Measuring coherence with Bayesian Networks

1 Motivations & introduction

The notion of coherence is often used in many philosophical, especially epistemological, discussions (for instance, in discussions about the truth-conduciveness of coherence, Olsson, 2001; Shogenji, 1999). An explication of the key notion involved seems desirable.

There is also a more practical reason to develop a better understanding of the notion: a plausible measure of coherence could be used to better evaluate the quality of some stories or narrations. For example in the legal context we would like to be able to assess the quality of a testimony in the court of law. Focusing only on the probability of a story is to some extent problematic, because from such a perspective, more detailed stories are penalized—they contain more propositions, so they (usually) have lower probabilities. Moreover, a plausible coherence measure could perhaps be used to obtain sensible priors.

Quite importantly, there is a disconnect between philosophical research on probabilistic coherence and the developments in Bayesian Network -based methods: the latter seems seemingly unaware of the philosophical discussion. In particular, in the context of Bayesian networks developed for stories and narrations in legal contexts (Fenton et al., 2013; Vlek, 2016; Vlek et al., 2013, 2014, 2015, 2016), an approach to coherence has been developed by Vlek. The proposal is to capture the coherence of the story by introducing a single narration root node.[1] Vlek then identifies coherence of a model with the prior probability of the narration node. We think this approach is too simplistic, as we want to capture the idea that coherence is distinct from probability, and the addition of a scenario node introduces probabilistic dependencies by fiat.

When we talk about the coherence of a set of propositions or about the coherence of a story, we seem to refer to how well their individual pieces fit together. How are we to understand and apply this notion systematically, though?

As with beliefs, we can use both a binary and a graded notion of coherence. The binary notion is not very exciting: a set is incoherent just in case it is logically inconsistent.[2] Intuitively, graded coherence should be a generalization of this requirement: logically incoherent sets should have minimal level of graded coherence or, at least, lower coherence than consistent ones. What other requirements should a coherence measure satisfy and how should it be explicated formally, if we want to massage this notion into a more general framework of probabilistic

[1] The root node becomes an ancestor node to all the other nodes such that the conditional probability of each dependent node given that the state of this root is 1 (that is, the corresponding proposition is assumed to be true), is also 1.

[2] There is a related notion in the neighborhood where an agent's degrees of beliefs are coherent just in case it they are probabilistic. We will not use this notion in this paper.

epistemology? Defining a measure of graded coherence in probabilistic terms turned out to be quite a challenge, which resulted in heaps of literature.

Our measure diverges from the known candidates in three important respects: (1) It is not a function of a probabilistic measure and a set of propositions alone, because it is also sensitive to the selection and direction of arrows in a Bayesian Network representing an agent's credal state. (2) Unlike in the case of quite a few coherence measures, it is sensitive the weakest links in the narration, (3) The key elements used in coherence score calculations are not not confirmation levels, but rather expected and weighted confirmation levels.

We first describe the main probabilistic explications of coherence present in the literature (Section 2). Then, we describe two philosophically motivated thought experiments and one real-life example that will serve as illustration in our Bayesian network approach (Section 3). Next we try to identify the key problems with the existing measures (Section 4), which leads us to our own positive proposal—that of *structured coherence* (Section 5, which we explain with a running example in the background. With this tool in hand we approach our real-life example (the Sally Clark case) and argue that our measures handles it better than the other measures (Section 6). We finish with the comparison of our measure with respect to the other examples we used, and a few closing comments (Section 7).

2 Measures

Quite a few different measures of coherence have already been developed. One of the first ones, the so-called deviation from independence measure, comes from Shogenji. Another early proposal, the relative overlap measure, was offered by Olsson & Glass. Unfortunately, both of these approaches are susceptible to various objections and counterexamples (Akiba, 2000; Bovens and Hartmann, 2004; Crupi et al., 2007; Koscholke, 2016; Merricks, 1995; Schippers and Koscholke, 2019; Shogenji, 1999, 2001, 2006; Siebel, 2004, 2006). We focus here on the most recent approach — a class of measures called *average mutual support*, which were developed to avoid these problems. Let's take a look at the general recipe for an *average mutual support* measure.

- Given that S is a set whose coherence is to be measured, let P indicate the set of all ordered pairs of non-empty, disjoint subsets of S.
- First, define a confirmation measure for the confirmation of a hypothesis H by evidence E: $Conf(H,E)$.
- For each pair $\langle X,Y\rangle \in P$, calculate $Conf(\bigwedge X, \bigwedge Y)$, where $\bigwedge X$ ($\bigwedge Y$) is the conjunction of all the elements of X (Y).
- Take the mean of all the results.

$$\mathscr{C}(P) = mean\left(\left\{Conf(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right)$$

Depending on the choice of a confirmation measure, we achieve different measures of coherence. One thing to keep in mind is that different measures use different scales and have different neutral points, if any (the idea is: the coherence of probabilistically independent propositions should be neither positive nor negative). Here are the key candidates present in the literature:

- Fitelson (2003) uses the following confirmation function (ranging from -1 to 1 with neutral point at 0):

$$F(H,E) = \begin{cases} 1 & E \models H, E \not\models \bot \\ -1 & E \models \neg H \\ \frac{P(E|H)-P(E|\neg H)}{P(E|H)+P(E|\neg H)} & \text{o/w} \end{cases}$$

$$\mathscr{C}_F(P) = mean\left(\left\{F(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right) \qquad \text{(Fitelson)}$$

- Douven and Meijs (2007) use the *difference* confirmation measure (also ranging from -1 to 1 with neutral point at 0):

$$D(H,E) = P(H|E) - P(H)$$

The resulting definition of coherence is:

$$\mathscr{C}_{DM}(P) = mean\left(\left\{D(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right) \qquad \text{(DM)}$$

- Roche (2013) starts with the absolute confirmation measure:

$$A(H,E) = \begin{cases} 1 & E \models H, E \not\models \bot \\ 0 & E \models \neg H \\ P(H|E) & \text{o/w} \end{cases}$$

which results in the following coherence measure (ranging from 0 to 1 with neutral point at 0.5):

$$\mathscr{C}_R(P) = mean\left(\left\{A(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right) \qquad \text{(Roche)}$$

