Let us illustrate this with Grissom's case. Consider new information that comes to light[6]: Mr. S and the victim turn out to be brothers, and Mr. S states that he had lent his shoes two weeks ago to his brother. The brother broke into his house wearing Mr. S's shoes. They ended up fighting each other and the shooting was self-defense. The relevant question shifts from "Whose toenail is it?"[7] to "Did Mr. S wear the victim's shoes only after the shooting or did the victim potentially wear Mr. S's shoes for the last two weeks?" As Grissom moves up a level in the hierarchy from source to activity, he needs more information than just the matching striation pattern between the toenail fragment and the toenail clipping of Mr. S. For example, he needs information about how likely a toenail fragment would shed from a toe and transfer to the inside of a shoe given a certain wearing duration and activity, and how likely it is that the toenail fragment would persist in a shoe over time.

Both in theory and practice, there is increasing awareness of the importance of addressing activity level issues (Prinz *et al.*, 2024; Taylor & Kokshoorn, 2023; Willis *et al.*, 2015). This trend is also evident in the growing number of requests

[5]This statement is being challenged in Chapter 7.

[6]From this point forward, the case study is no longer equal to the case scenario from the CSI episode.

[7]I do not intend to say that this question is useless to answer. This question is still of importance when addressing activity level. As Biedermann & Hicks (2016) explain: "but we now must realize that, increasingly often, the traditional perspective of sub-source level propositions, and the main focus on the rarity of the corresponding features (i.e., the so-called conditional genotype probability), may represent only a first step of the evaluative process. This does not make these evaluation approaches wrong, only less comprehensive. The fact is that the extrinsic characteristics of the trace material (i.e., low quantities of DNA) and the propositions of interest have changed, and it is important to realize that this represents the relevant starting point." (p. 3)

at the Netherlands Forensic Institute to evaluate findings given activity level propositions, particularly in the DNA department (Kokshoorn & Luijsterburg, 2023). While previous requests primarily focused on determining the source of a trace (e.g., "Whose DNA is this?"), there is now an increasing focus on additional questions about the alleged activity ("Who did what?"). Forensic science literature increasingly focuses on improving activity level evaluations, for example by providing results from case-specific experiments on the shedding, transfer, persistence, prevalence and recovery of traces under different possible circumstances (e.g., De Ronde *et al.* (2021); Taylor *et al.* (2017); Wieten *et al.* (2015)).

1.2 Why so certain?

Beyond Grissom's leap from source to offense level, his reasoning demonstrates another noteworthy aspect to discuss: drawing conclusions (the toenail fragment comes from Mr. S) from his findings with absolute certainty, also known as "categorical" conclusion. As Evett *et al.* (2000) argue, such definitive opinions cannot be "rationalised logically. " (p. 237)

Historically, expressing doubt as a forensic expert in court was not considered a sign of expertise (Dixon & Ward, 2023). In the late 19th century, science was viewed as a source of mechanical objectivity that could eliminate "individual perspective" (Daston, 2007). This "trust" in science eventually led to a new discipline in the legal context: *criminalistics* (the use of science in the criminal justice system). With this belief in science firmly established, the notions of uncertainty became associated with a lack of expertise, and experts consequently tended to avoid expressing any uncertainty (Dixon & Ward, 2023).

The forensic science community later recognized that the field needed more than just technical scientific methods—it required correct and balanced reasoning about evidence and dealing with uncertainties. This recognition led to the introduction of Bayesian reasoning in forensic science, particularly the

likelihood ratio (LR) as a way for forensic scientists to express their conclusions. Consequently, a new forensic discipline, *forensic inference and statistics*, emerged. With this development, forensic scientists are encouraged to be transparent about uncertainties—this is not a lack of expertise but rather a fundamental part of it.

Using the LR framework, experts no longer make direct claims about scenarios (hereafter also referred to as hypotheses or propositions), but instead evaluate their findings by weighing them against competing scenarios. Findings can update prior beliefs about a certain scenario, but without prior beliefs, making claims about a scenario is incorrect. This reasoning is captured in (the odds form of) Bayes' theorem:

$$\frac{P(H_1 \mid E,\ I)}{P(H_2 \mid E,\ I)} = \frac{P(E \mid H_1,\ I)}{P(E \mid H_2,\ I)} \times \frac{P(H_1 \mid I)}{P(H_2 \mid I)} \tag{1.1}$$

where the left-hand side represents the posterior odds and the right-hand side is the product of the LR and the prior odds. The prior odds represent our belief in propositions H_1 and H_2 before considering the findings (or evidence) E, while taking into account the "framework of circumstances" (or background information) I (Evett *et al.*, 2000), and the posterior odds reflect our "updated" belief about H_1 and H_2 after considering the findings E within the same framework of circumstances I. Bayes theorem is often compared to a scale with sides H_1 and H_2 (Jackson, 2000). The initial position of the scale represents the prior odds for H_1 and H_2 and the LR indicates how much weight is added to each side of the scale. The final position of the scale[8] then represents the posterior odds. When the LR equals one, the balance stays where it was, leaving our prior beliefs exactly where they were[9]. The "weight of the evidence" depends heavily on the specific circumstances of each case—in some

[8] We imagine a scale that can have any position.

