Article

THE INTERNATIONAL JOURNAL *of* EVIDENCE & PROOF

The International Journal of
Evidence & Proof
1–23
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1365712719864608
journals.sagepub.com/home/epj
SAGE

Probabilistic models of legal corroboration

Rafal Urbaniak
LoPSE, University of Gdansk, Gdansk, Poland

Pavel Janda
LoPSE, University of Gdansk, Gdansk, Poland

Abstract

The aim is to develop a sensible probabilistic model of legal corroboration in response to an attack on the probabilistic approach to legal reasoning due to Cohen. One of Cohen's arguments is that there is no probabilistic measure of evidential support which satisfactorily captures the situation in which independent witnesses testify to the truth of the same proposition (or independent pieces of evidence converge on a certain claim)—the phenomenon called corroboration (or convergence). We investigate the properties of several probabilistic measures discussed by Cohen, discuss Cohen's criticism of those measures, and develop our own. Finally, we offer a probabilistic measure of corroboration that evades the critical points raised against the ones discussed so far.

Keywords

Legal probabilism, corroboration, probabilistic evidence evaluation, evidence aggregation

Introduction

Corroboration, intuitively speaking, takes place when two independent witnesses testify to the truth of the same proposition. In his large-scale attack on the probabilistic approach to legal reasoning, Cohen (1977) devotes Chapter 10 of his book to what he calls *the difficulty about corroboration and convergence*. In that chapter he argues that the probabilistic model of weight of evidence and legal decision standards is incapable of adequately capturing the phenomenon. In this paper, we develop a probabilistic model of corroboration which handles Cohen's and other objections better than its predecessors. The model that we will offer follows quite naturally from general Bayesian considerations, and, following Cohen's desideratum, it incorporates the fact that witnesses can report multiple false stories.

Corresponding author:
Rafal Urbaniak, LoPSE, University of Gdansk, Bazynskiego 4, Gdansk, 80-309, Poland.
E-mail: rfl.urbaniak@gmail.com

Given corroborating testimonies, it seems that the probability of the target claim should increase significantly.[1] Cohen argues that no known probabilistic measure of confirmation results in a theorem that would somehow capture this intuition—and that is the main gist of his criticism. Cohen is not the only one who shares this critical view. As Hailperin (1986: 383) points out,

> Although at one time actively pursued, the combination of testimonies (or of evidence) is now no longer a standard item in the repertoire of probability applications.

This opinion is also confirmed by Zabell (1988):

> In the increasingly frequentist environment of the latter 19th century [. . .] the probabilistic analysis of testimony was viewed with increasing hostility and suspicion. In the end it became largely discredited, and today such attempts are often considered mere curiosities: naive, erroneous, or uninformative. [327]

For this reason, Cohen does not have much to go on in his search for probabilistic explications of the phenomenon in question.[2] He does, however, take the existing approaches for a ride, one by one.

The approaches Cohen discusses and finds lacking are: Boole's formula (Boole, 1857), Ekelöf's principle (Ekelöf, 1964) and his own theorem about the issue. The latter he finds unsatisfactory, because it only states that the joint confirmation level will be higher than the separate confirmation levels, but not that it will be much higher.

The plan is to evaluate Cohen's criticism of probabilistic measures of corroboration and then introduce the improved measure. In the next section we introduce Boole's formula, propose a modern derivation of it, discuss Cohen's criticism, and put forward our own. It turns out that from our perspective Boole's formula is in a sense correct, but it applies only to very specific scenarios, in which exactly two witnesses answer a 'yes'/'no' question as to the truth of a hypothesis with prior probability .5. In the following section we discuss Ekelöf's measure of corroboration based on the 17th century Hooper's formula and extend Cohen's criticism of this measure. Next, we briefly discuss another probabilistic measure of corroboration due to Lambert that Cohen didn't consider in the chapter and show that the measure is unfit for the task due to strong assumptions that it makes. Finally, we offer an improved probabilistic corroboration model.

Boole's formula

Deriving Boole's formula

Let's focus on a very simple case in which two witnesses testify regarding the truth of a single simple statement. Let p be the probability that the first witness tells the truth, and q the probability that the other one

[1] If, instead of witnesses' testimonies, we are faced with two independent pieces of (circumstantial) evidence, the analogous phenomenon is called 'circumstantial convergence'. Since as far as it is known there are no major formal differences here, in the paper we'll focus on corroboration, with the thought that everything that is said, applies, *mutatis mutandis*, to circumstantial convergence.

[2] We are aware of recent developments in theory of combining evidence by obtaining likelihood ratios for independent pieces of evidence and then multiplying the results to get the overall weight of combined evidence (Robertson et al., 2016: 69–71) or combining various pieces of evidence in a Bayesian Network (Fenton and Neil, 2013: 412–414). We want to, however, fill a gap in the literature by developing an explicit probabilistic approach to corroboration that answers Cohen's worries directly and is more closely connected to his criticism. This being said, our strategy and reply to Cohen's criticism is in line with these recent general developments. We can find a related approach to witness testimonies in (Bovens and Hartmann, 2003). Bovens and Hartmann concentrate on modelling the following Cohen's thought about surprising information: the lower the prior probability of the information provided by the witnesses, the higher the posterior probability that the information is true (Bovens and Hartmann, 2003). They generalize to multiple false stories and multiple witnesses, but they do not consider Cohen's requirement that the jump be significant and that the element of surprise should contribute to this significance. We come to a similar result from a slightly different angle, mostly focusing on Cohen's worry about the significance of the jump caused by a surprising agreement of unreliable witnesses. We do not focus on how unbelievable the story itself is, although its prior probability is a factor also in our analysis.

does.[3] Boole's formula tells us that the level of the confirmation provided by their joint testimony, w, is:

$$w = \frac{pq}{pq + (1-p)(1-q)} \tag{Boole}$$

Cohen doesn't explain why (Boole) should be used, and (Boole, 1857) contains an idiosyncratic derivation which Hailperin (1986: 385) called 'extensive and shallow'.[4] But deriving the formula from the current perspective isn't too difficult.[5] Let A be witness 1, B be witness 2. 'tA' and 'tB' stand for 'A tells the truth' and 'B tells the truth'. Finally, let 'Z' represent 'A and B agree' (for mnemonic purposes, think 'they agree' with a heavy German accent). So we have $\mathrm{P}(tA) = p$ and $\mathrm{P}(tB) = q$.

If A and B are independent, the probability that they both tell the truth can be obtained by multiplication:

$$\mathrm{P}(tA \wedge tB) = \mathrm{P}(tA) \times \mathrm{P}(tB) = pq \tag{1}$$

Since for any φ, $\mathrm{P}(\neg \varphi) = 1 - \mathrm{P}(\varphi)$, we have:

$$\mathrm{P}(\neg tA) = 1 - \mathrm{P}(tA) = 1 - p \tag{2}$$

$$\mathrm{P}(\neg tB) = 1 - \mathrm{P}(tB) = 1 - q \tag{3}$$

Given the independence of witnesses, we also take $\neg tA$ and $\neg tB$ to be independent:

$$\mathrm{P}(\neg tA \wedge \neg tB) = \mathrm{P}(\neg tA) \times \mathrm{P}(\neg tB) = (1-p)(1-q) \tag{4}$$

Now, the key move, and a point at which it is crucial that the witnesses are only testifying as to the truth or falsehood of a single claim, is that they agree only in the case where they both tell the truth or they both tell something false. Notice that if the testimonies could be more elaborate, multiple different false stories are possible and it no longer holds that if the witnesses tell something false, they both agree.

$$\mathrm{P}(Z) = \mathrm{P}((tA \wedge tB) \vee (\neg tA \wedge \neg tB)) \tag{5}$$

Moreover, the witnesses both telling the truth and the witnesses both being wrong exclude each other, and so by additivity we have:

$$\mathrm{P}(Z) = \mathrm{P}(tA \wedge tB) + \mathrm{P}(\neg tA \wedge \neg tB) \tag{6}$$

Now, them agreeing and telling the truth is equivalent to them telling the truth:

$$tA \wedge tB \wedge Z \Leftrightarrow tA \wedge tB$$

$$\mathrm{P}(tA \wedge tB \wedge Z) = \mathrm{P}(tA \wedge tB) \tag{7}$$

Our w is the conditional probability of their telling the truth on the assumption that they both agree:

$$w = \mathrm{P}(tA \wedge tB|Z) \tag{8}$$

By the definition of conditional probability we have:

$$\mathrm{P}(tA \wedge tB|Z) = \frac{\mathrm{P}(tA \wedge tB \wedge Z)}{\mathrm{P}(Z)}$$

[3] Taking Boole literally (Boole, 1857: 366), p and q should be conditional probabilities of testimonies being true given they have been provided. Since these would directly yield posterior odds, we proceed differently, by focusing on witness sensitivity and specificity as given, and calculating the posterior. Our goal is not historical adequacy, but formal and philosophical understanding.

