Journal of Machine Learning Research 24 (2023) 1-56 Submitted 9/21; Revised 12/22; Published 5/23

# Clustering with Tangles: Algorithmic Framework and Theoretical Guarantees

**Solveig Klepper**[1] SOLVEIG.KLEPPER@UNI-TUEBINGEN.DE
**Christian Elbracht**[2] CHRISTIAN.ELBRACHT@UNI-HAMBURG.DE
**Diego Fioravanti** [1] FIORAVANTI.DIEGO@GMAIL.COM
**Jakob Kneip**[2] JAKOBFKNEIP+JMLR@GMAIL.COM
**Luca Rendsburg**[1] LUCA.RENDSBURG@UNI-TUEBINGEN.DE
**Maximilian Teegen**[2] MAXIMILIAN.TEEGEN@UNI-HAMBURG.DE
**Ulrike von Luxburg**[1] ULRIKE.LUXBURG@UNI-TUEBINGEN.DE

[1] *Department of Computer Science and Tübingen AI Center, University of Tübingen, Germany*

[2] *Department of Mathematics University of Hamburg Germany*

**Editor:** Samory Kpotufe

## Abstract

Originally, tangles were invented as an abstract tool in mathematical graph theory to prove the famous graph minor theorem. In this paper, we showcase the practical potential of tangles in machine learning applications. Given a collection of cuts of any dataset, tangles aggregate these cuts to point in the direction of a dense structure. As a result, a cluster is softly characterized by a set of consistent pointers. This highly flexible approach can solve clustering problems in various setups, ranging from questionnaires over community detection in graphs to clustering points in metric spaces. The output of our proposed framework is hierarchical and induces the notion of a soft dendrogram, which can help explore the cluster structure of a dataset. The computational complexity of aggregating the cuts is linear in the number of data points. Thus the bottleneck of the tangle approach is to generate the cuts, for which simple and fast algorithms form a sufficient basis. In our paper we construct the algorithmic framework for clustering with tangles, prove theoretical guarantees in various settings, and provide extensive simulations and use cases. Python code is available on github.

**Keywords:** tangles, clustering framework, soft clustering, hierarchical clustering, interpretable clustering

## 1. Introduction

In this paper, we present tangles, a new tool that can be used for clustering, to the machine learning community. Tangles are an established concept in mathematical graph theory. They were initially introduced by Robertson and Seymour (1991) as a mechanism to study highly cohesive structures in graphs and have since become a standard tool in the analysis of other discrete structures (Diestel, 2018). Recently, Diestel (2019) suggested applying the

abstract notion of tangles beyond their original context to data clustering problems. The purpose of our paper is to make this suggestion come true. We translate abstract mathematical notions into practical algorithms, prove theoretical guarantees for the performance of these algorithms, and demonstrate the usefulness and flexibility of the new approach in diverse applications.

The mechanism of tangles is very different from all of the current clustering algorithms we know. To introduce this concept, we consider the example of a personality traits questionnaire, in which a group of persons answers a set of binary questions. Based on the answers, we would like to identify groups of like-minded persons and characterize their associated mindsets, such as being "narcissistic". One would expect that persons sharing a mindset agree on many relevant statements; for example, most narcissists would agree on the statement "I have a strong will to power". Accordingly, we would like to *softly characterize* a mindset by saying that most persons with this mindset answer similarly to most questions. We can formalize this idea using tangles. First, we interpret every question as a bipartition of all the persons who participated in the questionnaire. This bipartition (equivalently, cut) splits the set of persons into the ones answering "yes" versus the ones answering "no". Let us assume that most persons who share a mindset give the same answers to most questions. Visualized in terms of cuts, we can say that persons of the same mindset tend to lie "on the same side" of most of the cuts. We now assign an orientation to each of the cuts to identify one side of the respective bipartition: we orient the cut to "point towards" the group of persons. Assume for the moment that we already know the mindset that we want to describe. The description then consists of the chosen orientations, indicating the "typical way" of answering all the questions. Conversely, the orientations of all the cuts identify a group of persons: the persons that the cuts point towards. See Figure 1.
More generally, the tangle framework is as follows. Given a dataset, in the first step, we construct a set of bipartitions of the data. These cuts can be constructed in a quick and dirty manner; all we need is that they provide a little information regarding the cluster structure of the points. In a second step, we then find "consistent" orientations of these cuts. Typically, there will be several consistent orientations. Each of them is one particular "tangle" of the data. In a final step, the tangles can then be converted into meaningful output, for example, a hard or soft clustering of the dataset or even a soft dendrogram (see Figure 2).

**What are the benefits of this approach?** The tangle approach is very general and highly flexible. Instead of assigning cluster memberships to individual objects, tangles characterize a cluster indirectly by a set of pointers. This flexible representation mitigates the problem of dealing with ambiguous cases and naturally entails a hierarchical structure. Tangles require as input only a collection of cuts of the dataset. When choosing these cuts, we can incorporate prior knowledge that we might have about our problem. We do not require a particular data representation. Quite the contrary: tangles can be applied to many different scenarios such as feature-based data, metric data, graph data, and questionnaire data. For exemplary use cases see Sections 4, Section 5 and Section 6. From a conceptual point of view, tangles resemble the boosting approach for classification, where one aggregates many weak classifiers – slightly better than chance – to obtain a strong classifier.

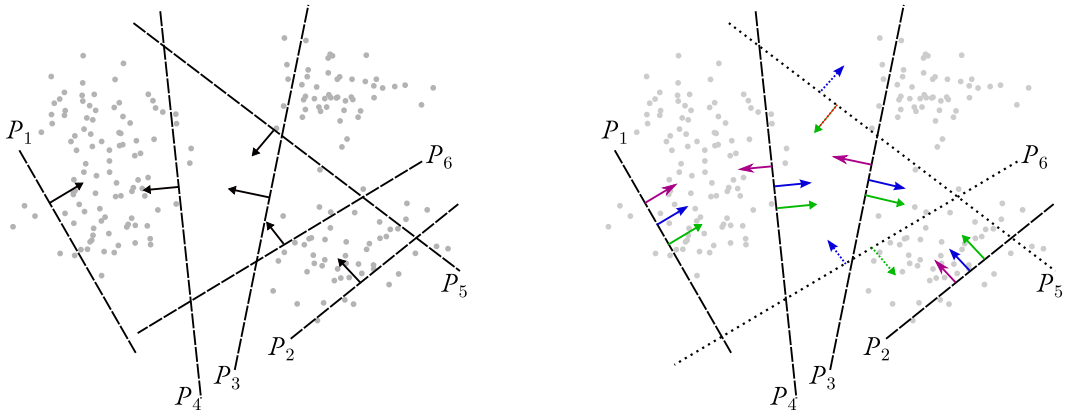



Figure 1: We consider a set of points and six cuts. The left image visualizes one possible tangle (consistent orientation). The right image visualizes three additional tangles, that exist on (sub)sets of the same cuts. When constructing the tangle search tree, we would first obtain two tangles on the set $\{P_1, ..., P_4\}$: the purple and the green/blue tangle. The green and the blue tangle share the orientations of the more informative cuts but differ in bipartition $P_5$ and $P_6$, indicated by dashed arrows. Lower down in the hierarchy, we get three tangles on the whole set of cuts $\{P_1, ..., P_6\}$: green, blue and the black tangle visualized in the left picture.

Tangles aggregate many "weak" cuts that contain a large chunk of a cluster on one side to obtain a holistic, "strong" view of the cluster structure of a dataset. The computational complexity of the tangle approach is composed of two parts: constructing cuts in the pre-processing phase and orienting the cuts in the central part of the algorithm. This central part of the algorithm is only linear in the number of data points. That means that given a simple way of constructing a set of cuts in the pre-processing phase, the whole approach is fast and works for large-scale datasets.

Our contributions are as follows:

- **Algorithmic framework.** We translate the abstract notion of tangles from the mathematical literature to a more practical version for machine learning in Section 2. We then develop a highly flexible algorithmic framework for clustering. We propose a basic version in Section 3 and refer to Appendix II for further extensions and details.

- **Simulations and experiments.** To demonstrate the flexibility of the tangle approach, we provide case studies in three different scenarios: a questionnaire scenario in Section 4, a graph clustering scenario in Section 5, and a feature-based scenario in Section 6. In each of these sections, we outline different properties of the tangle approach. Generally, we compare tangles to other state-of-the-art algorithms in the respective domains, for example, spectral clustering in the graph clustering domain or $k$-means in the feature-based domain.

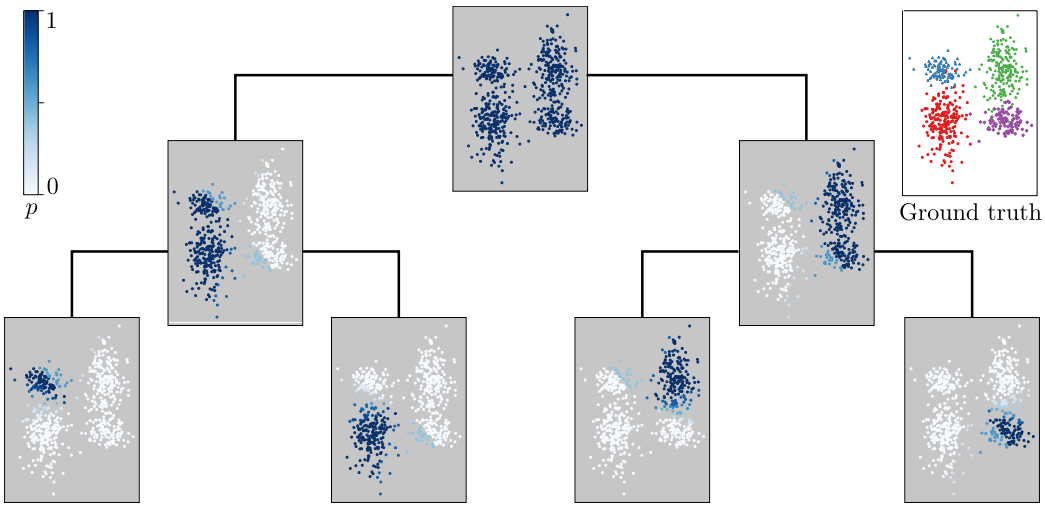



Figure 2: A soft dendrogram as possible post-processing of tangles (Appendix 3.4). The estimated probability that a point belongs to the respective cluster is given by $p$.

- **Theoretical guarantees.** In each of the three scenarios, we prove theoretical guarantees. Given a statistical model for the questionnaire setting, we prove that tangles always discover the ground truth under specific parameter choices. We prove the same for the graph clustering scenario in a stochastic block model. Finally, we investigate theoretical guarantees on feature-based clustering for interpretable clustering.
- **Python package.** We implemented the central part of the algorithm and different options for pre- and post-processing. The code and basic examples are publicly available at: https://github.com/tml-tuebingen/tangles/tree/vanilla.

The strength of tangles is not that they outperform all other algorithms; this would be pretty unrealistic. Instead, we are intrigued by how flexible and how generic the tangle approach turns out to be, while at the same time producing results that are comparable to many state-of-the-art algorithms in many domains. All in all, we consider this paper as a proof of concept for a completely new approach to data clustering.

## 2. Tangles: Notation and definitions

Tangles originate in mathematical graph theory, where they are treated in much more generality than what we need in our paper (cf. Diestel (2018) for an overview, and Section 7 for more discussion and pointers to literature). Through our joint effort between mathematicians and machine learners, we condensed the general tangle theory to what we believe is the essence of tangles needed for applications in machine learning. We present this condensed version below. For readers with a mathematical tangle theory background, we provide a translation dictionary of the essential terms in Appendix I.1.

Consider a set $V = \{v_1 \dots, v_n\}$ of arbitrary objects. A subset $A \subset V$ induces a **bipartition** or **cut** of the data into the set and its complement $P = \{A, A^{\complement}\}$. In order to construct tangles, we will consider a set of initial cuts $\mathcal{P} = \{\{A_1, A_1^{\complement}\}, \dots, \{A_m, A_m^{\complement}\}\}$. We consider

a single cut useful if it does not separate many similar objects. The more it cuts through dense regions, the less insight we get into the cluster structure. This intuition is being quantified in terms of a **cost function** $c : \mathcal{P} \to \mathbb{R}$, indicating the "quality" of a cut. This cost function needs to be chosen application-dependent; see later for examples. The set of bipartitions and associated costs hold all the necessary information for tangles to discover the dataset's structure. Tangles operate by assigning an **orientation** to all cuts. For a single cut $P = \{A, A^{\complement}\}$, an orientation simply "points" towards one of the sides. We denote the orientation pointing from $A$ to $A^{\complement}$ by $\vec{P} = (A, A^{\complement})$ or simply $\vec{P} = A^{\complement}$. For a set of cuts $\mathcal{P}$, we define an orientation $O_{\mathcal{P}}$ by choosing one side for each cut, giving an orientation to every $P \in \mathcal{P}$. We write $A \in O_{\mathcal{P}}$ if $\{A, A^{\complement}\} \in \mathcal{P}$ and $O_{\mathcal{P}}$ orients it towards $A$. The intuition is that orientations can characterize clusters, but not every orientation of cuts characterizes a cluster: the orientations need to be "consistent" in some way. For a meaningful orientation, we have to ensure that the chosen sides of all the cuts point to one single structure. This consistency is precisely the purpose of tangles and is captured in the following definition.

**Definition 1 (Consistency and Tangles)** *Let $\mathcal{P}$ be a set of bipartitions on a set $V$. For a fixed parameter $a \in \mathbb{N}$, an orientation $O_{\mathcal{P}}$ of $\mathcal{P}$ is consistent if all sets of three of oriented cuts have at least $a$ objects in common:*

$$\forall A, B, C \in O_{\mathcal{P}} : \quad |A \cap B \cap C| \geqslant a \,. \tag{1}$$

*We call Eq.* (1) *the* consistency condition *and $a$ the* agreement parameter*. A consistent orientation of $\mathcal{P}$ is called a* $\mathcal{P}$-tangle*. If clear from the context, we drop the dependency on $\mathcal{P}$ and say* tangle*.*

Note that neither the definition of an orientation nor the definition of a tangle is symmetric. Choosing the other set of each bipartition, i.e. the inverted orientation will point to a different set of points and will not correspond to the same structure in the data.

At this point, the reader might wonder why we consider an intersection of exactly three cuts in Eq. (1). The short answer is that there are good mathematical reasons for this choice. One can prove that considering the intersection of at least three cuts guarantees that there exist at most as many distinct tangles as there are data points — which makes perfect sense in the application of data clustering. If one uses the intersection of only two cuts in Eq. (1), then there might be up to $2^{2|V|}$ many tangles, which is undesirable both from a conceptual as well as a computational point of view. On the other hand, it turns out that choosing sets of more than three in Eq. (1) does not produce more powerful mathematical results, but considerably increases the computational complexity. We explain more details about the question of three in Appendix I.2.

In what follows, we will often sort the cuts according to their cost $c(P)$, and start orienting the (more useful) low-cost cuts before moving on to the (less useful) high-cost cuts. To this end, we sometimes introduce a parameter $\Psi \in \mathbb{R}$ that specifies the set of cuts we are interested in, namely the subset $\mathcal{P}_\Psi = \{P \in \mathcal{P} \mid c(P) \leqslant \Psi\} \subseteq \mathcal{P}$. We say a tangle on the set $\mathcal{P}_\Psi$ is of order $\Psi$.

In the following, we build intuition on the above definitions using the running example of the questionnaire that we already hinted in the introduction.

**Example (Questionnaire)** A set $V$ of $n$ persons takes a questionnaire of $m$ binary questions. The goal is to discover groups of persons who answer most questions similarly. If such a group exists, we say they share the same mindset. We interpret each question as a cut in $V$, separating persons based on their answer to this question. In this way, the questionnaire defines a set $\mathcal{P}$ of $m$ cuts. Generally, the cuts in $\mathcal{P}$ can split $V$ at different levels of granularity, depending on how general a question is. Some of the cuts might not be informative; for example, a person's hair color is mostly independent of other personality traits. To judge on an abstract level how useful a cut might be, we introduce a cost function $c\colon \mathcal{P} \to \mathbb{R}$. In our example, we judge the similarity of two persons, or rather their answered questionnaires $v, w \in \{0,1\}^m$, by counting how many questions they have answered the same way: $sim(v,w) = \sum_{i=1}^{m} \mathbb{1}\{v_i = w_i\}$. We then define the cost of a cut as the mean over the similarities over all possible pairs of separated persons: $c(A) = \frac{1}{|A|\cdot(n-|A|)} \sum_{v\in A, w\in A^\complement} \mathrm{sim}(v,w)$.
We now process the cuts in increasing order of costs: most useful cuts come first, and less useful cuts come later. This approach is equivalent to repeatedly setting the threshold $\Psi$ and restricting our attention to the set $\mathcal{P}_\Psi$. Increasing the order $\Psi$ enables us to discover a hierarchy of substructures. For a small order, we can only distinguish between coarse structures (such as extroverts and introverts), while for a larger order, we include cuts that further separate them into more fine-grained structures. For any given $\Psi$, we need to find an orientation of the cuts in $\mathcal{P}_\Psi$ that "points towards a cluster", as formalized in Definition 1. Concretely, we need to set the agreement parameter $a$ and invoke an algorithm that discovers consistent orientations of all orders of cuts. Once we have found consistent orientations, we need to post-process them to the final output. This output could consist of a description of all mindsets in terms of the typical way of answering questions; it could be a hard clustering of the persons or a soft hierarchical clustering. We will introduce the algorithm and all these notions in the next section.

## 3. Basic algorithms for tangle clustering

In this section, we present the basic algorithmic framework for clustering with tangles. On a high level, this requires the following three independent steps: finding the initial set of cuts (Section 3.1), orienting cuts to identify tangles (Section 3.3), and post-processing tangles to clusterings (Section 3.4). To allow for a deeper understanding of the framework we give intuition on parameters and how they interact in Sections 3.2 and 3.5. For more algorithmic details we refer to Appendix II. In Sections 4 – 6 we will then spell out all details in three different application settings. Python code of the basic version as well as examples can be found on github [1].

1. https://github.com/tml-tuebingen/tangles/tree/vanilla

### 3.1 Constructing the initial set of cuts

The first step for finding tangles is to construct a set of initial cuts $\mathcal{P}$. This construction is very much problem-dependent, and in our pipeline for finding tangles, it has the flavor of a pre-processing step. We can distinguish two principal scenarios that occur in different types of applications:

**Predefined cuts.** In our running example of a questionnaire, each question induces a natural cut of the data space: the persons who answered "yes" versus those who answered "no". The set of the cuts induced by all questions is a natural candidate for the desired set $\mathcal{P}$. In this case, we can interpret tangles as a typical way of answering the questionnaire. More generally, if the objects in $V$ are described by discrete, continuous, or ordinal features, we can consider a collection of half-spaces of the form $\{x_i \leqslant k\}$. The resulting cuts (and consequently, the tangles) have a simple form and result in interpretable output. See Section 4.1 for an example of interpretable clustering.

**Cuts by simple pre-processing.** If no natural choice for cuts exists or if they are not flexible enough, it is necessary to invoke another algorithm that produces the initial cuts in a pre-processing phase. In this case, we can view tangles as a boosting mechanism that allows us to use a fast, greedy heuristic for producing decent cuts, which then get aggregated to a tangle and can be processed to clustering. One example of such a setting is graph clustering. Here we could construct initial cuts by the Kernighan-Lin (KL) algorithm (Kernighan and Lin, 1970) and then use tangles to infer the cluster structure on the graph. The complexity of this approach is $\mathcal{O}(rn^2 log\ n)$, for $n$ nodes and $r$ iterations of the KL-Algorithm. Another example is clustering in Euclidean spaces, where we can quickly construct initial partitions with the help of random projections in a one-dimensional subspace. The complexity of this approach is $\mathcal{O}(n^2)$.

Below, we will study cut-finding strategies in three specific settings: binary questionnaires (Section 4), graphs (Section 5) and metric/feature data (Section 6). In Section 5.2 and 6.2 we additionally review the influence and trade-off between a large versus a small set of initial cuts, and in Appendix III.2.1 we discuss why purely random initial cuts are not a good idea.

### 3.2 Setting the key parameters

Once we have fixed a set of partitions, we need to find consistent orientations of these partitions, that is, the tangles. This first requires some parameter choices: we need to define a cost function $c$ of the cuts (to be able to order the cuts according to their usefulness) and choose the agreement parameter $a$ (which is related to the size and the number of clusters we expect to find). A natural choice for the cost function is the sum of similarities between separated objects $c(\{A, A^\complement\}) = \sum_{v\in A, w\in A^\complement} \mathrm{sim}(v, w)$. We often also normalize this cost function by dividing it by the number of pairs $|A|\,(n - |A|)$. We discuss the influence of normalizing in Appendix II.2.2. The agreement parameter $a$ roughly fixes the smallest size of the clusters that tangles discover. See Section 3.5 for a discussion of all parameters.

