Measuring coherence with Bayesian Networks

Abstract.

The notion of coherence is often used in many philosophical, especially epistemological, discussions (for instance, in discussions about the truth-conduciveness of coherence). An explication of the key notion involved seems desirable. We introduce the most prominent coherence measures and a number of counterexamples put forward against them. Then, we point out some common problems that underlie these counterexamples. These observations lead us to a new measure of coherence. 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 not obtained by taking a mean of some list of intermediate values (such as confirmation levels between subsets of a narration). It is sensitive also to the variance and the minimal values of the intermediate values. (3) The intermediate values used are not confirmation levels, but rather expected and weighted confirmation levels. We apply our measure to the existing counterexamples and compare its performance to the performance of the other measures. It does a better job.

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). 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. A plausible coherence measure could be used to asses an important aspect of a narration which so far seems to escape probabilistic analysis.

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] Intuitively, graded coherence should be a generalization of this requirement: logically incoherent sets should have minimal (or at least negative) 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 epistemology? Defining a measure of graded coherence in probabilistic terms turned out to be quite a challenge, which resulted in heaps of literature.

Also, in reasarch unconnected to and seemingly unaware of the philosophical discussion, in the context of Bayesian networks developed for stories and narrations in legal contexts (C. Vlek et al., 2013, C. Vlek et al. (2014), C. S. Vlek et al. (2015), C. Vlek (2016), Fenton et al. (2013), Fenton et al. (2013)), 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 which 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. This is defended by observing that in such a network previously independent nodes become dependent without any principled reason. This is true, but we don't think this is desirable: one shouldn't introduce probabilistic dependencies by fiat in a model. Moreover, Vlek then identifies coherence of a model with the prior probability of the narration node, and we specifically want to capture the idea that coherence is distinct from probability. For this reason, we think an account of coherence for such practical applications is still missing.

We first introduce the main existing coherence measures. Then we describe a lengthy list of counterexamples to these measures. A general discussion of certain common features and issues we observed follows, which leads us to the description of our own coherence measure.

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 not obtained by taking a mean of some list of intermediate values (such as confirmation levels between subsets of a narration). It is sensitive also to the variance and the minimal values of the intermediate values. (3) The intermediate values used are not confirmation levels, but rather expected and weighted confirmation levels (read on for details). Finally, we apply our measure to the existing counterexamples and compare its performance to the performance of the other measures. Spoiler alert: it does a better job.

The whole work has been made possible by all those who contributed to the development of `R` language, and Marco Scutari, the author of `bnlearn` package, who was kind enough to extend his package with additional features upon our requests. The use of these tools here is essential, because we used the environment to write Bayesian Networks (BNs) for all the counterexamples, all coherence functions as applicable to BNs (including ours), and automated performance analysis and BN visualisation, which otherwise wouldn't be manageable.[2]

[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] Our code can be found at: TODO-repo-link.

2 Measures

Let's take a look at different approaches to measuring 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).

2.1 Deviation from independence – Shogenji

The first measure we present was developed by Shogenji (1999, p. 340) and is often called *deviation from independence*. This measure 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.

$$C_s(A_1,\dots,A_n) = \frac{P(A_1 \& \dots \& A_n)}{P(A_1)\times\cdots\times P(A_n)} \tag{Shogenji}$$

scale: [0, ∞]
neutral point: 1
This measure was later generalized by Meijs & Douven (2007). According to this approach, (Shogenji) is applied not only to the whole set of propositions, but to each non-empty non-singleton subset of the set, and the final value is defined as the average of all sub-values thus obtained.

2.2 Relative overlap – Olsson & Glass

The second approach, a *relative overlap* measure, comes from Olsson (2001) and Glass (2002). This measure is defined as the ratio between the intersection of all propositions and their union. It was later generalized in a way analogous to the one used in the generalization of the Shogenji's measure.

$$C_o(A_1,\dots,A_n) = \frac{P(A_1 \& \dots \& A_n)}{P(A_1 \vee\cdots\vee A_n)} \tag{Olsson}$$

scale: [0, 1]
neutral point: NO

2.3 Average mutual support

Finally, the most recent approach — a class of measures called *average mutual support*. The general recipe for such a measure is this.

- 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.

2.3.1 Fitelson

Fitelson (2003a, 2003b) uses the following confirmation function:

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

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

For instance, Fitelson's coherence for two propositions boils down to this:

$$\mathscr{C}_F(X,Y) = \frac{F(X,Y)+F(Y,X)}{2} \tag{Fitelson pairs}$$

scale: [-1, 1]
neutral point: 0

2.3.2 Douven and Meij's

Another coherence measure of this type has been introduced by Meijs & Douven (2007, p. 412)

They use the *difference* confirmation measure:

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

The resulting definition of coherence is:

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

For two propositions, the coherence measure boils down to:

$$C_{DM}(X,Y) = \frac{P(X|Y) - P(X) + P(Y|X) - P(Y)}{2} \tag{DM pairs}$$

scale: [-1, 1]
neutral point: 0 (not explicit)

2.3.3 Roche

Yet another measure, due to Roche (2013, p. 69), 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:

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

For two propositions, the measure gives the following:

$$C_R(X,Y) = \frac{P(X|Y) + P(Y|X)}{2} \tag{Roche pairs}$$

scale: [0, 1]
neutral point: 0.5

3 Challenges

Here are the counterexamples put forward against various coherence measures in the literature. We ignored only a few where both we didn't share the authors' intuitions and the examples were not picked up in further discussion in the literature.

3.1 Penguins

The scenario. A challenge discussed in (Bovens & Hartmann, 2004, p. 50) and (Meijs & Douven, 2007) consists of the following set of propositions (instead of *letters* or *abbreviations*, we'll talk about *nodes*, as these will be used later on in Bayesian networks):

node	content
B	Tweety is a bird.
G	Tweety is a grounded animal.
P	Tweety is a penguin.

Desiderata. It seems that the set {B,G}, which doesn't contain the information about Tweety being a penguin, should be less coherent than the one that does contain this information: {B,G,P}.

(BG<BGP) {B,G} should be less coherent than {B,G,P}.

Another intuition about this scenario (Schippers & Koscholke, 2019) is that when you consider a set which says that Tweety is both a bird and a penguin: {B,P}, adding proposition about not flying (G) shouldn't really increase the coherence of the set. It's a well-known fact that penguins don't fly, and so one can deduce G from P. Therefore by adding G explicitly to the set, one wouldn't gain any new information – so if a set expresses the same information, its coherence shouldn't be different. However, as G is not a logical consequence of P, it can be argued that {B,P} and {B,P,G} represent different information sets, and a slight difference in their coherence is also acceptable.

(BP≈BGP) {B,P} should have similar coherence to {B,P,G}.