For comparison purposes, we will also bring up two coherence measures based on average mutual support:

- Shogenji's *deviation from independence* (Shogenji, 1999), defined as the ratio between the probability of the conjunction of all claims, and the probability that the conjunction would get if all its conjuncts were probabilistically independent (scaling from 0 to ∞ with neutral point 1):

$$C_s(A_1,\ldots,A_n) = \frac{P(A_1 \& \ldots \& A_n)}{P(A_1) \times \cdots \times P(A_n)} \qquad \text{(Shogenji)}$$

- *Relative overlap* coming from (Olsson, 2001) and (Glass, 2002), defined as the ratio between the intersection of all propositions and their union (scaling from -1 to 1 with no clear neutral point):

$$C_o(A_1,\ldots,A_n) = \frac{P(A_1 \& \ldots \& A_n)}{P(A_1 \vee \cdots \vee A_n)} \qquad \text{(Olsson)}$$

3 Scenarios and Bayesian Networks

Counterexamples to a coherence measure usually have the form of a few propositions formulated in natural language, such that intuitive judgments of coherence involved and the probabilistic calculations diverge. While such counterexamples are not in our focus in this paper, we will give a couple of examples. Due to space limitations we do not discuss other well-known examples, which we are aware of, but whose treatment is postponed to a different paper: Penguins (Bovens and Hartmann, 2004; Meijs and Douven, 2007), Dunnit (Merricks, 1995), Japanese swords (Meijs and Douven, 2007), Robbers (Siebel, 2004), and two similar ones: Depth and Dice (Akiba, 2000; Schippers and Koscholke, 2019; Shogenji, 2001).

However, later in this paper we will provide calculations for these scenarios, just to briefly mention how various coherence measures perform in those cases. As a sanity check, we will also include a real-life example of the famous Sally Clark case. These examples will be used to explain how we represent various narrations as Bayesian networks. Later on we'll look at the results that our coherence measure yields for these cases.

3.1 The Beatles

The scenario. The challenge has been offered by Shogenji (1999, 339) to criticize defining coherence in terms of pairwise coherence — it shows there are jointly incoherent pairwise coherent sets. The scenario consists of the following claims:

node	content
D	Exactly one of the Beatles (John, Paul, George and Ringo) is dead.
J	John is alive.
P	Paul is alive.
G	George is alive.
R	Ringo is alive.

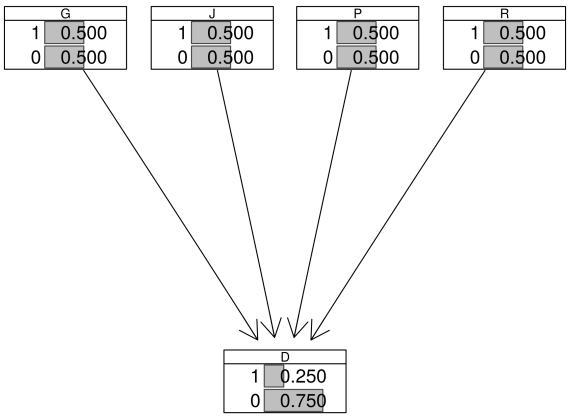

Figure 1: Beatles BN with marginal probabilities.

We assume the prior probability of each individual band member being dead to 0.5 (as in the above table), and the CPT for D is many-dimensional and so difficult to present concisely, but the method is straightforward: probability 1 is given to D in all combinations of the parents in which exactly one is true, and otherwise D gets conditional probability 0. The BN with marginal probabilities looks as in Figure 1.

3.2 The Witnesses

The scenario. This one comes from (Olsson, 2005, 391). Equally reliable witnesses try to identify a criminal. Consider the following reports (we extended the original scenario by adding W5). The problem might be seen as involving subsets of the following nodes:

node	content
W1	Witness no. 1: “Steve did it”
W2	Witness no. 2: “Steve did it”
W3	Witness no. 3: “Steve, Martin or David did it”
W4	Witness no. 4: “Steve, John or James did it”
W5	Wittness no. 5: “Steve, John or Peter did it”
D	Who committed the deed (6 possible values)

Note that each proposition has the structure “Witness no. X claims that . . . " instead of explicitly stating the witness’ testimony.

Two requirements are associated with this example: both $\{W1, W2\}$ and $\{W4, W5\}$ should be more coherent than $\{W3, W4\}$.

Each of these three sets is represented by a BN with three nodes: the root note D (who actually committed the deed), and its two binary children nodes corresponding to the propositions contained in a given set. In Figure 2 is the DAG for the first set, together with the marginal probabilities. The other two networks are analogous.

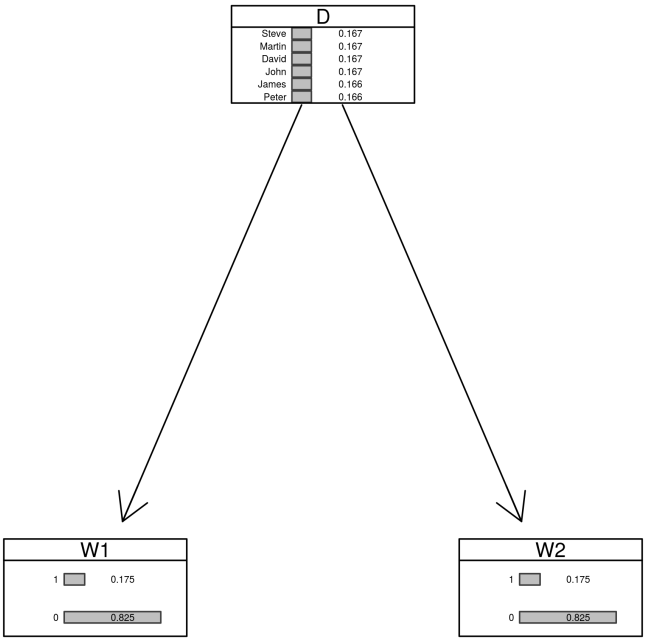

Figure 2: BN for the witnesses example (W1W2) with marginal probabilities.