### 3.3 Orienting cuts to identify tangles

Once we have the data and fixed all parameters, we face the following algorithmic challenge:

| Given a set of initial cuts $\mathcal{P}$ of $V$ and a cost function, for every $\Psi$ identify all orientations of $\mathcal{P}_\Psi$ that satisfy the consistency condition Eq. (1). |
|---|

The naive approach of testing every possible orientation for consistency is infeasible. Instead, we are now going to construct a tree-based search algorithm that achieves this task more efficiently. The algorithm proceeds by looking at one cut after the other, starting with the lowest cost cuts. It maintains a tree, the *tangle search tree*, of the possible orientations of all the cuts considered. The critical observation is that processing a tree branch can be stopped once a cut cannot be oriented consistently any more.

The algorithm's output is a labeled binary tree as depicted in Figure 3b. Each node in the tree corresponds to one specific orientation of one particular cut. We construct the tree in such a way that each of its nodes corresponds to exactly one tangle. Precisely, for a node $t$ on level $i$ the node labels on the path from the root to $t$ form a consistent orientation of $\{P_1, \dots, P_i\}$, that is, a $\mathcal{P}_\Psi$-tangle for $\Psi = c(P_i)$.

The tangle search tree algorithm proceeds as follows. We first sort all cuts in $\mathcal{P}$ by increasing cost and list them as $P_1 = \{A_1, A_1^\complement\}, \dots, P_m = \{A_m, A_m^\complement\}$. We now perform something like a breadth-first search on possible orientations. We initialize the tree with an unlabeled root on level 0. We now iterate over the $P_i$. In the $i$-th step, for both sides $A \in \{A_i, A_i^\complement\}$ and every node $t$ on level $i-1$, we check whether adding the orientation $A$ to the tangle identified with node $t$ is consistent. If it is, we add a child node to $t$, labeled with $A$ (see Algorithm 1 and Appendix II.1). In the resulting tree, each node represents a tangle, and each leaf represents a maximal tangle, one that cannot be extended to a tangle of a larger set $\mathcal{P}_\Psi$.

The algorithm has complexity $O(n\ell h^3)$ where $n$ is the number of objects in our dataset, $h$ is the height of the tangle search tree, and $\ell$ is the number of its leaf nodes. The number of leaf nodes is bound by the number of nodes or the height of the tree: $\ell \leqslant n$ and $\ell \leqslant 2^h$; usually we observe $\ell \ll n$. In practice, we find that the worst-case complexity is rarely attained (Figure 15). The height $h$ is upper bounded by the number of cuts $m$. The tangles at the leaf nodes correspond to the smallest clusters. The agreement parameter $a$ indirectly controls both $\ell$ and $h$. Increasing $a$ makes the Eq. (1) more restrictive and thus cuts the tree quicker.

### 3.4 Post-processing the tangles into soft or hard clusterings

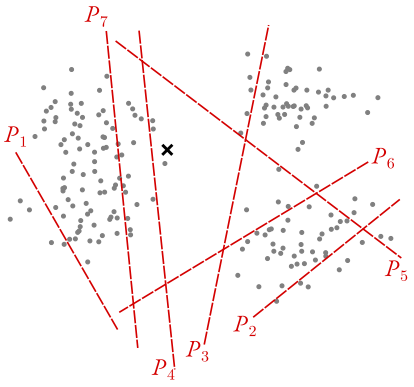
The output of Algorithm 1 is a tangle search tree, which reveals the cluster structure of a dataset from the cut point of view. Strictly speaking, it is inappropriate to think of tangles as subsets; instead, they "point towards a region" without making statements about individual objects. Nevertheless, traditional clustering objectives are concerned with assigning individual objects to clusters. In order to achieve this with tangles, we post-process the tangle search tree in different ways resulting in hierarchical, soft, and hard

**Algorithm 1:** tangle search tree

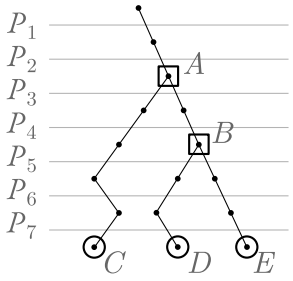
**Data:** Set of cuts $\mathcal{P} = \{P_i = \{A_i, A_i^{\complement}\}\}_i$ with cost function $c : \mathcal{P} \to \mathbb{R}$, agreement parameter $a$
**Result:** Tangle Search Tree $T$

$T \leftarrow$ empty tree with root;
sort $P_i$ increasing according to $\Psi_i = c(P_i)$;
**for** $P_i \in \mathcal{P}$ **do**
  **for** *tangle* $\tau \in$ *nodes of layer* $i$ *of* $T$ **do**
    **if** *consistent(* $\tau \cup \{A_i\}$*)* **then**
      add $A_i$ as right child of $\tau$ to $T$ ;
    **end**
    **if** *consistent(* $\tau \cup \{A_i^{\complement}\}$*)* **then**
      add $A_i^{\complement}$ as left child of $\tau$ to $T$ ;
    **end**
  **end**
**end**
**return** $T$

clustering. To this end, we propose a procedure that builds on the hierarchical nature of the tangle search tree to convert it into a "soft dendrogram".

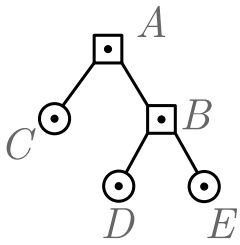



(a) Dataset $V$ and cuts $\mathcal{P} = \{P_1, \ldots, P_7\}$.




• tangle
⊡ splitting tangle
⊙ maximal tangle

(b) Tangle search tree $T(\mathcal{P})$.




(c) Condensed tangle search tree $T^*(\mathcal{P})$.

Figure 3: Example of tangles for a dataset in $\mathbb{R}^2$.

For a given set of partitions $\mathcal{P} = \{P_1, \ldots, P_m\}$ sorted increasingly by their costs $c(P_i)$, let $T = T(\mathcal{P})$ be the corresponding tangle search tree obtained from Algorithm 1, see Figure 3a and 3b for an example. The tangle search tree is constructed hierarchically on the cuts, which serves as a proxy for what we are eventually interested in, namely, a hierarchy of the cluster structure of the objects. As a further simplification, we will explain below how to transform the tangle search tree into a simplified, condensed tree. Like a dendrogram, the condensed tree $T^*$ indicates how a dataset organizes into substructures. We call every internal node a splitting tangle as its two subtrees correspond to tangles that point to dif-

ferent regions and thus split the data. However, for a single object, a splitting tangle does not induce a binary decision as to whether the object belongs to the left or right branch. Instead, we will assign a probability for belonging to a specific tangle for every node and tangle.

**Contracting the tree.** We first condense the tree to the splitting tangles and ignore bipartitions that do not give information about the cluster structure, for example, $P_1$. For every splitting tangle, we identify the cuts responsible for the split and thus 'characterizing' for separating the two dense structures. The intuition becomes clear from Figure 3a: For the first splitting tangle $\tau$ at the node $A$, $\tau_{|A}$, we see that the set of cuts $\{P_3, P_4, P_7\}$ gives information about the separation between the left and the right structure $\mathcal{P}(\tau_{|A}) = \{P_3, P_4, P_7\}$. For the splitting tangle $\tau_{|B}$, we get $\mathcal{P}(\tau_{|B}) = \{P_5, P_6\}$, separating the upper from the lower structure on the right side.
We derive this information from the tangle search tree as follows. For a cut $P$ to be characterizing for a splitting tangle $\tau$, we require every tangle corresponding to a leaf in one subtree to orient $P$ one way and every tangle corresponding to a leaf in the other subtree to orient $P$ the other way. Considering the splitting tangle at node A, $P_7$ is characterizing: it is oriented to the left in all paths in the left subtree and to the right in all paths in the right subtree. The same holds for cuts $P_3$ and $P_4$. In contrast, the cut $P_6$ is not characterizing as it is oriented both; to the left and right side within the right subtree. So $P_6 \notin \tau_{[\mathcal{B}]}$. In this sense, the cuts in $\mathcal{P}(\tau)$ are the ones that help in distinguishing between the subtrees of $\tau$. More formally, let $P(\tau)$ be the orientation of $P$ in a tangle $\tau$ and let $T_\tau^{(left)}$ be the left subtree and $T_\tau^{(right)}$ be the right subtree of the node at a tangle $\tau$. Then we define the set of characterizing cuts as

$$\mathcal{P}(\tau) := \{P \in \mathcal{P} \mid \forall \text{ leaves } \tau^l \in T_\tau^{(left)}, \text{ leaves } \tau^r \in T_\tau^{(right)} : \ P(\tau^l) \neq P(\tau^r)\}.$$

Based on this information, we condense the tree as shown in Figure 3c and track the set of characterizing cuts for each of the splitting tangles.

**Computing the soft clustering.**
We now use these cuts to determine how likely an object $v \in V$ belongs to the right subtree of $\tau$ (the left subtree is then implicit, so we focus on the right side). We chose the set such that all cuts in $\mathcal{P}(\tau)$ serve the same purpose of subdividing $\tau$ into two substructures. For every point $v$ and every splitting tangle $\tau$, we compute the fraction of characterizing cuts oriented towards the point $v$ by the overall number of characterizing cuts $|\mathcal{P}(\tau)|$. As not all cuts are equally fundamental as measured by their costs $c(P)$, we include a weighting of the cuts with a non-increasing function $h : \mathbb{R} \to \mathbb{R}, P \mapsto e^{-c(P)}$. $\{P \in \mathcal{P}(\tau) \mid v \in P^{(right)}\}$ is

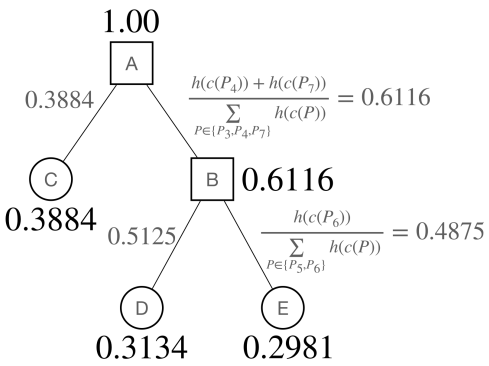



Figure 4: Tree $T^*$ with node and edge attributes for a fixed object $v_x \in V$.

the set of characterizing cuts that are oriented towards $v$ in the right side of the tree. We assign a probability $p_\tau^{(right)}$ of belonging to the right subtree at a tangle $\tau$ to every node $v$:

$$p_\tau^{(\text{right})}(v) = \frac{\sum_{P \in \{P \in \mathcal{P}(\tau) \mid v \in P^{(\text{right})}\}} h(c(P))}{\sum_{P \in \mathcal{P}(\tau)} h(c(P))}\,. \tag{2}$$

Based on these probabilities, we define the probability $p_t(v)$ that $v$ arrives at node $t$ as the product of the edge probabilities along the unique path from the root to $\tau$. For a single point $v_x$ marked with an $x$ in Figure 3a and the given characterizing cuts we get the tree $T^*$ with the probabilities as shown in Figure 4.

**Computing the hard clustering.** If desired, we can now assign points to a hard clustering: We assign each point to the tangle with the highest probability based on the soft clustering. For example, the point whose tree is shown in Figure 4 would be assigned to the tangle C represented by the leftmost leaf in the tree. The result is a hard clustering.

In our experiments, we sometimes apply heuristics, such as pruning "bad" branches of the tree, to avoid spurious tangles. We discuss algorithm improvements in Appendix II.2.1.

### 3.5 The ingredients and how they influence the output

**Initial cuts.** The utility of tangles depends strongly on the initial set of cuts $\mathcal{P}$ because the tangles' contribution is to aggregate information that is present in $\mathcal{P}$. If there is no cut in $\mathcal{P}$ that separates meaningful substructures, neither will the tangles. The better the cuts, the better the clustering we can derive from tangles. However, choosing the set of cuts is not as critical as one might think for two reasons: (1) Useless cuts do not interfere with tangles: very unbalanced cuts, such as $P = \{\{v\}, V \backslash \{v\}\}$, always get oriented towards their larger side and have little impact on the consistency condition; meaningless cuts, such as random cuts, have high costs and are considered last, quickly resulting in inconsistent orientations only. (2) It is not necessary that $\mathcal{P}$ contains high-quality cuts — otherwise, the whole approach would be somewhat pointless. It is enough to have some "reasonable" initial cuts. We will demonstrate this in experiments and partly in theory below and in the appendix.

**The parameter $\Psi$.** The parameter $\Psi$ controls the granularity of the tangles. Restricting our attention to a subset $\mathcal{P}_\Psi$ with a small parameter $\Psi$ will identify large subgroups in the data. As $\Psi$ increases, the corresponding $\mathcal{P}_\Psi$-tangles can identify smaller, less separate clusters, but at the same time, orientations towards larger, more separated clusters may become inconsistent. Eventually, when $\Psi$ gets too large, we might not find any consistent orientation anymore. We typically do not set $\Psi$ to a fixed value but generate a whole hierarchy of clusterings for increasing values of $\Psi$, as described in Section 3.3.

**Agreement parameter $a$.** The agreement parameter $a$ controls the minimal degree to which the sides of an orientation have to agree. When chosen too small, the consistency condition induced by $a$ may be too weak so that tangles identify substructures that we would not consider cohesive. On the other hand, we should not choose $a$ larger than the

smallest cluster we want to discover. Indeed, in practice, $a$ should be slightly smaller than the smallest cluster to allow for noise. The more the cuts in $\mathcal{P}$ respect the cluster structure and especially the richer the set $\mathcal{P}$ is, the more we can reduce $a$ without erroneously identifying incohesive structures as tangles.

## 4. Use Case: Binary Questionnaire

The most intuitive application for tangles is data coming from a binary questionnaire. In the following, we will give a better intuition about the different aspects of tangles in this practical setting using a simple real-world dataset.

### 4.1 Case study

As a simple instance, we chose the Narcissistic Personality Inventory questionnaire (Raskin, 1988), sometimes abbreviated npi in the following. Raskin and Hall developed the test in 1979, and it since then has become one of the most widely utilized personality measures for non-clinical levels of the trait narcissism. The dataset is accessible via `https://openpsychometrics.org/_rawdata/` and contains 40 binary questions answered by 11243 participants. Each question consists of a pair of statements, for example, "I am not sure if I would make a good leader" vs. "I see myself as a good leader". See Appendix III.3 for the full list of questions. Every participant is asked to choose the option that they most identify with. If a participant identifies with both equally, they should choose which statement is more important in their opinion. The developers handcrafted an evaluation score for the dataset: For every pair of statements, one statement gets assigned a score of 0, and the other one a score of 1. Each participant's final score $s_{npi}$ is defined as the sum of the scores of the answers, resulting in a number between 0 and 40. The higher the score $s_{npi}$, the more narcissistic a person is assumed to be. Figure 5 visualizes the frequencies of the participants over the score $s_{npi}$. We consider $s_{npi}$ as the baseline in the following.
For our experiments, we use each question as a natural bipartition of the persons $V$ into two sets $\{A, A^{\complement}\}$ where $A \in V$ is the set of persons choosing the first statement. This approach gives us one bipartition for each question, resulting in 40 cuts. To measure the similarity of two participants, we use the Hamming similarity between to answered questionnaires $u, v \in \{0,1\}^m$

$$s(u,v) = \sum_{i=1}^{m} \mathbb{1}\{u_i = v_i\}. \tag{3}$$

To assign a cost to a bipartition $\{A, A^c\}$ we then average this similarity over all pairs of persons of complementary sets:

$$c(\{A, A^{\complement}\}) = \frac{1}{|A| \cdot |A^{\complement}|} \sum_{u \in A, v \in A^{\complement}} s(u,v) \tag{4}$$

From Figure 5, it is evident that this data does not reveal a clear cluster structure when we only consider the score $s_{npi}$.

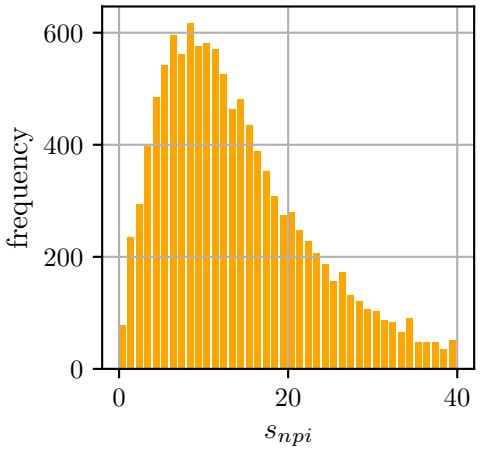



Figure 5: Frequencies of the hand-designed score $s_{npi}$ in the dataset.

Now we study the dataset from the tangle point of view. We naively apply tangles to the whole dataset without pre-processing: we use the data of all 11243 participants, consider all 40 questions as bipartitions and choose a small $a$ of 1500. We use the algorithm described in Section 3.3 to generate the tangle search tree. To avoid clustering on noise, we prune paths of length one, as described in Section 3.4.

The tangle algorithm returns exactly one tangle $\tau$ in this setting. This outcome is what we would expect from Figure 5, which already hints that the dataset does not contain a coarse cluster structure. The tangle $\tau$ orients 39 of the 40 questions towards one dense structure, and only one question does not get assigned an orientation by $\tau$ (question #1). The orientations specified in $\tau$ represent the "stereotypical way" by which persons of the corresponding mindset answer all the questions. Recap that this does not necessarily mean that a person in the dataset answered all the questions precisely this way. When we compare these orientations to the hand-crafted orientations by the inventors of the study, we find that $\tau$ discovers the "correct" assignments to all 39 questions! This outcome is remarkable: while the original study hand-designed the orientation of the questions (that is, which statement is 1 and which is 0), our algorithm discovers these orientations on its own. The only difference is that $\tau$ inverts all orientations: it points toward the larger group of people, which is the group of non-narcissistic persons, while in the original study, the authors oriented the question to point toward the minority group, the narcissistic people. So our first finding is that tangles reveal the same information as the authors hand-crafted into the data, but in a completely unsupervised manner.

We can now try to improve these results. Which questions are most important, and are there questions that we do not need to consider? We run a second experiment to demonstrate how tangles distinguish between different clusters. Based on the discovered tangle $\tau$, we assign a score $s_\tau$ to each of the participants: For every participant $u \in V$, we compute the Hamming distance of her answers $q_i$ to the stereotype answers $\tau_i$ given by tangle $\tau$: $s_\tau(u) = \sum_{i=1}^{m} \mathbb{1}\{q_i \neq \tau_i\}$. This score measures how much a participant's answers deviate

from the typical non-narcissistic person. $s_\tau$ takes values between 0 and 39, and the higher the score, the more narcissistic we believe a person to be. As expected by the fact that the tangle orientation essentially coincides with the hand-crafted orientation, the correlation coefficient between $s_{npi}$ and $s_\tau$ is very high, 0.996. We use our new score $s_\tau$ to sample a subset of participants that is balanced in terms of the score $s_\tau$: we randomly sample 18 participants that have score $s_\tau = 0$, another 18 participants that have score $s_\tau = 1$, and so forth. This results in a subset of $18 \cdot 40 = 720$ participants.
We now apply tangles to this new dataset. As before, we use all the 40 bipartitions given by the questions and the same cost function as before. We set our agreement parameter $a$ to 150 and prune paths of length one. On this balanced dataset, the tangle algorithm returns two tangles, indicating that within this balanced subset of the data, there is a cluster structure with two dense structures. If we wish to do so, we could now use our hard

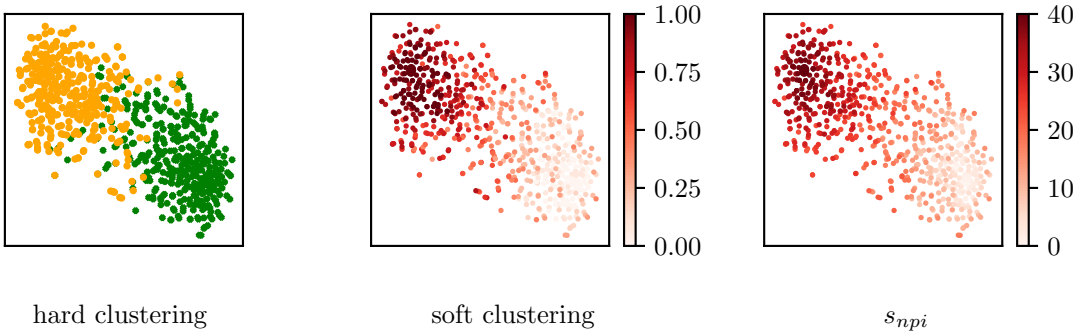



Figure 6: Hard and soft clustering output using the post-processings described in Section 3.4. We used tSNE to embed the questionnaire participants into two dimensions. The left figure shows the output of the hard clustering post-processing. The middle image shows the soft clustering for tangle A, and the right figure visualizes the hand-crafted scores: $s_{npi}$.

clustering output (Section 3.4) and assign each participant to one cluster, labeling them as either narcissistic or not, cf. Figure 6 left. However, this approach is very restrictive, and given our knowledge about the data, namely that there are scores on a large range and not binary classes, it seems inappropriate. Instead, we are interested in a soft output assigning a probability to each participant belonging to each cluster. We calculate these probabilities by our post-processing described in Section 3.4. The result can be seen in Figure 6. We plotted the sampled subset using tSNE (van der Maaten and Hinton, 2008) to embed the points into two dimensions. The two clusters in Figure 6 (left) correspond to one tangle each, and we assign points by their probability of belonging to one or the other tangle. Figure 6 (middle) visualizes our soft clustering output that indicates the probabilities of belonging to one tangle. In this case, we plot the probabilities of belonging to $\tau_A$, which points toward the upper left structure. In the right image of Figure 6 we visualize the score $s_{npi}$ as a reference. Figure 7 shows the correlation between the hand-crafted score $s_{npi}$ and the probability of being narcissistic based on the answers returned by the algorithm. The correlation coefficient is again very high, with a value of 0.944.
Looking at the tangle search tree, we can now calculate the characteristic cuts that help distinguish between the two clusters. The splitting tangle has eight characteristic cuts,

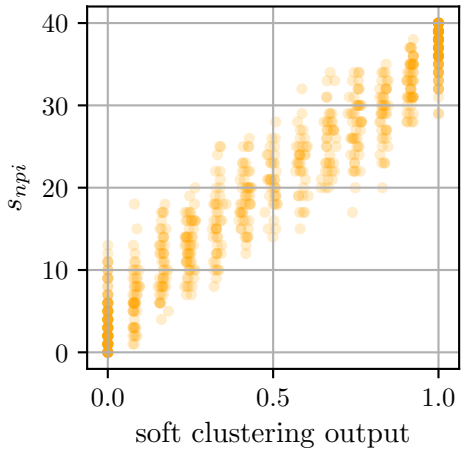



Figure 7: Correlation of the true score function and the deviation from the found tangle.

