Measuring coherence with Bayesian Networks

1 Motivations & introduction

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.[1] Our basic intuition is that graded coherence should satisfy a generalization of this requirement: logically incoherent sets should have minimal level of graded coherence or, at least, lower coherence than consistent ones. The goal of the paper is to provide an explication of the notion of coherence that satisfies this requirement and performs better than other measures. The key improvement is that once a narration is represented as a Bayesian network, its coherence is determined not only by the underlying probability measure, but also by the network structure.

Why do we need an explication of the notion of coherence, though? First, it is philosophically desirable, as the notion is often used in many philosophical, especially epistemological, discussions (for instance, in discussions about the truth-conduciveness of coherence, Olsson, 2001; Shogenji, 1999). Second, a plausible measure of coherence could be used to better evaluate the quality of stories or narrations. One example when it would be useful, is in the assessment of the quality of narations 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.

Interestingly, there is a disconnect between philosophical research on probabilistic coherence and the development of Bayesian networks–based methods. The latter seems unaware of the philosophical discussion, and the philosophical discussions of coherence do not refer to Bayesian networks. Our paper narrows this gap.

In a Bayesian networks approach to modeling stories and narrations in legal contexts. Vlek et al. (2013) proposed an account on which the coherence of a story (represented by a whole Bayesian network) is captured by the addition of a single narration root node.[2] Its prior probability is identified with coherence. 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.

Our measure diverges from the known candidates in three important respects: (1) It is not a function of a probability 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 to the weakest links in the narration. (3) It is not obtained by simply averaging confirmation levels between all possible combinations of elements.

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

[1] 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.

[2] 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. See (Vlek, 2016; Vlek et al., 2014, 2015, 2016) for more details, and (Fenton et al., 2013) for another take on Bayesian network representation of narrations.

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 measure 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. Two early proposals are the so-called deviation from independence measure and the relative overlap measure.

- Shogenji's *deviation from independence* (Shogenji, 1999), is 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):

$$\mathscr{C}_S(S) = \frac{P(\bigwedge S)}{\prod_{i=1}^{|S|}\{P(S_i)|i \in S\}} \tag{Shogenji}$$

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

$$\mathscr{C}_O(S) = \frac{P(\bigwedge S)}{P(\bigvee S)} \tag{Olsson}$$

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), To overcome them, more recent works developed a class of measures called *average mutual support*. Let's take a look at the general recipe for such a 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 $\bigwedge X$ is the conjunction of all the elements of X (and $\bigwedge Y$ is to be understood analogously).
- Take the mean of all the results.

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

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 (the resulting coherence measure ranges from -1 to 1 with neutral point at 0):

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

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

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

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

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

- Roche (2013) uses the absolute confirmation measure (the resulting coherence measure ranges from 0 to 1 with neutral point at 0.5):

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

$$\mathscr{C}_{R}(S) = mean\left(\left\{A(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i \rangle \in P\right\}\right) \tag{Roche}$$

3 Scenarios and Bayesian Networks

One important way to evaluate coherence measures is to look at how they behave in test scenarios.[3] Some of those come from philosophical literature, and were put forward as counterexamples: they usually have the form of a few propositions formulated in natural language, such that intuitive judgments of coherence involved and the formal coherence calculations diverge. Due to space limitations we will only bring up a couple of them to illustrate what we think the key problems with the existing measures are. We are aware of other cases (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)), but their treatment is postponed to a different paper.

More importantly, we will also include a real-life example of the famous Sally Clark case. As we intend our measure to have practical applications, we need a sanity check of taking a look at how it behaves in a real-life scenario.

In this section we introduce the scenarios we will use, describe how we capture them using Bayesian networks, and mention coherence-related intuitions that seem to accompany them.

3.1 The Beatles

The Beatles example has been offered by Shogenji (1999, 339) to criticize defining coherence of a set in terms of coherence of pairwise coherence. 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.

Table 1: Nodes in the Beatles scenario.

[3] Another way present in the literature is to formulate abstract formal requirements for a coherence measure and to investigate whether a given coherence measure satisfies them. Since there is no agreement in the literature on what such requirements should be, we decided not to take this path in this paper.