[9] In fact, the logarithmic LR is the "weighted" LR: an LR of one is equivalent to a log(LR) of zero (i.e., no weight is added to the scale).

situations, the evidence may add more or less weight to the scale given the framework of circumstances.

The role of forensic scientists is to derive an LR by assigning the probabilities within this framework of circumstances using their scientific knowledge, experimental data, published literature, and/or professional experience (see "Grissom's case: source level evaluations" below). They should report an LR that in essence says:

"The findings are X times more probable if H_1 is true than if H_2 is true,"

where X (times) represents either a numerical value or a verbal equivalent for the LR. The ENFSI Guideline for Evaluative Reporting in Forensic Science by Willis *et al.* (2015) provides a verbal scale for this purpose. For example, LRs of 50 and 5000 correspond to the verbal equivalents "more probable" and "much more probable," respectively.

Box 1 - Grissom's case: source level evaluations

Instead of Grissom's conclusion about the match between the toenail fragment and Mr. S's toenail clippings with absolute certainty, Grissom should report an LR, recognizing that his findings can provide different degrees of support for two competing propositions:

$$\frac{P(E \mid H_1)}{P(E \mid H_2)}$$

where the evidence E is the matching striation pattern[a] between the toenail fragment and Mr. S's toenail clippings and H_1 : The toenail fragment comes from Mr. S and H_2 : the toenail fragment comes from someone else[b].

Grissom needs to assess two probabilities to derive the LR: the probability of the matching striation pattern between the toenail fragment

and Mr. S's clippings given that the toenail fragment comes from Mr. S (the numerator of the LR) and the probability of the matching striation pattern given that the toenail fragment comes from someone else (the denominator of the LR). Assume the following assessment (this is not based on experimental data):

$$LR = \frac{P(E \mid H_1)}{P(E \mid H_2)} = 1/0.02 = 50$$

Grissom can express his conclusion based on this LR as:

"The matching striation pattern between the toenail fragment and the toenail clippings is more probable given that the toenail fragment comes from Mr. S than if it comes from someone else."

The above conclusion differs significantly from claiming that "given the matching striation pattern, the toenail fragment is about fifty times more probable to be from Mr. S than from someone else" (claim about the posterior odds—called a fallacy of "the transposed conditional" (Hicks *et al.*, 2022) or "the prosecutor's fallacy" (Thompson & Schumann, 1987)). It differs even more from declaring with absolute certainty that "the toenail fragment comes from Mr. S." (claim only about proposition H_1 and hence not "balanced").

[a] For the purpose of this introduction, we make some simplifying assumptions: the evidence is either a matching or non-matching striation pattern, the striation pattern does not change over time, and the striation pattern is the same whether a toenail is clipped or torn.

[b] The more relevant alternative proposition H_2 here is that the toenail fragment comes from the victim, as there is no information whatsoever that the toenail fragment could be from someone else other than the victim or Mr. S. For this case example, Grissom has no access to the victim's body nor any toenail clippings of the victim. The only physical material that is available to him is Mr. S's toenail clippings and the toenail fragment from the shoe of interest.

In summary, the LR framework offers benefits based on the three princi-

ples of evidence interpretation (Evett & Weir, 1998): it encourages balanced reasoning by considering at least two competing propositions (Principle I), it requires the forensic scientists to think about the probability of findings given the two competing propositions (Principle II) and it requires to evaluate our findings in a framework of circumstances (Principle III). Moreover, it promotes transparency about an expert's evaluation, and it acknowledges that evaluating findings may require both subjective judgment and scientific objectivity, demonstrating that these two can work together perfectly without diminishing an expert's expertise.

Pioneering work on the LR framework in forensic science began roughly around 1970, with contributions from for example Evett (1983, 1984); Finkelstein & Fairley (1970); Lindley (1977). Following a time period of many advancements in deriving LRs (by hand) for several problems (e.g., Curran *et al.* (1997); Dawid & Mortera (1996); Evett (1983, 1984, 1987); Taroni *et al.* (1996)), including standard works such as Aitken & Stoney (1991); Aitken *et al.* (2020); Evett & Weir (1998); Robertson & Vignaux (1995)[10], to the design of data-trained LR systems that yield numerical LRs (e.g., Gill *et al.* (2021b); Leegwater *et al.* (2017); Meuwly *et al.* (2017)).