The basic idea behind the CPTs we used is that for any particular witness we take the probability of them including the perpetrator in their list to be 0.8, and the probability of including an innocent to be .05. The CPT for D is uniform. The table for W1 provides the conditional probability of W1 listing (W1=1) or not listing (W1=0) a particular person given that the actual value of D is Steve/Martin/.... In the remaining the CPT for D remains thesame, and the CPTs for the witness nodes are analogous to the one for W1.

D	Pr
Steve	0.167
Martin	0.167
David	0.167
John	0.167
James	0.167
Peter	0.167

W1	D					
	Steve	Martin	David	John	James	Peter
1	0.8	0.05	0.05	0.05	0.05	0.05
0	0.2	0.95	0.95	0.95	0.95	0.95

Figure 3: BN for the W1W2 narration in the Witness problem. CPT for W2 is identical to the one for W1.

3.3 Sally Clark

Later on we take the existing coherence measures for a ride by testing them on a real-case based Bayesian network for the Sally Clark case. At this point, we introduce the case and the network. R. v. Clark (EWCA Crim 54, 2000) is a classic example of how the lack of probabilistic independence between events can be easily overlooked. Sally Clark's first son died in 1996 soon after birth, and her second son died in similar circumstances a few years later in 1998. At trial, the paediatrician Roy Meadow testified that the probability that a child from such a family would die of Sudden Infant Death Syndrome (SIDS) was 1 in 8,543. Meadow calculated that therefore the probability of both children dying of SIDS was approximately 1 in 1 in 73 million. Sally Clark was convicted of murdering her infant sons (the conviction was ultimately reversed on appeal). The calculation illegitimately assumes independence, as the environmental or genetic factors may predispose a family to SIDS. The winning appeal was based on new evidence: signs of a potentially lethal disease—contrary to what was assumed in the original case—were found in one of the bodies.

Fenton and Neil (2018) constructed a Bayesian Network to discuss the interaction of the key pieces of evidence in this case, and our Bayesian network is based on theirs. The network structure is in Figure 4.

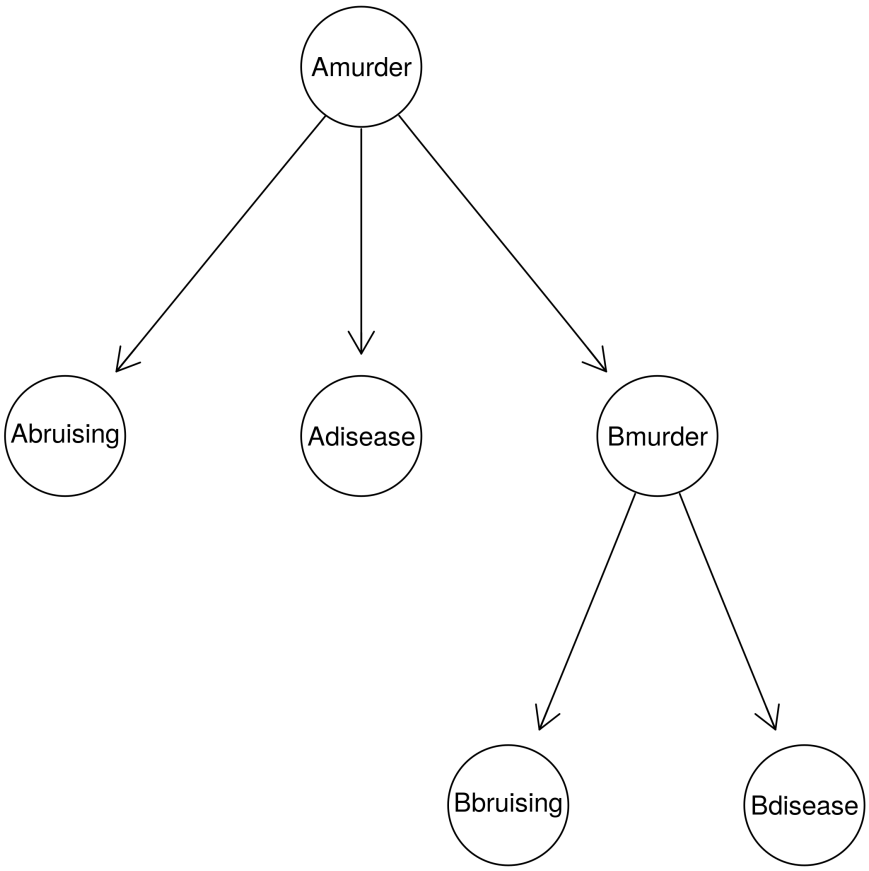

Figure 4: The directed acyclic graph for the Sally Clark BN

The arrows depict relationships of influence between variables. Amurder and Bmurder are binary nodes corresponding to whether Sally Clark's sons, call them A and B, were murdered. These influence whether signs of disease (Adisease and Bdisease) and bruising (Abruising and B.bruising) were present. Also, since son A died first, whether A was murdered casts some light on the probability of son B being murdered.

We employ the same probability tables as Fenton and Neil (2018), which incorporated Mr Meadow's statistical testimony. The prior probability of son A not being murdered is 0.921659, the probability of bruising in a son if he wasn't murdered is set to 0.01 and to 0.05 if he was. The probability of signs of a disease in a son if he wasn't murdered is set to 0.05 and to 0.001 if he was. Finally, the probability that the second son was murdered if the first was is set

to .9993604 and to $1 - 0.9998538$ if he wasn't. If these look too specific for the reader, the reader is welcome to re-run the analysis with a wide range of options. The resulting marginal probabilities are illustrated in Figure 5