[4] Supposedly, the the formula dates back to Condorcet and the 18th-century debate about the evidence for Christianity, especially the debate about the evidential support for miracles. Following Hume, Condorcet wanted to use a formula to show that no testimony is sufficient to establish a miracle (Zabell, 1988: 329). In 1837 Babbage argued against Hume's view on miracles, pointing out that the agreement of multiple independent witnesses bears evidential value that trumps concerns about the honesty and reliability of the witnesses, in fact employing a reasoning captured by Kruskal's formula (to be discussed further on in the paper). Kruskal quite correctly points out that Babbage's answer isn't too convincing because the assumption of independence is unjustified.

[5] Again, Boole's own statement of the problem makes it slightly different: 'Two witnesses, A and B, assert a fact. The probability of that fact, if we only knew of A's statement, would be p, if we only knew of B's, would be q; what is the probability when we know both?' (Boole, 1857: 365) See the previous footnote for our take on this.

And so by (7) we obtain:

$$\mathrm{P}(tA \wedge tB|Z) = \frac{\mathrm{P}(tA \wedge tB)}{\mathrm{P}(Z)}$$

Next, we apply (1):

$$\mathrm{P}(tA \wedge tB|Z) = \frac{\mathrm{P}(tA) \times \mathrm{P}(tB)}{\mathrm{P}(Z)}$$

and subsequently, (6):

$$\mathrm{P}(tA \wedge tB|Z) = \frac{\mathrm{P}(tA) \times \mathrm{P}(tB)}{\mathrm{P}(tA \wedge tB) + \mathrm{P}(\neg tA \wedge \neg tB)}$$

Two final moves lead us to the desired formula:

$$\mathrm{P}(tA \wedge tB|Z) = \frac{\mathrm{P}(tA) \times \mathrm{P}(tB)}{\mathrm{P}(tA) \times \mathrm{P}(tB) + \mathrm{P}(\neg tA) \times \mathrm{P}(\neg tB)}$$

$$w = \frac{pq}{pq + (1-p)(1-q)}$$

Cohen's criticism of Boole's formula

Cohen first points out that the formula gives a plausible result for P(tA), P(tB) > 0.5, in which case $w > p,q$. Again, he doesn't prove the claim, but one way we could argue for it is this. Suppose $p, q > 0.5$. We want to show that $p < \frac{pq}{pq+(1-p)(1-q)}$ (the argument is symmetric for q). Since $q > 0.5$, we have our starting point and proceed through rather simple algebraic moves:

$$\begin{aligned}
q &> 1-q \\
(1-p)q &> (1-p)(1-q) \\
q - pq &> (1-p)(1-q) \\
q &> pq + (1-p)(1-q) \\
1 &< \frac{q}{pq + (1-p)(1-q)} \\
p &< \frac{pq}{pq + (1-p)(1-q)}
\end{aligned}$$

It is also easy to show that for $w > p$ we need $q > 0.5$.

$$\begin{aligned}
w = \frac{pq}{pq + (1-p)(1-q)} &> p \\
pq &> p \times (pq + (1-p)(1-q)) \\
q &> pq + 1 - q - p + pq \\
2q &> 2pq + 1 - p \\
2q - 2pq &> 1 - p \\
2q(1-p) &> 1 - p \\
2q &> 1 \\
q &> 0.5
\end{aligned}$$

In fact, both arguments can be symmetrically run with p and q interchanged. These considerations entail that if $p, q < 0.5$, then $w < p, q$. This, however, Cohen has a problem with:

> One witness, for example, may seem rather unreliable because of his shifty demeanor, and another may seem rather unreliable because of his bad eyesight. Yet, perhaps, quite independently, they both testify to precisely the same set of propositions even though each could have told any number of other stories Similarly, the fact that he accused had a motive for murdering the victim may be only mildly incriminatory, since perhaps four other people also are known to have had motives, and the fact that he had an opportunity to commit the murder may also be only mildly incriminatory, since perhaps four other people are also known to have had opportunities. But the combination of motive and opportunity in one person out of the nine is more seriously incriminatory, and Boole's formula cannot represent this. (Cohen, 1977: 96)

As already mentioned, a key role in the argument for (Boole) is played by (5), which requires that the choice is between exactly one true scenario and exactly one false scenario. In such a case, it is far from obvious that the agreement of witnesses who are both more likely to deny a truth decreases the support for what they say.

Consider a scenario in which the prior of a given claim ϕ is 0.5. If each of the witnesses, independently testifying only as to whether φ is true, gets things right, say, 1 out of 10 times, the probability that ϕ and both witnesses testify that φ is $0.5 \times 0.1 \times 0.1 = 0.005$, the probability that $\neg\,\varphi$ and they both testify that φ is $0.5 \times 0.9 \times 0.9 = 0.405$. The probability that they agree is $0.005 + 0.405 = 0.41$. Now, the probability that φ given that they agree in saying φ is $0.005/0.41 \approx 0.012$ and the probability that ϕ given that they agree in saying φ is $0.405/0.41 \approx 0.99$. With the prior of ϕ lower than 0.5 this disproportion increases, with ϕ becoming even less likely.

Boole's formula in terms of conditional probabilities

Come to think of it, information about the probability that a witness tells the truth is information about two conditional probabilities: that they tell A if A, and that they say $\neg A$ if $\neg A$. In other words, it's information about the witness's sensitivity[6] and specificity on the assumption these are equal (in which case we'll call this probability the reliability of the witness, or the quality of a medical test). So, let's think of Boole's formula in terms of these conditional probabilities.

To see why in a restricted class of cases the result that the support is negative is as expected, consider two bad medical tests: the sensitivity and specificity of both of them is, say, 0.2, when the prior probability of the disease is 0.5. They are analogous to witnesses in a binary scenario: the disease either is or isn't present, and each tests only 'testifies' as to its presence or absence. If both tests are positive, regular Bayesian calculations lead to the right conclusion that the posterior probability of the disease is ≈ 0.06, which is definitely less than the reliability of each of the tests. This is not a bug, it's a feature. Cohen's argument about surprise and multiple potential false scenarios doesn't work here, because tests do not have too many options as to what to say, and so the probability that both tests agree will not be so low and their agreement will not suggest that they are both right.

As we can see from the Figures 1 and 2, it is normal that the posterior goes above the quality of the test (the reliability of the witnesses) only as the reliability is high enough, and the point at which this happens depends on the prior probability of the claim under consideration. Now, let's take a look at Figures 3 and 4 to see how the medical tests with prior 0.5 compare to Boole's formula with witness reliability $p = q$.

[6] In this context it is sometimes called *veracity* (Zabell, 1988: 332). We stick to the mainstream terminology.

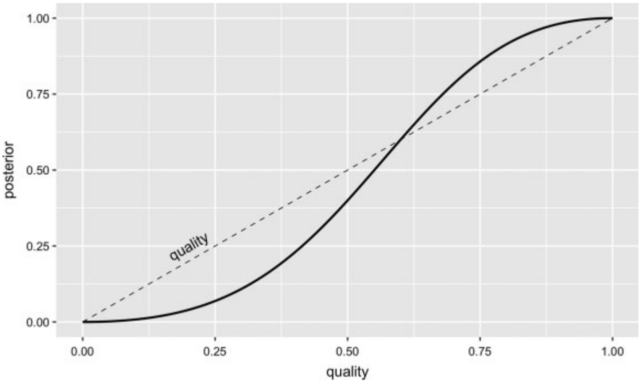

Figure 1. Posterior probability of a disease with prior 0.4 after two updates on positive results of independent medical tests given their (equal) quality. For simplicity, we assume their quality is the same and that sensitivity=specificity=quality.

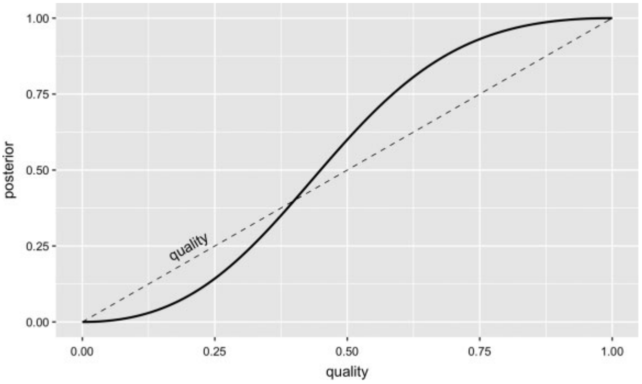

Figure 2. Posterior probability of disease with prior 0.6.

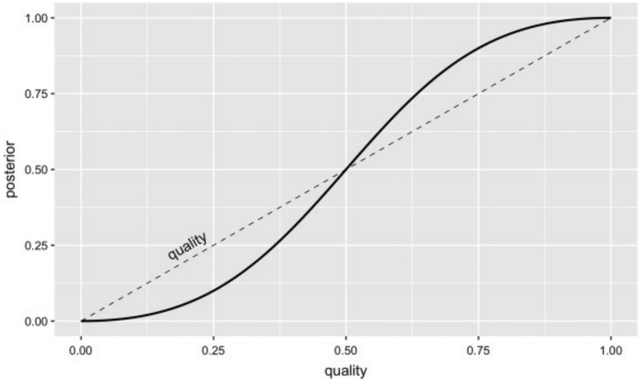

Figure 3. Posterior probability of a disease with prior 0.5 after two updates on positive results of independent medical tests given their quality.