3.2 Dunnit

The scenario. Another challenge, introduced by Merricks (1995) goes as follows: Mr. Dunnit is a suspect in the murder case. Detectives first obtained the following body of evidence:

node	content
I	Witnesses claim to have seen Dunnit do it (incriminating testimony).
M	Dunnit had a motive for the murder.
W	A credible witness claims to have seen Dunnit two hundred miles from the scene of the crime at the time of the murder.

In light of this information they try to assess whether Dunnit is responsible for the crime.

node	content
G	Dunnit is guilty.

Now, suppose the detectives learn Dunnit has a twin brother.

node	content
Tw	Dunnit has an identical twin which was seen by the credible witness two hundred miles from the scene of the crime during the murder.

and compare the coherence of {I,M,W,G} with the coherence of {I,M,W,G,Tw}.
Desideratum. It seems that adding proposition about a twin should increase the coherence of the set.

(Dunnit<Twin) {I,M,W,G} should be less coherent than {I,M,W,G,Tw}.

3.3 Japanese swords

The scenario. The next challenge comes from (Meijs & Douven, 2007, p. 414):

> We start by considering two situations in both of which it is assumed that a murder has been committed in a street in a big city with 10,000,000 inhabitants, 1,059 of them being Japanese, 1,059 of them owning Samurai swords, and 9 of them both being Japanese and owning Samurai swords. In situation I we assume that the murderer lives in the city and that everyone living in the city is equally likely to be the murderer. In situation II, on the other hand, we make the assumption that the victim was murdered by someone living in the street in which her body was found. In that street live 100 persons, 10 of them being Japanese, 10 owning a Samurai sword, and 9 both being Japanese and owning a Samurai sword. [...] [In situation III] we have 12 suspects who all live in the same house, and 10 of them are Japanese, 10 own a Samurai sword, and 9 are both Japanese and Samurai sword owners.

The nodes involved are as follows:

node	content
J	The murderer is Japanese.
O	The murderer owns a Samurai sword.

And we look at three separate scenarios: (1) The murderer lives in the city, (2) The murderer lives in the street popular amongst Japanese owners of Samurai swords, and (3) The murderer lives in the house with many Japanese owners of Samurai swords.
Desiderata. In all of the above situations the number of Japanese owners of Samurai swords remains the same. However, situations 1 and 2 differ in the relative overlap of J and O. Because J and O are more correlated in situation 2, it seems more coherent than situation 1.

(JO2>JO1) {J,O,2} should be more coherent than {J,O,1}

However, bigger overlap, supposedly, doesn't have to indicate higher coherence. In situation 3 J and O confirm each other to a lesser extent than in situation 2 (compare $P(J|O) - P(J)$ and

$P(O|J) - P(O)$ in both cases), and for this reason Douven and Meijs claim that situation 2 is more coherent than situation 3.

(JO2>JO3) {J,O,2} should be more coherent than {J,O,3}

We don't have clear intuitions about this desideratum. It seems to be in tension with the requirement offered by Siebel (2004, p. 336) which we'll discuss in the next subsection.

3.4 Robbers

The scenario. A challenge put forward by Siebel (2004, p. 336) goes as follows:

> Let there be ten equiprobable suspects for a murder. All of them previously committed at least one crime, two a robbery, two pickpocketing, and the remaining six both crimes. There is thus a substantial overlap: of the total of eight suspects who committed a robbery, six were also involved in pickpocketing, and conversely.

node	content
W	Real perpetrator status (three possible states).
P	The murderer is a pickpocket.
R	The murderer is a robber.

Desiderata. The first observation is that the set of propositions that corresponds to the situation in which a murderer committed both crimes should be regarded coherent. Most suspects committed both crimes, so this option is even the most probable one.

(PR>neutral) {P,R} should be regarded coherent.

According to Siebel (2004, p. 336) committing both crimes by the murderer should also be regarded more coherent than committing only one crime.

(PR>P¬R) {P,R} should be more coherent than {P,¬R} and {¬P,R}.

This requirement is slightly more controversial. Even though {P,R} is the most probable setup, P and R disconfirm each other ($Pr(P|R) < Pr(P)$ and $Pr(R|P) < Pr(R)$). Moreover, the intuition behind this desideratum seems to conflict with the intuition behind (JO2>JO3).

3.5 The Beatles

The scenario. The challenge has been offered by Shogenji (1999, p. 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.

Desiderata. The set consisting of all of these propositions is logically inconsistent (even though the propositions are pairwise consistent), so it seems quite intuitive that it should be incoherent.

(below neutral) $\{\mathsf{D,J,P,G,R}\}$ should be incoherent.

We can make this desideratum a bit stronger by requiring that the coherence score for $\{\mathsf{D,J,P,G,R}\}$ should be minimal.

(minimal) $\{\mathsf{D,J,P,G,R}\}$ should get the lowest possible coherence value.

One may argue that some coherence measures also measure the degree of incoherence, therefore logically inconsistent sets don't need to get the minimal score. We'll discuss this issue further in Section 4.2.

3.6 Alicja and books

Prima facie, at least some sets with low posterior probability can be quite coherent, and at least some sets with fairly high posterior probability can have low coherence. To keep track of how various measures perform with respect to this intuition, we developed the following example.

The scenario. Alicja reads (R) 10% of books she buys, but 15% of books she buys that Rafal advised (A) her to read.

Here, we just have two nodes:

node	content
A	Rafal adviced Alicja to read the book.
R	Alicja read the book.

Desiderata. At least *prima facie*, these conditions seem intuitive:

$(\mathsf{AR} > \mathsf{A}\neg\mathsf{R})$ Given that Alicja was advised to read, it's more coherent that she read the book than not.

$(\mathsf{AR} > \neg\mathsf{AR})$ Given that Alicja read the book, it's more coherent that she was advised than not.

$(\neg\mathsf{A}\neg\mathsf{R} > \mathsf{A}\neg\mathsf{R})$ Given that Alicja didn't read the book, it's more coherent that she wasn't advised than that she was.

($\neg$A$\neg$R $>$ $\neg$AR) Given that Alicja wasn't advised to read, it's more coherent that she didn't read the book than that she did.

3.7 The Witnesses

The scenario. This one comes from (Olsson, 2005, p. 391). Again, 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 this time each proposition has the structure "Witness no. *X* claims that ..." instead of explicitly stating the witness' testimony.

Desiderata. First, we can observe that W1 and W2 fully agree. Testimonies of W3 and W4 overlap only partially, therefore it seems that {W1,W2} is more coherent than {W3,W4}.

(W1W2>W3W4) {W1,W2} should be more coherent than {W3,W4}.

Similarly, there is a greater agreement between W4 and W5 than W3 and W4, so {W4,W5} seems more coherent than {W3,W4}.

(W4W5>W3W4) {W4,W5} should be more coherent than {W3,W4}.