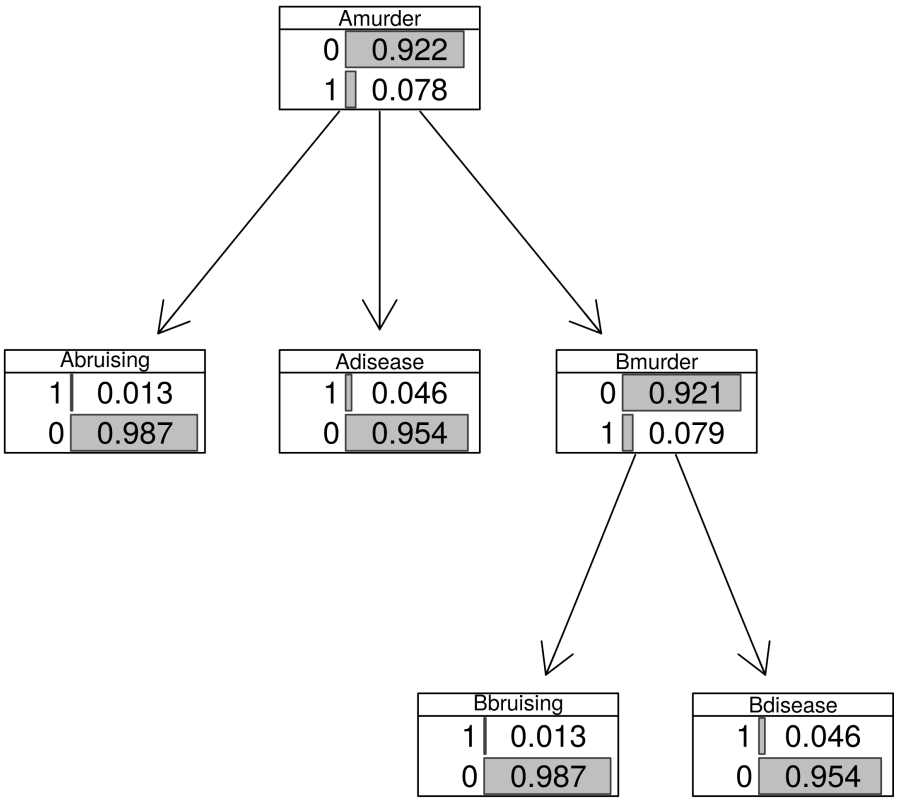

Figure 5: Marginal probabilities for the Sally Clark network

4 Challenges and ways out

The known challenges to the existing measures consist in discrepancies with intuitions in various thought experiments (Akiba, 2000; Bovens and Hartmann, 2004; Crupi et al., 2007; Koscholke, 2016; Merricks, 1995; Schippers and Koscholke, 2019; Shogenji, 1999, 2001, 2006;

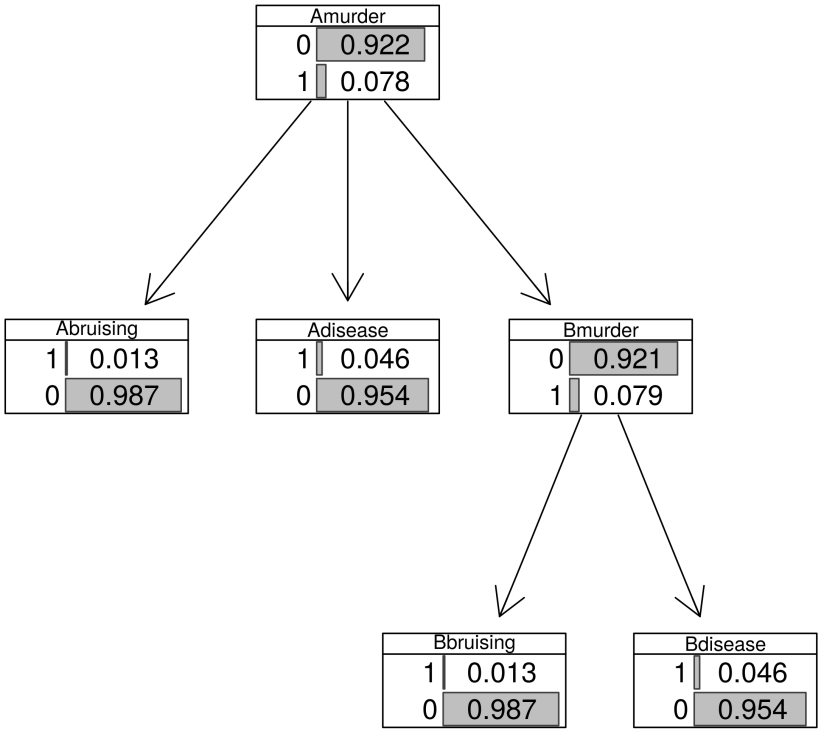

Figure 6: Bayesian network for the Sally Clark case, with marginal probabilities.

Siebel, 2004, 2006). We try a more principled approach, making a few more general conceptual points that suggest a way forward.

4.1 Mean

One thing to observe is that The Beatles is a logically inconsistent scenario. However, each measure considered so far gives it a higher score than to multiple logically consistent scenarios, such as W3& W4 in The Witnesses. This disagrees with our fundamental intuition that a coherence measure should keep track of logical consistency.

Our hypothesis is that the cause of this issue is as follows. Each measure that faces this problem uses subsets of a set (or pairs thereof) and then takes the average result calculated for these subsets or pairs of subsets. However, simply taking the mean of results so obtained

	Fitelson	Douven-Meijs	Roche
Witness W3W4 (11)	-0.2336	-0.1103	0.3147
Beatles (11111)	-0.0361	0.0247	0.3222

Table 1: Three coherence measures applied to a consistent (Witness, variant W3W4) and an inconsistent scenario (Beatles).

might be misleading, because a few low values (for the inconsistent subsets), which indicate inconsistency, might be mixed with many positive values (especially if a set is large), and taking the mean of all such results might give a relatively high score, despite involving an inconsistency. Therefore, we believe that a candidate for a coherence measure shouldn't simply take mean mutual confirmation scores.