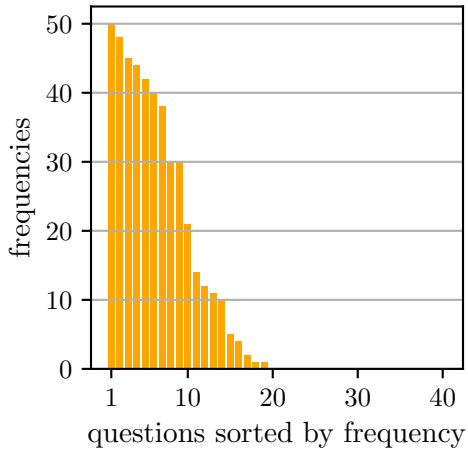



Figure 8: Frequencies of the important questions within 50 runs on random balanced sub-samples of the data.

meaning eight questions are essential to separate the two dense structures. Note that we can get variation in these results due to balanced sub-sampling, and the above shows one possible example. To support our claims, we ran the same procedure 50 times. Each time we sampled a random balanced subset of the data. The algorithm identifies between minimal five and maximal 11 important questions. We take their union, which results in an overall of 18 questions that seem to be important for splitting the data. This shows that the important questions overlap and underpins the claim that there are questions of little interest for the task. Figure 8 shows the frequencies of questions and Table 1 lists the 5 most important statements which where among the characterizing ones in at least 42 of the 50 runs. We list all statements of the dataset in Appendix III.3.

| # | question number | statement A | statement B |
|---|---|---|---|
| 50 | 12 | I like to have authority over other people. | I don't mind following orders. |
| 48 | 11 | I am assertive. | I wish I were more assertive. |
| 45 | 5 | The thought of ruling the world frightens the hell out of me. | If I ruled the world it would be a better place. |
| 44 | 31 | I can live my life in any way I want to. | People can't always live their lives in terms of what they want. |
| 42 | 32 | Being an authority doesn't mean that much to me. | People always seem to recognize my authority. |

Table 1: Characterizing question. Most left column gives the number of occurrences of the question within 50 runs. The question number is the one given by the dataset.

### 4.2 Theoretical guarantees: Binary Questionnaire

This section proposes a generative model to simulate mindsets in binary questionnaires. Based on this model, we prove that for suitable parameter choices, tangles *recover* the mindsets; that is, the set of all tangles coincides with the set of all mindsets with high probability. We refer to Appendix III.1 for the proofs.

#### 4.2.1 Generative model

We simulate $n$ persons that answer a questionnaire with $m$ questions. We start by generating $k$ ground truth mindsets $\mu_1, \dots, \mu_k \in \{0,1\}^m$. Each vector $\mu_i$ describes one specific way of answering all $m$ questions, so it represents the stereotype person with the corresponding mindset $i$. We generate the entries of each ground truth mindset vector by independent, fair coin throws. For every $\mu_i$, we generate a corresponding group $V_i$ of $n/k$ persons. We now choose a noise probability $p \in (0, 0.5)$ and let every person $v \in V_i$ answer question $s$ as indicated by $\mu_i(s)$ with probability $1-p$ and give the opposite answer with probability $p$ (independently across questions and persons). The union of the groups then forms the total population $V = \bigcup_{i=1}^{k} V_i$. Based on the answers, each question $s$ induces a cut of $V$ into the set $A_s^0 = \{v \in V \mid v(s) = 0\}$ and its complement $(A_s^0)^{\complement} = A_s^1 = \{v \in V \mid v(s) = 1\}$, where $v(s) \in \{0,1\}$ denotes the answer of person $v$ to question $s$; the collection of these cuts is denoted by $\mathcal{P}$. Since the questions induce the cuts, there is a natural one-to-one relationship between orientations of $\mathcal{P}$ and vectors in $\{0,1\}^m$. Using this relationship, we say that the tangles recover the mindsets if the set of all $\mathcal{P}$-tangles coincides with the set of all mindsets $\{\mu_1, \dots, \mu_k\}$.

When sampling the ground truth mindsets, we need to ensure that the vectors $\mu_i$ are not degenerate because the vectors by accident support more than $k$ tangles. We discuss the corresponding non-degeneracy-condition in Appendix III.1 (Assumption 1), where we also prove that this condition is satisfied with high probability.

#### 4.2.2 Main result in the questionnaire setting

The following theorem states that the orientations induced by the ground truth mindsets give rise to tangles and that all tangles on $\mathcal{P}$ correspond to mindsets with high probability.

**Theorem 2 (Tangles recover the ground truth mindsets)** *Assume that the model parameters $n, m, k$ and $p$ and the tangle parameter $a$ satisfy $p < 1/(k+3)$ and $a \in (pn, (1-3p)n/k)$. Let $\mathcal{P}$ be the set of cuts induced by questions in the questionnaire. Then with high probability, the mindsets correspond to tangles:*

1. *The probability that at least one of the mindsets does not induce a tangle is upper bounded by $km \exp\left(-2n(ka/n - 1 + 3p)^2/9k\right)$.*
2. *If the non-degeneracy Assumption 1 holds for the ground truth tangles, then the probability that there exists a spurious tangle that does not correspond to one of the mindsets is upper bounded by $km \exp\left(-2n(a/n - p)^2/k\right)$.*

In both statements, we take the probability over the random draw of the person's answers (and not over the randomness in generating the ground truth mindsets, which only play

a role regarding Assumption 1). In particular, the probability that the set of mindsets corresponds precisely to the set of tangles tends to 1 as $n$ tends to $\infty$ with fixed $a/n$.

The theorem is based on some conditions. The bound on $p$ ensures that the noise is not too large, considering the number of clusters. If the noise is too high, it becomes difficult to distinguish small clusters from spurious noise. The agreement parameter $a$ must not be too small (so that we do not cluster on noise) and not too large considering cluster size (otherwise, we cannot find the clusters anymore).

This theorem is what we would like to achieve: unless the parameters are so that they obfuscate the cluster structure, tangles provably find the ground truth clusters.

### 4.3 Experiments on synthetic data: Binary Questionnaire

We now run experiments on the generative model described in Section 4.2. We evaluate the influence of noise in the answers and the influence of irrelevant questions, that is, questions answered at random, and compare the performance of our post-processing to the output of the $k$-means algorithm.

If not stated otherwise, in our algorithmic setup, we use the bipartitions induced by all questions and choose $a$ to be a $1/3$ of the size of the smallest cluster. We choose the average Hamming similarity, stated in Equation (4), to assign a cost to the bipartitions. We use the normalized mutual information (nmi) between the ground truth mindsets and our discovered mindsets to measure the performance of our hard clustering output. The nmi assigns a score between 0 and 1; high scores indicate promising results. We average the results over ten random instances of the proposed model.

#### 4.3.1 Tangles discover the true mindsets and perform well on noisy data

One of the properties of tangles is their soft definition using sets of three orientations. As a result, they orient all bipartitions towards dense structures while the intersection of all cuts might be empty. As discussed above, we can interpret a tangle as one specific way of answering (all) questions. This scenario represents a stereotypical way of answering the questions, while no person in the dataset has to answer in this specific way. Thus tangles are inherently able to deal with noisy data. In Figure 9, we visualize the robustness of tangles on noisy data. In our model, we simulate noise by randomly flipping a percentage of each participant's answers individually. As a result, the respective person deviates more from the stereotypical answers, thus from its ground truth mindset. As a clustering baseline, we apply the $k$-means algorithm to the answer vectors of the participants, interpreting them as points in a Euclidean space. We give the actual number of mindsets $k$ to the $k$-means algorithm. In the left image of Figure 9, we observe that for balanced datasets, tangles perform comparably to $k$-means. Without fine-tuning any parameters, this is significant since tangles do not directly get the number of clusters as input; only a very rough lower bound on the size of the smallest cluster we want to discover. Tangles discover the correct number of clusters and the underlying structure even with high noise in the data.

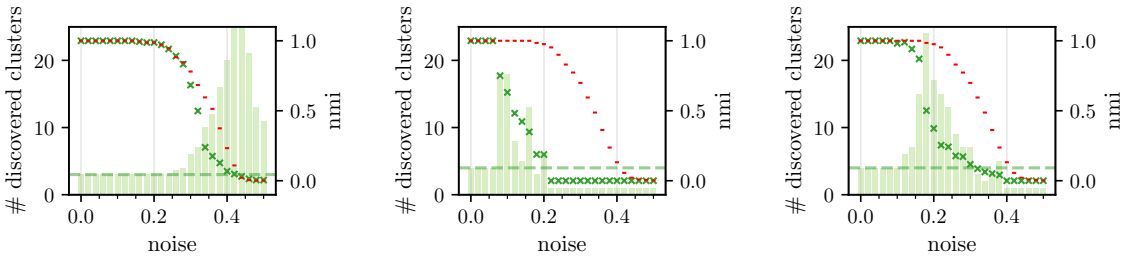



Figure 9: Influence of noise. We consider a questionnaire with 40 useful and no useless questions. We plot the performance depending on the noise for three clusters with 333 points each in the left image and show results for four unbalanced clusters with 100, 200, 300, and 400 data points in the middle and the right plot. In the two first images, we chose an agreement parameter $a$ of 1/3 of the smallest cluster size. On the right plot, we chose $a$ to be 2/3 of the smallest cluster size. In all three settings, we prune paths of length 1. Performance is measured by the nmi and plotted as markers. Red markers show the performance of $k$-means, and green markers the performance of tangles. Bars indicate the number of tangles in the data, and we evaluate both values for different noise levels in the x-axis.

#### 4.3.2 The tangle search tree holds all the information

For unbalanced datasets, we observe one of the open problems when translating tangles into practice; the gap problem discussed in Appendix II.2.1. In a nutshell, the gap problem arises from the fact that we never consider all possible bipartitions in practice. We get a sorted subset of all possible cuts that might not cover the set of data points uniformly. Therefore, we might have gaps or large jumps between the cost of cuts – for example, many unbalanced bipartitions followed by a random cut. This phenomenon becomes especially visible in datasets that consider highly unbalanced but non-hierarchical settings, where the clusters differ significantly in density. The middle image of Figure 9 shows the performance of tangles compared to $k$-means. We observe that, with increasing noise, the algorithm discovers significantly more tangles than there are clusters before the number of found clusters quickly drops to one.

We can reduce the influence of these gaps by adjusting the agreement parameter or the threshold $\Psi$ (see also Appendix II.2.1, Section 3.5). However, fine-tuning the parameters is not the goal in the end, and we believe there are other methods of post-processing the tangle search tree to avoid this, such as pruning. To highlight that tangles can also yield better results, and the tangle search tree holds all the information, we ran the same experiment with a tighter bound on the size of the smallest cluster. The larger agreement parameter results in the tangle search tree becoming inconsistent earlier and reassembles to early stopping the algorithm or choosing a smaller maximal order $\Psi$. In this case, we set the agreement parameter $a$ to $\frac{2}{3}$ the size of the smallest cluster. The left plot of Figure 9 shows the improvement when better estimating $a$, proving that the hierarchy of the tangles search tree contains the correct cluster structure.

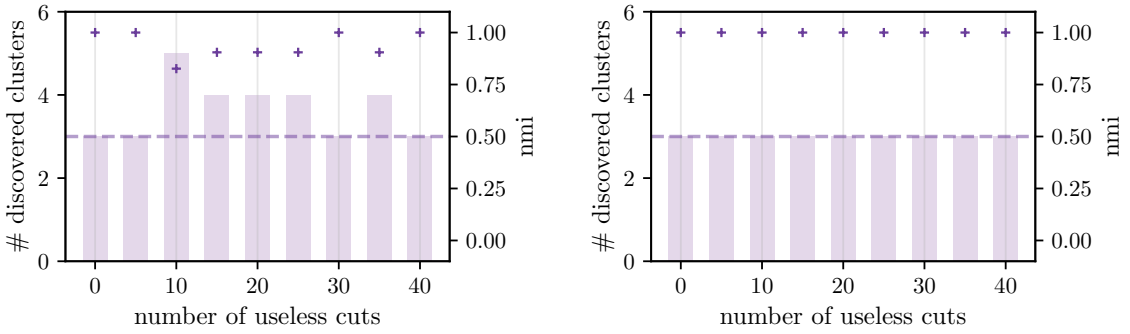



Figure 10: Influence of useless questions (answered at random) with and without pruning. We consider three clusters with 333 datapoints each. The noise $p$ is set to 0.15 and the questionnaire has 40 useful questions. The plot shows results on two axes. The left y-axis indicated by pluses visualizes the performance measured by the the nmi. Additional on the right y-axis we show the number of tangles as bars. Both are plotted over the number of useless questions. Results in the left image show output without any pruning. Random questions can result in more tangles than there are clusters. In the right image we visualize results when pruning path of length 1. Pruning can neutralize the effect of random questions.

#### 4.3.3 Useless questions do little harm

A bipartition is useless when it does not contain information relevant to separating the cluster structure; for example, a bipartition which roughly separates all the clusters in half. In this setting of a binary questionnaire, these would be questions that are irrelevant to the considered topic. We assume that for a given topic, such as narcissism in the npi data of Section 4.1, the question, "Do you wear glasses?" will not give any insight into a person's narcissism. However, the question is whether such useless bipartitions influence the quality of the tangle algorithm. In theory, such random bipartitions do not hurt tangles since the cut-cost will be high, so the cuts will be oriented late or not at all. They either get forced to one orientation by previous cheap bipartitions, or the tangle search tree gets inconsistent before orienting them.

As mentioned above, in practice, due to a lack of richness in the subset of cuts, we might find a large gap (cf. Appendix II.2.1) in the cost function between two cuts. Thus, even if the subsequent cut is useless, the cut will still be consistently orientable in both directions and result in two tangles splitting the large cluster into two smaller ones. Since these useless bipartitions often have a high cost, the following bipartitions will be even higher in cost and hold little to no information. We say such a split is random, and we can leverage this information to discover those splits and prune the tree along these branches. The exact procedure is described in Section 3.4. We show the result in Figure 10. On the left, we show the results when running the algorithm without pruning. The right image visualizes the results on the same data, but in this case, we prune paths to leaves of the depth of 1. We can significantly improve the output and thus find the correct number of tangles, each

corresponding to a cluster. Using the normalized mutual information score, we evaluate the performance based on our hard clustering (see Section 3.4).

### 4.4 Summary: Binary Questionnaire

We showed that tangles could automatically do what psychiatrists did by hand in a real-world example. They simultaneously discover the structure of the data and give insight into the questions we ask in the questionnaire. Theoretical guarantees reveal that tangles discover the ground truth with high probability in our questionnaire model. In experiments, we investigate different properties of tangles; we consider the effect of pruning the tangle search tree and the tangles' behavior on noisy data. Even though tangles do not need the correct number of clusters as an input, tangles often perform comparably to $k$-means, which we initialized with the correct $k$. The algorithm performs well for unbalanced datasets but seems more prone to noise. By sensitive sampling or adapted post-processing, we can extract more information from the data and enhance the performance. Stressing this, we show that the tangle search tree holds more information than we can currently leverage. This indicates rewarding research directions for developing and improving the hard and hierarchical clustering algorithm. Improving the post-processing or advancing the evaluation of the tangle search tree is promising.

## 5. Use Case: Graphs

In graph clustering, we are given a graph and want to divide the nodes of the graph into clusters such that sets of highly connected nodes are within the same clusters and there are only a few connections between different clusters. Tangles serve as an aggregation method for a set of cuts. We can generate these cuts by fast heuristic algorithms producing weakly informative cuts of the cluster structure.

### 5.1 Theoretical guarantees

In the following, we analyze the theoretical properties of tangles in graph clustering in the expected graph of a stochastic block model. We refer to Appendix III.2 for the proofs.

#### 5.1.1 Model

We consider a stochastic block model on a set $V$ of $n$ vertices that consists of two equal-sized blocks $V_1$ and $V_2$, which represent the ground truth clusters. Edges between vertices of the same block have weight $p$, and edges between blocks have weight $q$, where $0 \leqslant q < p \leqslant 1$. In a standard stochastic block model, we would now sample an unweighted, random graph from this model, where we would choose each edge with the probability given by $p$ and $q$ according to the ground truth model. In our case, we will perform the analysis just in expectation, as a proof of concept. This means we do not sample a random graph but consider the weighted graph described above.

We consider tangles induced by the set of all possible cuts $\mathcal{P}$ of the set $V$. We use the cost function

$$c(\{A, A^{\complement}\}) := \sum_{u \in A, v \in A^{\complement}} w(u, v) \tag{5}$$

where $w(u, v)$ denotes the weight of the edge between vertices $u$ and $v$. If we denote $\alpha_i = |A \cap V_i|/|V_i|$, this gives us, as $|V_1| = |V_2| = n/2$ the following explicit formula for the cost:

$$c(\{A, A^{\complement}\}) = \frac{n^2}{4} \left( p(\alpha_1 - \alpha_1^2 + \alpha_2 - \alpha_2^2) + q(\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2) \right). \tag{6}$$

Each of the two ground truth blocks $V_1$, and $V_2$ induces a natural orientation of the set of all cuts by picking from each cut $\{A, A^{\complement}\}$ the side containing the majority of that block's vertices. We find that for reasonable choices of $p$, $q$, and $a$, there is a range of costs in which these two orientations are indeed distinct tangles.

### 5.1.2 Main results in the graph clustering setting

The following Theorem states that in the graph clustering setting, tangles perfectly recover the ground truth: there exists a one-to-one correspondence between the tangles and the ground truth blocks.

**Theorem 3 (Tangles recover the ground truth blocks)** *Assume that the block model parameters $p, q, n$ and the tangle parameter $a$ satisfy $p > 3qn/(n-2a)$ and $a \geqslant 2$. Consider the set $\mathcal{P}$ of all possible graph cuts, and the set $\mathcal{P}_\Psi$ of those graph cuts with costs (cf. Equation 6) bounded by $\Psi$. Let $\xi = 1 + q/p$. If $\Psi$ satisfies*

$$q\left(\frac{n}{2}\right)^2 \leqslant \Psi < \frac{n^2}{4} p \left( \frac{1}{3}\xi\left(\xi - \frac{2a}{n}\right) - \frac{1}{9}\left(\xi - \frac{2a}{n}\right)^2 \right),$$

*then the two orientations of $\mathcal{P}_\Psi$ induced by the two ground truth blocks are distinct and exactly coincide with the $\mathcal{P}_\Psi$-tangles.*

If, on the other hand, $p < 2q$, there is no chance that tangles identify the two blocks as distinct clusters. The intuitive reason is that the within-cluster connectivity $p$ is smaller than two times the between-cluster connectivity $q$, the expected cost of a cut separating the two clusters is higher than one cutting through the clusters, which makes it impossible to recover the block structure. In this case, there will be precisely one tangle.

**Theorem 4 (Non-identifiability)** *If $a \geqslant 2$ and $p < 2q$, then for any value $\Psi$ and $\mathcal{P}_\Psi$ the set of all cuts of cost at most $\Psi$, there exists at most one $\mathcal{P}_\Psi$-tangle.*

Note that all our results are proved in the expected model, and they assume that tangles are constructed on the set of all possible graph cuts. In the experiments in the following section we complement these results with the cases where clusters are sampled from the model and where the tangles are constructed on a realistic, small subset of graph cuts.