Around 1995, the Netherlands adopted this work and began implementing the LR framework in practice. Researchers like Sjerps, Kloosterman, Broeders, and Van Koppen began to motivate practice to use the LR framework (e.g., Broeders (2000); Crombach *et al.* (2006); Sjerps & Biesheuvel (1999); Sjerps & Kloosterman (1999). Their work demonstrated to both the Dutch forensic and legal domain that the LR framework is the preferred way of reasoning about evidence for forensic scientists. A major milestone in Europe was the publication of the *ENFSI Guideline for Evaluative Reporting in Forensic Science* by Willis *et al.* (2015), which promoted the LR framework as a necessary component of logical, balanced, robust, and transparent evaluative reporting.

[10]First edition of Aitken *et al.* (2020) appeared in 2004.

Another achievement is the ISO 21403-4 *interpretation* standard, which provides requirements and recommendations across forensic science disciplines to harmonize practices (Berger, 2025).

Recall that propositions can be at different levels in the hierarchy of propositions (Cook *et al.*, 1998), and consequently, the LR can be derived at these different levels. Evaluating findings at activity level requires additional information, which becomes visible when using the LR framework (see "Box 2 - Grissom's case: activity level evaluation" below).

Box 2 - Grissom's case: activity level evaluation

The court asks Grissom to evaluate his findings given a set of activity level propositions and to report an LR. Grissom considers the alternative scenario for the presence of Mr. S's toenail[a] in the shoe from previous section: Mr. S wore the shoe a few weeks before the incident after which he lent the shoes to his brother (the victim). Grissom defines the propositions H_3 and H_4[b]:

H_3 : Mr. S wore the victim's shoes after the incident.
H_4 : Mr. S last wore his shoes two weeks prior to the incident.

Given H_3, the shoes of interest belong to the victim and Mr. S only wore the shoes after the shooting (to kick in the door to stage a home invasion). Given H_4, the shoes of interest belong to Mr. S and Mr. S wore his shoes for the last time two weeks prior to the shooting. The victim wore the shoes during and after the shooting.

The LR configuration and resulting conclusions follow a similar structure at both source and activity level. However, Grissom is now working within a different framework of circumstances. Grissom defines the evidence E as the presence of Mr. S's toenail fragment in the shoe[c]. He needs to set two probabilities in order to derive the LR for H_3 and H_4: the probability of finding Mr. S's toenail fragment if Mr. S wore the victim's shoe after the incident, and the probability of finding Mr. S's toenail fragment if Mr. S last wore his shoe two weeks before the incident.

To assess these probabilities, Grissom needs information about how toenail fragments shed (detach from the toenail) based on wearing

duration and activities, how these fragments transfer to shoes during different wearing activities, and how long such fragments persist in shoes over time.

Here is an example of reasoning under the propositions and the framework of circumstances: given H_3, Mr. S wore the victim's shoe for only a short time but with significant impact—such as kicking the door—which influences the probability of shedding and transfer. Under H_4, Mr. S presumably wore his shoes for a longer period but further in the past, which would likely affect persistence. Moreover, there is no information about specific wearing activities given H_4.

[a] For simplicity, the evidence E is now defined as the presence of Mr. S's toenail fragment in the shoe of interest. Therefore, the source level LR is excluded from this evaluation.

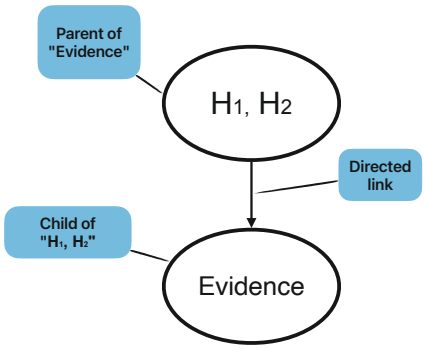
[b] I must emphasize that setting relevant propositions is both challenging and crucial, as different propositions will yield different LRs. The propositions must align closely with what is at stake in the case to ensure the LR remains relevant to the court and setting these propositions is an art on its own. Some good guidelines can be read in Hicks *et al.* (2022) and in Chapter 4 of Taylor & Kokshoorn (2023). More so, Chapter 7 of this thesis is fully dedicated to formulating propositions in digital forensic casework, specifically in Trojan horse defense cases.