4.2 Structure

In the existing discussion, each scenario was represented as a set of propositions. However, it seems that usually we do not face sets of propositions but rather scenarios with some more or less explicit narration, which also indicates how the propositions are supposed to be connected. In other words, agents not only report their object-level beliefs, but also have some idea about their structure: which are supposed to support which. This relation rarely is universal in the powerset of the scenario (minus the empty set of course), and so considering support between all possible pairs of propositions in the scenario in calculating coherence might be unfair towards the agent. We penalize her for lack of support even between those propositions which she never thought supported each other.

To notice that the selection and direction of support arrows matter, consider two agents whose claims are as follows:[3]

Agent 1 Tweety is a bird, more specifically a penguin. Because it's a penguin, it doesn't fly.

Agent 2 Tweety is a bird, and because it's a bird, it doesn't fly. Therefore Tweety is a penguin.

Even though both of them involve the same atomic propositions, the first narration makes much more sense, and it seems definitely more coherent. It is also quite clear that the difference between narrations lies in the explicitly stated direction of support. The approaches to coherence developed so far do not account for this difference.

Moreover, it seems that when we present challenges and our intuitions about the desiderata, we implicitly assume the narration involved is the one that best fits with our background knowledge (so, Agent 1 rather Agent 2 in the case of penguins). However, coherence measures developed so far do not make such a fine-grained distinction between narrations, and so the scenario which states that *Tweety is BGP* (bird, grounded, penguin) gets a lower score because, quite obviously, being a bird disconfirms being a grounded animal. In such a calculation it

[3]This example is inspired by a scenario discussed in (Bovens and Hartmann, 2004, 50) and (Meijs and Douven, 2007).

doesn't matter that no one even suggested this causal relationship. To illustrate this intuition, think about a picture puzzle. Just because a piece from the top right corner doesn't match a piece from the bottom left corner, it doesn't necessarily decrease the coherence of a complete picture. It just means you shouldn't evaluate how well the puzzle is prepared by putting these two pieces next to each other.

We believe that only those directions of support which are indicated by the reporting agent, or by background knowledge, should be taken into account when measuring coherence.

5 Structured coherence

Based on these observations we developed our own measure, which we call *structured coherence*. In this section we will describe how we manage to avoid the above mentioned problems.

In our calculations we use the Z confirmation measure (see Crupi et al., 2007, for a detailed study and defense). It results from a normalization of many other measures (in the sense that whichever confirmation measure you start with, after appropriate normalization you end up with Z) and has nice mathematical properties, such as ranging over $[-1,1]$ and preservation of logical entailment and exclusion. It is defined for hypothesis H and evidence E as follows:[4]

$$\begin{aligned} \mathsf{prior} &= \mathsf{P}(H) \\ \mathsf{posterior} &= \mathsf{P}(H|E) \\ \mathsf{d} &= \mathsf{posterior} - \mathsf{prior} \\ Z(\mathsf{posterior}, \mathsf{prior}) &= \begin{cases} 0 & \text{if } \mathsf{prior} = \mathsf{posterior} \\ \mathsf{d}/(1-\mathsf{prior}) & \text{if } \mathsf{posterior} > \mathsf{prior} \\ \mathsf{d}/\mathsf{prior} & \text{o/w} \end{cases} \end{aligned}$$

The running example employs the BN we constructed for the first scenario in the Witness problem for W1 & W2.

Now, a very general picture of how the calculations of structured coherence goes:

- Build a Bayesian network representing the scenario.
- For each child node which represents an element of the scenario, calculate the expected support it gets from its parent(s).
- Aggregate such expected support scores.

So say we have a Bayesian Network. How do we calculate the expected support? For any state s of a child node C included in a narration, we are interested in the support provided to $C = s$ by the combinations $\mathsf{pa}_1, \cdots, \mathsf{pa}_n$ of possible states of its parents. We ignore pa_i excluded by the narration (so, usually, if parents belong to narration as well, there is only one state to consider). For any remaining combination pa_r, the pair $\langle C = s, \mathsf{pa}_r \rangle$ is assigned Z score, where the prior is $\mathsf{P}(C = s)$, and the posterior is $\mathsf{P}(C = s|\mathsf{pa}_r)$.

In our running example, two child nodes, W1 and W2 correspond to the two testimonies and these are the parented narration nodes. The root node, D, represents the agent's initial

[4] Of course, it might be interesting to see what would happen with the coherence calculations if other confirmation measures are plugged in, but this is beyond the scope of this paper.

W1	D	priorC	post	priorN	weightN	Z	nZ
1	Steve	0.175	0.80	0.981	0.981	0.758	0.743
1	Martin	0.175	0.05	0.004	0.004	-0.714	-0.003
1	David	0.175	0.05	0.004	0.004	-0.714	-0.003
1	John	0.175	0.05	0.004	0.004	-0.714	-0.003
1	James	0.175	0.05	0.004	0.004	-0.714	-0.003
1	Peter	0.175	0.05	0.004	0.004	-0.714	-0.003

Table 2: ECS calculation table for W1 in the first scenario in the Witness problem.

uncertainty about who committed the deed (the prior distribution is uniform) and is not instantiated. For each parented node, we list all combinations of its states and the states of its parents not excluded by the narration. We do it for W1 in the first two columns of Table 1.

We only consider cases in which W1 holds, so we have 1s everywhere in the first column. However, the agent is not supposed to know who committed the deed, so all possible instantiations of D are listed. In our example the prior probability of W1 is in column priorC (prior for the **child**, it is constant here), and the posterior probability of W1 given different states of D is in column post. We then use these values to calculate the Z confirmation measures. In our example, these values are in column Z.

Now, to get from multiple Z scores to expected support levels we weight these scores by normalized marginal probabilities of pa_r *as perceived from the perspective of the narration*.

$$\mathsf{weightN}_i = \frac{\mathsf{priorN}_i}{\sum \mathsf{priorN}_j}$$
$$\mathsf{nZ}_i = \mathsf{weightN}_i \times \mathsf{Z}_i$$

(In the cases in which the parents also belong to a narration, each child has a single Z score.)