### 5.2 Experiments on synthetic datasets

To validate tangles in the graph clustering setup, we perform experiments on synthetic data where we randomly sample graphs from a standard stochastic block model.

#### 5.2.1 Setup of the simulation and baselines

As opposed to the questionnaire setting, in the graph clustering setting, there is no obvious choice for the initial partitions in the pre-processing step. Instead, we use the Kernighan-Lin-Algorithm (Kernighan and Lin, 1970) to generate a small set of initial cuts. This algorithm performs a local search for a cheap cut under fixed partition sizes. Starting with a randomly initialized cut, each iteration goes over all pairs of vertices and greedily swaps their assignment if this improves the current cut. In the original version, the algorithm stops when none of the possible pairs can improve the cut value. However, we found that it is enough to run the algorithm for just two iterations of the local search to speed up the pre-processing: a highly diverse set of initial cuts is essential for our purpose. We denote this version of the algorithm as the KL algorithm with early stopping. Given a graph with $n$ vertices, each pass of the algorithm runs in time $\mathcal{O}(n^2 \log n)$, and we run the algorithm for two iterations. We use the average cut value to assign a cost to each bipartition: $c(\{A, A^\complement\}) := \frac{1}{|A|\cdot(n-|A|)} \cdot \sum_{u\in A, v\in A^\complement} w(u,v)$ where $w(u,v)$ is one if there is an edge between the nodes $u$ and $w$, else 0. We then apply the tangle algorithm to the subset of bipartitions. We choose the agreement parameter for the algorithm to be $1/3$ of the size of the smallest cluster, which is a rough lower bound. We do not choose a threshold value for $\Psi$ for the tangle algorithm but use all bipartitions generated by the pre-processing. To derive a hard clustering from the tangle search tree, we apply the post-processing described in Section 3.4. To evaluate the output, we use the normalized mutual information score (nmi) and average the values over ten random instances of the stochastic block model.
As a baseline, we compare tangles to normalized spectral clustering in the *sklearn* implementation. It gets the correct number of clusters as input.

#### 5.2.2 Tangles meet the theoretical bound already with few, weak initial cuts

The theoretical results above show that tangles recover the correct blocks in expectation based on all possible graph cuts. This section explores how far the bounds hold when we only generate a few initial cuts in a pre-processing step.
Figure 11 shows the results for (top row) two and (bottom row) five different clusters and varying values for the within cluster connectivity $p$ and the between cluster connectivity $q$. In the left figures, we see the results for 20 cuts generated with the KL-Algorithm stopping after only two iterations. As indicated by the red line, tangles meet the theoretical bound in this setting. Improving the set of initial cuts by running the KL-Algorithm for 100 iterations (which usually is until convergence) and using a more significant number of cuts (100) improves the results but is barely visible. We visualize the results for this setting in the middle pictures of both rows. Tangles can only aggregate the information in the set of cuts. We perform better when the quality of the initial bipartitions increases. However, minimal improvement indicates that fast and simple algorithms usually suffice to achieve satisfying results.

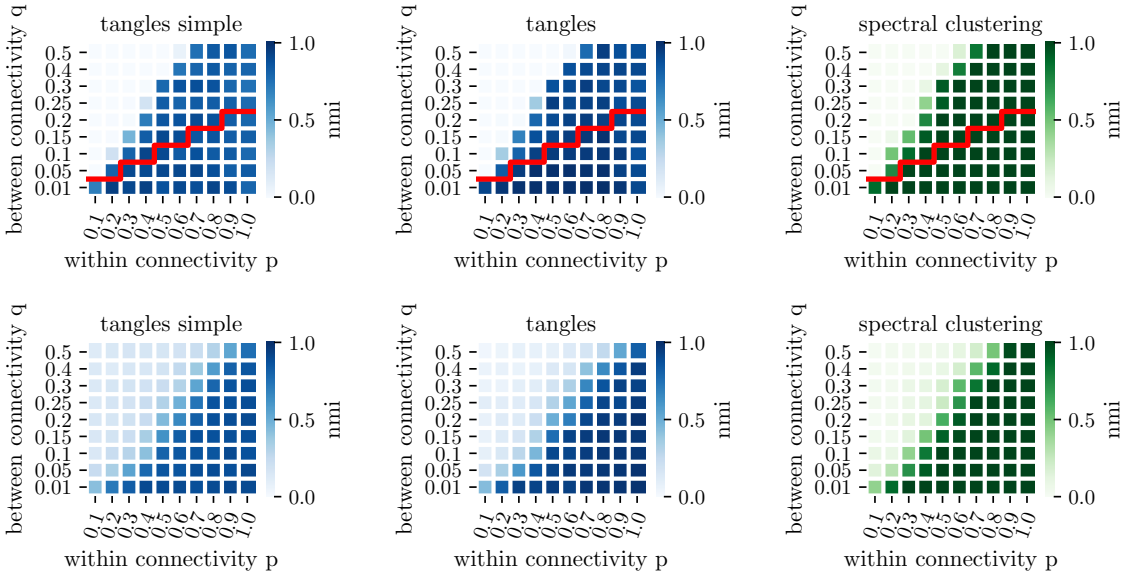



Figure 11: In the top row, we consider two clusters with 50 data points, each with an agreement parameter of $a = 16$. In the bottom row, we show results for five clusters and 20 datapoints each and an agreement parameter of $a = 6$. In the left and the middle figure, we use the KL-Algorithm to generate the initial set of cuts and apply the tangle algorithm. We achieve high performance with only 20 low-quality cuts (left). Increasing the number and the quality of the cuts to 100 and stopping after 100 iterations improves the overall performance of tangles only marginally (middle). In the right figures, we plot the results of normalized spectral clustering, which gets the correct number of clusters as a parameter. Tangles perform comparably. The red line indicates the theoretical bound derived in Section 5.1.2

Comparing the tangle results to spectral clustering, we can see that they perform comparably: they both recover the block structure under similar parameter settings and with comparable accuracy. We find this quite impressive, considering the "quick and dirty" pre-processing of generating only 20 cuts using a local search heuristic.

### 5.2.3 Performance of tangles saturates fast with increasing number of cuts

In the section above, we already saw that a small set of cuts slightly better than random is sufficient to yield satisfying results. In the next experiment, we investigate the number of cuts more closely. We show that the performance saturates fast with an increasing number of cuts. This observation is comforting: the number of cuts is no complex parameter to fine-tune. Figure 12 shows two simple examples for two and five clusters. With an increasing number of cuts, the performance increases fast before saturating. While for a small set of cuts, sometimes more tangles than clusters exist, with a more significant number of cuts, this number also stabilizes quickly.

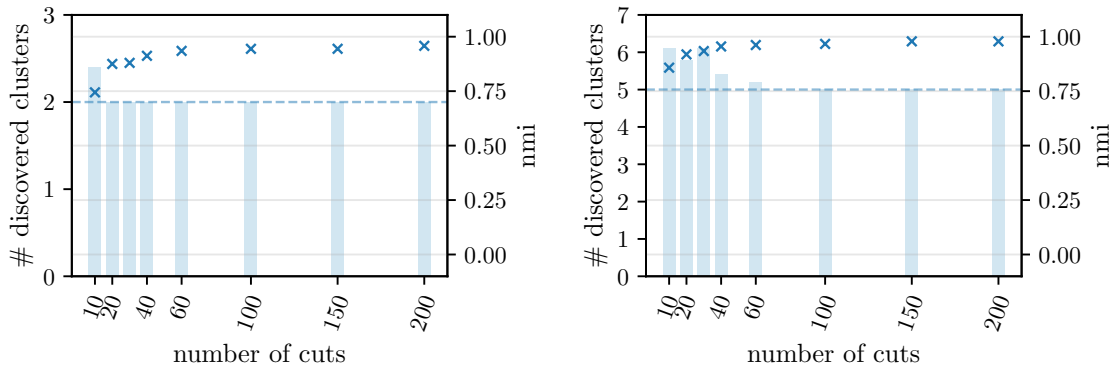



Figure 12: Performance saturates fast. (Left) Two clusters with 500 datapoints each and an agreement parameter of $a = 166$. (Right) Five clusters with 200 data points each and an agreement parameter of $a = 66$. On the y-axis, we plot the performance measured by nmi as markers, and we plot bars to visualize the number of tangles. The dashed line indicates the correct number of clusters. The number of cuts is on the x-axis. All results are averaged over ten random samples. With an increasing number of cuts, the performance increases fast before saturating.

### 5.3 Summary

This section demonstrates that tangles are well suited for a graph clustering setting. We provide theoretical performance guarantees and show that the algorithms work straightforwardly in practice. It is particularly encouraging to see that only a few initial cuts are necessary to achieve good performance. The number of considered cuts, which we initially believed to be the bottleneck of the computation (due to its cubic contribution to the running time), is not a limiting factor in practice. In Section 6.2, we will investigate the overall runtime of the algorithm and see that it behaves almost linearly in the number of cuts.

## 6. Use Case: Feature based data and interpretability

As our final use case, we consider a feature-based setup. Consider a set of data points $V$ described in terms of a vector of features. Each dimension represents one feature; these can be categorical or binary, or continuous features. The goal is to group points into clusters so that points that are featurewise similar to each other get assigned to the same cluster, while very dissimilar points are supposed to be in different clusters. Like in the graph setting, we can use fast and randomized algorithms to compute the initial set of cuts. One example of a cut-finding algorithm in a Euclidean setting is the following heuristic: randomly project the dataset on a one-dimensional subspace and generate a bipartition by applying the 2-means algorithm.

In order to explore yet another strength of tangles, we would like to focus on interpretable clustering algorithms in this section. To this end, we generate axis parallel data cuts in our pre-processing step. We then use the tangle mechanism to reveal clusters in the data.

**Algorithm 2:** Generate the initial set of cuts

**Data:** Set of points $V$, agreement parameter $a$
**Result:** Set of cuts $\mathcal{P}_a$

Choose $x_1$ such that $|A_{x,1}| = 1$ and $|A^{\complement}_{x,1}| = n - 1$.
$\mathcal{P}_a = \{A_{x,1}, A^{\complement}_{x,1}\}$
i = 1
**while** $A^{\complement}_{x,i} > a - 1$ **do**
    Choose $x_{i+1}$ such that $|A^{\complement}_{x,i} \cap A^{\complement}_{x,i+1}| = a - 1$
    $\mathcal{P}_a.append(\{A_{x,i+1}, A^{\complement}_{x,i+1}\})$
    i = i+1
**end**
**return** $\mathcal{P}_a$

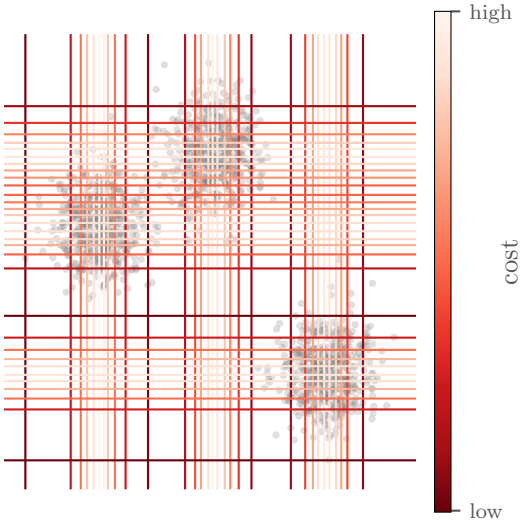



Figure 13: Visualization of resulting cuts.

These clusters then have a simple description in terms of features. Similar procedures have been used in interpretable clustering; see related work, Section 7.

#### 6.0.1 Setup of our interpretable tangle framework

Consider the set $V \subset \mathbb{R}^d$ (we assume all points are pairwise different). We generate axis-parallel cuts by a simple slicing algorithm. Moving along each axis, we select cuts exactly $a-1$ points away from each other. We outline the details in the pseudocode in Algorithm 2. Here $A_{x,i}$ represents the set of points smaller than some real value $x_i$ along the $x$-axis. $\{A_{x,i}, A^{\complement}_{x,i}\}$ is the cut along the $x$-axis at the real value $x_i$. The algorithm computes $\mathcal{O}(n/a)$ cuts for each dimension. The complexity is linear in the number of points and the number of dimensions: $\mathcal{O}(n \cdot d)$. As in the settings above, one possible post-processing is the one we describe in Section 3.4, which gives us a hard clustering output. If we have a low-dimensional embedding of our data, we can nicely visualize the soft output like in Figure 2.

### 6.1 Theoretical guarantees in the feature-based setting

We now prove theoretical guarantees for the tangle algorithm in the feature-based setting. As a ground truth model, we use a mixture of Gaussians. All theoretical results build on the pre-processing with axis-parallel cuts.

#### 6.1.1 Ground truth model: a simple Gaussian mixture

Suppose we are given two cluster centers $\mu = (\mu_1, \dots, \mu_d)$ and $\nu = (\nu_1, \dots, \nu_d)$ as points in the $d$-dimensional space. For ease of notation, let us assume that $\mu_i \leqslant \nu_i$ for all $i$. We

suppose that our data points $V$ are obtained by sampling $n$ points in total from a mixture of two Gaussian distributions $\mathcal{N}(\mu, \sigma^2 I)$ and $\mathcal{N}(\nu, \sigma^2 I)$ with equal weight, one with center $\mu$ and one with center $\nu$, and each with variance $\sigma^2 I$. Let us denote the bipartitions of $V$ obtained from Algorithm 2 along dimension $j$ as $(A_{j,1}, A_{j,1}^{\complement}), \ldots, (A_{j,m}, A_{j,m}^{\complement})$, where $A_{j,i} \subseteq A_{j,i+1}$. Let us, for the moment, assume that we sampled all axis-parallel bipartitions, that is, we used $a = 2$ for our sampling algorithm. Moreover, let us denote $x_{j,i}$ as the point in $\mathbb{R}$ for which we obtained $(A_{j,i}, A_{j,i}^{\complement})$ as $A_{j,i} = \{v \in V \mid v_j < x_{j,i}\}$. The set of all the bipartitions $(A_{j,i}, A_{j,i}^{\complement})$ for a fixed $j$ is denoted as $\mathcal{P}_j$.
For our proofs, we work in a scenario "in expectation": whenever we need to compute the volume of a set, we use the expected volume rather than the volume induced by the actual sample points.

#### 6.1.2 Main results

We will show that, under favorable conditions, there are two tangles, each pointing to one of the cluster centers $\mu$ and $\nu$, respectively. Here, the side of a cut $\{A_{j,i}, A_{j,i}^{\complement}\}$ that *points to* $y \in \mathbb{R}^n$ is $A_{j,i}$ if $x_{j,i} > y_j$ and $A_{j,i}^{\complement}$ if $x_{j,i} < y_j$; an orientation of cuts *points to* $y$ if all the sides of all cuts points to $y$.
The following theorem says that if the distance between the cluster centers is large enough and the agreement parameter $a$ is small enough, then we find at least two different tangles: one pointing to $\mu$ and one pointing to $\nu$.

**Theorem 5 (All cluster centers induce distinct tangles)** *Let $a < n/12$. If along some axis $j$ there is a local minimum $(A_{j,i}, A_{j,i}^{\complement})$, which is a global minimum and whose location $x_{j,i}$ has distance more than $\sigma$ to both $\mu_j$ and $\nu_j$, then there exist (at least) two tangles $\tau_\mu$ and $\tau_\nu$, where $\tau_\mu$ points to $\mu$ but not $\nu$ and $\tau_\nu$ points to $\nu$ but not $\mu$.*

The following theorem states that there are no spurious extra tangles if $a$ is chosen large enough, whereas this bound becomes lower the further apart $\mu$ and $\nu$ are.

**Theorem 6 (All tangles point to distinct cluster centers)** *Let $q$ be at most the fraction of points from $\nu$ at distance* $\mathrm{dist}$*; $q \leqslant (1 + \mathrm{erf}(-\mathrm{dist}/(2\sqrt{2\sigma^2})))/2$. If there exists a dimension $j$ where* $\mathrm{dist} = |\mu_j - \nu_j|$ *is large enough, that is* $\mathrm{dist} > 2\sigma$*, then for $a > n \cdot (0.42q + 0.056)$, every tangle points to either $\mu$ or $\nu$.*

The bound in Theorem 6 meets the one from Theorem 5 for $\mathrm{dist} \gtrsim 3,03\sigma$. In practice, we observe that the bounds are not tight, and the range in which we can choose the agreement parameter $a$ is much larger. We do not investigate the range for which the algorithm returns the perfect results; in practice we found the agreement parameter to be easy to choose. Usually, a rough estimate of the smallest cluster we want to discover suffices.

### 6.2 Experiments on synthetic datasets

In this section, we run experiments on a simple instance of a mixture of Gaussians, as the one shown in Figure 2. To generate an initial set of bipartitions, we use the slicing Algorithm 2 described in Section 6.0.1. As a cost function, we use

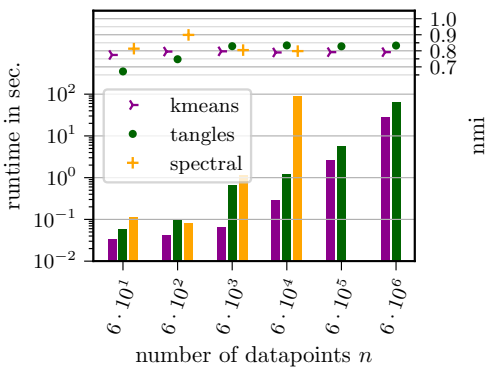



Figure 14: Performance and runtime of tangles with $m = 40$ cuts on a mixture of 4 Gaussians in two dimensions are competitive to two baselines. Runtime is shown as bar plots, and nmi performance with markers. We ran each algorithm for at most 1 hour; spectral clustering was too slow for more than 60.000 data points.

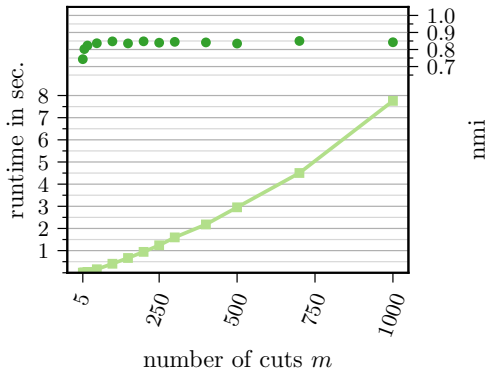



Figure 15: Computation time for the tangle search tree on a mixture of 4 Gaussians in two dimensions with $n = 6,000$ data points. Time grows less than cubic in the number of cuts, and performance quickly saturates.

$$c(\{A, A^{\complement}\}) = \sum_{v \in A, u \in A^{\complement}} -||v - u||. \tag{7}$$

To make the different methods comparable, we consider a challenging clustering task in which no method predicts the ground truth clusters. We assess their performance via the normalized mutual information between predicted and actual clusters. We experiment on simple instances of a mixture of Gaussians in $\mathbb{R}^2$ with $n = 6,000$ points and $k = 4$ clusters as the one visualized in Figure 2. All results are averaged over ten random instances of the model. We compare tangles to the $k$-means clustering algorithm as implemented in *sklearn*. We consider the agglomerative method of average linkage and a divisive method as hierarchical methods. For the latter, we iteratively use spectral clustering to split the cluster with the largest number of points into two clusters. All baseline algorithms get the correct number of ground truth clusters as an input parameter; tangles do not need this — they only get a rough lower bound on the size of the smallest cluster, specified by the agreement parameter $a$.

### 6.2.1 Computing tangles is fast

On the mixture of Gaussians, tangles perform comparable to $k$-means — although we used a straightforward cut-finding algorithm. Spectral clustering performs consistently well but is significantly slower, as shown in Figure 14.

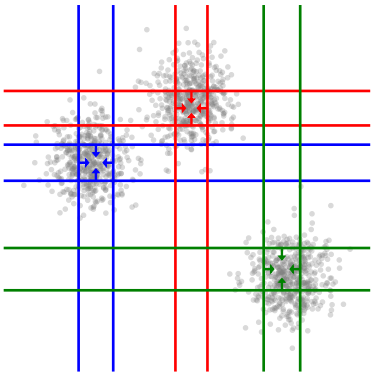


Figure 16: Visualisation of the core of all three tangles. Each color indicates one tangle. All other orientations are induced by the orientation of the core by the consistency condition.

Figure 17: Visualization using the soft clustering output for two dimensional data.

