Cambridge University Press & Assessment
978-1-009-47331-6 — Tangles
Reinhard Diestel
Excerpt
More Information

Part I

Tangles

A new paradigm for clusters and types

This first of the four main parts of the book offers a gentle and informal introduction to tangles.

We set out, however, not from what tangles are, but from where they might take us: what difference they might make to some fundamental methodological approaches in the natural sciences, in the social sciences, and in data science. Chapter 1 offers three separate introductions for readers from these three backgrounds. They can be read independently, and in any order. Readers are encouraged to read them all. For not only do they reflect the diversity of potential tangle applications, but through this diversity also highlight seemingly unrelated aspects of the notion of a tangle, the concept central to this book.

The notion of a tangle is then developed informally in Chapter 2. There will be ample reference back to the introductions from Chapter 1, to relate the rather abstract concept of tangles, as it is slowly developed, to their potential applications right away. Readers keen to see some concrete examples of how tangles might be applied, but less curious about the various aspects of the notion as such, may skip ahead to Part II after Section 2.3.

Chapter 3 offers a glimpse of what tangle *theory* has to offer in addition to just the notion of a tangle. The latter, however, goes a long way towards many tangle applications already. Chapter 3 can therefore be skipped by readers who feel sufficiently dedicated to read about tangle theory and its uses in Part III, where both the notion and the theory of tangles are developed rigorously at their simplest mathematical level.

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-47331-6 — Tangles
Reinhard Diestel
Excerpt
More Information

1
The idea behind tangles

This chapter offers three introductions to the concept and purpose of tangles: one for the natural sciences, one for the social sciences, and one for data science. These introductions can be read independently, and readers may choose any one of them as an entry point to this book, according to their own background.

However as all three introductions illuminate the same concept, readers from any background are likely also to benefit from the other two viewpoints. Indeed, while each of them may seem plausible enough on its own, they are rather different. The fact that they nevertheless describe the same concept, that of a tangle, illustrates better than any abstract discussion the breadth of this concept and its potential applications, including in fields not even touched upon here. Moreover, even in a given context where one of the three viewpoints seems more fitting than the other two, switching to one of those deliberately for a moment is likely to add insight that would otherwise be easy to miss.

1.1 Tangles in the natural sciences

Suppose we are trying to establish a possible common cause of some set of similar phenomena. To facilitate this, we may design a series of measurements to test various different aspects of each of these phenomena.[1]

If we already have an overview of all the potential causes, we might try to design these measurements so that each potential cause would yield a list of expected readings, one for each of these different measurements, so that different potential causes differ in at least one measurement.

Then only the true cause would be compatible with all the readings we get from our actual measurements performed on the phenomena we are trying to explain.

In our less-than-ideal world, it may not quite work like this. For a start, we might simply not be aware of all the potential causes – not to mention the fundamental issue of what, if anything, is a ‘cause’. Similar phenomena may have different causes, or no single cause. Even potentially single causes need not be mutually exclusive but may be able to co-exist; then we shall not be able to design experiments that will exclude all but one of them with certainty. And finally, measurements may be corrupted, but we may not know which ones were.

We usually try to compensate for this by building in some redundancy: perhaps by taking more measurements, or by measuring more different aspects. Or we might resign ourselves to making claims only in probability – which will protect us from being disproved by any single event, but which may also increase immensely the overheads needed to justify precise quantitative assertions (of probabilities).

> *Tangles offer a structural, rather than probabilistic, way to afford the redundancy needed in such cases. They allow us to derive predictions from our data as we would expect them from identifying causes, while sidestepping the philosophical issue of what constitutes a cause.*

The idea, at a high level, is to replace the search for ‘causes’ with a search for something we can observe directly: structure in our data that occurs *as a result* of the presence of an underlying cause – a kind of structural footprint in observable data that we find whenever phenomena have some common cause, no matter what that cause may be. Different causes have different footprints, but all causes have the same *structural type* of footprint: tangles. The idea is that the structural footprint of each particular cause should carry enough information to replace any reference in the scientific process to that cause, e.g. in making predictions, with a reference to this structure, the tangle that reflects this cause in observable data. If desired, we may think of tangles as an extensional definition of ‘cause’ that achieves precision and observability at the expense of the intuitive appeal of our informal notion of ‘cause’.

In our generic example, a tangle would be a set of *hypothetical* readings for the measurements we have performed on our phenomena, a set of one possible reading per measurement. It would not be just any such set, but one that is typical for the actual readings we got on our phenomena:

> *A tangle is a typical set of measurement readings, one for each measurement, such as a set of readings due to a particular cause.*

It may happen that one, or several, of our phenomena produced exactly this set of readings. But it can also happen that an ‘abstract’ set of readings is typical, and hence a tangle, for our collection of phenomena without occurring exactly in any one of them. This might be the case, for example, for a set τ of measurement readings produced by some given cause under laboratory conditions. This set of readings will be typical for our phenomena if they were indeed triggered by this cause, even if none of them produced exactly τ under our measurements.