In our example (Table 2), priorN gives the distribution of D that we would obtain if we updated the BN with W1 = W2 = 1, that is, with the narration in question. weightN is the result of normalizing priorN (in this case, the probabilities already add up to 1, so this move don't change anything). Now, weight the Z score by the normalized probability, and sum these weighted Z scores, obtaining what we call the *Expected Connection Strength* of the parented node under consideration. In our example, the last column weights Z using weightN. The ECS for W1 is the sum of nZ, 0.728.

As the result of applying this procedure to all nodes that belong to a narration, we get a list of *expected connection strengths*. What do we do with the list of ECS scores thus obtained? For the reasons already discussed, we don't want to simply take the mean.[5] We should pay

[5]The problem is a particular case of a common problem in statistics: how to represent a set of different values in a

special attention to mean and minimum. The mean gives us an idea of how strong the average support between the elements is.[6] We also look at the minimum, because special attention should be paid to weaker links: the weaker such links are, the less trust should be placed in a narration. The presence of strong links doesn't have to make up for the impact of weak links — after all, adding information to a fairly incoherent scenario shouldn't increase its coherence much.[7]

So here's our stab at a mathematical explication of a coherence measure that satisfies the desiderata we just discussed. We are not deeply attached to its particularities and clearly other ways of achieving this goal may we worth pursuing.

- If all values are non-negative, i.e. each relation between parents and a child is supportive, then even the weakest point of a story is high enough not to care about it. In such cases we take mean as the final result.
- If, however, some values are negative, we need to be more careful. We still look at mean, but the lower the minimum, the less attention we should pay to it, and the more attention we should pay to the minimum. If the minimum is -1, we want to give it full weight, $1 = |\mathsf{min}| = -\mathsf{min}$ and ignore (weight by 0) mean. In general, we propose to use $|\mathsf{min}|$ as the weight assigned to the minimum, and $1 - |\mathsf{min}|$ to weight mean. For instance, if the minimum is -0.8, the weight of mean should be $-0.8 + 1 = 0.2$, while if it is -0.2, this weight is 0.8.[8] Note that $1 - |\mathsf{min}| = 1 - |\mathsf{min}| = 1 - (-\mathsf{min}) = 1 + \mathsf{min}$, and so the formula is:

Thus, the full formula is as follows:

$$\mathsf{Structured}(\mathsf{ECS}) = \begin{cases} \mathsf{mean}(\mathsf{ECS}) \times (\mathsf{min}(\mathsf{ECS}) + 1) - \mathsf{min}(\mathsf{ECS})^2 & \text{if } \mathit{min}(\mathsf{ECS}) < 0 \\ \mathsf{mean}(\mathsf{ECS}) & \text{o/w} \end{cases}$$

This function has a desired property which was missing in most of the other coherence measures. Whenever we encounter a logically inconsistent story, i.e. a story with the lowest possible minimum (in our measure it is -1), we'll end up with -1 also as the final score. The achieved results are also plausible if the minimum is close to the lowest possible value.

Thus, our coherence measure is be a function of a series of expected confirmation scores. This is somewhat in line with average mutual support measures which took the coherence of a set to be a function of confirmation scores. One key difference, however, is that average mutual support measures used confirmation scores for all disjoint pairs of non-empty subsets of a given set, and our measure will only rely on the confirmation scores for the support relations indicated

simple way without distorting the information too much? One easy and accurate solution is to plot all values. The problem is, it gives us no unambiguous way to compare different sets. For such tasks, a single score is desirable.

[6] This might seem in line with the average mutual support measures. However, on our approach we only care about specific directions of support.

[7] To take the simplest example, if two elements are logically inconsistent, the whole narration is incoherent, even if some of its other elements cohere to a large degree. Imagine two narrations. In the first one, you have a case where all parent-child links except one get the maximal positive score. The remaining one gets the score of -1. We submit that the overall score should be -1. In the second narration all the relations take a value close to -1. We share the intuition that the narration still should have a higher overall score than -1. The presence of an element with the posterior that equals 0 (which is needed for Z confirmation being -1) means that the probability of the whole scenario itself is null, which is clearly lower than whatever low posterior the other scenario might have.

[8] Again, there are other ways to mathematically capture the intuition that the lower minimum, the more attention is to be paid to it, but we decided to take the most straightforward way of doing so for a ride.

by the BN representing a narration. This is because we don't think a narration should be punished for the lack of confirmation between elements that were never intended to be related. Another difference is that while average mutual support measure take simply confirmation levels, we take them as expected from the perspective of a given narration.

5.1 Updated weights?

We think that what this example illustrates is that we should really carefully think about whose cognitive perspective is taken when we represent a narration using a BN, focusing on whether the BN involves nodes which are not part of the narration whose coherence is to be evaluated. In particular, the probabilistic information about the uniform distribution of guilt probability is not part of any of the three involved narrations, but rather a part of a third-person set-up prior to obtaining any evidence.

To evaluate the coherence of a narration, at least for unmentioned assumptions that one doesn't have strong independent reasons to keep, one should think counterfactually, granting the consequences of the narration and asking what would happen if it indeed was true.

In our case, a judge who evaluates the coherence of witness testimonies once she has heard them, no longer thinks that the distribution of D is uniform. And this agrees with the counterfactual strategy we just described: it is a consequence of the probabilistic set-up and the content of W1 and W2 that if W1 and W2 were true, the distribution for D no longer would be uniform, and so it is unfair to judge the coherence of this scenario without giving up this assumption and updating one's assumptions about D.

In such a case, we think, we should update D to what it would be had W1 and W2 be instantiated with 1s:

	Steve	Martin	David	John	James	Peter
Pr	0.981	0.004	0.004	0.004	0.004	0.004

and use these updated probabilities to build the weights used in our coherence calculations for this narration (and proceed accordingly, instead updating on another set of narration nodes in the coherence evaluation of other narrations).[9] Once this strategy is taken, the problem turns out to be not that challenging for any of the coherence measures under discussion.

6 Coherence in Sally Clark

First, we are interested in the coherences assigned to different scenarios in the Sally Clark case prior to any evidence regarding bruising or disease. Call this Stage 0. Calculations for the coherence measures mentioned in the paper are as in Table 3.