While the overall runtime of the algorithmic framework also depends on the pre-processing, computing the tangles itself is linear in the number of data points. In Figure 15, we investigate the complexity concerning the number of cuts. As discussed in Section III.2.1, the worst-case complexity is cubic in the number of cuts $m$. However, our experiments show that this bound is quite pessimistic because many branches quickly become inconsistent, and additionally, the performance already saturates after a few cuts. Hence, this experiment demonstrates that the number of cuts is not the limiting factor of tangles, similar as we have seen it in the graph cut setting as well.

| Dataset | Mixture of Gaussians |
|---|---|
| Linkage | 0.664 |
| Divisive | 0.820 |
| $k$-Means | 0.797 |
| Spectral | 0.805 |
| Tangles | **0.829** |

Table 2: Performance across all methods measured by normalized mutual information and averaged over ten runs.

### 6.2.2 Interpretability of cuts translates to interpretable clustering

If the initial bipartitions are interpretable, so are the resulting clusters. The small number of necessary cuts enhances this effect. We have already seen this in the questionnaire setting, where tangles find a small number of essential questions that can characterize the clustering sufficiently. Now we look at interpretability in the feature-based setting. Axis parallel cuts are interpretable: "all patients with temperature larger than 39 Celsius". We can carry over such explanations to tangles in the following way.

A tangle gives a consistent orientation to a set of bipartitions. If every cut is interpretable, we can combine these interpretations with an "and". In a tangle, there will likely be redundant bipartitions in the sense that two bipartitions point towards the same direction along the same axis, but one is more restrictive. For the explanation, we only consider the most restrictive bipartitions along the axes. We then end up with an interpretation that gives us an interval on each of the dimensions: dimension $d_1$ between $x_1$ and $y_1$ **and** dimension $d_2$ between $x_2$ and $y_2$ and so on. This interpretation represents the core of our tangle and points to the center of the respective cluster. Based on this, we can explain the cluster; there is a cohesive structure with these properties. Note that it does not follow that points outside of this 'core' do not belong to the same structure. This fact arises from the soft definition of tangles. Based on the tangle search tree, we can develop different approaches to interpret the resulting tangles depending on what we aim to characterize. Using the hard clustering algorithm, we can define the boundary cuts of each of the clusters or find the intervals of indistinct points, that is, points that belong to two (or more) clusters with comparable probabilities, as well as the characterizing cuts that distinguish between two clusters. Figure 16 visualizes an interpretation of the core in the 2-dimensional setting. For data embedded in two dimensions, a heat-map of our soft clustering already gives a visual interpretation of the tangles. Figure 17 shows an example for the used mixture of Gaussians.

### 6.3 Summary of the feature-based scenario

We used a naive pre-processing in the feature setting: axis parallel cuts. Even with these simple cuts, tangles perform comparably to the baseline clustering algorithms in terms of clustering accuracy while at the same time predicting the number of clusters in the data. As we can see, tangles can be computed fast, similarly to the $k$-means algorithm, and far faster than spectral clustering. We also showed that tangles could allow for natural interpretations in some cases.

## 7. Related work

Clustering is a vast field that comes with a multitude of algorithms. Conceptually, there are some related lines of work that we briefly want to touch on in order to position our framework and the tangles background into the landscape of clustering algorithms.

**Clustering ensembles.** Clustering ensemble methods first generate a set of initial clusterings using multiple clustering algorithms and then combine them with a consensus function to form a final clustering (Vega-Pons and Ruiz-Shulcloper, 2011). Even though this roughly resembles the tangle framework, there are key differences. In particular, ensemble methods typically use "strong" clustering algorithms that provide close-to-perfect clusters already. The ensemble mechanism is then only invoked to make the result more robust against the biases of the individual methods. The tangle philosophy is quite different: we use "quick and dirty" heuristics to produce the initial cuts, which then get boosted to high-quality clusterings.

**Soft clustering.** Soft clustering relaxes the degree to which objects belong to clusters from unique assignments (hard clustering) to distributions over the clusters, see, for example, Hastie et al. (2009). Similar to soft $k$-means, we can interpret tangles as representing clusters by abstract cluster centers, which we can convert into soft clustering. While soft clustering algorithms specify the number of clusters in advance, tangles indirectly specify a cluster's minimum size through an agreement parameter.

**Hierarchical clustering.** In hierarchical clustering, we cluster a dataset at different granularity levels, often represented by a dendrogram (Murtagh and Contreras, 2017). The tangle search tree also encodes a hierarchical structure, and its generation strongly resembles divisive approaches. Fluck (2019) even showed that in one specific setup considering all possible cuts, tangles and single linkage hierarchical clustering coincide. In general, however, there are fundamental differences between the two: the subdividing cuts are given and not computed recursively, the nodes represent tangles and not subsets, and the consistency condition provides a natural stopping criterion.

**Interpretable and Explainable Clustering.** To date, there is little work on explainable unsupervised learning as clustering. While there are papers considering decision trees for explainable clustering (Fraiman et al., 2013; Bertsimas et al., 2018), most work is empirical without theoretical analysis of the performance. Recently Dasgupta et al. (2020) and Frost et al. (2020) developed an algorithm for explaining the $k$-means clustering, approximating the performance in practice and theory. They do so by combining unsupervised and supervised learning first to construct a clustering and then, using the output as ground truth, explain the result using decision trees. In contrast, tangles come with an inherently interpretable output as long as the initial cuts are interpretable. For geometric data, using only axis-parallel cuts, tangles will directly return an interpretable output without further post-processing or approximating the clustering.

**Theoretical results on clustering**. Few clustering algorithms and generative models admit consistency guarantees: spectral clustering is among them, and its behavior on stochastic block models is well understood, see Abbe (2018) and references therein. Linkage algorithms typically are not statistically consistent (Hartigan, 1981) and thus do not necessarily discover the ground truth hierarchy, but some of them admit guarantees on approximations (A. Rinaldo, 2010; Chaudhuri et al., 2014; Moseley and Wang, 2017; Cohen-Addad et al., 2017). Guarantees for $k$-means are complex because of local optima, but with careful initialization, some approximation guarantees exist (Arthur and Vassilvitskii, 2007). Moreover, a large bulk of theory literature exists on Gaussian mixture clustering. To only mention a part of it, see Dasgupta (1999); Genovese and Wasserman (2000); Ghosal and van der Vaart (2001); Arora and Kannan (2005); Li and Schmidt (2015) for learning algorithms, convergence rates, and theoretical performance guarantees, see Banks et al. (2017); Ashtiani et al. (2018) for theoretical, and complexity bounds. See Vankadara and Ghoshdastidar (2020) for optimality guarantees of kernel methods in high dimensions.

**Mathematical background on tangles**. Tangles were initially conceived in the Graph Minors Project of Robertson and Seymour (1991) as a tool to measure how 'tree-like' a

graph is. In the original sense, Tangles are orientations of bipartitions of the *edge set*, which represents a hard-to-separate area inside a graph, making it very unlike a tree. This interpretation led to an abstraction of this notion, introducing the concept of tangles first to other contexts such as matroids (Geelen et al., 2009), and later resulted in the development of an abstract framework that unifies the notion of tangles from various contexts (Diestel and Oum, 2021; Diestel et al., 2019). Grohe (2016) performed a detailed survey on tangles for connectivity functions, a large and essential subclass of the more general tangles mentioned above. Grohe and Schweitzer (2015) created a sophisticated algorithmic framework and data structure for efficiently computing these tangles along with the corresponding tree of tangles. These works developed orthogonally to the ideas of Diestel and Whittle (2016), leading up to Diestel (2019) and Diestel (2020). Diestel focuses on making the notion of tangles applicable to as wide a range of settings as possible. To do so, he suggests softening some of the mathematical requirements of tangle theory. Their rigorous mathematical results may no longer apply to Diestel's tangles, which nevertheless aim to capture the notion of clusters. Our work in this paper follows the impulse of these latter ideas, taking them into a machine-learning context. In particular, the approach of sampling cuts, where the mathematical theory demands to consider all cuts up to a specific order, fits into this picture of approximating an underlying, more rigorous mathematical object.

## 8. Conclusion

In this paper, we introduce tangles to the machine learning community. This required a significant effort to simplify general concepts to convert the mathematical theory of tangles to a practical framework. We provide a first framework that works in practice and give provable guarantees in three statistical models. The general concept of "pointing towards a cluster" is a flexible formulation of a generic clustering problem. It only requires a set of cuts of the dataset and some notion of similarity between the objects. Thus tangles are directly applicable to many datasets without a workaround like building a nearest-neighbor graph or embedding the nodes. Although we convert the output to a hard clustering for the numeric evaluation, it is of a more general soft and hierarchical nature.

Note that we do not claim that tangles outperform every other algorithm out there. However, we are intrigued by their flexibility and potential. We proved performance guarantees in three very different setups: the questionnaire setting, the stochastic block model setting, and a Gaussian mixture setting. We are aware that stronger guarantees can be proved for individual algorithms in each setting. However, we are unaware of any algorithm for which guarantees can be proved in many different scenarios. A similar statement holds for our experiments. What is impressive is that the tangles framework combines many desirable properties that none of the baseline methods can provide at the same time: it is accurate without making assumptions on the shape of the clusters (as spectral clustering), it is fast (as $k$-means), it generates a hierarchy (as average linkage) and can be post-processed to a soft clustering, and it entails natural explanations.

We consider this work the first proof of concept that establishes tangles as a promising tool for clustering. More future work is needed to explore the full potential. There are many open questions for future research. On the algorithmic side, what is the optimal interplay between the initial cuts, the tangle algorithm, and the post-processing? On the theoretic

side, the most intriguing question is whether it is possible to formalize the intuition that tangles provide a generic tool to convert many "weak" cuts to "strong" clusters, as is the case for boosting in classification.

## Acknowledgments

This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645) and the International Max Planck Research School for Intelligent Systems (IMPRS-IS)

## References

L. Wasserman A. Rinaldo. Generalized density clustering. *Annals of Statistics*, pages 2678–2722, 2010.

E. Abbe. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research (JMLR)*, 18(177):1–86, 2018.

S. Arora and R. Kannan. Learning mixtures of separated nonspherical Gaussians. *The Annals of Applied Probability*, 15(1A):69–92, 2005.

D. Arthur and S. Vassilvitskii. K-Means++: The Advantages of Careful Seeding. ACM-SIAM Symposium on Discrete Algorithms (SODA), 2007.

H. Ashtiani, S. Ben-David, N. Harvey, C. Liaw, A. Mehrabian, and Y. Plan. Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes. 2018.

J. Banks, C. Moore, N. Verzelen, R. Vershynin, and J. Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, 2017.

D. Bertsimas, A. Orfanoudaki, and H. Wiberg. Interpretable Clustering via Optimal Trees. *preprint arXiv:1812.00539*, 2018.

K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. von Luxburg. Consistent Procedures for Cluster Tree Estimation and Pruning. *IEEE Transactions on Information Theory*, 60 (12):7900–7912, 2014.

V. Cohen-Addad, V. Kanade, and F. Mallmann-Trenn. Hierarchical Clustering Beyond the Worst-Case. Neural Information Processing Systems (NeurIPS), 2017.

S. Dasgupta. Learning Mixtures of Gaussians. Foundations of Computer Science (FOCS), 1999.

S. Dasgupta, N. Frost, M. Moshkovitz, and C. Rashtchian. Explainable k-Means and k-Medians Clustering. *preprint arXiv:2002.12538*, 2020.

R. Diestel. Abstract separation systems. *Order*, 35(1):157–170, 2018.

R. Diestel. Tangles in the social sciences. *preprint arXiv:1907.07341*, 2019.

R. Diestel. Tangles - A new paradigm for clusters and types. *preprint arXiv:2006.01830*, 2020.

R. Diestel and S. Oum. Tangle-tree duality in abstract separation systems. *Advances in Mathematics*, 377:107470, 2021.

R. Diestel and G. Whittle. Tangles and the Mona Lisa. *preprint arXiv:1603.06652*, 2016.

R. Diestel, F. Hundertmark, and S. Lemanczyk. Profiles of separations: in graphs, matroids, and beyond. *Combinatorica*, 39(1):37–75, 2019.

E. Fluck. Tangles and Single Linkage Hierarchical Clustering. Mathematical Foundations of Computer Science (MFCF), 2019.

R. Fraiman, B. Ghattas, and M. Svarc. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7:125–145, 2013.

N. Frost, M. Moshkovitz, and C. Rashtchian. ExKMC: Expanding Explainable k-Means Clustering. *preprint arXiv:2006.02399*, 2020.

J. Geelen, B. Gerards, and G. Whittle. Tangles, tree-decompositions and grids in matroids. *Journal of Combinatorial Theory, Series B*, 99(4):657–667, 2009.

C.R. Genovese and L. Wasserman. Rates of Convergence for the Gaussian Mixture Sieve. *The Annals of Statistics*, 28(4):1105–1127, 2000.

S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263, 2001.

M. Grohe. Tangled up in blue (a survey on connectivity, decompositions, and tangles). *preprint arXiv:1605.06704*, 2016.

M. Grohe and P. Schweitzer. Computing with Tangles. *Symposium on Theory of Computing (STOC)*, pages 683–692, 2015.

J.A. Hartigan. Consistency of Single Linkage for High-Density Clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2009.

F. Helguerro. Sui Massimi Delle Curve Dimorfiche. *Biometrika*, (3):85–98, 1904.

B.W. Kernighan and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell Systems Technical Journal*, 49(2), 1970.

J. Li and L. Schmidt. A Nearly Optimal and Agnostic Algorithm for Properly Learning a Mixture of k Gaussians, for any Constant k. *preprint arXiv:1506.01367*, 2015.

B. Moseley and J. Wang. Approximation Bounds for Hierarchical Clustering: Average Linkage, Bisecting K-means, and Local Search. pages 3094–3103. Neural Information Processing Systems (NeurIPS), 2017.

R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6):e1219, 2017.

T. Raskin. A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. 1988.

N. Robertson and P.D. Seymour. Graph minors. X. Obstructions to tree-decomposition. *Journal of Combinatorial Theory, Series B*, 52(2):153–190, 1991.

M.F. Schilling, A.E. Watkins, and W. Watkins. Is Human Height Bimodal? *The American Statistician*, 56(3):223–229, 2002.

L. van der Maaten and G. Hinton. Visualizing Data Using T-SNE. *Journal of Machine Learning Research (JMLR)*, 9(11):2579–2605, 2008.

L.C. Vankadara and D. Ghoshdastidar. On the optimality of kernels for high-dimensional clustering. International Conference on Artificial Intelligence and Statistics (AIStats, 2020.

S. Vega-Pons and J. Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, 2011.

D. Zwillinger and S. Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

# Appendix

## I. Background on tangles

### I.1 Translation between our terminology and the one used in graph theory

Tangles originate in mathematical graph theory, where they are treated in much more generality than what we need in our paper (cf. Diestel (2018) for an overview). For our work, we condensed the general theory to what we believe is the essence of tangles needed for machine learning applications. Since our setting is much simpler than the general one in mathematics, we also opted for more straightforward terminology closer to the machine-learning community's language. Table 3 provides a glossary of the correspondences between our terminology and that of Diestel (2018).

| **This paper** | | **Mathematical literature, e.g. Diestel (2018)** |
|---|---|---|
| cuts/partitions | instance of | separations |
| side of a cut/oriented cut | instance of | oriented separation |
| cost of a cut | instance of | order of a separation |
| set $\mathcal{P}$ | instance of | separation system $S$ |
| $O$ is consistent | corresponds to | $O$ avoids $\mathcal{F} = \{\{A_1, A_2, A_3\} \mid \lvert\bigcap_{i=1}^{3} A_i\rvert < a\}$ |
| $O$ is a $\mathcal{P}_\Psi$-tangle | corresponds to | $O$ is an $\mathcal{F}$-tangle of $\mathcal{P}_\Psi$ (for $\mathcal{F}$ as directly above) |
| *not used: condition* $A \cap B \neq \emptyset\ \forall A, B \in O$ | — | $O$ is consistent |
| tangle search tree *(the output of our algorithm)* | — | *not used* |
| *not used* | — | tree of tangles *(a tree-like set of cuts that partitions the tangles; similar to a space-partition-tree)* |

Table 3: Translation table between this paper and the mathematical literature

Furthermore, whereas we denote by $\mathcal{P}_\Psi$ the set of all cuts from $\mathcal{P}$ of costs *at most* $\Psi$, in the mathematical literature, $S_k$ usually denotes the set of all separations from $S$ of order *less than* $k$.

Note that throughout the literature, there are multiple ways to denote an oriented cut. Diestel (2018), for example, comes at the topic from the angle of graph theory, closely following Robertson and Seymour (1991), and has the convention of always listing both sides $(A^\complement, A)$, which is interpreted as *pointing from* $A^\complement$ *towards* $A$. On the other hand,

coming from matroid theory, Grohe and Schweitzer (2015), among others, have the custom of representing a tangle as a set of all the *small sides*, the $A^{\complement}$s in our terminology, which means the element-wise complements of the tangles in this paper would represent tangles in their sense.

### I.2 The magic number three in the consistency condition.

The consistency condition in Eq. (1) is defined with exactly three cuts. Three is the lowest number which makes the *profile property* in mathematical tangle theory come true: if $O$ is a tangle of $\mathcal{P}_\Psi$ then $A, B \in O$ implies that $(A \cap B)^c \notin O$. This property is central to tangle theory; for example, in the proof of one of the two central theorems from tangle theory, the tree-of-tangles theorem (Diestel et al., 2019). As a corollary, this theorem gives us a bound on the number of tangles. If $\mathcal{P}$ is the set of all cuts of some $V$, and we consider sets of less than three, we could potentially find up to $2^{2^{|V|}}$ many tangles. If we consider sets of three, then there can be at most $2a^{-1}|V|$ many $\mathcal{P}_\Psi$-tangles for any $\Psi$. Also, the tangle-tree-duality theorem (Diestel and Oum, 2021), which gives a witnessing dual, treelike structure for the absence of a tangle of a fixed set of separation, is richer when we consider sets of three instead of just pairs: If we consider pairs, this treelike structure of the separations system can only take the shape of a path, which is very restrictive. When considering sets of three, the nodes in this structure tree can have degrees of up to three; that is, they may branch. Lastly, increasing the size of the considered sets to more than three, increases computational complexity without introducing new mathematical behavior.

## II. Algorithmic Details

We introduced our basic algorithmic framework in Section 3. In the following, we give details on the tangle algorithm and one specific post-processing. We additionally give insight into some algorithmic questions arising when applying tangles in practice and how these influence algorithmic decisions.

### II.1 Details on Tangle Algorithm

In Algorithm 1, we present the main loop for the algorithm. Supposed we are given an agreement parameter $a$ and a family of cuts $\mathcal{P} = \{P_i = \{A_i, A_i^{\complement}\}\}_i$ each of which has cost $\Psi_i \in \mathbb{R}$. We order $\mathcal{P}$ according to their cost $\Psi_i$, and then we iteratively try to add all the cuts. We terminate when either we cannot add a cut consistently or when we run out of cuts.

Consistency check: For each tangle $\tau$ that we have identified as non-maximal, we try to add both orientations of a new cut $P$ to $\tau$. Each orientation produces a potential new tangle $\tau'$. We now need to check if $\tau'$ is consistent. We can check the consistency of $\tau'$ by checking the consistency of the *core* of $\tau'$, that is the set of most restricting bipartitions: the set of all the inclusion minimal sets in $\tau'$. This is sufficient since if we have $A, A', B, C$ such that $A' \subseteq A$, then

$$\left|A' \cap B \cap C\right| \geqslant a \implies |A \cap B \cap C| \geqslant a \tag{8}$$

Therefore, if a tangle's core is consistent, so is the tangle. If $\tau'$ is consistent, we add it as a left or right child in the tree. Which side depends on the orientation that created $\tau'$, left for $A$ and right for $A^{\complement}$. Subsequently, we marked that it was possible to extend $\tau$. If neither orientation can be added, we cannot extend $\tau$ and return.

In Algorithm 3 we show how to add a new orientation $A$ to a tangle $\tau$. To do so, we first add $A$ to the specification of $\tau$, then update the core of $\tau$ if necessary.

**Algorithm 3:** add orientation to tangle

```
Data: Node τ and new orientation A
Result: Child node τ_new
τ_new = {A} ∪ τ;
for C in core(τ) do
    if C ⊆ A then
        core(τ_new) = core(τ) ;
        return τ_new
    else if A ⊂ C then
        remove C from core(τ);
    end
end
core(τ_new) = {A} ∪ core(τ) ;
return τ_new
```

We postprocess the tree as described in Section 3.4. The final output of the post-processing approach is the attributed tree $(T^*, (p_t)_t)$, which is a "soft" version of a dendrogram, pseu-

**Algorithm 4:** post-processing the tangle search tree

**Data:** Tangle search tree $T$, weighting function $h : \mathbb{R} \to \mathbb{R}$, prune depth $\rho \in \mathbb{N}_0$
**Result:** Condensed tangle search tree $T^*$
`/* delete every node that has less than two children */`
**for** *node in* $T$ **do**
  **if** *internal node has less than two children* **then**
    delete node from $T$ ;
  **end**
**end**
`/* prune branches that are shorter than` $\rho$ `*/`
**while** *not done* **do**
  **for** *leaf in* $T$ **do**
    **if** *original distance to parent is shorter than* $\rho$ **then**
      remove leaf;
      remove or contract parent if necessary ; `// every node needs to have exactly two children`
    **end**
  **end**
**end**
`/* bottom up propagate tangle information to get set` $\mathcal{P}$ `for each tangle` $\tau$ `*/`
**while** *root is not reached* **do**
  **if** *all children of a node* $\tau$ *orient a cut* $P$ *the same* **then**
    track this information for the parent node ;
  **else**
    add $P$ to $\mathcal{P}(\tau)$ ;
  **end**
**end**
compute $p_\tau^{(\text{right})}(v)$ according to Eq. 2 for every $v$ and every $\tau$ ;
`/* derive probabilities of belonging to tangle` $\tau$ `by summing up the probabilities along the path from root to` $\tau$ `*/`
**return** $(T^*, (p_\tau)_\tau)$

docode is given in Algorithm 4. Since every node attribute $p_t \in [0,1]^n$ consists of probabilities for every object, the tree can be visualized with heatmaps as done in Figure 2.