[c] See Footnote *a* of this box.

1.3 Bayesian networks

Grissom's LRs are pretty straightforward to derive. However, deriving an LR (by hand) can become very complex when multiple findings need to be combined give the propositions and background information, or when the framework of circumstances becomes so complex that it involves an overwhelming amount of information to consider. This brings us to the use of Bayesian networks (BNs) in forensic science, a powerful tool for evaluating forensic findings in such complex casework. A BN provides a graphical representation of propositions, information, findings, and other factors considered during evi-

dence evaluation. These factors are represented as nodes in the network, with directed links showing probabilistic relationships between them. In its simplest form, a BN consists of just two elements (see Fig. 1.2): the propositions (usually at the top, denoted as "H_1, H_2" in Fig. 1.2) and the evidence (at the bottom, denoted as "Evidence" in Fig. 1.2). Since the directed link goes from the proposition node to the evidence node, we refer to the proposition node as the "parent" of its "child" node EVIDENCE. The proposition variable has two possible states: H_1 and H_2, and so does the evidence variable: either "Present" and "Absent".



Evidence	H_1	H_2
Present	p	q
Absent	$(1-p)$	$(1-q)$

Figure 1.2: A simple Bayesian network of the relationship between propositions and evidence. The LR equals p/q when the evidence is present.

This network shows that findings and case propositions are connected, though their relationship is not certain but probabilistic. The directed link between the evidence and propositions can vary in strength: some findings strongly supports one proposition over another, while other findings only provide weak support. We can infer the strength of this relationship from the BN because of the conditional probability tables (CPTs). Each child node has a CPT that specifies the probabilities of the child node being in a particular

state, given the state of its parent node(s). For nodes without parent nodes (called "root nodes"), the table simply lists the prior probability of each possible state. Fig. 1.2 also shows these conditional probability tables for H_1, H_2 and EVIDENCE. The name "Bayesian network" reflects its underlying principles: the LR follows from the CPT of EVIDENCE (marked in grey, while the prior odds come from the CPT of H_1, H_2. Using Bayes' rule, we can obtain the posterior odds from the network[11].

Because this network consists of only two nodes, the LR can be read directly from the CPT. However, in practice, such networks contain multiple nodes between the proposition node and evidence node. We can model BNs and derive LRs from them using software like Hugin (www.hugin.com)[12] by setting H_1 and H_2 to true consecutively[13]. We then divide the probability of the findings being true given H_1 (the numerator of the LR) by the probability of the findings being true given H_2 (the denominator of the LR). For more theory on BNs I refer to Taroni *et al.* (2014).

Grissom's case: source level evaluation using a Bayesian network

Let us examine how BNs could have helped Grissom. First, we will address the source level question: "Whose toenail fragment is it?"

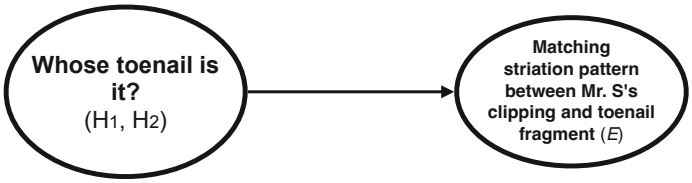
Fig. 1.3 shows our BN for evaluating the matching striation pattern between Mr. S's clipping and the toenail fragment given two propositions: H_1 (Mr. S's) and H_2 (someone else's)[14]. The structure resembles Fig. 1.2: the probability of observing the matching striation pattern is of course dependent on whose

[11]The role of forensic scientists is to obtain an LR, not the posterior odds, because the prior probabilities of the propositions are usually unknown.

[12]Throughout this thesis, we use Hugin Researcher 9.1. to model the BNs (Hugin Expert A/S, 2021).

[13]This thesis presents numerous examples of Bayesian networks and demonstrates how to derive likelihood ratios.

[14]See footnote b in "Box 1 - Grissom's case: source level evaluation" on a comment about the more relevant propositions here.



Matching striation pattern (*E*)	H_1 : Mr. S's	H_2 : Someone else's
True	p	q
False	$(1 - p)$	$(1 - q)$

Figure 1.3: A Bayesian network of the relationship between "Whose toenail is it?" and Grissom's observation of the matching striation pattern and the CPT of the evidence node. The grey area in the CPT reflect the entries of the LR: p/q. p is assumed (near) 1 and q is the probability of observing the matching striation pattern given that the toenail is from someone else than Mr. S. The question Grissom should ask himself here: "how rare are the striation characteristics I am looking at?"

toenail fragment it is. When the toenail fragment comes from Mr. S, the probability of observing the matching striation pattern is assumed (near) 1. However, when the fragment is from someone else, the probability of observing a matching striation pattern between Mr. S's clippings and the toenail fragment lies between 0 and 1. The key question Grissom should ask himself is: "what is the probability of observing the matching striation pattern between Mr. S's clippings and a toenail fragment from a random person?"