The key observation here is that each proper subset of the set is consistent, the whole set is not. So the coherence of the whole set should be lower than the coherence of each proper subset thereof.

In our representation of the scenario by means of a Bayesian network, we assume the prior probability of each individual band member being dead to be 0.5 (as in the above table), and the conditional probability table (CPT) for node D is many-dimensional and so difficult to present concisely, but the method is straightforward: probability 1 is given to node D in all combinations of the parents in which exactly one is false, and otherwise D gets conditional probability 0. The BN with marginal probabilities looks as in Figure 1.

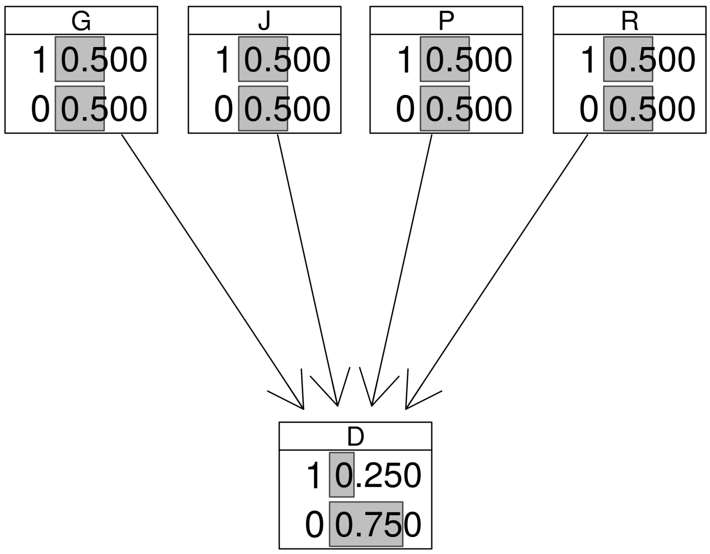

Figure 1: Beatles BN with marginal probabilities.

3.2 The Witnesses

The witnesses scenario comes from (Olsson, 2005, 391). Equally reliable witnesses try to identify a criminal. Consider the potential reports (we extended the original scenario by adding W5) listed in Table 2.

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)

Table 2: Potential witness reports in the witness scenario.

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 $\{\mathsf{W3}, \mathsf{W4}\}$. The underlying intuition is the more suspects the witnesses agree on, the more coherent the evidence.

In our BN representation, each of these three sets is represented by a network 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. Figure 2 illustrates the DAG for the first set, together with the marginal probabilities. The other two networks are analogous.

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 (Table 3) 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 Bayesian networks for other witness scenarios, the CPT for D remains the same, and the CPTs for the witness nodes are analogous to the one for W1.

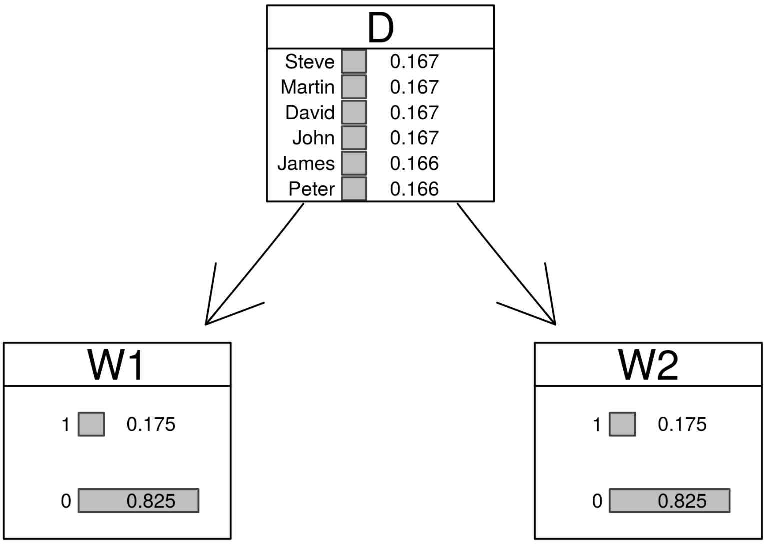

Figure 2: BN for the witnesses example (W1W2) with marginal probabilities.