### II.2 Translating theory to practice

For transparency we want to mention discrepancies between the theoretical object and tangles as algorithmic pipeline. There are two main changes when moving from the theoretical object to practise.

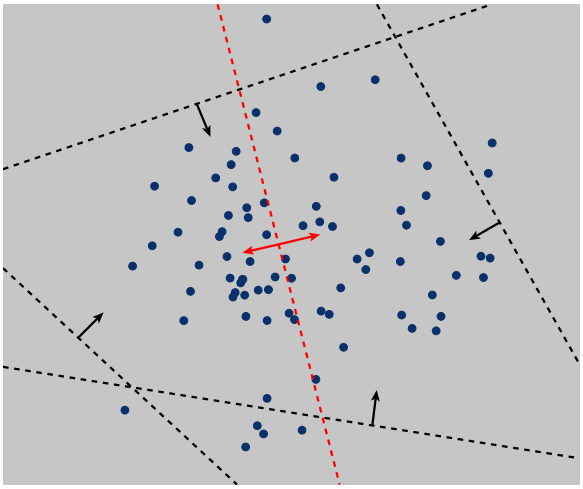


Figure 18: Gap problem. This figure depicts the intuition behind a large gap (here between the cheap bipartitions in black and the expensive bipartition in red). The red bipartition is orientable in both directions and would result in two tangles implying two clusters.

### II.2.1 The gap problem

In theory, we consider all possible cuts for the set of initial cuts, and the agreement parameter is chosen small, only to assure tangles point to *something*. Let us consider a complete graph in which tangles are supposed to detect the existence of only one cluster (the whole dataset).

We then have the following intuition for tangles on $\mathcal{P}_\Psi$ (with suitable $a$): For small $\Psi$, all considered cuts are unbalanced, and the consistency condition forces every cut to orient towards the larger side. As $\Psi$ increases, so does the balance of the cuts. At some point, $\mathcal{P}_\Psi$ starts to include cuts $P = (A, A^\complement)$ with a balance of $|A| = a$ and $\left|A^\complement\right| = n - a$ points that can be oriented both ways. However, the consistency condition prevents an orientation towards the smaller side $A$ because $\mathcal{P}_\Psi$ also contains cuts that subdivide $A$. Eventually, even an orientation towards the larger side will become inconsistent. The procedure stops, and the tangle search tree is a path, as desired. However, in practice, it is infeasible to consider all possible cuts. Instead, some procedure generates a small subset of initial cuts $\mathcal{P}$, which is supposed to contain all relevant cuts for the cluster structure. One particular problem arises when there is a large **gap in the cost** of the cuts, indicating the next cut through more dense regions. The extreme case of random cuts illustrates this: Most random cuts split a dataset roughly in half. Although they cannot contain information about the cluster structure, we can orient the first random cut in the sorted list of cuts both ways for reasonable $a$. The second random cut roughly halves both sides of the first cut, yielding four subsets of about $n/4$ data points, which are consistent orientations for $a < n/4$. The critical difference between considering all possible cuts and just a subset is the presence or absence of other cuts: although both sets of cuts contain these high-cost random cuts, the former contains additional cuts that prevent them from being oriented consistently. When we consider all possible cuts, we rely on the fact that previous cuts (and the consistency condition) force specific orientations. In practice, we wish to derive an algorithm that re-

turns tangles pointing towards the same structure as the ones given all possible cuts.

**Fine tuning parameters is an undesired solution.** We could avoid problems caused by the gap problem by tuning the parameters $a$ and $\Psi$. Setting $\Psi$ low enough will avoid using expensive cuts, and we will immediately discard random or 'bad' cuts. By tuning the agreement parameter $a$, we can indirectly influence the size of the clusters we can discover. Setting this parameter large enough enforces larger clusters and avoids splitting into several smaller ones. We can also stop searching the tangle search tree as soon as we discover the desired number of clusters which can also avoid subdividing large structures randomly. However, usually, we know little about our data; we might not know how small the smallest cluster is nor the number of clusters we are looking for. We also consider this as one of the strengths of tangles; they require few parameter choices, so in practice, we try to avoid options that require guessing the properties of the data.

**Dealing with a gap by pruning.** We figured that often in practice, these "spurious" consistent orientations, as shown in Figure 18, become inconsistent after only a few additional cuts. To give some intuition, consider a set of random bipartitions. We expect each bipartition to roughly split our data in half, and these random samples are likely to be maximally dissimilar. So for each new cut, we quickly reduce the number of points within the intersection of every set of three cuts to drop below the agreement parameter. This phenomenon is something we can easily detect. We implement and explain this method in Section 3.4. While this proved to work well in practice in some cases, for a very unbalanced cluster structure, we might still get long paths in the tangle search tree even for a set of random cuts, and it requires us to set a parameter, the pruning depth. Luckily, this parameter often is easy to choose after looking at the tangle search tree.

**Sensible sampling to avoid a gap.** Random sampling is a naive approach to generating the initial set of bipartitions. Especially considering the gap problem as described above, random cuts are not sufficient. Trivially the algorithm works well if we only consider cuts that already split the data well, and the aggregation using tangles then yields good results. We need to be more careful if we consider an initial set with bad cuts. Ideally, we want to sample a rich set of initial cuts that, in the end, get oriented the way they would be if we considered all the cuts in between. We want a set of initial cuts that covers the set of all possible bipartitions in a way such that significant gaps do not appear (or are very unlikely). In this case, we try to mimic the theoretical setting but reduce the number of bipartitions dramatically, making it computationally feasible, even fast. In Section 6.1, we introduce one naive approach to generate such an initial set and prove that tangles resemble clusters in the considered setting.

### II.2.2 The practical cost function

In tangle theory, the cost function does not consider the balance of the sides. For the cut value in the graph setting, the cheapest cuts are often very unbalanced, only splitting individual objects from the whole set and trivially get oriented towards their larger side. The same thing happens in practice. This does not prevent tangles from detecting clusters

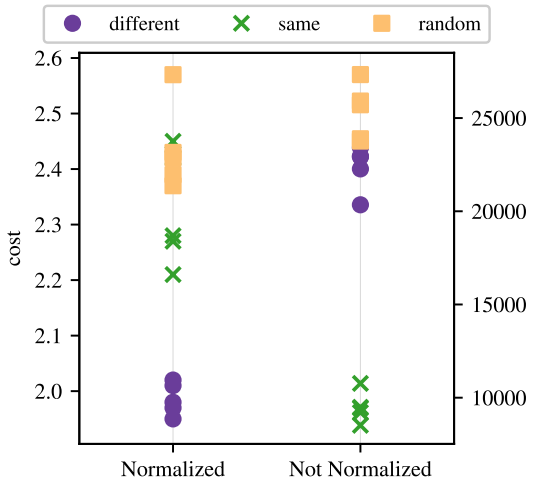



Figure 19: Order of the questions indicated by the costs with and without normalization. Normalizing the cost makes trivial questions (same answers) more expensive than informative questions (different answers) and starts with informative questions. The questionnaire consists of $k = 2$ mindsets and $m = 15$ questions. Five are answered differently by the mindsets, five are answered the same, and the remaining five are answered randomly by everyone. The marker indicates the question type.

but places an unnecessary computational burden on the tangle algorithm. Depending on the cost function, one natural countermeasure is to consider the balance of a cut $P = \{A, A^C\}$ by normalizing the cost with the factor $1/(|A|(n - |A|))$. Doing so makes the unbalanced cuts more expensive; tangles are built on informative cuts immediately.

For an illustrative example, consider the questionnaire model from Section 4.2 with $k = 2$ mindsets and small noise $p > 0$. Five questions are answered differently by the mindsets, five questions are answered the same, and another five questions are answered randomly by every person independently of the mindset. The questions with the same answer represent trivial, unbalanced cuts because they distinguish between persons that answer like their mindset (probability $1 - p$) and persons that answer differently (probability $p$). Only the questions with different answers are informative of the cluster structure. Figure 19 shows the cost of the questions with and without normalization. As expected, the unnormalized cost function places the trivial questions first. On the other hand, the normalized cost function increases their cost, such that the informative questions come first.

A stochastic block model gives another simplified example: two equal-sized blocks with 50 vertices each, $p = 0.3$, and $q = 0.1$. Figure 20 plots the (normalized) cost of a cut against its quality as measured by the normalized mutual information with the ground truth blocks. Since the basic idea is to use the cost of a cut as a proxy for its quality in separating clusters, a monotonic relationship between the two would be ideal. Normalizing increases the monotonic relationship strongly as measured by Spearman's rank correlation coefficient $\rho$ (Zwillinger and Kokoska, 1999). This measure is equal to the Pearson correlation between the rankings induced by the variables and ranges from -1 to +1, where the extremal values

indicate a perfect monotonic relationship. Normalizing the cost in Figure 20 changes this coefficient from $\rho = -0.25$ (left plot) to a significantly stronger correlation of $\rho = -0.90$ (right plot).

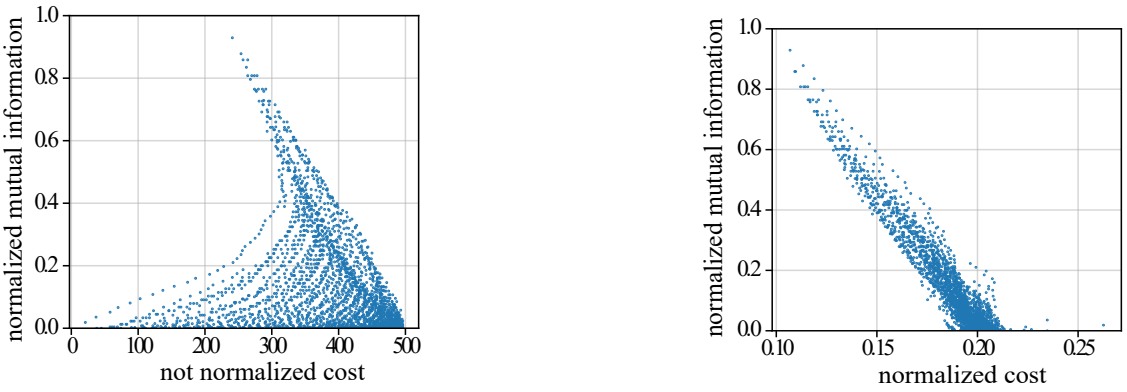



Figure 20: Normalized mutual information, of the cut given the ground truth, plotted against the (normalized) cost of a cut. Normalizing strongly increases the monotonic relationship between cost and quality of a cut.The dataset is a graph sampled from a stochastic block model with two blocks $V_1$ and $V_2$ with 50 vertices each, $p = 0.3$, and $q = 0.1$. For every $j, l \in \{0, \dots, 50\}$, one cut $P = \{A, A^{\complement}\}$ is randomly generated such that $|A \cap V_1| = j, |A \cap V_2| = l$. Each point in the plot corresponds to one cut.

## III. Proofs

### III.1 Binary Questionnaire

The proof of Theorem 2 is based on the following technical assumption, which excludes that in sampling the random mindsets $\mu_i$ we obtain a degenerate situation:

**Assumption 1 (Mindsets are not degenerate)** *If a tangle $\tau \in \{0,1\}^m$ satisfies that for all sets of three $x, y, z \leqslant m$ there exists a mindset $\mu_i$ such that tangle and mindset agree on all three: $\tau(x) = \mu_i(x)$ as well as $\tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, then the tangle $\tau$ is a single mindsets in the data, that is, $\tau = \mu_j$ for some $j$.*

Intuitively, Assumption 1 means we cannot trivially partition the ground truth into more than $k$ tangles, for $k$ ground truth mindsets. We now prove that for fixed $k$ and growing $m$ the probability of Assumption 1 being satisfied tends to 1:

**Proposition 7 (Assumption 1 satisfied with high probability)** *For $\beta = m/(k2^k)$, Assumption 1 is true with probability at least $1 - 2^{(1-\beta)k}$.*

**Proof**
We use a common bound for the coupon collectors problem (Motwani and Raghavan, 1995) to show that every partition of the mindsets is induced by one of the $m$ questions with probability at least $1 - 2^{(1-\beta)k}$.
For a contradiction we suppose further, that there is some $\tau \in \{0,1\}^m$ with the property that for all questions $x, y, z \leqslant m$ there is some mindset $\mu_i$ such that $\tau(x) = \mu_i(x), \tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, but $\tau$ itself is not a mindset.
Let $x, y, z \leqslant m$ such that there are as few mindsets $\mu_i$ as possible with $\tau(x) = \mu_i(x), \tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$ and suppose without loss of generality that $\mu_1$ does so. Since every partition of the mindsets is induced by some question, there is a question $x'$ such that $\mu_i(x') = \mu_i(x)$ for all $i \neq 1$ and $\mu_1(x') \neq \mu_1(x')$.
Now if $\tau(x') \neq \mu_1(x')$, then the set $\{x', y, z\}$ would contradict the fact that minimally many mindsets answered $x, y, z$ the same way as $\tau$.
Hence we may assume that $\tau(x') = \mu_1(x')$. Now, since $\tau \neq \mu_1$ there is some question $w \leqslant m$ with $\tau(w) \neq \mu_1(w)$. But then there is no mindset $\mu_i$ which answered the triple $x, x', w$ in the same way as $\tau$, which likewise contradicts the choice of $x, y, z$. ■

**Proof of Theorem 3.** We will prove the two directions of this statement separately, starting with showing that, with high probability, every mindset gives rise to a tangle:

**Lemma 8 (Mindsets give tangles)** *Let $k, m$ be fixed and let $\alpha = a/n$ be the agreement parameter as fraction of $n$. If $p < (n - ka)/(3n)$, then the probability that there is one of the $k$ mindsets which does not induce a tangle is bounded above by*

$$km \exp\left(-\frac{2n}{9k}(1 - k\alpha - 3p)^2\right),$$

*which tends to 0 as $n$ goes to $\infty$.*

Intuition: For more questions and more mindsets it gets less likely that every ground truth mindset induces a tangle. The larger the number of questions $m$ we consider the more likely that for one specific question, many (significantly more than $p \cdot n_i$) persons do not agree with their ground truth mindset. In this case the data might not represent the ground truth anymore, so we might not be able to recover it. If we consider a larger number of mindsets, we increase the number of possibilities to make an error. Also the number of persons per mindset decreases and thus for fixed $\alpha$ will get inconsistent if they become too small. It is intuitive that the probability of mindsets inducing tangles also decreases with increasing noise $p$. For a large number of persons $n$ answering the questionnaire the probability of a statistical oilier decreases and so does the probability of mindsets not inducing tangles.

**Proof** If some mindset $\mu_i$ is not a tangle then it contains a subset of three questions with intersection at most $a$. Hence by the pigeon-hole principle for at least one of these three questions at most $(2n_i + a)/3$ of the $n_i$ persons in $V_i$ answered that question as $\mu_i$ does. Each person in $V_i$ answers a question as $\mu_i$ does with probability $(1 - p)$. We can apply Hoeffding's tail bound on the binomial distribution with success percentage $(1 - p)$. With $X$ beeing the random variable counting the number of persons in $V_i$ that answer a fixed question as $\mu_i$ does, we find that

$$\mathbb{P}\left[X \leqslant \frac{2n_i + a}{3}\right] \leqslant \exp\left(-2\frac{(\frac{2n_i+a}{3} - (1-p)n_i)^2}{n_i}\right) = \exp\left(-\frac{2n}{9k}(1 - k\alpha - 3p)^2\right),$$

Note that we used $a = n\alpha = kn_i\alpha$ here.
Since there are $k$ mindsets and $m$ questions, by the union bound we obtain that the probability of the event that there is some question for which few persons answer following their mindset is at most

$$km \exp\left(-\frac{2n}{9k}(1 - k\alpha - 3p)^2\right).$$

Consequently so is the probability that some mindset is not a tangle. ■

**Lemma 9 (No degenerate tangles)** *If $a = \alpha n$ and $p < a/n$, then, given Assumption 1, the probability that there is a tangle which is not a mindset is bounded above by $km \exp\left(-2(\alpha - p)^2 n/k\right)$.*

Intuition: Increasing the number $m$ of questions or the number $k$ of mindsets makes it more likely that there are tangles which do not completely agree with one of the mindsets. Like for Lemma 8, the larger the number $m$ of questions or the number $k$ of mindsets we consider, the more likely it is that for one specific question and one mindset, significantly more than $p \cdot n_i$ persons from that mindset do not agree with the ground truth of that mindset. This may then result in the existence of a tangle which agrees with one of the mindsets on all but this question, where the tangle chooses the opposite orientation, and thus corresponds to no existing mindset. For increasing $k$ this is especially true, since increasing $k$ with fixed $n$ results in smaller values of $n_i$.

**Proof** By Assumption 1; for a tangle $\tau$ which does not correspond to a mindset there need to be a set of three $x_1, x_2, x_3$ such that for $y_i = \tau(x_1)$ we have that for $A_1 := A_{x_1}^{y_1}$, $A_2 := A_{x_2}^{y_2}$

and $A_3 := A_{x_3}^{y_3}$ that $|A_1 \cap A_2 \cap A_3| \geqslant a$. However, no mindset $\mu$ can incorporate $A_1, A_2, A_3$, that is there is no $\mu$ with $\mu(x_i) = y_i \; \forall i \in \{1,2,3\}$.
So suppose that there are such $A_1, A_2, A_3$ with $|A_1 \cap A_2 \cap A_3| > a_Q$. Then there is, by the pigeon-hole principle, a mindset $\mu_i$ for which at least $a/k$ many persons from $V_i$ lie in $A_1 \cap A_2 \cap A_3$. If this mindset does not incorporate this triple, then it does not incorporate one of the questions, say $\mu_i(x_1) \neq y_1$.
We are going to compute the probability that there is some such $A_x^y$ and a mindset $\mu_i$ where more than $n_i - a_Q/k$ many persons from $V_i$ lie in $A_x^y$, but $\mu_i(x) \neq y$. Each person in $V_i$ answers a question as $\mu_i$ does with probability $(1-p)$. Since it follows from $p \leqslant \alpha$ that $n_i - a/k \leqslant (1-p)n_i$, we can apply Hoeffding's tail bound on the binomial distribution with success probability $(1-p)$. With $X$ beeing the random variable counting the number of persons in $V_i$ that answer a fixed question as $\mu_i$ does, we find that

$$\mathbb{P}\left[X \leqslant n_i - \frac{a}{k}\right] \leqslant \exp\left(-2\frac{(n_i - \frac{a}{k} - (1-p)n_i)^2}{n_i}\right) = \exp\left(-\frac{2n}{k}(\alpha - p)^2\right) ,$$

Note that we used $a_Q = n\alpha_Q = kn_i\alpha_Q$ here.
Since there are $k$ mindsets and $m$ questions, by the union bound we obtain that the probability of the event that there is some such $A_x^y$ and $\mu_i$ is at most

$$mk \exp\left(-\frac{2n}{k}(p - \alpha_Q)^2\right) . \tag{9}$$

Consequently the probability that there exists a triple $A_1, A_2, A_3$ as above is also at most 9 since – as argued above – the $A_1$ in such a triple is such an $A_x^y$, witnessed by $\mu_i$. ■

**Proof of Theorem 2.** It is easy to see that in the setting of Theorem 2 the requirements of the Lemmas 8 and 9 are satisfied. ■

### III.2 Stochastic Block Model

We will show Theorem 3 in two stages. The first step is to calculate the maximum value of $\Psi$ for which the two orientations of $\mathcal{P}_\Psi$ induced by $V_1$ and $V_2$ respectively are indeed tangles.
For this we shall compute the minimum cost of a *forbidden triple* in these orientations that is, a triple which violates $a$-consistency, that is a triple $A, B, C$ of sides with the property that $|A \cap B \cap C| < a$. This we will do as follows: suppose that $A, B, C$ is a forbidden triple in the orientation given by $V_1$. We will sequentially manipulate an existing such triple, while only reducing its cots, until we obtain a triple of a specific form, whose cost is easy to calculate.
The following lemma says that (for a certain range of parameters) the cost of cutting through both blocks is larger than cutting through only one block and moving the other block entirely to one side of the cut.