This probability is denoted as q in the CPT in Fig. 1.3 and the LR equals $1/q$ (marked in grey). A higher probability of a "random match" leads to a lower LR, resulting in weaker support for proposition 1 (the toenail fragment comes from Mr. S) over proposition 2 (the toenail fragment comes from someone else). Therefore, the more common the striation pattern is, the weaker the evidence becomes.

The LR depends on what Grissom's findings (the matching striation pattern) are and on the case propositions. If Grissom had performed a DNA comparison between the clippings and the toenail fragment instead of a physical comparison and found a match, the evidential strength would likely be

different.

Grissom's case: activity level evaluation using a Bayesian network

In Section 1.1, we recognized that the more relevant question in this case example concerns whether Mr. S wore the (victim's) shoes after the incident or for the last time two weeks prior to the incident. As previously established, evaluating the findings against propositions about whether Mr. S wore the victim's shoe after the incident requires more information beyond the matching striation pattern between the toenail fragment and Mr. S's clippings.

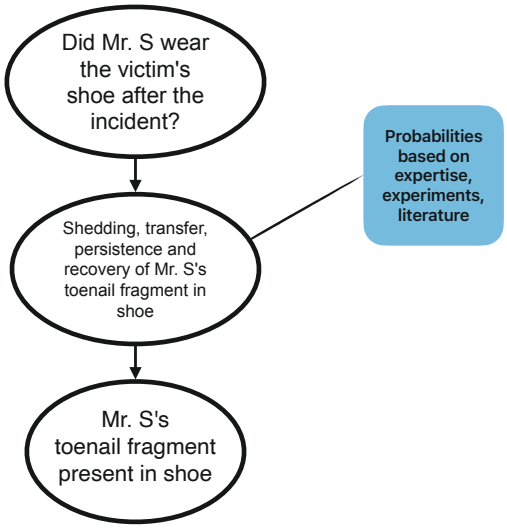
This additional information can be modeled in a BN, as shown in Fig. 1.4. The network maintains a path from the proposition node to the evidence node. The proposition node now contains the activity level propositions H_3 and H_4, while the evidence is defined as the presence of Mr. S's toenail in the shoe[15]. The additional information is modeled as an extra node ("Shedding, transfer, persistence and recovery of toenail fragment in shoe") between the proposition and evidence node.

Grissom needs to assign probabilities for two scenarios: how likely a toenail fragment would shed, transfer to the shoe, persist there, and be recovered if Mr. S wore the victim's shoe after the incident (probability r), or if he wore it up to two weeks prior (probability s)[16]. These probabilities are incorporated in the CPT of SHEDDING, TRANSFER, PERSISTENCE AND RECOVERY OF THE TOENAIL FRAGMENT IN SHOE, as illustrated in Fig. 1.4.

To assign these probabilities, Grissom can draw from his expert knowledge, consult relevant literature, or conduct case-specific experiments. The LR can be derived using software like Hugin (Hugin Expert A/S, 2021). Deriving an

[15] Here we assume that the toenail fragment found in the shoe comes from Mr.S, and hence the observations are now defined as the presence of Mr.S's toenail in the shoe. Therefore, strictly speaking, only the probabilities of "Shedding, transfer, persistence and recovery" given H_3 and given H_4 affect the LR. See also Box 2, Footnote *a*.

[16] Probability s is often called the probability of prevalence (or "background") of a certain trace on an item.



STPR of the toenail fragment in shoe	H_3 : Mr. S wore victim's shoe after incident	H_4 : Mr. S wore his shoe two weeks prior to incident
True	r	s
False	$(1-r)$	$(1-s)$

Figure 1.4: A Bayesian network of the relationship between the activity level propositions, shedding transfer, persistence and recovery, and the observation (here defined as the presence of Mr.S's toenail fragment in the shoe) and the CPT of the "STPR" node.

LR using software highlights one of the key advantages of BNs. But there are advantages that I find even more notable: it is now much clearer what Grissom considers in their evaluation of findings ("a picture says more than a 1000 words"), and this transparency can only promote discussion within areas of expertise and between Grissom and the court.

Previous work

Since 1989, when Aitken & Gammerman first documented the potential of using probabilistic graphical models in forensic science, researchers have been publishing on the use of BNs in this field. To illustrate the development of

the literature on BNs in forensic science, I will categorize the research on them into three distinct eras:

- "Early pioneers": from 1989 - 2000.