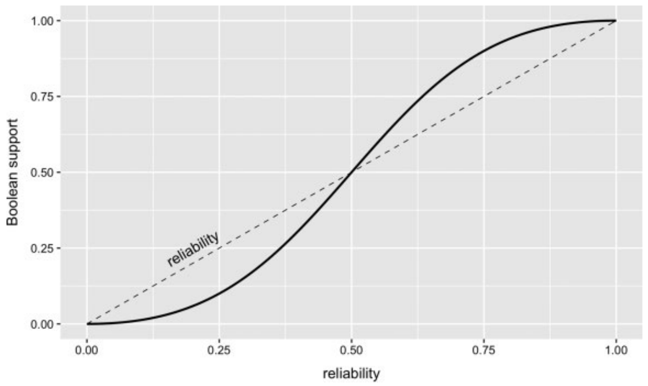

Figure 4. Posterior probability of a claim resulting from applying (Boole) to $p = q$, a.k.a. the same function.

The phenomenon illustrated in Figures 3 to 4 can be proven as follows. Let $+_1, +_2$ be the positive results of the (independent) tests, D be the presence of a disease. Let $\mathrm{P}(+_1|D) = \mathrm{P}(\neg +_1|D) = p$, $\mathrm{P}(+_2|D) = \mathrm{P}(\neg +_2|D) = q$, $\mathrm{P}(D) = 0.5$.

$$\begin{aligned}
\mathrm{P}(D|+_1 \wedge +_2) &= \frac{\mathrm{P}(+_1 \wedge +_2|D)\mathrm{P}(D)}{\mathrm{P}(+_1 \wedge +_2)} \\
&= \frac{\mathrm{P}(+_1|D)\mathrm{P}(+_2|D)\mathrm{P}(D)}{\mathrm{P}(+_1 \wedge +_2|D)\mathrm{P}(D) + \mathrm{P}(+_1 \wedge +_2|\neg D)\mathrm{P}(\neg D)} \\
&= \frac{\mathrm{P}(+_1|D)\mathrm{P}(+_2|D)\mathrm{P}(D)}{\mathrm{P}(+_1|D)\mathrm{P}(+_2|D)\mathrm{P}(D) + \mathrm{P}(+_1|\neg D)\mathrm{P}(+_2|\neg D)\mathrm{P}(\neg D)} \\
&= \frac{0.5pq}{0.5pq + 0.5(1-p)(1-q)} \\
&= \frac{0.5pq}{0.5(pq + (1-p)(1-q))} \\
&= \frac{pq}{pq + (1-p)(1-q)}
\end{aligned}$$

which gives us a second derivation of Boole's formula.[7]

Substantial raise?

One of Cohen's worries is that intuitively the agreement in testimonies raises the probability of their truth substantially (Cohen, 1977: 95), which doesn't seem to be illustrated by Boole's formula or by Cohen's own more limited result, according to which the posterior is greater than either of the reliabilities (Cohen, 1977: 103–107). Is the support described by (Boole) substantial? For one thing, it's hard do convert this requirement into a clearly stated mathematical claim (and thus hard to require a theorem to this effect). But let's take a look at a three-dimensional plot of the function, first for the whole range of values of p and q, then focusing on the upper quarter where $p, q > 0.5$ (see Figures 5 and 6).

[7] The observation that Boole's formula assumes the prior of 0.5 is already due to Keynes (1921: 210).

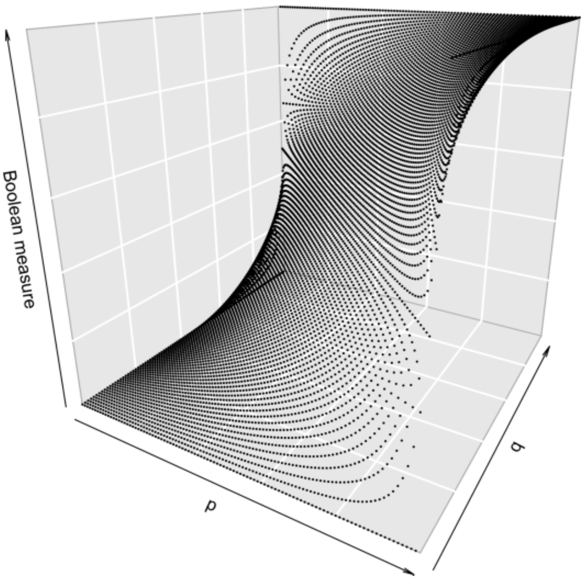

Figure 5. Boolean support depending on the values of *p* and *q*.

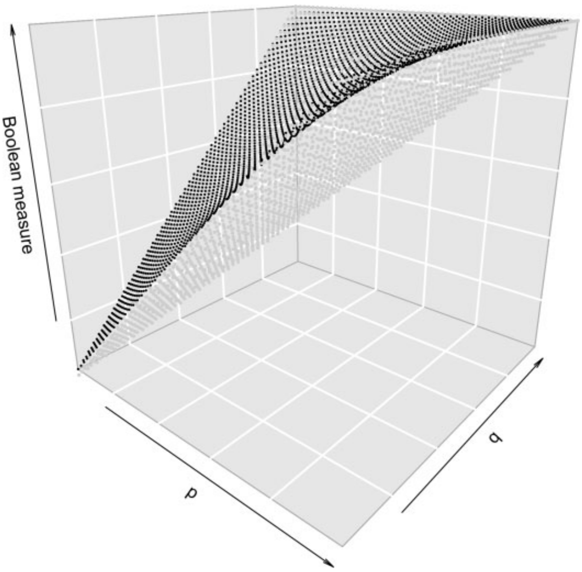

Figure 6. Boolean support for *p*,*q* > 0.5. The grey shade is the maximum of *p*,*q*. Rotated to improve visibility.

We see that as soon as the reliabilities are above 0.5, the support is usually higher than the maximum of the reliabilities. But how much higher? One way to measure this 'jump' is to take mean squared difference from $\max(p, q)$ for evenly distributed points of p and q above 0.5.[8] Let's call this value jumpb.

[8] In this case, we take 2500 points of the form $(0.li, 0lk)$ where l is a digit above 4, and i,k are digits. Later on we look at other numbers of points, but the spread remains the same.

When we look only at the histogram, the impression is that jumpb is rather small. Let's provide some perspective, though (see Figure 7). In fact, we should look at a scaled version of jumpb, when compared to the maximal value it could take, squared distance of $\max(p, q)$ from 1 (also, keep in mind that it having the maximal value clearly isn't optimal either). In Figure 8 you can observe what the jump function looks like after rescaling. Later on, we will discuss the jumps of other functions to compare their performance in this respect.

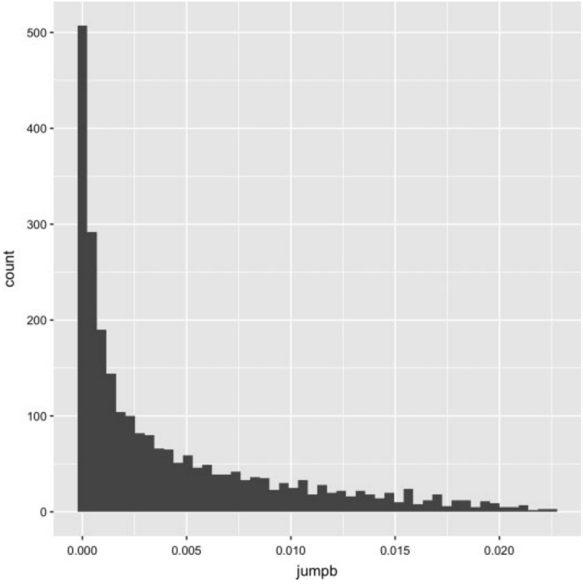

Figure 7. Histogram of jumpb with mean $\approx$ 0.0043 and sd $\approx$ 0.0051.

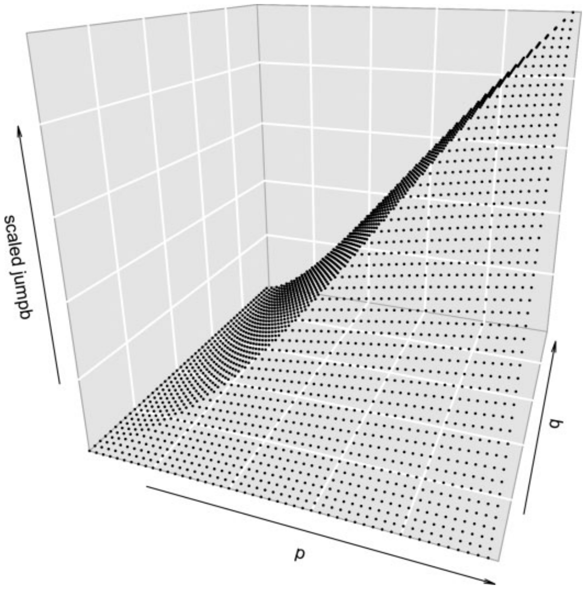

Figure 8. Scaled jumpb for Boolean support for $p,q > 0.5$ with mean $\approx$ 0.248 and sd $\approx$ 0.242 (NaN removed).