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

Table 3: CPT for W1 in the W1W2 scenario. CPTs for other witnesses are analogous.

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 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.

We will be interested in tracking probability and coherence in three stages of the case:

- In Stage 0, it is known that the children have died, but no evidence regarding bruising or traces of disease is available.
- In Stage 1, which corresponds to the original case, bruising is found in two children, but no trace of disease in either.
- In Stage 2, bruising was found in both sons, but signs of disease are also present in the first son.

Here are some intuitions one might have about the coherences involved. Let's call the scenario in which both children died of SIDS 00, the one in which both were murdered 11, the one in which only the first child was murdered 10, and the one in which only the second one was murdered 01.

11 and 00 $>$ 10 and 01	In all stages, we would expect 11 and 00 to be more coherent than either 10 or 01. The intution is that the claim that both sons died in the same way sounds more coherent than the alternative scenarios.
11 Stage 1 $>$ 11 Stage 0	When moving from Stage 0 to Stage 1, the coherence of 11 should increase. After all, we include evidence in support of 11.
00 Stage 1 $<$ 00 Stage 0	For the same reason—we now consider evidence in support of 11—we would also expect the coherence of 00 to decrease in Stage 1 as compared to Stage 0.
10 Stage 2 $<$ 01 Stage 2	Given that in Stage 2 we include evidence supporting the claim that son A was not murdered, we would expect the coherence of 01 to be larger than the coherence of 10 in Stage 2.
00 Stage 2 $>$ 00 Stage 1	Once evidence in support of innocence is obtained, we would expect the coherence of 00 to increase.
11 Stage 2 $<$ 11 Stage 1	When evidence against 11 is obtained, the coherence of 11 is expected to decrease.

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 3.

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). The CPTs for the key nodes are in Table 4 (conditional probabilities for Bbruising and Bdisease are the same as for Abruising and Adisease).[4]

[4]Note that one might have somewhat different view on what these should be. For one thing, the probability that the second child has been killed if the first died of SIDS is extremely low. For another, the probability of sings of bruising in case of murder increase from .01 to only .05. Moreover, the probabilities might look too specific for the reader. Analysis with a range of alternative CPTs might indeed be worthwile.

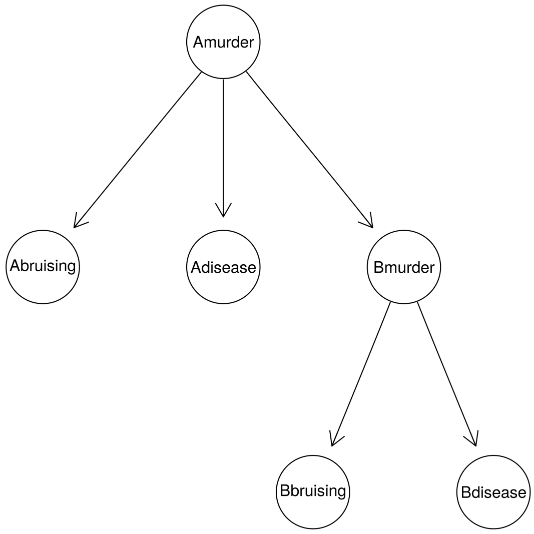

Figure 3: The directed acyclic graphs for the Sally Clark BNs.

Amurder	Pr
0	0.922
1	0.078

Bmurder	Amurder 0	Amurder 1
0	0.999	0
1	0.001	1

Abruising	Amurder 0	Amurder 1
1	0.01	0.05
0	0.99	0.95

Adisease	Amurder 0	Amurder 1
1	0.05	0.001
0	0.95	0.999

Table 4: CPTs for key nodes in the Sally Clark BN.

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; Siebel, 2004, 2006). We try a more principled approach, making a few more general conceptual points that suggest a way forward. We will critize taking the mean support in all possible directions between the elements of a narration and that attention needs to be paid to the structure of a narration.

4.1 Mean

To illustrate the problem with taking the mean of all confirmation levels, note the results in Table 5. The Beatles is a logically inconsistent scenario and yet 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.