  This era marks the beginning of the use of BNs in the context of forensic science. In 1989, Aitken & Gammerman were the first to document the potential of probabilistic graphical models in forensic science. Several influential researchers in the field, including Aitken, Evett, and Dawid, published follow-up articles (Aitken *et al.*, 1996; Dawid & Evett, 1997). At that time, no (easily accessible) computer programs existed for creating these models, so they were all modeled by hand, though Dawid & Evett (1997) noted that computer assistance would be beneficial.

- "March of the BNs": from 2000 - 2017.

  After 2000, I recognize an era I call the "March of the BNs." This era is characterized by growing interest in using BNs in forensic science, including the emergence of commercial software like Hugin for BN modeling. The focus lies on improving formal and logically sound evaluations using BNs, particularly when LR computation becomes complex. Complex LRs are derived from networks such as in (Aitken *et al.*, 2003), and earlier quantitative work such as the "two-trace problem" (Stoney, 1994) is repeated using BNs (Gittelson *et al.*, 2012). Moreover, first models are published for common inference problems concerning scientific evidence (Biedermann *et al.*, 2012a; Cowell *et al.*, 2007; Dawid *et al.*, 2002; Garbolino & Taroni, 2002) and first review papers appear discussing BN applications in forensic science (Biedermann & Taroni, 2012). The use of BNs is explored across all levels in the hierarchy (Cook *et al.*, 1998; Evett *et al.*, 2002). BNs also attracts more interest from scholars such as Fenton, Neil, Lagnado (Fenton & Neil, 2000; Neil *et al.*, 2000), Vlek (Vlek *et al.*, 2013a) and de Zoete (De Zoete *et al.*, 2017). Research has explored integrating BNs into digital forensic science (Kwan *et al.*, 2008,

2011; Lee *et al.*, 2009), which aligns with this discipline's growing presence in forensic casework during this period.

It seems that practical adoption by forensic practitioners remains low. However, such a gap between theoretical research and practical implementation is not unusual and scholars foresee a greater usage of BNs among forensic practitioners (Evett *et al.*, 2002).

- "Post-Taylor et al.-era": from 2017 - now.

  This era differs from the previous one by focusing more on practical applications of BNs in forensic casework. While earlier periods laid the fundamental foundation for formal and logically sound evidence evaluation using BNs, this era emphasizes the forensic practitioner's perspective through the publication of more ready-to-use and case-specific models, see for example De Ronde *et al.* (2019); Fonneløp *et al.* (2022); Mayuoni-Kirshenbaum *et al.* (2020); Onofri *et al.* (2023); Szkuta *et al.* (2018); Uitdehaag *et al.* (2022); Van Dijk *et al.* (2022); Volgin (2019). Research explores BNs not only as tools for LR calculation but more so as means of communication between practitioners and recipients. Even entire legal cases, including both legal and forensic evidence, are being modeled using BNs (Fenton *et al.*, 2020). I call this the "Post-Taylor et al.-era" because the template model for evaluating DNA evidence given activity level propositions by Taylor *et al.* (2018) has inspired numerous subsequent publications. Their template builds on fundamental work by the earlier scholars and relies on intuitive backward reasoning from findings on an item of interest to activities proposed in the propositions ("Follow the trace"). This demonstrates the importance of both theoretical foundations and practical implementation. While DNA evidence remain the primary focus, more disciplines arise that follow the template model in their casework. From my experience, Taylor et al.'s template has become an important factor in BN training at institutions like the NFI and the University of Amsterdam.

In summary, the literature on BNs continues to grow, as does their practical application. However, most advances remain focused on the evaluation of DNA findings and on monodisciplinary evaluations. Several approaches exist to model BNs in forensic science, though in the "post-Taylor et al.- era", most case-specific publications use the Taylor et al.'s template to construct their BNs. Currently, there is no standardization, neither for modeling BNs in casework nor for the way they should be implemented in forensic casework. Moreover, while BNs are being used in traditional forensic disciplines dealing with physical evidence, they have, to the best of my knowledge, not been adopted in digital forensic casework, despite it being one of the most prominent forensic disciplines today.

1.4 My contribution to today's discourse

The LR framework is well-established in most disciplines—it is widely accepted as the proper way for forensic scientists to report findings, and in the Netherlands, it is (close to) standard practice. Moreover, the advantages of BNs within LR practice have been recognized for nearly 30 years. And they are becoming especially valuable as casework grows more complex with increased requests to evaluate various findings given activity level propositions.