**Lemma 10 (Not cutting through a block reduces the cost)** *Let $P = \{A, A^\complement\}$ be a cut, and let $\alpha_i = |A \cap V_i|/|V_i|$ and $\beta_i = |A^\complement \cap V_i|/|V_i| = 1 - \alpha_i$. If $\alpha_2 \geqslant (1 - 2\alpha_1)q/p$, then*

$c(\{A \cup V_2, A^\complement \setminus V_2\}) \leqslant c(P)$. *Similarly, if* $\beta_2 \geqslant (1-2\beta_1)q/p$, *then* $c(\{A \setminus V_2, A^\complement \cup V_2\}) \leqslant c(P)$.

In words: Given a cut $P$ which cuts through one of the blocks, say with a fraction $\alpha_2$, if $\alpha_2 > (1-2\alpha_1)q/p$, then we can prove that the cut that results from moving all these $\alpha_2 n$ many points to the other side is cheaper.

**Proof** For $\alpha_2 = (1-2\alpha_1)q/p$ we have by (6), that

$$\begin{aligned} c(P) =& \frac{n^2}{4}\left(p(\alpha_1 - \alpha_1^2 + \alpha_2 - \alpha_2^2) + q(\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2)\right) \\ =& \frac{n^2}{4}\left(p(\alpha_1 - \alpha_1^2) + q(1-\alpha_1)\right) = c(\{A \cup V_2, A^\complement \setminus V_2\}) \end{aligned}$$

Now if we consider (6), for fixed $n, p, q, \alpha_1$ as a real-valued function $f$ mapping $\alpha_2$ to $c(\{A, A^\complement\})$, then this is a quadratic function with a unique maximum. Thus by $f(1) = f((1-2\alpha_1)q/p)$ we have $f(x) \geqslant f(1)$ for every $(1-2\alpha_1)q/p \leqslant x \leqslant 1$, meaning for $\alpha_2 \geqslant (1-2\alpha_1)q/p$ we have that $c(\{A \cup V_2, A^\complement \setminus V_2\}) \leqslant c(P)$.
The argument for $\beta_2 \geqslant (1-2\beta_1)q/p$ is analogous. ■

We will now compute the lowest possible cost of a forbidden triple $A, B, C$ of which $B$ and $C$ both completely contain $V_2$ but $A$ does not. We will later use Lemma 10 to see that the minimum cost forbidden triples are of this form.

**Lemma 11 (Bound on the costs of a forbidden triple not cutting through a block)** *For* $q/p \leqslant 1/2$, *if* $A, B, C$ *are such that they each contain more than half of* $V_1$, *their intersection has size* $\leqslant a$, *and* $V_2$ *is a subset of* $A^\complement$, $B$, *and* $C$, *then the minimum possible value of* $\max\{c(\{A, A^\complement\}), c(\{B, B^\complement\}), c(\{C, C^\complement\})\}$ *is*

$$\frac{n^2}{4} p \left( \frac{1}{3}(1+r-x)(1+r) - \left(\frac{1+r-x}{3}\right)^2 \right), \text{ where } r = \frac{q}{p},\ x = \frac{2a}{n}.$$

*with some such triple attaining this bound (up to discretization).*

**Proof** We first calculate the cost $\{A, A^\complement\}$ would have if $|A^\complement \cap V_1| = |V_1|/2$.

$$c(\{A, A^\complement\}) = \frac{n^2}{4}\left(p(1 - 1^2 + \frac{1}{2} - \frac{1}{2}^2) + q(1 + \frac{1}{2} - 2 \cdot 1 \cdot \frac{1}{2})\right) = \frac{n^2}{4} p \left(\frac{1}{4} + \frac{r}{2}\right).$$

We observe for later that

$$\left(\frac{1}{3}(1+r-x)(1+r) - \left(\frac{1}{3}(1+r-x)\right)^2\right) \leqslant \left(\frac{2}{9}(1+r)^2\right) \leqslant \frac{1}{4} + \frac{r}{2} \tag{C}$$

for $0 < r \leqslant 1/2$. Moreover, by the calculations in the proof of Lemma 10, for fixed $p, q, n$ and $\alpha_1 = 0$ the quadratic function given by (6) has its maximum at

$$\frac{1 + \frac{(1-2\alpha_1)q}{p}}{2} > \frac{1}{2}.$$

Thus this function is monotone in the interval between 0 and 1/2. Therefore, if $A, B, C$ would be chosen such that $\max\{c(\{A, A^\complement\}), c(\{B, B^\complement\}), c(\{C, C^\complement\})\}$ is smaller than $n^2 p\left(\alpha(1+r) - \alpha^2\right)/4$, both $|B^\complement \cap V_1|/|V_1|$ and $|C^\complement \cap V_1|/|V_1|$ need to be smaller than $\alpha$. Therefore, in order for $A, B, C$ to be a forbidden triple, $|A^\complement \cap V_1|/|V_1|$ needs to be larger than $(\alpha - r)|V_1|$. Since the function given by (6) is quadratic, and, by (C), its value at $1/2$ is larger than its value at $(\alpha - r)$, this would imply that the cost of $\{A, A^\complement\}$ is larger than $n^2 p\left(\alpha(1+r) - \alpha^2\right)/4$, contradicting the assumption. ■

**Proof of Theorem 3.** We first show that $\tau^1_\Psi$, and similarly $\tau^2_\Psi$ are indeed tangles. For this, suppose for a contradiction that $A, B, C \in \tau^1_\Psi$ such that $|A \cap B \cap C| < a$ and such that the maximum over the cost of the three is as small as possible. Then, by pigeon hole principle one of $A, B, C$ must contain at least $(|V_2| - a)/3$ vertices from $V_2$.
By Lemma 10, we may now suppose without loss of generality, that$A^\complement \supseteq V_2$, say. Again by Lemma 10, we may additionally suppose that $B^\complement \cap V_2 = \varnothing$, $C^\complement \cap V_2 = \varnothing$. Thus by Lemma 11, the maximal cost of $A, B, C$ is minimized by

$$\frac{n^2}{4} p \left( \frac{1}{3}(1 + r - x)(1 + r) - (\frac{1 + r - x}{3})^2 \right) .$$

The argument for $\tau^2_\Psi$ is analogous.
Moreover, since the cost of the cut $\{V_1, V_2\}$ is $qn^2/4 < \Psi$, $\tau^1_\Psi$ and $\tau^2_\Psi$ are distinct as $V_1 \in \tau^1_\Psi$ and $V_2 \in \tau^2_\Psi$.
It remains to show that $\tau^1_\Psi$ and $\tau^2_\Psi$ are the only tangles. So suppose for a contradiction that there is an orientation $\tau$ of $\mathcal{P}_\Psi$ which is a tangle, but $\tau \notin \{\tau^1_\Psi, \tau^2_\Psi\}$. Then there is some $A \in \tau$ such that $|A^\complement \cap V_1| \geqslant |A \cap V_1|$. Suppose that $|A \cap V_1|$ is as small as possible; we will show that $|A \cap V_1| = 0$.
If $A \cap V_2 \neq \varnothing$, it is easy to see that the costs of the cut $\{A^\complement \cup V_2, A \setminus V_2\}$ are at most the cost of the cut $\{A^\complement, A\}$. If $\tau$ contains $A \setminus V_2$ we could chose the cut $\{A^\complement \cup V_2, A \setminus V_2\}$ instead of $\{A^\complement, A\}$, so suppose that $\tau$ contains $A^\complement \cup V_2$.
As the costs of the cut $\{A \cap V_2, A^\complement \cup V_1\}$ are again at most the costs of $\{A, A^\complement\}$, our tangle needs to contain either $A \cap V_2$ or $A^\complement \cup V_1$. However the latter is not possible as $(A^\complement \cup V_1) \cap (A^\complement \cup V_2) \cap A = \varnothing$ contradicting the consistency of $\tau$. Thus $A \cap V_2 \in \tau$ which already verifies that $\tau$ contains a set disjoint from $V_1$.
On the other hand, if $A \cap V_2 = \varnothing$ and there is is some $v \in A \cap V_1$ we can consider the cut $\{A^\complement \cup \{v\}, A \setminus \{v\}\}$. Since $p \geqslant 2q$ the cost of this cut is, as $A \subseteq V_1$, strictly lower than that of $\{A, A^\complement\}$, hence $\tau$ contains an orientation of this cut. Since $\tau$ is a tangle with $a \geqslant 2$, and $A \cap (A^\complement \cup \{v\}) = \{v\}$, $\tau$ cannot contain $A^\complement \cup \{v\}$ and must therefore contain $A \setminus \{v\}$. However, $A \setminus \{v\}$ now contradicts the choice of $A$.
Thus $\tau$ contains an $A$ disjoint from $V_1$, and similarly, as $\tau \neq \tau^2_\Psi$, the tangle $\tau$ contains an $B$ disjoint from $V_2$. But since $A \cap B = \varnothing$ this contradicts the assertion that $\tau$ is a tangle. ■

To prove Theorem 4 we introduce the concept of locally minimal cuts. A cut $\{A, A^\complement\} \in \mathcal{P}$ is a *local minimum* if moving any single $v \in V$ to the other side does not decrease the cost. In generality for tangles for $a \geqslant 2$ every cut which is of minimum cost such that a given pair of tangles disagrees on it is a local minimum given the cost function.
However, all these local minimum cuts need to respect the blocks.

**Lemma 12 (Local minimum cuts do not cut through the blocks)** *Independently of the choice of parameters, as long as $p > 0$, every local minimum cut respects the blocks, that is if $\{A, A^\complement\}$ is a local minimum cut, then for $i = 1, 2$ either $V_i \subseteq A$ or $V_i \subseteq A^\complement$.*

**Proof** Suppose that $\{A, A^\complement\}$ is a local minimum cut but that, without loss of generality, both $V_1 \cap A$ and $V_1 \cap A^\complement$ are non-empty. Pick some arbitrary $v \in V_1 \cap A$ and $v' \in V_1 \cap A^\complement$. Since $\{A, A^\complement\}$ is locally minimal, the cut $\{A \setminus \{v\}, A^\complement \cup \{v\}\}$ has at least the cost of $\{A, A^\complement\}$, hence, as $w(v, v) = w(v', v') = 0$, we have

$$\sum_{x \in A} w(v, x) \geqslant \sum_{x \in A^\complement} w(v, x) \quad \text{and, similarly,} \quad \sum_{x \in A} w(v', x) \leqslant \sum_{x \in A^\complement} w(v', x)\,.$$

However, since $v, v'$ both lie in $V_1$, we have that $w(v, x) = w(v', x)$ for all $x \neq v, v'$, thus:

$$\sum_{x \in A} w(v, x) \geqslant \sum_{x \in A^\complement} w(v, x) = \sum_{x \in A^\complement} w(v', x) + w(v, v') \geqslant \sum_{x \in A} w(v', x) + w(v, v') \geqslant \sum_{x \in A} w(v, x) + 2w(v, v')$$

Which is a contradiction as, $w(v, v') = p > 0$. ■

**Proof of Theorem 4.** Suppose there exist two tangles. Then there exists a lowest cost cut on which they disagree. This cut is a local minimum since $a \geqslant 2$. By Lemma 12 the only local minimum cuts are $\{\varnothing, V\}$ and $\{V_1, V_2\}$. So the cut in question must be $\{V_1, V_2\}$. As this cut has cost $qn^2/4$, our two tangles have to be $\mathcal{P}_\Psi$-tangles for some $\Psi \geqslant qn^2/4$.
Let $\tau$ be the tangle with $V_1 \in \tau$. Pick any two disjoint $X_1, X_2 \subseteq V_1$ with $|X_1| = |X_2| = \lfloor |V_1|/2 \rfloor$. We have that

$$c(\{X_1, X_1^\complement\}) = c(\{X_2, X_2^\complement\}) = p \left\lfloor \frac{n}{4} \right\rfloor \left\lceil \frac{n}{4} \right\rceil + q \left\lfloor \frac{n}{4} \right\rfloor \frac{n}{2} \leqslant p\frac{n^2}{16} + q\frac{n^2}{8} \leqslant q\frac{n^2}{4}.$$

So $\tau$ contains one of each $X_1$ or $X_1^\complement$ and $X_2$ or $X_2^\complement$. As $|V_1 \cap X_1^\complement \cap X_2^\complement| \leqslant 1 < a$, the tangle $\tau$ cannot contain both $X_1^\complement$ and $X_2^\complement$.
Without loss of generality we may assume that $X_1 \in \tau$ and that $X_1$ is of minimum size such that $X_1 \in \tau$. Now for any $x \in X_1$ the cut $\{X_1 \setminus \{x\}, (X_1 \setminus \{x\})^\complement\}$ has lower cost than $\{X_1, X_1^\complement\}$. Thus by the minimal choice of $X_1$ the tangle $\tau$ contains $(X_1 \setminus \{x\})^\complement$, but $\left|(X_1 \setminus \{x\})^\complement \cap X_1\right| = |\{x\}| = 1 < a$, which contradicts our assumption that $\tau$ is a tangle. ■

### III.2.1 Identifying tangles from random cuts is hard

Note that all our results for the stochastic block model rely on the fact that $\mathcal{P}_\Psi$ contains *all* cuts of $G$ up to cost $\Psi$ rather than just a sample of those cuts. In practice $\mathcal{P}_\Psi$ might consist of many cuts, and so one usually would try to perform some sampling strategy to obtain a 'sensible' subset of these cuts. The following result shows that sampling from $\mathcal{P}_\Psi$ **uniformly at random is not a useful sampling strategy** for this purpose, as one would still be required to draw a sample of size exponential in $n$.
Observe that by Theorem 3 the blocks define $\mathcal{P}_\Psi$-tangles for *any* collection $\mathcal{P}$ of cuts. Thus, in order for the two blocks to define distinct tangles it suffices for the sampled set of cuts

to contain a single cut which *distinguishes* them – that is one cut $\{A, A^\complement\}$ such that more than half of $V_1$ lies in $A$ and more than half of $V_2$ lies in $A^\complement$. Unfortunately, the number of these *good* cuts is exponentially small compared to the number of cuts up to any given cost $\Psi$:

**Theorem 13 (Number of silly cuts)** *Let $p, q$ be fixed and let $\Psi, a$ be chosen dependent on $n$ such that the orientations induced by the blocks are distinct $\mathcal{P}_\Psi$-tangles for $\mathcal{P}_\Psi$ the set of all cuts up to cost $\Psi$. Then, asymptotically, the number of cuts not distinguishing these tangles is exponentially larger than the number of cuts distinguishing those tangles, for $n$ going to infinity.*

This theorem implies that, if sampling cuts from $\mathcal{P}_\Psi$ uniformly at random, one would need to sample exponentially many of the cuts in order for the two blocks to define distinct tangles, since our sample needs to include one cut for this which distinguishes the blocks.

**Proof of Theorem 13.** We can construct all good partitions $\{A, A^\complement\}$ as follows: Starting with the partition $\{V_1, V_2\}$, we pick any subsets $G_1 \subseteq V_1$ of size $g_1|V_1| < |V_1|/2$ and $G_2 \subseteq V_2$ of size $g_2|V_2| < |V_2|/2$ and let $\{A, A^\complement\} = \{(V_1 \smallsetminus G_1) \cup G_2\,,\, (V_2 \smallsetminus G_2) \cup G_1\}$. We observe that the cost of $\{A, A^\complement\}$ depends only on $g_1$ and $g_2$. For fixed $g_1$ and $g_2$ there are exactly $\binom{|V_1|}{g_1 n} \cdot \binom{|V_2|}{g_2 n}$ many distinct cuts realizing this $g_1$ and $g_2$. In this case we say that the good cut $\{A, A^\complement\}$ *corresponds to* $(g_1, g_2)$.
Similarly, we can construct all bad partitions $\{C, C^\complement\}$ as follows: Starting with the partition $\{\varnothing, V_1 \cup V_2\}$, we pick subsets $B_1 \subseteq V_1$ of size $b_1|V_1| < |V_1|/2$ and $B_2 \subseteq V_2$ of size $b_2|V_2| < |V_2|/2$ and let $\{C, C^\complement\} = \{B_1 \cup B_2, (V_1 \cup V_2) \smallsetminus (B_1 \cup B_2)\}$. We observe that again, the cost of $\{C, C^\complement\}$ depends only on $b_1$ and $b_2$. Moreover for fixed $b_1$ and $b_2$ there are exactly $\binom{|V_1|}{b_1 n} \cdot \binom{|V_2|}{b_2 n}$ many distinct cuts realizing this $b_1$ and $b_2$. In this case we say that the bad cut $\{C, C^\complement\}$ *corresponds to* $(b_1, b_2)$.
Therefore if we can show that for $b_1 = g_1 < 1/2$ and $b_2 = g_2 < 1/2$ the cost of $\{A, A^\complement\}$ as constructed above is always as least as big as the cost of $\{C, C^\complement\}$, this would imply that there are at least as many bad partitions as good ones, since we can construct an injective function from the set of good into the set of bad partitions of cost $< k$. The cost of a bad cut for $b_1, b_2$ is

$$\frac{n^2}{4}\left(p(b_1 - b_1^2 + b_2 - b_2^2) + q(b_1 + b_2 - 2b_1b_2)\right)$$

and the cost of a good cut for $g_1$ and $g_2$ is

$$\begin{aligned}&\frac{n^2}{4}\left(p(g_1 - g_1^2 + (1-g_2) - (1-g_2)^2) + q(g_1 + (1-g_2) - 2g_1(1-g_2))\right)\\ =&\frac{n^2}{4}\left(p(g_1 - g_1^2 + g_2 - g_2^2) + q(1 - g_1 - g_2 + 2g_1g_2)\right)\,.\end{aligned}$$

However, for $b_1, b_2 < 1/2$ we have that

$$2b_1 + 2b_2 - 4b_1b_2 < 1$$

and thus

$$b_1 + b_2 - 2b_1b_2 < 1 - b_1 - b_2 + 2b_1b_2\,,$$

meaning for $g_1 = b_1$ and $g_2 = b_2$ the cost of $\{C, C^\complement\}$ is indeed at most the cost of $\{A, A^\complement\}$. Therefore up to any given cost there are at least as many bad cuts as there are good ones. Moreover: For $\delta = 1 + q/p$ we can show that, under the condition that the cost of a good cut corresponding to $(g_1, g_2)$ is at most $p\left(2\left(1+\frac{q}{p}\right)^2/9\right)$ (for larger $\Psi$, $\tau_k^1$ is not a tangle by Lemma 11), then also the cost of a bad cut corresponding to $(\delta g_1, \delta g_2)$ is at most the cost of a good cut corresponding to $(g_1, g_2)$.
Thus in this case we have, given a fixed $\Psi$ and the set $A_\Psi$ of all pairs $(g_1, g_2)$ for which the cost of a good cut corresponding to $(g_1, g_2)$ is at most $\Psi$, that there are exactly

$$\sum_{\substack{(i,j) \\ (\frac{i}{|V_1|}, \frac{j}{|V_2|}) \in A_\Psi}} \binom{|V_1|}{i} \cdot \binom{|V_2|}{j}$$

good cuts. However, there are at least

$$\sum_{\substack{(i,j) \\ (\frac{i}{\delta|V_2|}, \frac{j}{\delta|V_2|}) \in A_\Psi}} \binom{|V_1|}{i} \cdot \binom{|V_2|}{j}$$

bad cuts. Using these numbers we can see that for fixed $\delta > 1$ the number of bad cuts grows exponentially faster in the number of nodes than then number of good cuts, as $n$ goes to infinity. ■

### III.3 Feature based data

The following computations are in expectation, which we incorporate by assuming two things. First, we assume that the cost function behaves like the density function of the marginal distributions. Concretely, we assume the following:

**Assumption 2** *The cost function $c$ is such that if $(A_{j,i}, A_{j,i}^\complement)$ and $(A_{k,l}, A_{k,l}^\complement)$ are two axis-parallel bipartitions obtained from Algorithm 2, then $c(\{A_{j,i}, A_{j,i}^\complement\}) \leqslant c(\{A_{k,l}, A_{k,l}^\complement\})$ if and only if the density at $x_{j,i}$ of the marginal distribution on the $j$-axis is smaller than the density at $x_{k,l}$ of the marginal distribution on the $k$-axis.*

Secondly, we assume that the intersection of any three sides of bipartitions contains the fraction of the points that is expected from the density functions, that is, we assume the following:

**Assumption 3** *Given three bipartitions $\{A, A^\complement\}, \{B, B^\complement\}, \{C, C^\complement\}$ in $\mathcal{P} = \bigcup_j \mathcal{P}_j$, where $A = \{v \in V \mid v_i < a\}$, $B = \{v \in V \mid v_j < b\}$, $C = \{v \in V \mid v_k < c\}$ for some $i, j, k \in \{1, \ldots, d\}$. Then the intersection of three sides of the bipartitions contains as many points as we would expect, that is, as large as the integral of the density over the intersection of the induced half-spaces.*

This means that we can use the quantiles of normal distributions in our arguments. In particular, we will use that for $\mathcal{N}(0,\sigma^2)$ less than 16% of the points are below $-\sigma$.
Let us say that a cut $\{A_{j,i}, A^{\complement}_{j,i}\}$ in $\mathcal{P}_j$ is a *local minimum* if and only if

$$c(\{A_{j,i-1}, A^{\complement}_{j,i-1}\}) > c(\{A_{j,i}, A^{\complement}_{j,i}\}), \qquad c(\{A_{j,i}, A^{\complement}_{j,i}\}) < c(\{A_{j,i+1}, A^{\complement}_{j,i+1}\}).$$

If we are given Assumption 2, we immediately obtain the following from the fact that the marginal densities of a Gaussian have exactly one local mimimum:

**Observation 14** *If Assumption 2 holds, then there is, in every dimension, at most one separation which is a local minimum.*

We prove Theorem 5 by showing that, for the smallest possible order for which we find a local minimum, this local minimum is oriented differently by $\mu$ and $\nu$. We show further, that this order is small enough such that the orientations induced by $\mu$ and $\nu$ cannot contain a triple of bipartitions with too small intersection.