3.8 Depth

The scenario. There are eight equally likely suspects $1,\ldots,8$, and three equally reliable witnesses a,b,c, each trying to identify the person responsible for the crime. Compare two different situations – X1 and X2:

$$X_1 = \{a : (1 \vee 2 \vee 3), b : (1 \vee 2 \vee 4), c : (1 \vee 3 \vee 4)\}$$
$$X_2 = \{a : (1 \vee 2 \vee 3), b : (1 \vee 4 \vee 5), c : (1 \vee 6 \vee 7)\}$$

Desiderata. In X1 witnesses' testimonies have bigger overlap, between each pair of the witnesses 2 suspects are the same, and in X2 only 1 suspect is always the same. Following Schupbach (2008), one may have an intuition that the first situation is more coherent.

(X1>X2) X_1 should be more coherent than X_2.

3.9 Dice

The scenario. This scenario was offered by Schippers & Koscholke (2019). You're either tossing a regular die, or a dodecahedron, X is the result (there is nothing particular about this choice of dice; *mutatis mutandis* this should hold for other possible pairs of dice as well). Consider the coherence of:

$$D = \{X = 2, (X = 2 \vee X = 4)\}.$$

Desiderata. In this scenario posterior conditional probabilities are fixed: getting 2 or 4 logically follows from getting 2 ($P(X = 2 \vee X = 4|X = 2) = 1$), and you always have 50% chance to get 2 given that the outcome was 2 or 4 ($P(X = 2|X = 2 \vee X = 4) = 0.5$). Therefore, according to Schippers & Koscholke (2019), the coherence of the set D shouldn't change no matter which die you use.

(D=const) the coherence of D should not change.

4 Observations

4.1 Coherence scores and outcomes

To be able to clearly see how well the existing measures of coherence deal with the mentioned desiderata, we decided to put all the results together. Our analysis extends the work of Koscholke (2016). In total we have analyzed 17 different desiderata and 7 candidates for a coherence measure (8 including ours, to be discussed later on). This required quite a lot of calculations, so we used programming language R and Marco Scutari's bnlearn package to build on. Our code can be found at: TODO-repo-link.

We represented all counterexamples as Bayesian networks. If all probabilities were defined by the author(s) of the counterexample, we used those values, otherwise we had to come up with some common-sense values. Each particular scenario is represented by a set of nodes — usually binary ones, because they correspond to propositions which can be either true or false — and their appropriate instantiations. Finally, we wrote general functions for each of the measures to calculate the coherence scores for all the scenarios we were interested in.

In the following tables you can find coherence scores for various scenarios and measures, a summary of how the measures handle the desiderata, and their success rate for this list of challenges (OG stands for Olsson-Glass, OGen for Olsson-Glass generalized, Sh for Shogenji, ShGen for Shogenji generalized, Fit for Fitelson, DM for Douven-Meijs, R for Roche).

	OG	OGen	Sh	ShGen	Fit	DM	R
Penguins: BGP 111	0.010	0.015	4.000	2.010	0.453	0.255	0.255
Penguins: BG 11	0.010	0.010	0.040	0.040	-0.960	-0.480	-0.480
Penguins: BP 11	0.020	0.020	2.000	2.000	0.669	0.255	0.255
Dunnit: MGWI 1111	0.000	0.087	4.294	11.012	0.169	0.167	0.167
Dunnit: MTGWI 11111	0.000	0.042	73.836	13.669	0.385	0.150	0.150
Japanese Swords 1: JO 11	0.004	0.004	80.251	80.251	0.976	0.008	0.008
Japanese Swords 2: JO 11	0.818	0.818	9.000	9.000	0.976	0.800	0.800
Japanese Swords 3: JO 11	0.818	0.818	1.080	1.080	0.286	0.067	0.067
Robbers: MIsPMIsR 11	0.600	0.600	0.937	0.937	-0.143	-0.050	-0.050
Robbers: MIsPMIsR 10	0.250	0.250	1.250	1.250	0.571	0.125	0.125
Robbers: MIsPMIsR 01	0.250	0.250	1.250	1.250	0.571	0.125	0.125
Beatles: JPGRD 11111	0.000	0.202	0.000	1.423	-0.036	0.025	0.025
Books: AR 11	0.014	0.014	1.493	1.493	0.212	0.027	0.027
Books: AR 10	0.009	0.009	0.945	0.945	-0.127	-0.025	-0.025
Books: AR 01	0.100	0.100	0.995	0.995	-0.101	-0.003	-0.003
Books: AR 00	0.892	0.892	1.001	1.001	0.016	0.001	0.001
Witness: W1W2 11	0.451	0.451	3.551	3.551	0.771	0.446	0.446
Witness: W3W4 11	0.187	0.187	0.740	0.740	-0.234	-0.110	-0.110
Witness: W4W5 11	0.365	0.365	1.260	1.260	0.218	0.110	0.110
DepthA: T123T124 11	0.664	0.664	1.014	1.014	0.280	0.012	0.012
DepthB: T123T145 11	0.331	0.331	0.996	0.996	-0.047	-0.003	-0.003
Regular: TTF 11	0.500	0.500	3.000	3.000	0.833	0.500	0.500
Dodecahedron: TTF 11	0.500	0.500	6.000	6.000	0.917	0.625	0.625

	OG	OGen	Sh	ShGen	Fit	DM	R
Penguins: BG<BGP	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Penguins: BP≈ BGP	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
Dunnit: Dunnit<Twin	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE
Swords: JO2>JO1	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
Swords: JO2>JO3	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
Robbers: PR>P¬R	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Robbers: PR>neutral	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE
Beatles: below neutral	NA	NA	TRUE	FALSE	TRUE	FALSE	TRUE
Beatles: minimal	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
Books: AR>A¬R	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: AR> ¬AR	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: ¬A¬R>A¬R	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: ¬A¬R> ¬AR	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Witness: $W_1W_2 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Witness: $W_4W_5 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Depth: $X_1 > X_2$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Dodecahedron: Regular = Dodecahedron	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE

OG	OGen	Sh	ShGen	Fit	DM	R
0.733	0.733	0.706	0.647	0.706	0.647	0.706

Unfortunately, no measure was able to deal with all challenges. Note that the more recent measures, which were developed to improve on the previous ones, didn't really achieve a much higher success rate. Analysing these tables and various counterexamples, we noticed a few general issues.

4.2 Mean

While the challenges have been discussed separately, considering some of them jointly also leads to an insight. Only one counterexample, The Beatles, is logically inconsistent. However, for each measure (except for the basic versions of Olsson's and Shogenji's, which are the earliest measures and unfortunately face other serious difficulties), there is a scenario that scored lower even though it was logically consistent. For example, the generalized Olsson's measure gives a lower value to the scenario that Tweety is a bird and a penguin than to The Beatles scenario. Other measures give lower values to the scenario with a murderer who committed both pickpocketing and robbery than to The Beatles. This result 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. When you consider the measures that face this problem, you can notice that all of them use subsets of a set (or pairs thereof) and then take the average result calculated for these subsets or 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 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.