However, four challenges remain in practice: 1) despite widespread acceptance, most forensic disciplines rarely adopt BNs; 2) most advances in the field focus primarily on DNA evidence; 3) the evaluation of evidence given activity level propositions in interdisciplinary casework using BNs is not yet fully explored and; 4) there exists a significant gap between digital forensic science and traditional physical evidence disciplines regarding evidence interpretation—both in the use of the LR framework and the use of BNs.

Regarding point (1): Constructing BNs requires significant time and training, and getting additional time during casework is often impractical. I feel that this is one of the reasons that withhold forensic scientists to use BNs in their casework. Therefore, we are seeking an approach that reduces time in-

vestment of the scientists when modeling BNs. This approach must be specific enough to save forensic practitioners valuable time, while remaining flexible enough to adapt to multiple types of cases.

Regarding point (2): The advances in the literature regarding BNs in forensic science remain predominantly focused on DNA evidence evaluation. Similarly, at the Netherlands Forensic Institute, the DNA department shows the quickest uptake of BNs in casework. In this thesis, we aim to make BNs more accessible to other disciplines. One of our research objectives is to develop a template BN that can evaluate all types of transfer evidence, not just DNA evidence, given a variety of possible sets of activity level propositions. This template would ideally follow a modeling structure similar to Taylor *et al.* (2018), where *traces are followed* from activities to an item of interest.

Regarding point (3): Activity level propositions offer a key benefit: they allow various findings to be evaluated using the same set of propositions. As forensic evidence evaluation increasingly centers around activity level propositions, interdisciplinary casework becomes more prominent. In interdisciplinary casework, case scenarios consist of multiple subscenarios, each requiring evaluation. For example, we have an item of interest—a sweater—that is connected to a crime scene via fiber evidence, and we have a suspect who can be linked to the sweater through DNA evidence. The real issue is: can the suspect be linked to the crime scene? For this, we will have two LRs, one for the fiber evidence and one for the DNA evidence, and we want to combine these evaluations into a single combined LR for both the fiber and DNA evidence given propositions concerning the question "did the suspect perform the alleged activity?" We want to explore whether BNs are suitable tools in the evaluation of findings in interdisciplinary casework.

Regarding point (4): The LR framework and BNs are already being applied—with the former more widely used than the latter—in several physical forensic disciplines. However, in digital forensic science both the LR framework and BNs are rarely used. This is a unfilled potential because digital traces have

a particular advantage: they can often be used to directly infer someone's activities (Van Zandwijk & Boztas, 2023). Take health app data from someone's smartphone as an example. This data contains traces such as step count or distance recordings which has a relation to the actual distance someone has walked at a certain time. But in digital forensics too, conclusions cannot be drawn with absolute certainty. Questions we should ask ourselves are: how accurate are these step count and distance data? More importantly, how can we evaluate this information given a set of propositions that are relevant in a specific case in a logical, balanced, robust and transparent manner? We still need a method that can accomplish this.

In Grissom's case we have seen that a trace can be the result of multiple possible activities. This is no different in digital forensic casework. Consider cases with suspected possession of illegal digital material on a computer. Trojan horse defenses are common in these types of cases (Bowles & Hernandez-Castro, 2015): "I did not place the material on my computer, a virus did." In such cases, the LR framework and specifically BNs may help to reason in a structured way about how likely it is that a particular digital trace results from user activities versus automated processes from a virus.

Previous work demonstrates the role digital forensic science plays or should play in evaluating evidence given activity level propositions (Henseler & De Poot, 2020). Research specifically on evaluating digital findings—with or without the LR framework - has emphasized that digital forensic science needs to keep pace with traditional forensic science practices (Pollitt *et al.*, 2018). This aligns with feedback from digital forensic experts at the Netherlands Forensic Institute. Kwan *et al.* (2008) and Lee *et al.* (2009) advocate using BNs to enhance structured reasoning in digital forensic science, though research on their use diminished after these publications, to the best of my knowledge. By now, we hopefully developed a toolbox for forensic practitioners containing template BNs applicable across various cases and forensic disciplines involving physical evidence. We are going to explore how to align digital forensic science with

the more traditional forensic science disciplines in terms of interpretation by promoting the use of the LR framework and introducing BNs there as well.

This project aims to develop a uniform approach to modeling BNs for evaluating evidence given activity level propositions across all forensic disciplines. The uniformity of BNs offers significant advantages within each discipline: it saves time, promotes understanding and discussion among experts, enhances transparency in expert evaluations, and provides guidance for reasoning in (complex) casework. Furthermore, uniformity between disciplines brings additional benefits—it facilitates knowledge exchange and creates opportunities for a multidisciplinary approach in forensic casework.