Note that scaled jumbp grows fairly steadily up to the expected maximum with $p = q = 1$, and it doesn't seem unusually low (as we will later see, it doesn't seem unusually low compared to other options).

Real issues with Boole's formula

What do we make of Cohen's objection involving a witness with shifty demeanour and a witness with bad eyesight? On the one hand, this counter-example doesn't involve a binary situation in which witnesses testify as to the truth or falsehood of a single claim, so strictly speaking, it's not a counter-example to (Boole) as applied in a restricted context.[9] On the other hand, the counter-example describes a phenomenon that indeed should be modelled, but falls outside of the scope of (Boole). The large number of potential false stories in such contexts, intuitively, should have impact on how we interpret the testimonies, and the agreement that Cohen describes, intuitively, should prove strong support for the truth of the testimonies. We'll return to this issue later on.

The second derivation of (Boole) points to another problem, which hasn't been raised by Cohen in this context.[10] Boole's formula behaves like Bayesian updating with prior = 0.5. However, in many circumstances such a high prior is not sensible.[11] In a civil case if you start with prior 0.5, a minute piece of incriminating evidence would satisfy the preponderance standard (that the plaintiff needs to prove that their claim is more probable than not) and decide the case. In a criminal case, on the other hand, the presumption of innocence seems to disagree with such a high prior. Having taken a deeper look at Boole's proposal and its criticism by Cohen, let's move to the next candidate by Ekelöf.

Ekelöf's corroboration measure

Ekelöf's measure

Another approach to measuring corroboration levels is that of Ekelöf (1964). Let us introduce the idea behind Ekelöf's measure using his own example:

> The length of the braking marks prove that the speed exceeded 60 miles per hour in 12 out of 16 similar cases; at the same time, this is proved by witness's statement in 3 out of 4 remaining cases. The convincing force of the combined evidential facts would thus be 15/16 (Ekelöf, 1964: 58).

Accordingly, Cohen (1977: 99) suggests Ekelöf's formula has the following form: $w = p + q - p \times q$. And indeed, this seems to fit the example: we do have 12/16 + 3/4 − (12/16 × 3/4) = 15/16.[12]

Ekelöf's formula corresponds to Hooper's rule for concurring testimony (Shafer, 1986: 168). An implementation of Hooper's rule gives us, for example, that if two independent witnesses have certain probability of testifying accurately and faithfully (say p and q), then the probability that at least one will be faithful and accurate is

$$1 - (1 - p)(1 - q) = p + q - pq \tag{9}$$

[9] Another aspect in which (Boole) is restricted is that it doesn't tell us what to do if witness sensitivity and specificity are allowed to differ. However, it seems that these values can be plugged in appropriately in our second derivation with prior $= 0.5$ to give a more or less Boolean estimate of the support.

[10] He does worry about setting priors in general at some points (Cohen, 1977: 109, 110, 112), but this is a whole different general problem that lies beyond the scope of this paper.

[11] For example, Posner's use of the prior of guilt set at 0.50 (Posner, 1999) has been heavily criticised (Friedman, 2000; Dahlman, 2018). Here, we only briefly mention key reasons to avoid such a prior.

[12] Ekelöf's wording is somewhat different. He first takes p, and then he adds $(1 - p) \times q$. These formulae are equivalent.

Intermezzo: Evidentiary mechanisms

Strictly speaking, when Ekelöf talks about probability in this context, he doesn't mean the probability of a hypothesis, e.g. that the blue bus ran the red light given evidence such as eyewitness testimony, but the probability that the evidentiary mechanism involved has worked properly. For example, if the sensory mechanism of the witness is an evidentiary mechanism, then he is interested in the probability that that mechanism worked correctly on the basis of the present piece of evidence (Sahlin, 1986).

From a modern perspective, however, introducing the notion of a well-working evidentiary mechanism and the probability that it worked, instead of talking about the hypothesis H directly, comes at a price. For one thing, it is slightly blurry what evidentiary mechanism is. Hallden's original idea was that the mechanism is a hypothesis about the causal connection between a hypothesis such as that a blue bus ran the red light and pieces of evidence, for example, an eyewitness testimony, and these are the only two examples discussed (Sahlin, 1986: 92). But what do you take the causal mechanism to be in the case of, say, DNA evidence, or some other medical tests which we know give reliable results but whose full underlying mechanisms remain unknown?[13] The difficulties with ensuring that relevance tracks causality in the contexts of using Bayesian Networks for modelling reasoning with evidence are rather clearly acknowledged in the literature:

> In the context of evidential reasoning, it can turn out to be very difficult, if not impossible, to show genuine causal relationship… there is no obligation to try to figure out causal explanations… for any argument in evidential reasoning. Evidential analysis should take a neutral stance, as long as it is possible, on the controversies about what a causal relationship is and what constitutes a causal explanation. (Taroni et al., 2006: 62)

Secondly, the strategy seems to insist on evaluating the probability that a given evidential mechanism worked properly given the evidence instead of the probability of the hypothesis given the same evidence. It is, however, not clear why this requirement doesn't lead us into a regress problem. After all, the claim that the evidentiary mechanism worked is another hypothesis, and so it seems only fair to treat it on a par and ask what evidentiary mechanisms ensure the adequacy of our assessment of the probability that it worked, and what is the probability that they worked, and so on.

Moreover, if evidence has value only insofar the underlying mechanism worked (that is, there is a causal link through which knowledge has been obtained), we seem to employ causal theory of knowledge (Goldman, 1967) which isn't a light assumption. For instance, the causal theory of knowledge was criticised in the philosophy of mathematics, since it is difficult to explain what causal connection exists between abstract entities, such as numbers and agents. Causal theory of evidence doesn't perform too well in the legal context either, even putting aside the controversy about the definition of causality in such contexts. For instance, Thomson argues that individualised evidence is one which is causally connected to the target claim (Thomson, 1986), but Gardiner (2019) presents convincing counterexamples. The discussion so far seems to lead to the conclusion that thinking that causality provides some and only guarantee that the target claim is true is too hasty. Firstly, a causal relation doesn't always supply the epistemic value to evidence. For example, there might be a causal link between some political views and rates of domestic violence, but the causal link doesn't seem enough to satisfy the burden of proof. Also, non-causal evidence plays a role in courts. For example, logical entailment can lead to establishing a hypothesis, so it is valuable without any need for a causal relation.

Last but not least, the proponents of the view did provide some elucidations of what they have in mind, but they seem rather unhelpful. Denote the hypothesis that the mechanism worked as A, denote the original hypothesis as H and the evidence as E. Then (Sahlin, 1986: 99) proposes that evidence E tracks

[13] For instance, work by Zajenkowski et al. (2011) suggests that early diagnosis of schizophrenia can be made based on a subject's performance when processing natural language quantifiers (patients with early schizophrenia take more time to solve the problems and are significantly less accurate only with proportional quantifiers, such as *more* and *half*). Yet, our understanding of human brain doesn't allow us to say that we know what the underlying mechanism of this phenomenon is.

the truth of hypothesis H just in case both (i) $H \rightarrow E$ and (ii) $\neg H \rightarrow \neg E$. Thus, on this proposal, the content of A is that E tracks the truth of H, so that (i) and (ii) hold.

Let's think about the proposal for a moment. These two conditions can be written as $H \leftrightarrow E$, and then the chain of reasoning can be deployed:

$$\begin{aligned} \mathrm{P}(A|E) &= \mathrm{P}(H \leftrightarrow E|E) \\ &= \mathrm{P}[(H \wedge E) \vee (\neg H \wedge \neg E)|E] \\ &= \mathrm{P}(H \wedge E|E) + P(\neg H \wedge \neg E|E) \\ &= \mathrm{P}(H \wedge E|E) \\ &= \mathrm{P}(H|E) \end{aligned}$$

So, if conditions (i) and (ii) really explicate the working of the mechanism, then we should ask what value there is in replacing H with A.

Another somewhat hasty claim made by the proponents is that $\mathrm{P}(A|E)$ is more conservative, as it is a lower bound on $\mathrm{P}(H|E)$ (Sahlin, 1986: 93). This doesn't seem to agree with fairly well-established claims about evidentiary value of various sources of information. For instance, suppose we have a medical test with high positivity and sensitivity (say around 0.99). If the result is positive, then the probability that the mechanism works well is high, but one's degrees of belief in H should not have that probability as a lower bound—if the base rate of a given illness is sufficiently low, the posterior of H should be much lower.