4.3 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:

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. 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 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:

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

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.

We use BNs to model the directions of support indicated in the story. A scenario is represented as a selection of instantiated nodes. We'll follow an example as we proceed. The running example employs the BN we constructed for the first scenario in the Witness problem (Figure 10 on page 26). Both witnesses testify that Steve is the murderer. Two child nodes, W1 and W2 correspond to the two testimonies, and as part of the narration, are to be instantiated to 1. The root node, D prior to any update represents the agent's initial uncertainty about who committed the Deed (the prior distribution is uniform) and is not instantiated.

- To calculate the coherence of a scenario, first find all nodes that have at least one parent. In the example, these are W1 and W2.
- For each parented node, list all combinations of its states and the states of its parents not excluded by the narration. We do it for W1 in the table below this list (Table 1), in the first two columns. 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.
- For each parented node and for each combination of possible states: get the prior probability of the child node and get the posterior probability of this child given the parent nodes with their fixed states. In our example the prior probability of W1 is in column priorC (it is constant here), and the posterior probability of W1 given different states of D is in column post.
- Use these values to calculate the Z confirmation measures. In our example, these values are in column Z.
- Get the joint probability of these parent node states. In typical cases, this is simply the prior probability. In cases in which narration nodes are actually pieces of evidence that one learns, these should be posterior probabilities obtained by instantiating the narration nodes in the BN and propagating. Such unusual cases will be discussed when we address the Witness problem. In our example, priorA gives the prior probabilities of various states of D, whereas 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.
- Normalize these joint probabilities of the parent(s) so that all joint parent probabilities in the variant list add up to 1. In the example, weightA is the result of normalizing priorA and weightN is the result of normalizing priorN (in this case, the probabilities already add up to 1, so these moves don't change anything).
- Weight the Z score by this 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 two columns weight Z using weightA and weightN respectively. In normal circumstances, ECS of W1 would now be the sum of aZ, but — as we discuss further on — in this particular case we should use the updated weights, and so the ECS for W1 is the sum of nZ.

As you can see, our approach is a bit similar to *average mutual support* measures, but instead of calculating confirmation of each pair of disjoint subsets, we calculate it only for parents-child pairs. As a result we get a list of *expected connections strengths* of each child in the BN.

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. 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

W1	D	priorC	post	priorA	priorN	weightA	weightN	Z	aZ	nZ
1	Steve	0.175	0.80	0.167	0.981	0.167	0.981	0.758	0.126	0.743
1	Martin	0.175	0.05	0.167	0.004	0.167	0.004	-0.714	-0.119	-0.003
1	David	0.175	0.05	0.167	0.004	0.167	0.004	-0.714	-0.119	-0.003
1	John	0.175	0.05	0.167	0.004	0.167	0.004	-0.714	-0.119	-0.003
1	James	0.175	0.05	0.167	0.004	0.167	0.004	-0.714	-0.119	-0.003
1	Peter	0.175	0.05	0.167	0.004	0.167	0.004	-0.714	-0.119	-0.003

Table 1: ECS calculation table for W1 in the first scenario in the Witness problem.

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.

In the context of measuring coherence of a scenario it seems that we should pay special attention to mean, sd, and minimum. The mean gives us an idea of how strong the average support between the elements is. Standard deviation informs us how stable the support between elements is (the higher sd, the more cautious we need to be about the mean). The minimum represents the weakest point in the narration. Note that it can be far from the mean — this is important, because even a single very weak point destroys the coherence of a whole story.[3]

We developed a function which uses these 3 values:

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

The intuitions behind it are as follows. 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 $\mathsf{mean} - \mathsf{sd}$ as the final result.

One way to look at the first case is to think of it as a weighted average of $\mathsf{mean} - \mathsf{sd}$ and $\mathsf{min}(\mathsf{ECS})$, min for brevity. The weight assigned to min is $|\mathsf{min}|$. This is fairly natural: if the minimum is -1, we want to give it full weight, $1 = |\mathsf{min}| = -\mathsf{min}$. The weight assigned to $\mathsf{mean} - \mathsf{sd}$ is the rest, $1 - |\mathsf{min}| = 1 - (-\mathsf{min}) = 1 + \mathsf{min}$.

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.

Another desired feature is that if there are some negative ECS values, we need to be cautious

[3] This is a conservative score in the following sense. Imagine two narrations. In the first one, you have a case where all parent-child relations except one get the maximal positive score. The remaining one gets the score of -1. Then the overall score will be -1. In the second narration all the relations take a value close to -1. It will have a higher overall score than -1. On this approach, the second narration is more coherent than the first one. We find this property desirable: the presence of an element with the posterior that equals 0 (which is needed for Z confirmation being -1) kills the coherence of an otherwise strong narration, and a narration containing such an element is in worse standing than simply a very unlikely one.

about the weakest point. The lower the minimum, the less attention we should pay to $\mathsf{mean} - \mathsf{sd}$. So, for instance, if the minimum is -0.8, the weight of $\mathsf{mean} - \mathsf{sd}$ should be $-0.8 + 1 = 0.2$, while if it is -0.2, the weight is 0.8. Now, let's take this coherence measure for a ride.

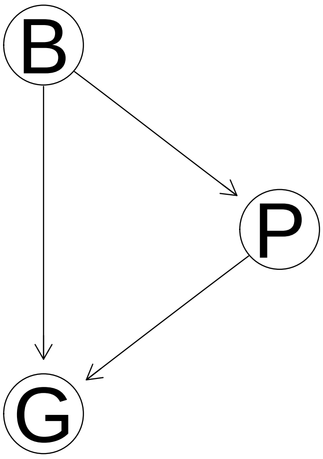
6 Handling counterexamples

Using the notation and desiderata already introduced, let's go over BNs for the counterexamples involved, and use the counterexamples to evaluate the performace of the new measure.

6.1 Penguins

We used the distribution used in the original formulation to build three BNs corresponding to the narrations at play (Fig. 1-3).[4]

[4] Not without concerns. There are around 18 000 species of birds, and around 60 of them are flightless. We couldn't find information about counts, but it seems the probability of being a penguin if one is grounded is overestimated by philosophers. Also, there are many things that are not grounded but are not birds, mostly insects, and there's plenty of them. We did spend some time coming up with plausible ranges of probabilities to correct for such factors, and none of them actually makes a difference to the main point. So, for the sake of simplicity, we leave the original unrealistic distribution in our discussion.