1.5 How to read this thesis

Most chapters in this thesis are based on published manuscripts[17], except for the introduction, Chapter 8[18] and Chapter 9. Chapters are not arranged in order of publication as I believe that the current arrangement presents the reader with a clearer and more logical representation of the project's story and progression. Nevertheless, each part can be read independently and in any order. While the manuscripts are primarily written for forensic practitioners, this work may offer valuable insights for other professionals in the legal and forensic fields—even if only to better understand the work of forensic practitioners. I hope this research will therefore reach a broader audience.

The thesis is divided into four main parts. The first part presented the research objectives and situated the research within current developments in forensic science. I provided a brief introduction to the main research concepts, which readers can refer to for clarification. I hope this concise introduction

[17] A full list of these publications and author contributions are available on Page 285.
[18] Currently under review (September 19, 2025).

sufficiently familiarizes readers with the topics and sparks their interest. Each subsequent part or chapter also contains its own detailed introduction to the specific concepts relevant to that paper.

Part II is called "starting small: from idioms to case models" and describes a bottom-up approach to modeling case-specific BNs given activity level propositions in forensic science. We aimed to address the complexity of designing BNs by using the idiom-based modeling approach (Neil *et al.*, 2000) in forensic science. This approach combines smaller qualitative networks into larger networks (hence, "starting small") to break down the modeling into smaller tasks. Just as building a Lego model becomes easier when all the pieces are available, modeling a BN becomes more manageable when all building blocks are accessible to the forensic practitioner. We created a collection of these "idioms" that can be used—with minimal adjustments—across multiple cases. Forensic practitioners can select and combine these idioms to build case-specific BNs that evaluate transfer evidence when the activity and/or actor is disputed. We demonstrated the practical application of this collection through a template model that is based on the work of Taylor *et al.* (2018) and Kokshoorn *et al.* (2017), using a fictive case example to illustrate its use.

Part III introduces the application of the template model from Part II in interdisciplinary casework. This research emerged from conversations and brainstorming sessions with Jan de Koeijer MSc, an expert in interdisciplinary forensic casework. Jan de Koeijer and colleagues have developed a method for combining evidence strength across multiple disciplines into a single combined likelihood ratio (De Koeijer *et al.*, 2020)—an approach that has proven highly effective for both understanding and communication in court. For reasons described in Section, we wanted to explore whether BNs could further improve practices within interdisciplinary forensic casework.

We selected a common interdisciplinary casework scenario to test the application of our template model. We focused on cases involving multiple disputes: questions about the actor, the activity, or the relationship between an item of

interest and the alleged activity. Using a concrete case example, we show that the template model can be used to evaluate two or more disputes when multiple forensic disciplines are involved.

Finally, Part IV explores the field of digital forensic science—an emerging forensic discipline I knew little about four years ago. This part, consisting of three chapters, forms the largest part of this thesis as well as the research within the PhD course. Each chapter is based on a different paper, as opposed to the chapters in Part II and Part III. Instead of a general introduction to and final remarks to/on that part, each chapter has their own. Chapter 6 presents the first project of the PhD, which focused on interpreting iPhone Health app data, particularly step count data. Using experimental data collected by Van Zandwijk & Boztas (2019), we developed a numerical LR method applicable in cases where an individual's walking distance is disputed. This method represents a contribution toward the application of the LR framework in digital forensic science.

We also recognized the need for a more structured approach to reasoning about uncertainty in digital forensic casework in general. This was based on a specific request by a digital forensic expert working at the Netherlands Forensic Institute, Ruud Schramp MSc: evaluating evidence in Trojan horse defense cases. We began by applying our template model from Parts II and III to existing cases where a Trojan horse defense was proposed. This revealed that formulating propositions—essential for conducting relevant evaluations—was particularly challenging in practice. Therefore, as a first step, we focused on formulating relevant propositions, which is detailed in Chapter 7. Building on these propositions, we then applied our template models to create BNs for Trojan horse defense cases, which we present in Chapter 8.

These chapters are concluded with a "Final thoughts" chapter where I reflect personally on the research conducted, discuss subjects of debate, and explore potential future applications of BNs in forensic casework.

The end matter starts at Page 279 and consists of: 1) a Dutch summary, 2)

a list of publications with contributing authors' credentials, 3) a summary of modifications made to the published manuscripts to fit them into this thesis, 4) acknowledgement, 5) a bibliography. The appendices (referenced throughout the text) begin on Page 315. Most BNs resulting from this research are openly available and accompany the journal publications. These can be accessed through the QR codes in Appendix A "Supplementary material".