Having said all this, let's put the question of the nature of evidentiary mechanisms aside, and let's think about the formulas in terms of probabilities involving evidence and the hypothesis itself. What can be said about the Ekelöf's measure from this perspective?

The jump of Ekelöf's measure

One question that we may ask is whether this measure does better than Boole's in terms of how much stronger the conjunctive support is, as compared to the strongest of the individual supports. That is, we want to compare the jump provided by the Ekelöf measure (jumpe) with that provided by Boole. As it turns out, for cases in which the measure is positive (that is, all cases except for the extreme ones), the jump is lower than that provided by Boole's measure (see Figures 9 and 10). The differences, however, don't seem to be large.

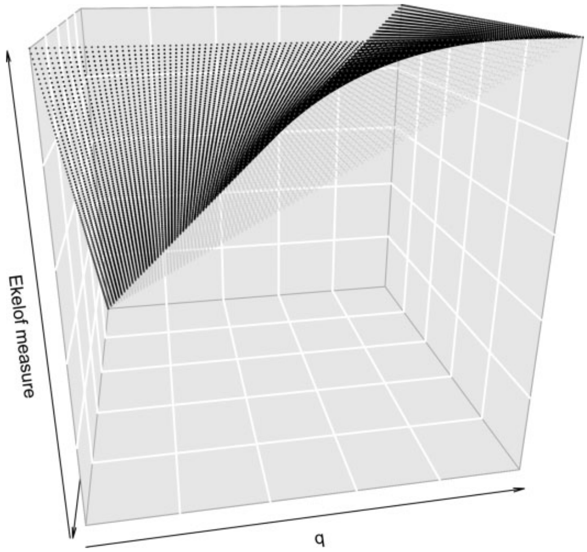

Figure 9. Plot of Ekelöf measure vs. $max(p, q)$ (shadow). jumpe has mean $\approx$ 0.0108 and sd $\approx$ 0.013.

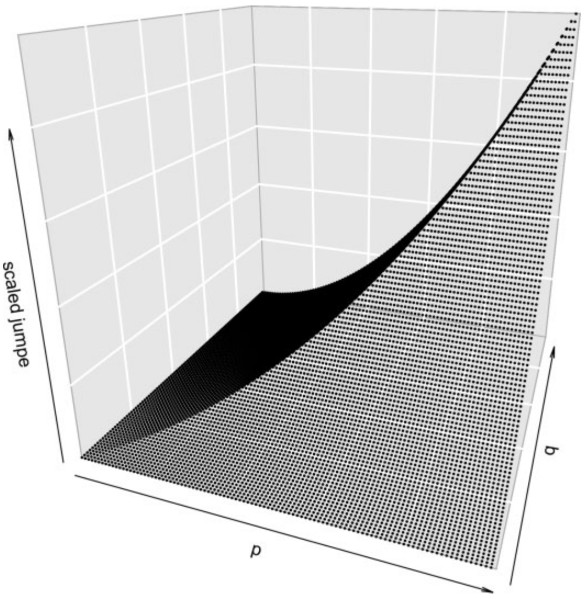

Figure 10. Scaled jumpe for Ekelöf measure for all cases in which it is higher than $max(p, q)$ (almost always), with mean $\approx$ 0.163 (less than the Boolean measure) and sd $\approx$ 0.1949 (NaN removed).

Ekelöf's measure: What is it a measure of?

Ekelöf's rule is concerned in general with concurrent evidential mechanisms and with the event that such a mechanism is working correctly. A witness is an evidential mechanism that is working correctly if she is testifying faithfully and accurately (Shafer, 1986: 168). We have seen that introducing a working mechanism is an interesting but a problematic step. Let us briefly consider whether we could read Ekelöf in a more traditional way. One thing that could be meant by saying 'the braking marks prove that the speed exceeded 60 miles per hour in 12 out of 16 similar cases' is that on average, for each 16 cases in which the braking marks are indicative of speeding, in 12 speeding actually took place (and similarly for witness's testimony). Let's call this the posterior reading. Another reading might be that for each 16 cases in which speeding did take place, the breaking marks (or witness's testimony) would indicate this. Let's call this the sensitivity reading.[14]

Let test A be the length of the breaking marks, test B the witness's testimony, and X the event that the car speeded over 60 mph. The fact that a test indicates X (doesn't indicate X) will be marked by adding a $+$ (a $-$) in the superscript.

Now, how do those readings fare when faced with Ekelöf's informal explanation? Let's think about the case in terms of the posterior reading. Suppose we have sixteen cases in which the items of evidence agree, that is, cases with both A^+ and B^+. Given that A^+ is correct 3 out of 4 times, X holds in 12 of these cases. Now, Ekelöf, on this reading, would seem to invite us to think that the general rate of error of B should apply also to cases in which A was incorrect. But this

[14] This ambiguity in the discussion has already been observed by Keynes (1921, 183), who used *credibility* to mean what we will mean by the value in the posterior reading, and *veracity* for the sensitivity reading. We prefer our terminology because it better aligns with current terminology.

would mean that in the remaining four cases (in which $\neg X$!) B^+ is correct 3 out of four times, that is, that in three out of four cases in which $\neg X$, also X. This doesn't seem like an attractive option.

So let's try thinking about this in terms of sensitivity reading instead. This time, imagine we have 16 cases in which X, and in 12 of them A^+ holds. Among the remaining four cases of X, Ekelöf suggests, we should expect B^+ three times. So far so good. This also means that we should expect B^- in 3 out of 12 cases in which A^+, that we should expect both A^+ and B^+ in 9 cases out of 16, and that we should expect both A^- and B^- in 1 out of 16 cases. A possible distribution of test results of this sort is displayed in Figure 11.

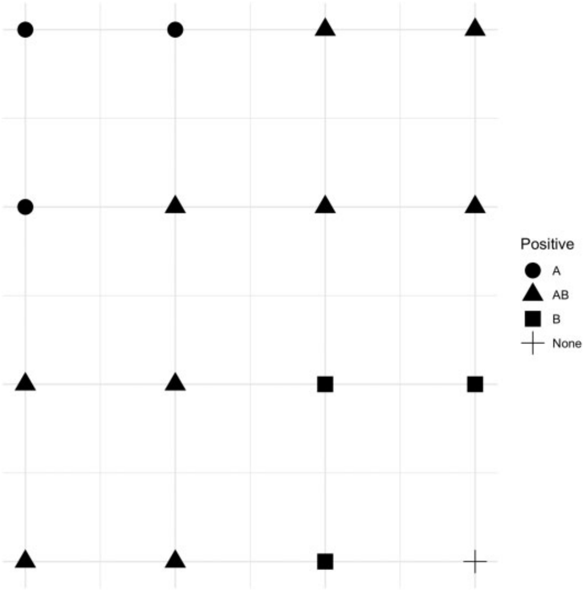

Figure 11. Example of a distribution fitting Ekelöf's example.

But what is 15/16 the frequency of? Figure 11 suggests that in 15 out of 16 cases we should expect at least one test to be positive. And indeed, if $p = \mathrm{P}(A^+|X)$ and $q = \mathrm{P}(B^+|X)$, on the assumption that A and B are independent conditional on X, the exclusion-inclusion formula yields:

$$\mathrm{P}(A^+ \vee B^+|X) = \mathrm{P}(A^+|X) + \mathrm{P}(B^+|X) - \mathrm{P}(A^+ \wedge B^+|X)$$

$$= \mathrm{P}(A^+|X) + \mathrm{P}(B^+|X) - \mathrm{P}(A^+|X)\mathrm{P}(B^+|X) = p + q - pq$$

So, Ekelöf's measure, if p and q are sensitivities of two tests, is simply the sensitivity of running the two tests in parallel, and considering the joint test positive just in case at least one of the two components is. In a sense this is an indicator of the strength of having two tests instead of one. However, this method of joining tests might not be terribly attractive, because the increase in sensitivity takes place at the price of decreasing the specificity, which on the assumption of independence conditional on $\neg X$ is:

$$\mathrm{P}(\neg A^+ \wedge \neg B^+|\neg X) = \mathrm{P}(\neg A^+|\neg X)\mathrm{P}(\neg B^+|\neg X)$$

and is bound to be lower than individual specificities as soon as these are <1. More importantly, this seems to indicate that Ekelöf's measure is not the right one for our purpose. Our question was: how does the fact that tests agree improve our epistemic situation? Ekelöf's formula instead of agreement, on the sensitivity reading (and we failed come up with another one that would make sense), pertains not to tests that agree, but to tests at least one of which is positive.

Even if the reasons discussed so far failed to convince you that Ekelöf measure should be used to measure the evidential strength of two tests agreeing, there are other reasons to think that in such contexts it is a bit problematic, to say the least.

For one thing, no matter how unreliable B is, as long as she is not fully reliable, it will always increase the combined support for X (unless the first test is completely unreliable). This is because if and $p, q < 1$, we have $p \times q < p$, and it follows that $w = p + q - q \times p > 0$. This is independent on whether we go with the posterior or the sensitivity reading.