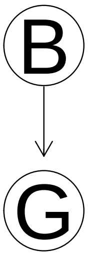


B	Pr
1	0.5
0	0.5

P	B	
	1	0
1	0.02	0
0	0.98	1

	B	P	
G			Pr
1	1	1	1.00
0	1	1	0.00
1	0	1	0.00
0	0	1	1.00
1	1	0	0.00
0	1	0	1.00
1	0	0	0.98
0	0	0	0.02

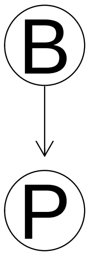
Figure 1: Bayesian network for the BGP scenario.



B	Pr
1	0.5
0	0.5

G	B	
	1	0
1	0.02	0.98
0	0.98	0.02

Figure 2: Bayesian network for the BG scenario.



B	Pr
1	0.5
0	0.5

P	B	
	1	0
1	0.02	0
0	0.98	1

Figure 3: Bayesian network for the BP scenario.

Now, let's calculate the coherences and see if the desiderata are satisfied (the abbreviations we already used are as before, and S stands for *Structured*, our coherence measure):

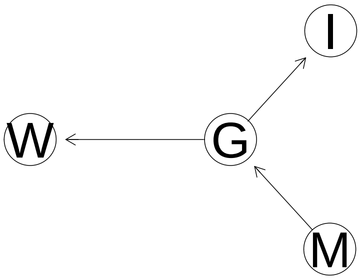
	OG	OGen	Sh	ShGen	Fit	DM	R	S
Penguins: BGP 111	0.01	0.015	4.00	2.01	0.453	0.255	0.255	0.01
Penguins: BG 11	0.01	0.010	0.04	0.04	-0.960	-0.480	-0.480	-0.96
Penguins: BP 11	0.02	0.020	2.00	2.00	0.669	0.255	0.255	0.01

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Penguins: BG<BGP	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Penguins: BP≈ BGP	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE

6.2 Dunnit

Here, we deal with two separate BNs. One, before the Twin node is even considered (Fig. 4), and one with the Twin node (Fig. 5).

The CPTs for the no-twin version are in agreement with those in the ones in the Twin case. Since the original example didn't specify exact probabilities, we came up with some plausible values.



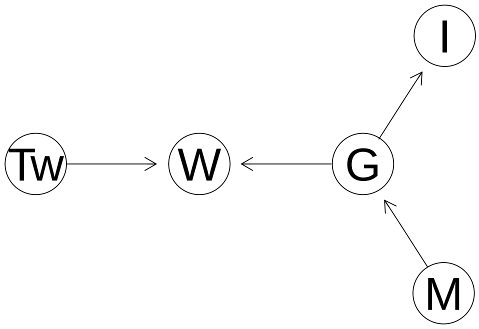
M	Pr
1	0.4
0	0.6

I	G	
	1	0
1	0.8	0.005
0	0.2	0.995

G	M	
	1	0
1	0.05	0.005
0	0.95	0.995

W	G	
	1	0
1	0.012	0.207
0	0.988	0.793

Figure 4: Twin-less BN for the Dunnit problem.



	G	Tw	
W			Pr
1	1	1	0.200
0	1	1	0.800
1	0	1	0.400
0	0	1	0.600
1	1	0	0.005
0	1	0	0.995
1	0	0	0.200
0	0	0	0.800

Figure 5: BN for the Dunnit problem. The key difference for the twin version lies in the construction of the CPT for W. The table gives conditional probabilities for W given various joint states of Tw and G.

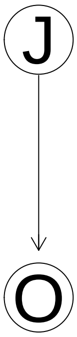
Coherence calculations result in the following:

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Dunnit: MGWI 1111	0	0.087	4.294	11.012	0.169	0.167	0.167	-0.932
Dunnit: MTGWI 11111	0	0.042	73.836	13.669	0.385	0.150	0.150	-0.100

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Dunnit: Dunnit<Twin	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE

6.3 Japanese swords

There is a common DAG for the three scenarios, but the CPTs differ (Fig. 6).



J	Pr
1	0
0	1

O	J	
	1	0
1	0.008	0
0	0.992	1

(a) Scenario 1.

J	Pr
1	0.1
0	0.9

O	J	
	1	0
1	0.9	0.011
0	0.1	0.989

(b) Scenario 2.

J	Pr
1	0.833
0	0.167

O	J	
	1	0
1	0.9	0.5
0	0.1	0.5

(c) Scenario 3.

Figure 6: A common DAG and three sets of CPTs for the Japanese Swords problem.

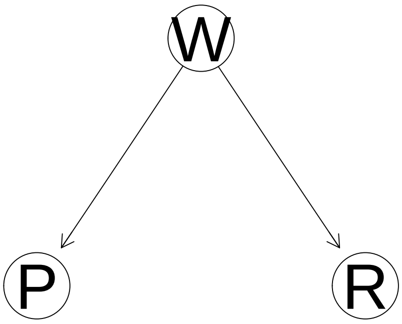
Coherence calculations yield:

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Japanese Swords 1: JO 11	0.004	0.004	80.251	80.251	0.976	0.008	0.008	0.008
Japanese Swords 2: JO 11	0.818	0.818	9.000	9.000	0.976	0.800	0.800	0.889
Japanese Swords 3: JO 11	0.818	0.818	1.080	1.080	0.286	0.067	0.067	0.400

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Swords: JO2>JO1	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Swords: JO2>JO3	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

6.4 Robbers

The robbers counterexample involves a phenomenon we've already seen: it is not clear whether the information about the prior probabilities is supposed to be part of the narration or not. If we want to include this information in our coherence assessment, we can do this employing a single BN.



	Pr
OnlyP	0.2
OnlyR	0.2
Both	0.6

MisP	WhoMurdered		
	OnlyP	OnlyR	Both
1	1	0	1
0	0	1	0

MisR	WhoMurdered		
	OnlyP	OnlyR	Both
1	0	1	1
0	1	0	0

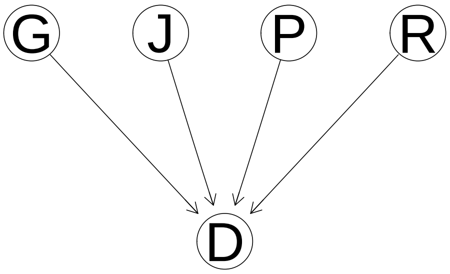
Figure 7: BN for the Robbers problem.

Coherence calculations yield the following results:

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Robbers: MIsPMIsR 11	0.60	0.60	0.937	0.937	-0.143	-0.050	-0.050	0.6
Robbers: MIsPMIsR 10	0.25	0.25	1.250	1.250	0.571	0.125	0.125	-0.6
Robbers: MIsPMIsR 01	0.25	0.25	1.250	1.250	0.571	0.125	0.125	-0.6

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Robbers: PR>P¬R	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
Robbers: PR>neutral	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

6.5 The Beatles



G	Pr
1	0.5
0	0.5

Figure 8: Bayesian network for the Beatles scenario.