One source of the problems might be that 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

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

Table 5: Three coherence measures applied to a consistent (Witness, variant W3W4) and an inconsistent scenario (Beatles).

pairs of subsets. However, simply taking the mean of results so obtained might be misleading, because a few low values (for the inconsistent subsets), which indicate inconsistency, might be masked by many positive values, and taking the mean of all such results might give a relatively high score, despite the set being inconsistent. Therefore, we believe that a candidate for a coherence measure shouldn't simply take the mean of all confirmation scores. Which confirmation scores should count, we will argue, depends on the structure of a narration.

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:[5]

- 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.

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. For this reason, from the perspective of these measures, being a bird disconfirms being a grounded animal, and this will decrease the coherence of the scenario that *Tweety is a bird, grounded, and a penguin*. In such a calculation it 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]This example is inspired by a scenario discussed in (Bovens and Hartmann, 2004, 50) and (Meijs and Douven, 2007).

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

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:[6]

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

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, 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, 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 nodes. The root node, D, represents the agent's initial 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

[6] 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.

the **child**, it does not depend on the state of D, so it is constant, repeated to facilitate row-wise calculations), 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 6), 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 doesn'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? Our discussion of logical inconsistency suggests that special attention should be paid to weakest links in a narration, and so our measure will not only average the ECS scores, but also take the minimum into consideration.[7] The mean gives us an idea of how strong the average support between the elements is.[8] 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.[9]

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.[10] Note that $1-|\mathsf{min}| = 1-|\mathsf{min}| = 1-(-\mathsf{min}) = 1+\mathsf{min}$, and so the formula is:

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

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

[9] 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.

[10] 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.

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

This function has a desired property which was missing in average mutual support 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.

Our measure might resemble average mutual support measures in being a fuction of confirmation scores. However, there are at least two key differences. One is that while the latter take simply confirmation scores, we take their *expected values* from the perspective of a given narration. Another is that, the latter use confirmation scores for all disjoint pairs of non-empty subsets of a given set, and our measure only relies on the (expected) confirmation scores for the support relations indicated 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.

5.1 Updated weights?

One insight that we would like to propose 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. For instance, in The Witnesses, 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, one should think counterfactually: granting the consequences of the narration and asking what would happen if it indeed was true. In The Witnesses, 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 consider an updated conditional probability table for node 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.[11]

6 Results

Now, let's see what results the structured coherence calculations give for the examples and requirements discussed earlier. We start with The Beatles and The Witnesses (Table 7).

First, note that once the updating strategy we discussed is taken, the desiderata in The Witnesses turn out to be not that challenging for any of the coherence measures under discussion.

Second, our measure assigns the minimal coherence to The Beatles scenario, which is as desired. The only two other measures that do so are Shogenji and Olsson-Glass, but they

[11] 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.

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

Table 8: Coherence scores in the Sally Clark scenarios in three stages, with priors and posteriors given evidence.

Stage	States	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass	Priors	Posteriors	Evaluation
Stage 0	00	0.1984	0.9924	0.0783	0.9997	1.085	0.9993	0.9211	NA	0.9211
Stage 0	11	0.2	0.9993	0.9176	0.9962	12.67	0.9924	0.0783	NA	0.0783
Stage 0	01	-0.9855	-0.9998	-0.4996	1e-04	2e-04	0	0	NA	0
Stage 0	10	-0.9997	-0.9919	-0.4962	0.0041	0.0081	6e-04	6e-04	NA	6e-04
Stage 1	00	0.0192	-0.1243	-0.0355	0.3491	0.2075	2.77e-05	8.313e-05	0.2981	0.2981
Stage 1	11	0.6062	0.4974	0.1428	0.288	3.824	3.7e-06	0.0001954	0.7009	0.7009
Stage 1	01	-0.9842	-0.0442	-0.0249	0.216	6.176	7.06e-05	2.797e-07	0.001	0.001
Stage 1	10	-0.9997	-0.058	-0.0221	0.2184	0.0247	3e-07	5.435e-09	2e-05	2e-05
Stage 2	00	0.0205	-0.1243	-0.0355	0.2324	0.2257	1.458e-06	4.375e-06	0.9541	0.9541
Stage 2	11	-0.9525	0.2443	0.0889	0.1591	4.159	2.04e-07	1.956e-07	0.04266	0.04266
Stage 2	01	-0.9842	-0.1213	-0.0188	0.1175	6.716	3.719e-06	1.472e-08	0.003211	0.003211
Stage 2	10	-0.9998	-0.1564	0.0088	0.1463	0.0268	1.5e-08	5.44e-12	1e-06	1e-06