Now, let p' and q' be measures or support that A^+ and B^+ separately provide to $\neg X$ (on the sensitivity reading, we'd have $p^{'} = 1 - p, q^{'} = 1 - q$, but we don't need this assumption in this argument). The previous paragraph entails that if $p^{'} > 0$ and $0 < q^{'} < 1$, which seems to be the case in many natural situations, the Ekelöf support that A^+ and B^+ jointly provide to $\neg X$ will also be positive. This is clearly an undesirable feature.[15]

Lambert's rule

Cohen considered only Boole's and Ekelöf's measures of corroboration, but those options do not complete the list of possibilities. One example we could consider is Lambert's rule. In Lambert's picture, there is a probability p that the witness will be faithful and accurate, probability q that she will be mendacious, and probability $1 - p - q = c$ that she will be careless. Lambert's rule then says if the two witnesses agree, the credibility of their testimony is:

$$\frac{1 - (1 - p_1)(1 - p_2) + k}{1 - k} \tag{10}$$

In the formula, indexed numbers indicate the probabilities for corresponding witnesses and $k = p_1 q_2 + p_2 q_1$.

Since we have two witnesses and three options (faithful, mendacious and careless), there are nine possible situations represented in the table below. The idea behind the construction is that if both witnesses agree, then, on Lambert's view, it is impossible that one is truthful and the other one is mendacious (options indicated by 'out' in the table) (Shafer, 1986: 161). We then assume that if one witness is faithful, then the testimony is true:

q_1	out		
c_1	true	?	
p_1	true	true	out
	p_2	c_2	q_2

The options in which the claim is true, $p_1 p_2$, $p_1 c_2$ and $p_2 c_1$, summed up, give the numerator of (10). The numerator is formed by summing up the probabilities of all the options except for the two that are 'out'.

The approach is not without issues. For one thing, depending on what you mean by mendacious, it seems quite possible that one witness is faithful and the other is mendacious. After all, an ill-willed witness might be mistaken about what the truth is, and tell the truth by accident. In this

[15] An extension of Ekelöf's approach has been developed by Edman and Halldén. We don't find it too relevant to our discussion, because the assumptions that they used are very strong. See (Shafer, 1986: 171) for a discussion.

sense, the approach assumes that the witnesses indeed do know the truth, and it's only their willingness to tell it that's at issue—an assumption that would severely limit the applicability of the formula.

According to the construction, the agreement of witnesses lends credibility (is accounted for in the numerator) only if at least one witness is faithful. So, on this approach, if they are both careless and agree, this has no impact on the outcome of our evaluation. But this assumes that witnesses who don't care whether they tell the truth are equally likely to tell the truth as to tell the falsehood—a local application of the principle of indifference, so to speak, and one that is not really justified. If one thinks about being careless this way, the division of witnesses is between the faithful, who always tell the truth, the mendacious, who always tell the falsehood, and the careless, who answer yes/no questions randomly with even probabilities—and this is quite an unrealistic idealisation.

Corroboration with multiple false stories

PI measure with 3 hypotheses and 2 witnesses

Let's start with a simple example, in which there are three possible hypotheses h_1, h_2, h_3, exactly one of which is true, and each of two witnesses w^1 and w^2 can testify as to the truth of exactly one of these hypotheses. We denote the fact that witness w^i testified that h_k is true by $t^i{}_k$.

Further, for the sake of simplicity, suppose that the reliability of each of the witnesses—$\mathrm{P}(w^i_k|h_k)$—is 0.6, the prior probability of h_1 is 0.5, and that if a witness is wrong, he is equally likely to testify as to the truth of any of the false hypotheses. The third assumption is an idealisation, as in reality different false stories might have different probability of being told (further on, we'll discuss dropping it). For now, the point will approximately hold as long as there are sufficiently many possible false stories with more or less similar probability. For the conditional probabilities for each witness w^i we obtain the following table of probabilities of obtaining various testimonies, conditional on h_1, h_2 or h_3

	h_1	h_2	h_3
$t^i{}_1$	0.6	0.2	0.2
$t^i{}_2$	0.2	0.6	0.2
$t^i{}_3$	0.2	0.2	0.6

What we are interested in is $\mathrm{P}(h_1|t_1^1 \wedge t_1^2)$. With the assumption that the witnesses' reliability is independent conditional on each of the hypotheses and their negations, standard Bayesian calculations yield:

$$\mathrm{P}(h_1|t_1^1 \wedge t_1^2) = \frac{\mathrm{P}(t_1^1 \wedge t_1^2|h_1)\mathrm{P}(h_1)}{\mathrm{P}(t_1^1 \wedge t_1^2)} \tag{11}$$

$$= \frac{\mathrm{P}(t_1^1|h_1)\mathrm{P}(t_1^2|h_1)\mathrm{P}(h_1)}{\mathrm{P}(t_1^1 \wedge t_1^2|h_1)\mathrm{P}(h_1) + \mathrm{P}(t_1^1 \wedge t_1^2|\neg h_1)\mathrm{P}(\neg h_1)} \tag{12}$$

$$= \frac{\mathrm{P}(t_1^1|h_1)\mathrm{P}(t_1^2|h_1)\mathrm{P}(h_1)}{\mathrm{P}(t_1^1|h_1)\mathrm{P}(t_1^2|h_1)\mathrm{P}(h_1) + \mathrm{P}(t_1^1|\neg h_1)\mathrm{P}(t_1^2|\neg h_1)\mathrm{P}(\neg h_1)} \tag{13}$$

$$= \frac{0.6 \times 0.6 \times 0.5}{0.6 \times 0.6 \times 0.5 + 0.2 \times 0.2 \times 0.5} \tag{14}$$

$$= 0.9 \tag{15}$$

That is, in this particular situation, despite fairly low witness reliability (0.6), already when three different testimonies are possible, agreement in testimonies results in the posterior probability of 0.9. Given that some sort of principle of indifference is involved in obtaining the values, let's call this measure the *PI support measure*. Keeping the prior at 0.5, the level of PI support for the whole range of p and q, and the scaled jump for the measure (jumpp) looks as shown in Figures 12 and 13.

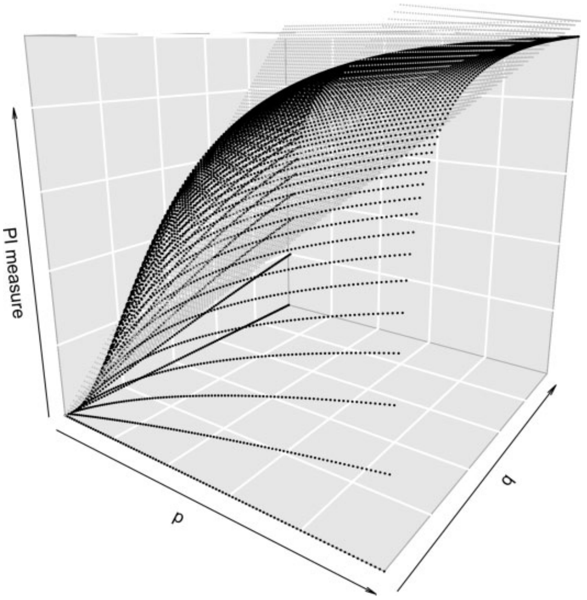

Figure 12. PI support depending on the values of p and q with prior = 0.5. The grey shade is the maximum of p,q.

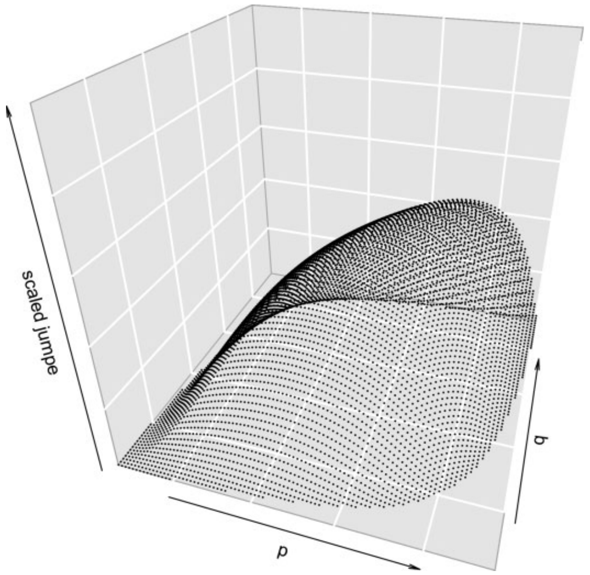

Figure 13. Scaled jumpp for prior = 0.5 and cases in which the measure is above $max(p, q)$, mean 0.246 and sd 0.17.

Note that the value of the measure sometimes is below $\max(p, q)$, which is to be expected. For this reason, scaling to the maximal possible jump exactly in the way we proceeded before doesn't always produce results in [0,1]. For instance, if $p = 0.91$ and $q = 0.08$, the PI support is 0.645, squared distance from $\max(p, q) = 0.91$ is $(0.645 - 0.91)^2 \approx 0.07$ while the maximal possible jump from $\max(p, q)$ is 0.0081, and so scaled PI jump is 8.64. This is the reason why we present the graph of jumpp for cases where the measure is above $\max(p, q)$ (Figures 14 and 15).