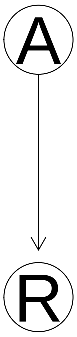
We assume the prior probability of each individual band member being dead to 0.5 (as in the above table), and the CPT for D is many-dimensional and so difficult to present concisely, but the method is straigtforward: probability 1 is given to D in all combinations of the parents in which exactly one is true, and otherwise D gets conditional probability 0. Coherence calculations give the following results:

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Beatles: JPGRD 11111	0	0.202	0	1.423	-0.036	0.025	0.025	-1

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Beatles: below neutral	NA	NA	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
Beatles: minimal	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE

6.6 Alicja and books

The BN is fairly straightforward (Fig. 9) and the results are as follows:



A	Pr
1	0.01
0	0.99

R	A	
	1	0
1	0.15	0.1
0	0.85	0.9

Figure 9: Bayesian network for the Books problem.

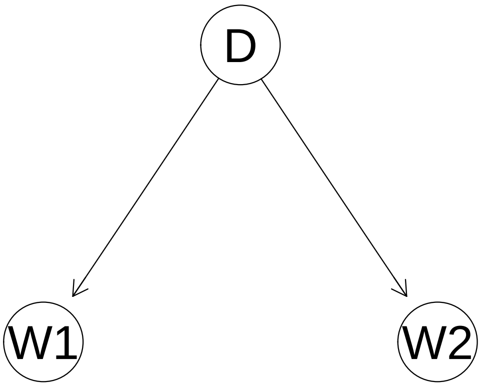
	OG	OGen	Sh	ShGen	Fit	DM	R	S
Books: AR 11	0.014	0.014	1.493	1.493	0.212	0.027	0.027	0.055
Books: AR 10	0.009	0.009	0.945	0.945	-0.127	-0.025	-0.025	-0.055
Books: AR 01	0.100	0.100	0.995	0.995	-0.101	-0.003	-0.003	-0.005
Books: AR 00	0.892	0.892	1.001	1.001	0.016	0.001	0.001	0.005

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Books: AR>A¬R	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: AR> ¬AR	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: ¬A¬R>A¬R	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: ¬A¬R> ¬AR	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

6.7 The witnesses

Two requirements are associated with this example: both {W1, W2} and {W4, W5} should be more coherent than {W3, W4}. 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. Of course, the example can be run with different conditional probability tables. Let's first take a look at the BN for the first scenario (Fig. 10).

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/.... The underlying rule is: if someone is guilty, a witness will mention them



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

with probability .8, and if they aren't, they will be listed with probability .05. In the remaining two BNs for the problem the CPT for D remains the same, and the CPTs for the witness nodes are analogous to the one for W1. The remaining BNs have the following obvious DAGs (Fig. 11).

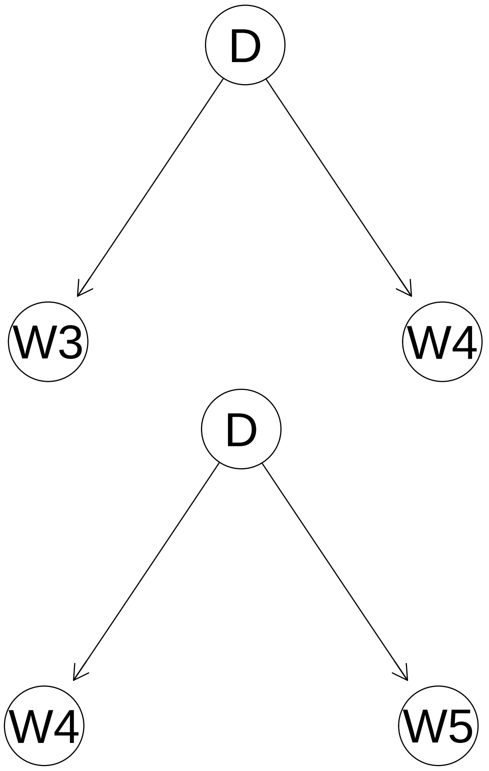


Figure 11: Two remaining DAGs for the Witness problem.

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).[5] Once this strategy is taken, the problem turns out to be not that challenging for any of the coherence measures under discussion.

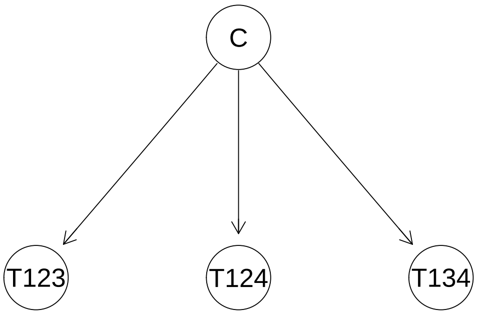
	OG	OGen	Sh	ShGen	Fit	DM	R	S
Witness: W1W2 11	0.451	0.451	3.551	3.551	0.771	0.446	0.446	0.729
Witness: W3W4 11	0.187	0.187	0.740	0.740	-0.234	-0.110	-0.110	0.494
Witness: W4W5 11	0.365	0.365	1.260	1.260	0.218	0.110	0.110	0.602

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Witness: $W_1W_2 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Witness: $W_4W_5 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

[5] 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.

6.8 Depth

We start with representing the two scenarios with two fairly natural BNs (C stands for who Committed the crime, TXYZ stands for Testimony that $X \vee Y \vee Z$), see Fig. 12 and 13.



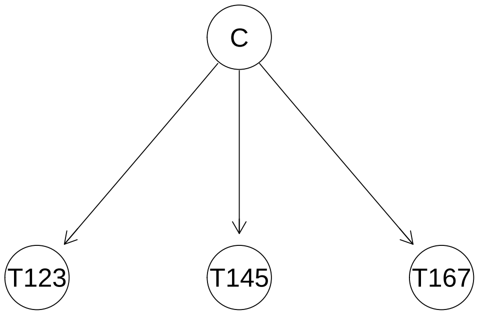
C	Pr
1	0.125
2	0.125
3	0.125
4	0.125
5	0.125
6	0.125
7	0.125
8	0.125

T123	C							
	1	2	3	4	5	6	7	8
1	1	1	1	0	0	0	0	0
0	0	0	0	1	1	1	1	1

T124	C							
	1	2	3	4	5	6	7	8
1	1	1	0	1	0	0	0	0
0	0	0	1	0	1	1	1	1

T134	C							
	1	2	3	4	5	6	7	8
1	1	0	1	1	0	0	0	0
0	0	1	0	0	1	1	1	1

Figure 12: BN for X1 in the Depth problem.



C	Pr
1	0.125
2	0.125
3	0.125
4	0.125
5	0.125
6	0.125
7	0.125
8	0.125

T123	C							
	1	2	3	4	5	6	7	8
1	1	1	1	0	0	0	0	0
0	0	0	0	1	1	1	1	1

T145	C							
	1	2	3	4	5	6	7	8
1	1	0	0	1	1	0	0	0
0	0	1	1	0	0	1	1	1

T167	C							
	1	2	3	4	5	6	7	8
1	1	0	0	0	0	1	1	0
0	0	1	1	1	1	0	0	1

Figure 13: BN for X2 in the Depth problem.