**Lemma 15** *If $\tau, \tau'$ are distinct tangles, then there is a local minimum which is oriented differently by $\tau$ and $\tau'$*

**Proof** Let $\{A_{j,i}, A^{\complement}_{j,i}\}$ be a cut of minimal possible cost distinguishing $\tau$ and $\tau'$, say $A_{j,i} \in \tau$ and $A^{\complement}_{j,i} \in \tau'$. If $\{A_{j,i}, A^{\complement}_{j,i}\}$ is not a local minimum, say because $\{A_{j,i-1}, A^{\complement}_{j,i-1}\}$ has lower cost (the other case is analogous), then both $\tau$ and $\tau'$ would need to orient $\{A_{j,i-1}, A^{\complement}_{j,i-1}\}$ similarly, and thus, since one of the two contains $A^{\complement}_{j,i}$, they would both need to contain $A^{\complement}_{j,i-1}$ by our consistency condition. However, since $\left|A_{j,i} \cap A^{\complement}_{j,i-1}\right| < a$ this contradicts the consistency of the tangle $\tau$. ■

**Lemma 16** *There exists a local minimum in $\mathcal{P}_j$ if and only if $|\mu_j - \nu_j| > 2\sigma$. Moreover if $(A_{j,i}, A^{\complement}_{j,i})$ is a local minimum, then either $\nu_j < x_{j,i} < \mu_j$, or $\mu_j < x_{j,i} < \nu_j$.*

**Proof** There is a local minimum in $\mathcal{P}_j$ if and only if the density function has a local minimum by Assumption 2. This is the case precisely if $|\mu_j - \nu_j| > 2\sigma$, see Helguerro (1904); Schilling et al. (2002). The second part is then also immediate, since the density function is monotone on the intervals $(-\infty, \mu_j)$ and $(\nu_j, \infty)$. ■

**Proof of Theorem 5.** We consider the set of all cuts in $\bigcup \mathcal{P}_j$ up to and including the cost of $\{A_{j,i}, A^{\complement}_{j,i}\}$, that is, $\{A_{j,i}, A^{\complement}_{j,i}\}$ is the only local minimum that we consider and it is also the most expensive cut that we consider.
Since the $j$-axis is the only axis on which we consider a local minimum, each cut which we are considering except $(A_{j,i}, A^{\complement}_{j,i})$ needs to have $\mu$ and $\nu$ on the same side, which we call the 'middle'. We define $\tau_\mu$ by orienting $\{A_{j,i}, A^{\complement}_{j,i}\}$ as $A_{j,i}$, and all other cuts towards this middle.
Let us now show that this orientation is consistent; a tangle. Observe that, by Assumption 3, each side in $\tau_\mu$ contains at least $n/2$ points. Further if we consider some three sides of cuts in

$\tau_\mu$ which are along the same axis, then at most two of them are relevant to the intersection, since one of the sides will always be a superset of one of the others. So, considering three sides of any cuts in $\tau_\mu$, we need to consider just two cases:
*Case I: All cuts are along distinct axes.* Each of the three sides is expected to contain at least $n/2$ points, as observed. As the distributions along the distinct axes are independent, we thus have that the intersection of the three contains at least $n/8 > a$ points by Assumption 3.
*Case II: Two cuts $\{A_{k,l}, A^{\complement}_{k,l}\}, \{A_{k,l'}, A^{\complement}_{k,l'}\}$ are along the same axis, the third is along another axis.* We first determine the size of the intersection of the cuts along the same axis $k$. We claim that one of the two is chosen at distance more than $\sigma$ from $\mu_k$. Indeed, if one of the two cuts equals $\{A_{j,i}, A^{\complement}_{j,i}\}$ then this is true by the assumption of this theorem. Otherwise we claim that for one of the two cuts the point $x_{k,l}$ or $x_{k',l}$ has lower distance from $\nu_k$, respectively $\nu_{k'}$, than from $\mu_k$, respectively $\mu_{k'}$. Indeed, if for both of the two cuts the points $x_{k,l}$ and $x_{k',l}$ would have larger distance from $\nu_k$ and $\nu_{k'}$ than from $\mu_k$ and $\mu_{k'}$, respectively, then the size of the intersection of the sides of the two cuts contained in $\tau_\mu$ would be equal to the size of the smaller of the two respective sides, hence we do not need to consider this triplet of cuts. Therefore, indeed we have for one of the two cuts, say for $\{A_{k,l}, A^{\complement}_{k,l}\}$, that $x_{k,l}$ has smaller distance from $\nu_k$ than from $\mu_k$. Now if $x_{k,l}$ would have distance less than $\sigma$ from $\mu_k$, it would also need to have distance less than $\sigma$ from $\nu_k$, but this implies that this cut has higher costs than $\{A_{j,i}, A^{\complement}_{j,i}\}$, by Assumption 2, as $x_{j,i}$ has distance at least $\sigma$ from both, $\mu_j$ and $\nu_j$.
Thus, indeed one of the two cuts, say $\{A_{k,l}, A^{\complement}_{k,l}\}$, satisfies $|x_{k,l} - \mu_k| > \sigma$.
By Assumption 3, in expectation, at least 84% of the points belonging to $\mu$ – that is at least $0.42n$ points – lie on the same side of $\{A_{k,l}, A^{\complement}_{k,l}\}$ as $\tau_\mu$ chooses. At least half of the points belonging to $\mu$ – that is at least $0.25n$ points – lie on the chosen side of the other cut along the same axis, $\{A_{k,l'}, A^{\complement}_{k,l'}\}$. Thus in their intersection there are expected to be at least $0.17n$ points.
The third cut is along an independent axis and, by the same argument as in Case I, at most halves this number. So there are already at least $0.085n > n/12 > a$ points in the intersection, alone from the points belonging to $\mu$. ■

To prove Theorem 6 we will use the following simplification: Given a sampled cut $\{A_{j,i}, A^{\complement}_{j,i}\}$ there exists a whole interval in $\mathbb{R}$ in which we can chose an $x_{j,i}$ such that $A_{j,i} = \{v \in V \mid v_j < x_{j,i}\}$. To simplify the arguments, we will assume that our Assumptions 2 and 3 hold, regardless of how we have chosen $x_{j,i}$ for our given cut. This gives us the technical possibility to consider the cut $\{B_{j,x}, B^{\complement}_{j,x}\}$ sampled at $x \in \mathbb{R}$, that is, $B_{j,x} := \{v \in V \mid v_j < x\}$, for any $x \in \mathbb{R}$ and any dimension $j$. The result of this technicality is, that we do not have to take into account that, while we know that this cut is contained in $\mathcal{P}_j$, we may have put it into that set for a slightly different $x$. This simplifies the arguments a lot. We are aware, that as a formal statement this is an unrealistic assumption. However as the number of points tends to $\infty$, we will converge to this assumption, and thus we expect that the consequence of this theorem still holds.

**Proof of Theorem 6.** Suppose there is a tangle $\tau$ which points neither to $\mu$ nor to $\nu$. Let us suppose that $j$ is chosen such that $|\mu_j - \nu_j|$ is maximal. The general strategy of this proof is as follows: We will show that there are two local minima cuts in $\tau$, one along the $j$-axis and one along another axis, such that one of the two points away from $\mu$,

and the other points away from $\nu$. The cost of these local minima cuts will allow us to find a set of three cuts which $\tau$ needs to orient a certain way, but which together violate the consistency condition on $\tau$, due to the quantiles of the gaussian distribution. This will result in a contradiction to the assumption that $\tau$ is a tangle.
So let us first deal with the task of finding these local minima cuts.
As $\tau$ does not point towards $\mu$, there exist a cut $\{B_{i,x}, B^{\complement}_{i,x}\}$ whose orientation in $\tau$ points away from $\mu$, say $B^{\complement}_{i,x} \in \tau$ but $\mu_i < x$. We want to show that in this case we may suppose that $\{B_{i,x}, B^{\complement}_{i,x}\}$ is a local minimum.
So suppose that we cannot choose $\{B_{i,x}, B^{\complement}_{i,x}\}$ as a local minimum. We claim that there is an $x' \geqslant x$ and some $\epsilon > 0$ such that $B^{\complement}_{i,x'} \in \tau$ and $\{B_{i,x'+\epsilon'}, B^{\complement}_{i,x'+\epsilon'}\}$ is not oriented by $\tau$ for any $\epsilon' < \epsilon$. If we do not find such an $x'$ and $\epsilon$, then $\tau$ would need to orient every $\{B_{i,x'}, B^{\complement}_{i,x'}\}$ for any $x' > x$. However, since $B^{\complement}_{i,x} \in \tau$ we can conclude by the consistency condition that also $B^{\complement}_{i,x'} \in \tau$ for all $x' > x$ where $|x' - x|$ is small enough. Applying this argument inductively, we can conclude that in fact $B^{\complement}_{i,x'} \in \tau$ for all $x' > x$. But, for large enough $x'$ we have that $\left|B^{\complement}_{i,x'}\right| < a$ which is a contradiction to the fact that $\tau$ is a tangle.
Thus we may suppose without loss of generality that our point $x$ with $B^{\complement}_{i,x} \in \tau$ but $\mu \in B_{i,x}$ is chosen such that there is some $\epsilon > 0$ with the property that the cut $\{B_{i,x+\epsilon'}, B^{\complement}_{i,x+\epsilon'}\}$ is not oriented by $\tau$ for any $\epsilon' < \epsilon$. This means that $x$ was chosen as large as possible with the property that $\{B_{i,x}, B^{\complement}_{i,x}\}$ is still oriented differently by $\tau$ and $\mu$.
Now this choice of $x$ implies that $\mu_i < x \leqslant \nu_i$ as otherwise, if $\mu_i \leqslant \nu_i < x$, the cut $\{B_{i,x+\epsilon}, B^{\complement}_{i,x+\epsilon}\}$ would, for any $\epsilon > 0$, have lower costs than $\{B_{i,x}, B^{\complement}_{i,x}\}$, contradicting the choice of $x$. Now, if $|\mu_i - \nu_i| > 2\sigma$, for $x' = \mu_i + \frac{|\mu_i - \nu_i|}{2}$ the cut $\{B_{i,x'}, B^{\complement}_{i,x'}\}$ is a local minimum and it is easy to see that $\tau$ contains $B^{\complement}_{i,x'}$ and thus this is a local minimum cut of the type that we assumed does not exists. Hence we have that $|\mu_i - \nu_i| \leqslant 2\sigma$.
Thus the density function along the $i$-axis has a unique local maximum between $\mu_i$ and $\nu_i$, and the cost of $\{B_{i,\mu_i}, B^{\complement}_{i,\mu_i}\}$ is less than the cost of $\{B_{i,x}, B^{\complement}_{i,x}\}$. Now for any $y \in \mathbb{R}$, the cut $\{B_{j,y}, B^{\complement}_{j,y}\}$ along the $j$-axis (recall that $j$ was chosen such that $|\mu_j - \nu_j|$ is maximal) has lower cost than $\{B_{i,\mu_i}, B^{\complement}_{i,\mu_i}\}$ as the cost of such a cut is maximal if $y$ equals either $\mu_j$ or $\nu_j$ and in that case we still have that $|y - \nu_j| > 2\sigma > |\mu_i - \nu_i|$ or $|y - \mu_j| > 2\sigma > |\mu_i - \nu_i|$. However, this implies that $\tau$ needs to orient $\{B_{j,y}, B^{\complement}_{j,y}\}$ for every $y \in \mathbb{R}$ which is not possible without violating the consistency condition. Hence the assumption that we do not have a local minimum cut in $\tau$ pointing away from $\mu$ results in a contradiction to the fact that $\tau$ is a tangle.
So, there indeed needs to be a local minimum cut in $\tau$ which points away from $\mu$. Similarly, we find a local minimum cut in $\tau$ which points away from $\nu$. As $j$ was chosen such that $|\mu_j - \nu_j|$ is maximal, $\tau$ in particular needs to orient the locally minimal cut $\{B_{j,l}, B^{\complement}_{j,l}\}$ in dimension $j$, since this cut is the local minimum cut of lowest possible cost. So let us suppose wlog. that $\tau$ orients this cut the same way as $\mu$, that is, $\tau$ contains $B_{j,l}$ and $\mu_j < l < \nu_j$.
Further let $\{B_{i,x}, B^{\complement}_{i,x}\}$ be a local minimum cut that is oriented differently by $\tau$ and $\mu$, so that $B^{\complement}_{i,x} \in \tau$ but $\mu_i < x$.
Since $\{B_{i,x}, B^{\complement}_{i,x}\}$ is a local minimum, we know that $|x - \mu_i| > \sigma$, as $|x - \mu_i| = |x - \nu_i|$ and $|\mu_i - \nu_i| > 2\sigma$ by the fact that there is a local minimum cut along the $i$-axis. Moreover, we can consider the cut $\{B_{j,z}, B^{\complement}_{j,z}\}$ corresponding to the point $z := \mu_j - |x - \mu_i|$ in dimension

$j$, which, by Assumption 2, is also oriented by $\tau$ since the density along the $j$-axis at $z$ is less than the density along the $i$-axis at $x$. Since $\tau$ is a tangle, it needs to be the case that $B^{\complement}_{j,z}$ is contained in $\tau$: If $B_{j,z} \in \tau$ this contradicts consistency, as $\tau$, by Assumption 2, also orients all the cuts $\{B_{j,y}, B^{\complement}_{j,y}\}$ for $y \leqslant z$.

Furthermore let us consider the point $z' := \mu_j + |x - \mu_i|$. As $|x - \mu_i| < \frac{|\nu_j - \mu_j|}{2}$ we have that $|z' - \mu_j| = |x - \mu_i|$ and $|z' - \nu_j| \geqslant |x - \nu_i|$, and thus Assumption 2 again ensures that the cut $\{B_{j,z'}, B^{\complement}_{j,z'}\}$ in dimension $j$ corresponding to $z'$ has lower cost than $\{B_{i,x}, B^{\complement}_{i,x}\}$ and is thus oriented by $\tau$. Since $B_{j,l} \in \tau$ it needs to be the case that $B_{j,z'} \in \tau$ as otherwise, $\tau$ would by Assumption 2 also need to orient all the cuts $\{B_{j,y}, B^{\complement}_{j,y}\}$ for $z' \leqslant y \leqslant l$ which would then contradict consistency.

We can now use Assumption 3 to calculate how many points are contained in $B^{\complement}_{j,z} \cap B_{j,z'} \cap B^{\complement}_{i,x}$. Let us first consider the points from $\mu$. Let us denote the fraction of points from $\mu$ contained in $B^{\complement}_{i,x}$ as $p$, so $p = \left|V_\mu \cap B^{\complement}_{i,x}\right| \frac{2}{n}$. As $|x - \mu_i| > \sigma$, we have that $p < 0.16$.

By the choice of $z$ we have that $B^{\complement}_{j,z}$ contains $(1-p)\frac{n}{2}$ points from $\mu$ and similarly, $B_{j,z'}$ contains a fraction of $(1-p)$ of the points from $\mu$, namely $(1-p)\frac{n}{2}$ points.

So in total, by Assumption 3, the set $B^{\complement}_{j,z} \cap B_{j,z'} \cap B^{\complement}_{i,x}$ contains $p(1-p)^2\frac{n}{2}$ points from $\mu$.

For the points from $\nu$ we observe that $B^{\complement}_{i,x}$ contains, by symmetry and the choice of $p$ exactly $(1-p)\frac{n}{2}$ points from $\nu$. $B_{j,z'}$ contains fewer points from $\nu$ than $B_{j,l}$ and $B_{j,l}$ contains $q \leqslant \frac{1}{2} \cdot \left(1 + \operatorname{erf}\left(\frac{-|\mu_j - \nu_j|}{(2\sqrt{2\sigma^2})}\right)\right)$ of the points from $\nu$. So, again using Assumption 3, we can bound the total number of points contained in $B^{\complement}_{j,z} \cap B_{j,z'} \cap B^{\complement}_{i,x}$ by $\frac{n}{2}(p(1-p)^2 + q(1-p))$. As $p \leqslant 0.16$ and $q < p$, this is maximized for $p = 0.16$, where we obtain a value of about $n \cdot (0.42q + 0.056)$, hence if $a > n \cdot (0.42q + 0.056)$, $\tau$ cannot be a tangle, as claimed. ■

**Questions in the Narcissistic Personality Inventory Dataset**

| | statement A | statement B |
|---|---|---|
| 1 | I have a natural talent for influencing people. | I am not good at influencing people. |
| 2 | Modesty doesn't become me. | I am essentially a modest person. |
| 3 | I would do almost anything on a dare. | I tend to be a fairly cautious person. |
| 4 | When people compliment me I sometimes get embarrassed. | I know that I am good because everybody keeps telling me so. |
| 5 | The thought of ruling the world frightens the hell out of me. | If I ruled the world it would be a better place. |
| 6 | I can usually talk my way out of anything. | I try to accept the consequences of my behavior. |
| 7 | I prefer to blend in with the crowd. | I like to be the center of attention. |
| 8 | I will be a success. | I am not too concerned about success. |
| 9 | I am no better or worse than most people. | I think I am a special person. |
| 10 | I am not sure if I would make a good leader. | I see myself as a good leader. |
| 11 | I am assertive. | I wish I were more assertive. |
| 12 | I like to have authority over other people. | I don't mind following orders. |
| 13 | I find it easy to manipulate people. | I don't like it when I find myself manipulating people. |
| 14 | I insist upon getting the respect that is due me. | I usually get the respect that I deserve. |
| 15 | I don't particularly like to show off my body. | I like to show off my body. |
| 16 | I can read people like a book. | People are sometimes hard to understand. |
| 17 | If I feel competent I am willing to take responsibility for making decisions. | I like to take responsibility for making decisions. |
| 18 | I just want to be reasonably happy. | I want to amount to something in the eyes of the world. |
| 19 | My body is nothing special. | I like to look at my body. |
| 20 | I try not to be a show off. | I will usually show off if I get the chance. |

| | statement A | statement B |
|---|---|---|
| 21 | I always know what I am doing. | Sometimes I am not sure of what I am doing. |
| 22 | I sometimes depend on people to get things done. | I rarely depend on anyone else to get things done. |
| 23 | Sometimes I tell good stories. | Everybody likes to hear my stories. |
| 24 | I expect a great deal from other people. | I like to do things for other people. |
| 25 | I will never be satisfied until I get all that I deserve. | I take my satisfactions as they come. |
| 26 | Compliments embarrass me. | I like to be complimented. |
| 27 | I have a strong will to power. | Power for its own sake doesn't interest me. |
| 28 | I don't care about new fads and fashions. | I like to start new fads and fashions. |
| 29 | I like to look at myself in the mirror. | I am not particularly interested in looking at myself in the mirror. |
| 30 | I really like to be the center of attention. | It makes me uncomfortable to be the center of attention. |
| 31 | I can live my life in any way I want to. | People can't always live their lives in terms of what they want. |
| 32 | Being an authority doesn't mean that much to me. | People always seem to recognize my authority. |
| 33 | I would prefer to be a leader. | It makes little difference to me whether I am a leader or not. |
| 34 | I am going to be a great person. | I hope I am going to be successful. |
| 35 | People sometimes believe what I tell them. | I can make anybody believe anything I want them to. |
| 36 | I am a born leader. | Leadership is a quality that takes a long time to develop. |
| 37 | I wish somebody would someday write my biography. | I don't like people to pry into my life for any reason. |
| 38 | I get upset when people don't notice how I look when I go out in public. | I don't mind blending into the crowd when I go out in public. |
| 39 | I am more capable than other people. | There is a lot that I can learn from other people. |
| 40 | I am much like everybody else. | I am an extraordinary person. |