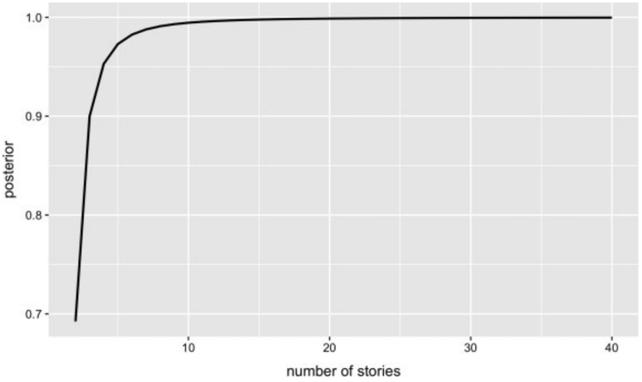

Figure 14. PI support for $p = q = 0.6$, prior $= 0.5$, depending on the number of hypotheses.

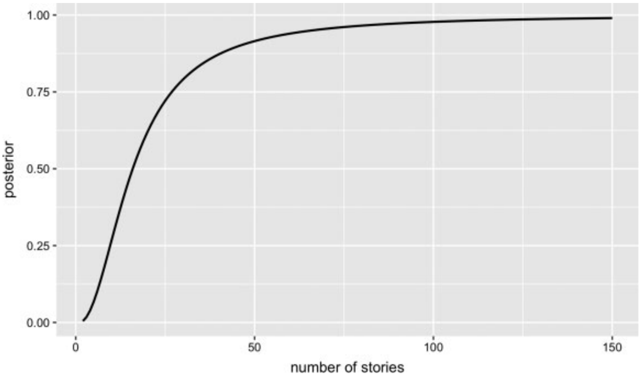

Figure 15. PI support for $p = q = 0.4$, prior $= 0.01$, depending on the number of hypotheses.

Notice that it looks quite different than the jumps of other measures. Come to think of it, it isn't unnatural that the jump decreases as p or q get closer to 1: if one piece of evidence is very reliable, adding information from a less reliable source will provide you with less of a jump than in the case of adding information from a moderately reliable witness to information from another one.

PI measure with varying number of hypotheses

For now, keeping the number of witnesses fixed, we still can generalise, and investigate how the support changes with the increase of the number of possible hypotheses. The general format of such calculations can be described as follows. Let the number of possible testimonies available to any given witness (which is identical to the number of possible hypotheses) be k. Let the reliability of w^i be $\mathrm{rel}^\mathrm{i} = \mathrm{P}(t^i_k|h_k)$ for any $i = 1{,}2$ and any finite k. Say the prior of $\mathrm{h_j}$ is $\mathrm{prior_j}$.[16]

$$\mathrm{P}(h_j|w_j^1 \wedge w_j^2) = \frac{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \mathrm{prior_j}}{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \mathrm{prior_j} + \left[\frac{(1-\mathrm{rel}^1)}{k-1} \times \frac{(1-\mathrm{rel}^2)}{k-1} \times (1-\mathrm{prior_j})\right]}$$

Notice that as soon as $p, q > 0$ and $\mathrm{prior_j} < 1$, as the number of hypotheses goes to infinity, the PI measure tends to 1. Let $rel^1 \times rel^2 \times prior_j = c$. Then we have:

$$\begin{aligned}
&\lim_{k\to\infty}\left[\frac{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \mathrm{prior_j}}{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \mathrm{prior_j} + \left[\frac{(1-\mathrm{rel}^1)}{k-1} \times \frac{(1-\mathrm{rel}^2)}{k-1} \times (1-\mathrm{prior_j})\right]}\right] \\
&= c \Bigg/ \left[c + lim_{k\to\infty}\left(\frac{(1-\mathrm{rel}^1)}{k-1} \times \frac{(1-\mathrm{rel}^2)}{k-1} \times (1-\mathrm{prior_j})\right)\right] \\
&= c \Bigg/ \left[c + lim_{k\to\infty}\frac{(1-\mathrm{rel}^1)}{k-1} \times lim_{k\to\infty}\frac{(1-\mathrm{rel}^2)}{k-1} \times (1-\mathrm{prior_j})\right] \\
&= c/[c + 0 \times 0 \times (1-\mathrm{prior_j})] \\
&= c/c = 1
\end{aligned}$$

Now, this doesn't mean that in practical contexts corroborating witnesses make the hypothesis extremely well-supported simply because they could've said so many different things. This really depends on what kind of question they were answering. Were they answering a yes/no question? Where they asked to pick a suspect from a group? Were they asked to tell what colour a car was? In other words, normally one expects a witness to testify as to a fairly restricted class of elements of the case, and for those, options always can be quite limited.

Moreover, as Kruskal (1988) points out, the assumption of independence is quite strong—and in some cases, agreement might be such that it indicates collusion rather than confirmation. In practice, we'd expect truthful witnesses to agree about most of the main elements of a scenario, but if they agree too much about things that they're quite unlikely to remember (if, for instance, their reports are almost identical in their wording or in their choice of irrelevant details that are described), this raises our suspicion about their credibility. Most of these issues, however, are to be studied empirically. The mathematical aspect—modelling shifts in witnesses' credibility depending on how many not-too-relevant facts they agree on, is beyond the scope of this paper.

Generalisation to multiple witnesses

Intuitively, the level of support should increase not only with the number of potential false stories, but also with the number of witnesses that agree. This factor can be taken into account by means of the

[16] The idea that false stories are equally likely to be told can be found in Laplace. Laplace, however, requires that the witnesses do know which hypothesis is true, takes the prior of the hypothesis to be equal to $1/n$ where n is the number of hypotheses, and assumes that the reliability of all witnesses is the same. See (Shafer, 1986) for a discussion. Another formula (for n witnesses) similar to ours is due to Kruskal (1988), but it doesn't incorporate the idea that the number of possible false scenarios has an impact, and it assumes that all witnesses have the same sensitivity and the same specificity, which we don't assume in our formulation.

following formula (let u be number of witnesses):

$$\mathrm{P}(h_j|w_j^1 \wedge w_j^2 \wedge \cdots \wedge w_j^u) = \frac{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \cdots \times \mathrm{rel}^{\mathrm{u}} \times \mathrm{prior}_{\mathrm{j}}}{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \cdots \times \mathrm{rel}^{\mathrm{u}} \times \mathrm{prior}_{\mathrm{j}} + \left[\frac{(1-\mathrm{rel}^1)}{k-1} \times \frac{(1-\mathrm{rel}^2)}{k-1} \times \cdots \frac{(1-\mathrm{rel}^{\mathrm{u}})}{k-1}(1-\mathrm{prior}_{\mathrm{j}})\right]}$$

As expected, the posterior grows quite quickly with the number of witnesses that agree. Lower reliability of witnesses and lower prior slow down the growth a bit at the beginning, but not too much (see Figures 16 and 17).

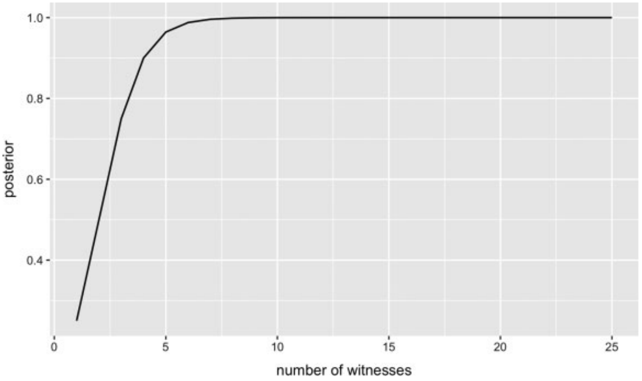

Figure 16. PI support for $p = q = 0.6$, prior $= 0.1$, 3 possible stories, depending on the number of witnesses.

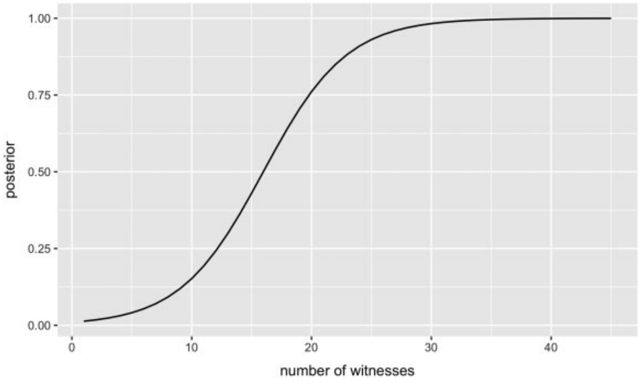

Figure 17. PI support for $p = q = 0.4$, prior $= 0.01$, 3 possible stories, depending on the number of witnesses.