One effect of dropping the "the witness testified that" and using the testimony contents themselves is that the CPTs for the narration nodes are deterministically connected with the root node. In result, the coherence calculations give in the following:

	OG	OGen	Sh	ShGen	Fit	DM	R	S1	S2
Depth: T123T124T134 111	0.250	0.438	2.37	1.926	0.382	0.198	0.198	-0.25	1
Depth: T123T145T167 111	0.143	0.186	2.37	1.259	0.343	0.188	0.188	-0.25	1

	OG	OGen	Sh	ShGen	Fit	DM	R	S1	S2
Depth: $X_1 > X_2$	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE

Note that this time we listed two values for our measure. Structured 1 shows the values obtained if we do not update the weighting of the node not included in the narration, and Structured 2 is the result of such an updated weighing (analogous to the updating involved in the Witness problem). Now, what are we to make of this?

Structured 1 is negative. This isn't too surprising: after all, this is the coherence of the narration with the probabilistic assumption that the distribution for C is uniform, and this probabilistic assumption undermines the narration. Why, however, does Structured 2 equal 1, and why are the results identical for both narrations? This, upon reflection, isn't too suprising either. If the BN and the narration is supposed to represent a single agent's credal state, there is only one state of C in which the whole narration X_1 is true – trivially, it is the one in which suspect 1 is guilty, and it is the same unique state of C in which the whole narration X_2 is true. Since seen as narrations these sets have exactly the same truth conditions, there is no surprise in them being equally coherent.

What if the sentences in the set are not claims made by one agent and there is no single underlying credal state? We aren't convinced that our tool is optimal for measuring the agreement of multiple witnesses. Instead, there already exists a working measure of such an agreement — Cohen's κ – which already gives the desired results.

To illustrate, let's think of a simplified situation (devoid of three-dimensional tables) with two witnesses $w1$ and $w2$, where the respective sets are $A = \{1 \vee 2 \vee 3, 1 \vee 2 \vee 4\}$ and $B = \{1 \vee 2 \vee 3, 1 \vee 4 \vee 5\}$ and in each set the first proposition comes from $w1$ and the second from $w2$. The information for these two sets can be tabulated as follows:

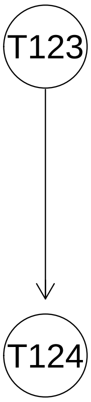
	w2: suspect	w2: innocent
w1:suspect	2	1
w1: innocent	1	4

	w2: suspect	w2: innocent
w1: suspect	1	2
w1: innocent	2	3

Standard calculations using the vcd package results in the following unweighted values of Cohen's κ.

	A	B
value	0.467	-0.067

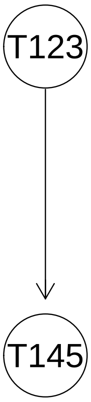
Let's further illustrate our point about the requirement that the BN should represent a single agent's cognitive state. For instance, you can represent, the situation in A from the perspective of the first witness. This suggests we should focus only on the nodes involved in the narration, and on the fact that from the witness' perspective the suspects are not equally likely. The example doesn't provide us enough information to build a table for C. In fact, no information about the wintess attitude towards this node is given, but given they say what they say, it's unlikely they think the distribution is uniform. So let's take one of the witness' own statements as the root (which ones we choose doesn't change the outcome). Clearly (or, at least, hopefully, if we talk about witnesses), the agent thinks her own claim is very likely and evaluates the probability of the other statements in A or B from its perspective. This gives us two different BNs, and when we calculate the respective coherences we actually do get the desired result, which isn't too hard for the other measures either.



T123	Pr
1	0.98
0	0.02

T124	T123	
	1	0
1	0.667	0.2
0	0.333	0.8

Figure 14: A witness perspective for the agreement problem, set *A*.



T123	Pr
1	0.98
0	0.02

T124	T123	
	1	0
1	0.333	0.4
0	0.667	0.6

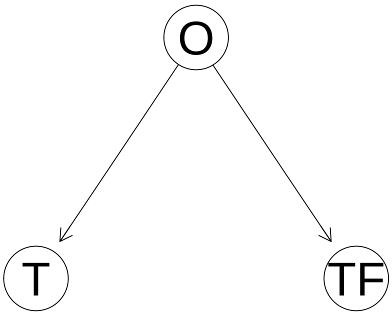
Figure 15: A witness perspective for the agreement problem, set *B*.

	OG	OGen	Sh	ShGen	Fit	DM	R	S
DepthA: T123T124 11	0.664	0.664	1.014	1.014	0.280	0.012	0.012	0.027
DepthB: T123T145 11	0.331	0.331	0.996	0.996	-0.047	-0.003	-0.003	-0.004

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Depth: $X_1 > X_2$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

6.9 Dice

We'll follow the strategy similar to the one we already used. Since neither the example nor the narrations involve information about how probable it is that we're dealing with a regular die, as opposed to a dodecahedron, we avoid using a node representing this. Moreover, if at a given time the agent claims that the result is both two and (two or four), their cognitive situation at that time cannot be represented using uniform distribution for possible toss outcomes. Instead, we start with initial separate BNs for a regular die and a dodecahedron which do have uniform distributions for the O (outcome) node (Fig. 16), but when weighing the antedecent nodes which are not strictly speaking part of the narration, we use the probabilities updated in light of the narration content itself.



O	Pr
1	0.167
2	0.167
3	0.167
4	0.167
5	0.167
6	0.167

(a) Root CPT for the regular die.

O	Pr
1	0.083
2	0.083
3	0.083
4	0.083
5	0.083
6	0.083
7	0.083
8	0.083
9	0.083
10	0.083
11	0.083
12	0.083

(b) Root CPT for the dodecahedron.

T	O					
	1	2	3	4	5	6
1	0	1	0	0	0	0
0	1	0	1	1	1	1

TF	O					
	1	2	3	4	5	6
1	0	1	0	1	0	0
0	1	0	1	0	1	1

(c) Conditional probabilities for the regular die.

T	O											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	0	0	0	0	0	0	0	0	0	0
0	1	0	1	1	1	1	1	1	1	1	1	1

TF	O											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	0	1	0	0	0	0	0	0	0	0
0	1	0	1	0	1	1	1	1	1	1	1	1

(d) Conditional probabilities for the dodecahedron.

Figure 16: BNs for the dice problem.

Calculation of coherences of the scenarios in the respective BNs yield the following result:

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Regular: TTF 11	0.5	0.5	3	3	0.833	0.500	0.500	1
Dodecahedron: TTF 11	0.5	0.5	6	6	0.917	0.625	0.625	1

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Dodecahedron: Regular = Dodecahedron	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

The measure that we think is appropriate here yields the same result, 1, for both situations. Come to think of it, we don't find this counterintuitive. These are two sentences one of which is a trivial consequence of the other.