[9] Note however that you should not simply instantiate the BN with W1 and W2, propagate and run the coherence calculations on the updated BN. Then both these nodes would get 1s in their respective CPTs and coherence calculations would make all confirmation measures involved in such calculations based on posterior probability equal 1. If narration members have probability one, no other information will be able to confirm it.

States	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass	Probability
00	0.1984	0.9924	0.0783	0.9997	1.0850	0.9993	0.9211
11	0.2000	0.9993	0.9176	0.9962	12.6694	0.9924	0.0783
01	-0.9855	-0.9998	-0.4996	0.0001	0.0002	0.0000	0.0000
10	-0.9997	-0.9919	-0.4962	0.0041	0.0081	0.0006	0.0006

Table 3: Coherence scores and probabilities for the Sally Clark BN prior to evidence regarding bruising or disease.

Compare it to Stage A, when bruising was considered found in both sons but signs of disease in none (Table 4).

States	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass	Probability
00	0.1990	0.9985	0.6997	0.9983	3.3430	0.9966	0.2981
11	0.2000	0.9966	0.2979	0.9993	1.4247	0.9985	0.7009
01	-0.9914	-0.9999	-0.4995	0.0000	0.0001	0.0000	0.0000
10	-0.9998	-0.9952	-0.4981	0.0024	0.0048	0.0010	0.0010

Table 4: Coherence scores and probabilities for the Sally Clark BN after evidence of bruising in both sons and of no disease in both.

Finally, let's also inspect the coherence scores for Stage B, in which bruising was found in both, but signs of disease were present in son A (Table 5)

States	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass	Probability
00	0.1854	0.9346	0.0426	0.9983	1.0446	0.9966	0.9541
11	0.2000	0.9967	0.9207	0.9650	21.7984	0.9300	0.0427
01	-0.8727	-1.0000	-0.4984	0.0000	0.0000	0.0000	0.0000
10	-0.9999	-0.9312	-0.4649	0.0367	0.0731	0.0032	0.0032

Table 5: Coherence scores and probabilities for the Sally Clark BN after evidence of bruising in both sons and of disease in son A.

Here are some observations with respect to what we would intuitively expect.

- We would expect both 01 and 10 to have much lower coherence than 11 or 00. This condition is satisfied by all measures.
- In Stage 0 we would expect the coherences of 00 and 11 to be close. This requirement fails for Douven-Meijs and Shogenji.
- Perhaps, we would not expect 11 and 00 to be nearly maximally coherent. If we buy into this intuiton, we have reasons to dislike Fitelson, Roche and Olsson-Glass.
- When moving from Stage 0 to Stage 1, the coherence of 11 should not decrease. After all, we include evidence in support of 11. This condition fails for Douven-Meijs, Shogenji and to a slight extent for Fitelson.

- For similar reason, we would not expect the coherence of 11 to be much less than the coherence of 00 in Stage A. This condition fails for Douven-Meijs and Shogenji.
- Given that in Stage B we include evidence supporting the claim that son A was not murdered, we would not expect the coherence of 01 to be smaller than the coherence of 10 in Stage B. This condition fails radically for Fitelson (01 becomes maximally incoherent, even though the evidence is not conclusive!), and to a small degree for Douven-Meijs.

Another interesting observation arises when we take all the table rows as data points and think about the truth-conduciveness of coherence. While scenarios with high coherence, presumably, can have various probabilities, we might have the intuition that coherence is at least a negative criterion: scenarios with very low coherence should tend to have low probability. What's the situation with the four scenarios in the three stages we discussed, with respect to all the measures we discussed? Well, this condition holds for all measures except for Douven-Meijs and Shogenji, where scenarios with higher coherence tend to have much lower probability than scenarios with relatively neutral coherence, which we find somewhat surprising (Figure 7).

7 Some results & discussion

	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass
Beatles: JPGRD 11111	-1	-0.0361	0.0247	0.3222	0	0

Table 6: Coherence scores in the Beatles scenario.

	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass
Witness: W1W2 11	0.7294	0.7711	0.4464	0.6214	3.5510	0.4508
Witness: W3W4 11	0.4944	-0.2336	-0.1103	0.3147	0.7405	0.1867
Witness: W4W5 11	0.6016	0.2183	0.1103	0.5353	1.2595	0.3655

Table 7: Coherence scores for the witness scenarios.

In light of conceptual problems with the existing coherence measures we have developed a Bayesian-network based coherence measure that relies not only on the probability measure, but also on the underlying network structure.

One question is whether it handles the philosophical counterexamples present in the literature. The answer is, it does, but these issues will be covered in a different paper, as this discussion is quite extensive.

A potential, more practice-oriented application of this tool is to investigate coherence in Bayesian networks based on real-life legal cases— our treatment of the Sally Clark case is only a small step in this direction. This remains a project for the future.

Another line of investigation concerns the potential use of coherence to estimate model priors for Bayesian model averaging, which might be useful in legal contexts might be more fair than