The limit of the posterior as the number of witnesses goes to ∞, for similar reasons, is 1. Keep k constant, let $\mathrm{rel}^1 \times \mathrm{rel}^2 \times \cdots \times \mathrm{rel}^{\mathrm{u}} > 0$ and $\mathrm{prior}_{\mathrm{j}} < 1$. What happens when the number of witnesses goes to infinity? Let's write v instead of $k-1$. We're interested in:

$$\lim_{u\to\infty}\left[\frac{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \cdots \times \mathrm{rel}^{\mathrm{u}} \times \mathrm{prior}_{\mathrm{j}}}{\mathrm{rel}^1 \times \mathrm{rel}^2 \times \cdots \times \mathrm{rel}^{\mathrm{u}} \times \mathrm{prior}_{\mathrm{j}} + \left[\frac{(1-\mathrm{rel}^1)}{v} \times \frac{(1-\mathrm{rel}^2)}{v} \times \cdots \frac{(1-\mathrm{rel}^{\mathrm{u}})}{v}(1-\mathrm{prior}_{\mathrm{j}})\right]}\right]$$

We can now multiply the whole fraction inside the brackets by

$$1 = \frac{1/\mathrm{rel}^1 \times \mathrm{rel}^2 \times \cdots \times \mathrm{rel}^{\mathrm{u}} \times prior_j}{1/rel^1 \times rel^2 \times \cdots \times rel^u \times prior_j}$$

so we get:

$$\frac{1}{1 + \frac{1}{v} \times \frac{1-\mathrm{prior_j}}{\mathrm{prior_j}} \times \lim_{u\to\infty}\left[\frac{(1-\mathrm{rel}^1)\times(1-\mathrm{rel}^2)\times\cdots\times(1-\mathrm{rel}^{\mathrm{u}})}{\mathrm{rel}^1\times\mathrm{rel}^2\times\cdots\times\mathrm{rel}^{\mathrm{u}}}\right]}$$

If we suppose that witnesses are more reliable than unreliable, then the bracket will tend to 0 and the whole denominator to 1 (constants $1/v$ and $1 - \mathrm{prior_j}/\mathrm{prior_j}$ will change nothing about it). Since the numerator is also 1, the whole formula tends to the limit 1 as $u \to \infty$. So, as expected, the posterior grows quite quickly with the number of witnesses that agree.

The growth of the posterior with the greater number of witnesses is an issue already pointed out by Babbage in the context of his criticism of Hume's claim that that testimony cannot establish a miracle (Zabell, 1988). According to Babbage, if we have enough witnesses, the testimony can support the occurrence of a miracle. This might be a valid theoretical point, but in practical cases such as judicial decisions, the number of concurring, independent witnesses is rarely high enough for the worry to be relevant.

A short remark on Venn's paradox

Before we leave the topic, let's briefly address the so-called Venn's paradox, closely related to corroboration. Kruskal's formula has it that the corroboration level is:

$$\frac{p^n\theta}{p^n\theta + (1-p)^n(1-\theta)}$$

where θ is the prior probability of the hypothesis, and p is the common reliability (sensitivity = specificity) of the witnesses (Kruskal, 1988: 932).

Now, this formula (which can be thought of a very particular instantiation of our formula) seems to have the consequence that given $m + n$ witnesses of the same veracity, m of whom attest to an event, n of whom deny it, with $m > n$, the recommended posterior degree of belief is the same as if simply $m - n$ witnesses would testify unanimously as to the event. This came to be known as Venn's paradox, as Venn in *The Logic of Chance* commented:

> . . . it would be hard to find a case in which love of consistency has prevailed over common sense to such an extent as in the admission of the conclusion that it is unimportant what are the numbers for and against a particular statement, provided the actual majority is the same. [404–405]

Now, does this problem arise for our approach? Well, in a sense, yes—Kruskal's formula is a particular instance of ours, so perhaps in a context in which it is appropriate, the result would be the same. However, the formula applies if the witness reliability is kept fixed and assumed not to change. This, however, doesn't seem to be assumption that should be used in a context described by Venn. After all, if witnesses disagree, this clearly has impact on our evaluation of the reliability of the witnesses. This means that it is not the formula that fails, but, rather, that this is a context it which it fails to apply. How our assessment of witness reliability should change in light of witnesses' disagreement is a question that lies beyond the scope of this paper, devoted to contexts in which witnesses do agree.

Discussion and summary

Let's review what we've done. We first looked at Boole's formula, which turned out to result from a fairly specific case of independent tests with sensitivity p and q with prior probability 0.5. As far as modelling corroboration goes, the assumption about prior being 0.5 and the restriction to a situation in which there are only two potential hypotheses makes the formula inapplicable in most contexts.

Then, we investigated Ekelöf' s measure. It turned to be a measure of quite a different probability: that at least one of the tests is positive if the hypothesis is true. As such, it is positive for pretty much the whole range of p and q, and for both a hypothesis and its negation. This makes it inapplicable to the problem that we set out to handle.

We also looked at Lambert's formula, and argued that the assumptions needed for its applicability are too artificial and too strong.

Next, we developed a Bayesian analysis that takes the idea that the number of potential false scenarios has impact on the increase in the posterior. This lead us to the formulae and graphs that indicate that our method is not only well-founded in existing Bayesian methods, but also yields fairly sane results. Our solution is also quite general: any values of the prior, any (positive) number of witnesses and any (positive) number of hypotheses can be plugged in.

Now, there is at least one aspect in which idealisation is involved in our resolution. We assumed that each false hypothesis is equally likely to be told by a given witness, and we distributed the probability that the witness says something untrue evenly among these cases. This, of course, isn't strictly speaking justified.

The worry can me somewhat mitigated, though. What in fact plays a role in the formula is the probability that a given story is told if it is false. One way to try to obtain this probability is to use some sort of principle of indifference, as we suggested above. But the applicability of our approach doesn't depend on this essentially. In (11) it is clear that what gives the desired result is simply the conditional probability of a given story being told if it is false—and a low estimate of this probability doesn't have to be grounded in an application of the principle of indifference. One can abandon the latter and keep the former. It is Cohen who insisted that the number of potential stories matters, and our formulas employing the principle of indifference were meant to capture this intuition to some extent.

Authors' Note

For more information visit http://lopsegdansk.blogspot.com/

Acknowledgements

The authors would like to express their gratitude to Malgorzata Stefaniak, Alicja Kowalewska, Stephan Hartmann, and the anonymous referee.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Science Center, SONATA BIS research grant number: 2016/22/E/HS1/00304.

ORCID iD

Rafal Urbaniak https://orcid.org/0000-0002-6321-2866

References

Boole G (1857) On the application of the theory of probabilities to the question of the combination of testimonies or judgments. *Transactions of the Royal Society of Edinburgh*, 21(4): 597–653.

Bovens L and Hartmann S (2003) *Bayesian epistemology*. Oxford, UK: Oxford Publications.

Cohen J (1977) *The Probable and the Provable*. Oxford: Oxford University Press.

Dahlman C (2018) Determining the base rate for guilt. *Law Probability and Risk* 17(1): 15–28.

Ekelöf P (1964) Free evaluation of evidence. *Scandinavian Studies in Law* 8: 47–66.

Fenton N and Neil M (2013) *Risk Assessment and Decision Analysis with Bayesian Networks*. Boca Raton, FL: CRC Press.

Friedman RD (2000) A presumption of innocence, not of even odds. *Stanford Law Review* 52(4): 873–887.

Gardiner G (2019) Legal burdens of proof and statistical evidence. In: Coady D and Chase J (eds) *Routledge Handbook of Applied Epistemology*. Abingdon: Routledge.

Goldman AI (1967) A causal theory of knowing. *The Journal of Philosophy* 64(12): 357–372.

Hailperin T (1986) *Boole's Logic and Probability: A Critical Exposition from the Standpoint of Contemporary Algebra, Logic and Probability Theory*. Amsterdam: Elsevier.

Keynes JM (1921) *A Treatise on Probability*. London: Macmillan.

Kruskal W (1988) Miracles and statistics: The casual assumption of independence. *Journal of the American Statistical Association* 83(404): 929–940.

Posner RA (1999) An economic approach to the law of evidence. *Stanford Law Review* 51: 1477–1546.

Robertson B, Vignaux GA and Berger CEH (2016) *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. London: John Wiley & Sons, Ltd.

Sahlin N-E (1986) How to be 100% certain 99.5% of the time. *The Journal of Philosophy* 83(2): 91–111.

Shafer G (1986) The combination of evidence. *International Journal of Intelligent Systems* 1(3): 155–179.

Taroni F, Biedermann A, Bozza S, Garbolino P and Aitken C (2006) *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*. London: John Wiley & Sons.

Thomson JJ (1986) Liability and individualized evidence. *Law and Contemporary Problems* 49(3): 199–219.

Zabell S (1988) The probabilistic analysis of testimony. *Journal of Statistical Planning and Inference* 20(3): 327–354.

Zajenkowski M, Styła R and Szymanik J (2011) A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders* 44(6): 595–600.