7 Conclusions

Ultimately, all the coherence results and desiderata yield the following two tables and success rates:

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Penguins: BGP 111	0.010	0.015	4.000	2.010	0.453	0.255	0.255	0.010
Penguins: BG 11	0.010	0.010	0.040	0.040	-0.960	-0.480	-0.480	-0.960
Penguins: BP 11	0.020	0.020	2.000	2.000	0.669	0.255	0.255	0.010
Dunnit: MGWI 1111	0.000	0.087	4.294	11.012	0.169	0.167	0.167	-0.932
Dunnit: MTGWI 11111	0.000	0.042	73.836	13.669	0.385	0.150	0.150	-0.100
Japanese Swords 1: JO 11	0.004	0.004	80.251	80.251	0.976	0.008	0.008	0.008
Japanese Swords 2: JO 11	0.818	0.818	9.000	9.000	0.976	0.800	0.800	0.889
Japanese Swords 3: JO 11	0.818	0.818	1.080	1.080	0.286	0.067	0.067	0.400
Robbers: MIsPMIsR 11	0.600	0.600	0.937	0.937	-0.143	-0.050	-0.050	0.600
Robbers: MIsPMIsR 10	0.250	0.250	1.250	1.250	0.571	0.125	0.125	-0.600
Robbers: MIsPMIsR 01	0.250	0.250	1.250	1.250	0.571	0.125	0.125	-0.600
Beatles: JPGRD 11111	0.000	0.202	0.000	1.423	-0.036	0.025	0.025	-1.000
Books: AR 11	0.014	0.014	1.493	1.493	0.212	0.027	0.027	0.055
Books: AR 10	0.009	0.009	0.945	0.945	-0.127	-0.025	-0.025	-0.055
Books: AR 01	0.100	0.100	0.995	0.995	-0.101	-0.003	-0.003	-0.005
Books: AR 00	0.892	0.892	1.001	1.001	0.016	0.001	0.001	0.005
Witness: W1W2 11	0.451	0.451	3.551	3.551	0.771	0.446	0.446	0.729
Witness: W3W4 11	0.187	0.187	0.740	0.740	-0.234	-0.110	-0.110	0.494
Witness: W4W5 11	0.365	0.365	1.260	1.260	0.218	0.110	0.110	0.602
DepthA: T123T124 11	0.664	0.664	1.014	1.014	0.280	0.012	0.012	0.027
DepthB: T123T145 11	0.331	0.331	0.996	0.996	-0.047	-0.003	-0.003	-0.004
Regular: TTF 11	0.500	0.500	3.000	3.000	0.833	0.500	0.500	1.000
Dodecahedron: TTF 11	0.500	0.500	6.000	6.000	0.917	0.625	0.625	1.000

	OG	OGen	Sh	ShGen	Fit	DM	R	S
Penguins: BG<BGP	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Penguins: BP≈ BGP	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
Dunnit: Dunnit<Twin	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE
Swords: JO2>JO1	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Swords: JO2>JO3	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Robbers: PR>P¬R	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
Robbers: PR>neutral	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
Beatles: below neutral	NA	NA	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
Beatles: minimal	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
Books: AR>A¬R	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: AR> ¬AR	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: ¬A¬R>A¬R	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Books: ¬A¬R> ¬AR	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Witness: $W_1W_2 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Witness: $W_4W_5 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Depth: $X_1 > X_2$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Dodecahedron: Regular = Dodecahedron	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

OG	OGen	Sh	ShGen	Fit	DM	R	S
0.733	0.733	0.706	0.647	0.706	0.647	0.706	1

Let's recap what we've done. We introduced the most prominent coherence measures and a number of counterexamples put forward against them. Then, we pointed out some common problems they face. These observations helped us develop our own measure. It improves on the existing approaches by using the structure of BNs, and by doing something a bit more sophisticated than taking means. Finally, we argued that this way we managed to avoid many counterexamples that were problematic for other measures. This, in fact turned out to be a balancing act: we agreed with many intuitions behind the counterexamples had doubts about some of them, and a few of the cases needed somewhat more elaborate reflection before our measure gave the desired outcome. We end with a list of tasks for further reearch.

One issue that needs further study is whether the structured coherence measure yields desired results in more straightforward cases as compared with empirical results on how real agents assess coherence. Another question is how the measure handles legal cases for which BNs have already been developed (C. Vlek et al., 2013, C. Vlek et al. (2014), C. S. Vlek et al. (2015), C. Vlek (2016), Fenton et al. (2013), Fenton et al. (2013)). It might also be worthwile to investigate what happens if confirmation measures other than Z are plugged in. Finally, a more general study of the properties of the structured coherence measure would be useful.

References

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

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential Support: theoretical and empirical Issues. *Philosophy of Science*, *74*(2), 229–252. https://doi.org/10.1086/520779

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, *37*(1), 61–102. https://doi.org/10.1111/cogs.12004

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

Fitelson, B. (2003b). Comments on jim franklin's the representation of context: Ideas from artificial intelligence (or, more remarks on the contextuality of probability). *Law, Probability and Risk*, *2*(3), 201–204.

Glass, D. H. (2002). Coherence, Explanation, and Bayesian Networks. In G. Goos, J. Hartmanis, J. van Leeuwen, M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton, & N. J. L. Griffith (Eds.), *Artificial Intelligence and Cognitive Science* (Vol. 2464, pp. 177–182). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45750-X_23

Koscholke, J. (2016). Evaluating Test Cases for Probabilistic Measures of Coherence. *Erkenntnis*, *81*(1), 155–181. https://doi.org/10.1007/s10670-015-9734-1

Meijs, W., & Douven, I. (2007). On the alleged impossibility of coherence. *Synthese*, *157*(3), 347–360. https://doi.org/10.1007/s11229-006-9060-x

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. https://doi.org/10.1007/s10670-005-4007-z

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

Schippers, M., & Koscholke, J. (2019). A General Framework for Probabilistic Measures of Coherence. *Studia Logica*. https://doi.org/10.1007/s11225-019-09848-3

Schupbach, J. N. (2008). On the alleged impossibility of bayesian coherentism. *Philosophical Studies*, *141*(3), 323–331. https://doi.org/10.1007/s11098-007-9176-y

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

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

Vlek, C. (2016). *When stories and numbers meet in court: Constructing and explaining bayesian networks for criminal cases with scenarios*. Rijksuniversiteit Groningen.

Vlek, C. S., Prakken, H., Renooij, S., & Bart Verheij. (2015). Representing the quality of crime scenarios in a bayesian network. In A. Rotolo (Ed.), *Legal knowledge and information systems* (pp. 133–140). IOS Press.

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