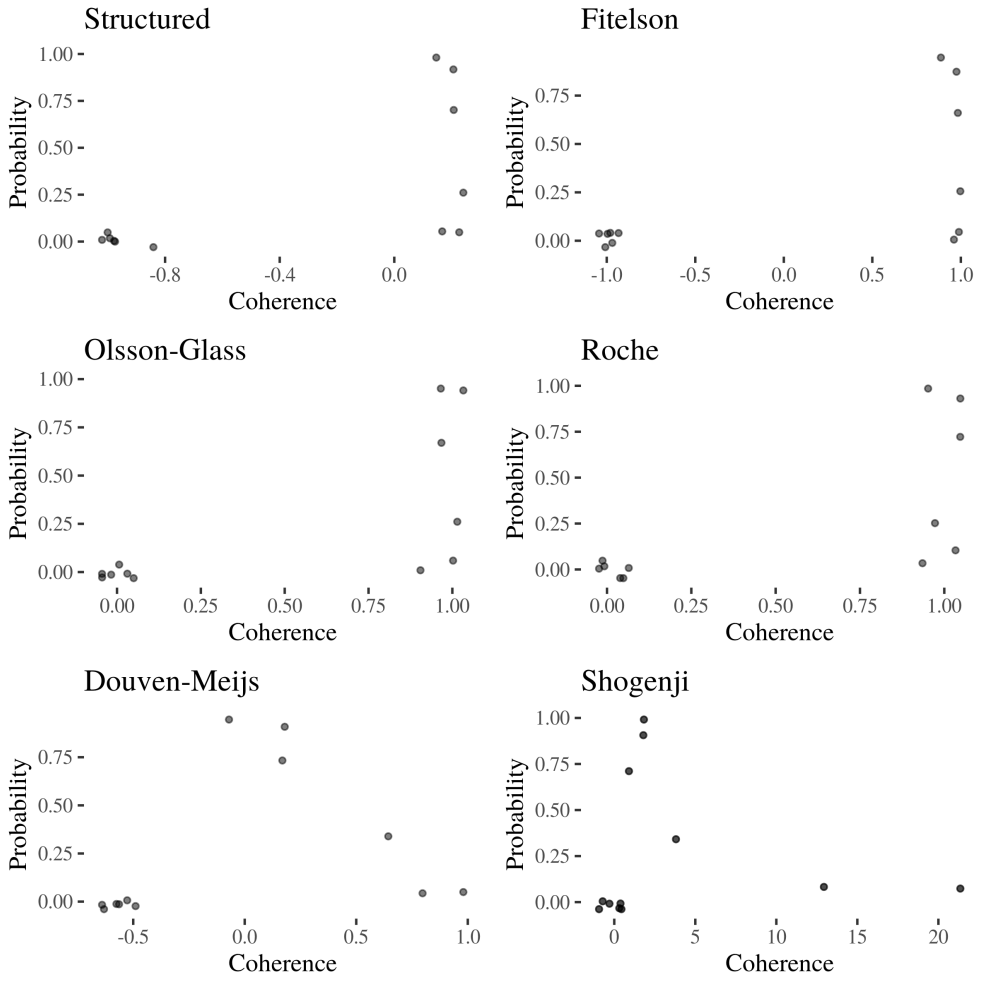

Figure 7: Probability vs. coherence by coherence measure. Note the unusal patterns for Douven-Meijs and Shogenji, and agreement in patterns for the other measures.

using equal or fixed priors or unprincipled intuitive assessment of priors.

References

Akiba, K. (2000). Shogenji's probabilistic measure of coherence is incoherent. *Analysis*, 60(4):356–359.

Bovens, L. and Hartmann, S. (2004). *Bayesian epistemology*. Oxford University Press.

Crupi, V., Tentori, K., and Gonzalez, M. (2007). On Bayesian measures of evidential Support: theoretical and empirical Issues. *Philosophy of Science*, 74(2):229–252.

Douven, I. and Meijs, W. (2007). Measuring coherence. *Synthese*, 156(3):405–425.

Fenton, N. and Neil, M. (2018). *Risk Assessment and Decision Analysis with Bayesian Networks*. Chapman and Hall.

Fenton, N., Neil, M., and Lagnado, D. (2013). A general structure for legal arguments about evidence using bayesian networks. *Cognitive science*, 37(1):61–102.

Fitelson, B. (2003). A Probabilistic Theory of Coherence. *Analysis*, 63(3):194–199.

Glass, D. H. (2002). Coherence, Explanation, and Bayesian Networks. In Goos, G., Hartmanis, J., van Leeuwen, J., O'Neill, M., Sutcliffe, R. F. E., Ryan, C., Eaton, M., and Griffith, N. J. L., editors, *Artificial Intelligence and Cognitive Science*, volume 2464, pages 177–182. Springer Berlin Heidelberg, Berlin, Heidelberg.

Koscholke, J. (2016). Evaluating Test Cases for Probabilistic Measures of Coherence. *Erkenntnis*, 81(1):155–181.

Meijs, W. and Douven, I. (2007). On the alleged impossibility of coherence. *Synthese*, 157(3):347–360.

Merricks, T. (1995). Warrant entails truth. *Philosophy and Phenomenological Research*, 55:841–855.

Olsson, E. J. (2001). Why Coherence Is Not Truth-Conducive. *Analysis*, 61(3):236–241.

Olsson, E. J. (2005). The Impossibility of Coherence. *Erkenntnis*, 63(3):387–412.

Roche, W. (2013). Coherence and probability: A probabilistic account of coherence. In Araszkiewicz, M. and Savelka, J., editors, *Coherence: Insights from philosophy, jurisprudence and artificial intelligence*, pages 59–91. Dordrecht: Springer.

Schippers, M. and Koscholke, J. (2019). A General Framework for Probabilistic Measures of Coherence. *Studia Logica*.

Shogenji, T. (1999). Is Coherence Truth Conducive? *Analysis*, 59(4):338–345.

Shogenji, T. (2001). Reply to akiba on the probabilistic measure of coherence. *Analysis*, 61(2):147–150.

Shogenji, T. (2006). Why does coherence appear truth-conducive? *Synthese*, 157(3):361–372.

Siebel, M. (2004). On Fitelson's measure of coherence. *Analysis*, 64:189–190.

Siebel, M. (2006). Against probabilistic measures of coherence. In *Coherence, Truth and Testimony*, pages 43–68. Springer.

Vlek, C. (2016). *When Stories and Numbers Meet in Court: Constructing and Explaining Bayesian Networks for Criminal Cases with Scenarios*. Rijksuniversiteit Groningen.

Vlek, C., Prakken, H., Renooij, S., and Verheij, B. (2013). Modeling crime scenarios in a bayesian network. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 150–159. ACM.

Vlek, C., Prakken, H., Renooij, S., and Verheij, B. (2014). Building bayesian networks for legal evidence with narratives: a case study evaluation. *Artificial Intelligence and Law*, 22:375–421.

Vlek, C., Prakken, H., Renooij, S., and Verheij, B. (2015). Representing the quality of crime scenarios in a bayesian network. In Rotolo, A., editor, *Legal Knowledge and Information Systems*, pages 133–140. IOS Press.

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