are problematic for other reasons (Akiba, 2000; Koscholke, 2016; Merricks, 1995; Schippers and Koscholke, 2019; Shogenji, 2001, 2006; Siebel, 2004, 2006). Moreover, they are also problematic when it comes to the Sally Clark case, to which we now turn.

We are interested in the coherences assigned to different scenarios in the Sally Clark case in the three stages, together with the prior probability of a given combination of node values (we include combinations of evidence nodes in Stage 1 and Stage 2). The results of the calculations for the coherence measures mentioned in the paper are as in Table 8.

Table 9 contains the results that the coherence measures yield for the intuitions we discussed in Subsection 3.3. Structured coherence is the only measures that satisfies all of them.

Table 9: Satisfaction of intuitions about the Sally Clark problem.

	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass
11 and 00 > 10 and 01	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
11 Stage 1 > 11 Stage 0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
00 Stage 1 < 00 Stage 0	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10 Stage 2 < 01 Stage 2	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
00 Stage 2 > 00 Stage 1	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
11 Stage 2 < 11 Stage 1	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE

Another interesting observation arises when we take all the table rows as data points and think about the truth-conduciveness of coherence. Let's introduce another variable, Evaluation, which simply collects the priors for Stage 0 and the posteriors for Stages 1 and 2— it corresponds to "probabilities given the evidence", where the evidence in Stage 0 is null. 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 (i.e. Evaluation, in this context). 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, it holds to a low extent for Olsson-Glass and Roche, and to a larger extent for Fitelson and Structured (Figure 4).[12]

[12]The Spearman correlation test p-values are fairly low in most of the cases. Here is a table of rounded P-values for

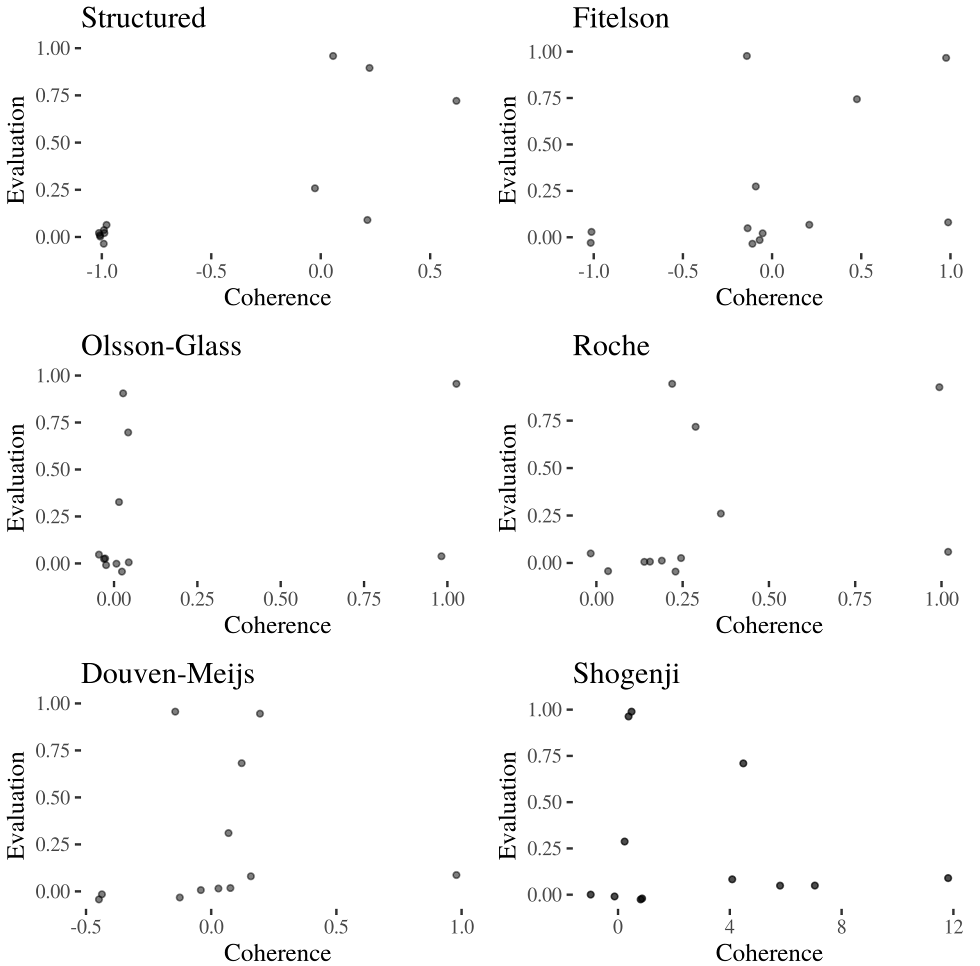

Figure 4: Probability vs. coherence by coherence measure. Note the unusal patterns for Douven-Meijs and Shogenji, and agreement in patterns for the other measures.

Another interesting perspective is gained by looking at Spearman correlations.[13] On the one hand, we look at pairwise correlation between various coherence measures to see to what extent they are order-equivalent. On the other hand, we inspect correlation with both the prior probability and the Evaluation variable (Figure 5), tentatively thinking of it as a measure of the extent to which coherence is truth-conducive for the scenarios at hand. Note that structured coherence, Fitelson and Roche's measure are the only ones which have correlation coefficients above .5 with both priors and posteriors, and that the correlation with posteriors is significantly higher for structured coherence.

Spearman correlation tests for the Sally Clark scenarios:

	Structured	Fitelson	Douven-Meijs	Roche	Shogenji	Olsson-Glass
Structured	0.000	0.010	0.081	0.004	0.044	0.136
Fitelson	0.010	0.000	0.001	0.006	0.004	0.097
Douven-Meijs	0.081	0.001	0.000	0.079	0.013	0.519
Roche	0.004	0.006	0.079	0.000	0.183	0.063
Shogenji	0.044	0.004	0.013	0.183	0.000	0.159
Olsson-Glass	0.136	0.097	0.519	0.063	0.159	0.000
Priors	0.014	0.060	0.291	0.014	0.265	0.000
Evaluation	0.000	0.060	0.249	0.003	0.118	0.106

[13] Spearman correlation is simply Pearson correlation run on ranks instead of raw values. The connection is not linear, so Pearson correlation would be misleading in this context.

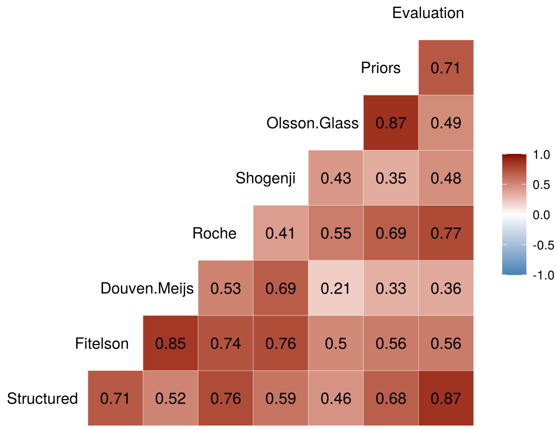

Figure 5: Spearman correlations for the Sally Clark coherence measures and probabilities.

7 Some results & discussion

Having compared the performance of structured coherence with that of the other measures, let's wrap it up with a quick summary and a discussion of further research directions.

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. We illustrated and investigated its pefromance in relation to a list of scenarios and evidential stages involved in the Sally Clark case. This opens a path to further research.

For one thing, a potential, more practice-oriented application of structured coherence is to investigate coherence in other 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 further question we did not discuss is whether the measure handles all 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.

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 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.