What, then, does it mean that τ is typical for our phenomena? We shall address this question in detail below. Its most important aspect, however, is that our definition of ‘typical’ will *not* stipulate the existence of some common cause for our phenomena. It will be such that a single cause, or some fixed combination of causes, will produce a set of readings that satisfies our definition of ‘typical’. But the definition itself will refer only to our data: the measurement readings we obtained on the phenomena we are seeking to explain. This will allow us to identify tangles directly in the data, without guessing at possible causes, and then *from* these tangles infer that, perhaps, some known cause is present.

Just as a set of similar phenomena can have several possible causes, our measurements might indicate the presence of a single tangle, or several, or none. Given any one of these tangles, we may try to find a common cause for this typical set of readings, or choose not to try. If there *is* a common cause for sufficiently many of the phenomena investigated, it will show up as a tangle and can thus be identified.

But there can also be tangles that cannot, or not yet, be ‘explained’ by a common cause. Such tangles are just as substantial, and potentially useful, as those that can be labelled by a known common cause; indeed perhaps more so, since the absence of an obvious common cause may have left them unidentified in the past. In this sense, identifying tangles in large sets of similar phenomena can lead to the discovery of new metaphenomena that had previously gone unnoticed and might, henceforth, be interpreted as a ‘cause’ for the group of phenomena that gave rise to this tangle.

So when is a set τ of hypothetical measurement readings deemed 'typical' for the actual measurements taken on our phenomena, and is therefore a tangle? There are two notions of 'typical' that are important in tangle theory: a strong one that is satisfied by many tangles but not required in their definition, and a weaker one that is required in their definition, and which suffices to establish the main theorems about tangles.

The strong notion, which we might call *popularity-based*, is that our set of phenomena has a subset X – not too small – such that, for every measurement made, some healthy majority – more than two thirds – of the phenomena in X yielded the reading specified by τ for that measurement.[2] Note that these may be different two thirds of X for different measurements: every phenomenon, even one in X, may for *some* measurements produce readings different from the readings that τ specifies for those measurements. But every entry in τ reflects some aspect of what we measured that many of the phenomena in X have in common, and is in this sense typical for all the phenomena in X. Clearly, there can be several such tangles τ, witnessed by different sets X of phenomena.

The weaker notion of when our set τ of hypothetical measurement readings is 'typical' for the actual readings obtained from our measurements, and hence constitutes a tangle, might be called *consistency-based*. It requires of τ that, for every set of up to three of our measurements, at least one of the phenomena we measured gave the readings specified by τ for these three phenomena.[3] Note that if τ is typical in the popularity-based sense it will also be typical in this consistency-based sense,[4] but not conversely.

We often strengthen these consistency requirements of a tangle by asking a little more: that, for some 'agreement parameter' n we chose for the given context, every three measurement readings specified by τ must be shared not only by one of the phenomena we measured but by at least n of them. Such sets of hypothetical measurement readings, one for every measurement, will thus be even more typical for our phenomena, the more so the larger n is.

All three of these notions of 'typical' are robust against small changes in our data. This makes tangles well suited to 'fuzzy' data with the kind of inherent variation indicated earlier. But the definition of tangles will be completely precise: a formal description of the structure of our data *including* any aspects of its fuzziness. We shall therefore be able to use tangles in rigorous mathematical analysis of our data as it comes.

1.2 Tangles in the social sciences

Suppose we run a survey S of political questions on a population P of a thousand people. If there exists a group of, say, a hundred like-minded people among these, there should be a way of answering all the questions in S that is typical for those one hundred people: that, after all, is what 'like-minded' means. Let us write X for this set of a hundred people, and τ for the way of answering S that is typical for them. Thus, τ is one typical set of answers to all the questions in S.

Let us try to quantify this last assertion, that our answer set τ is typical for the people in X. One way to do this is to require that, for every question s in S, some healthy majority – more than two thirds, say – of the people in X agree with the answer to s that τ specifies. Note that *which* two thirds of X these are may differ for different questions s. Every answer specified by τ reflects the views of a large majority of the people in X, and is typical for all of X in this way. But we may not be able to pin down anyone in P, let alone two thirds of the people in X, that answered all of S as specified by τ.

We shall call such a complete collection τ of views that is typical for some of the people in P a *mindset tangle*. There may be more than one mindset tangle for S, or none, just as there may be several groups of like-minded people in P, or none, each with their own typical way of answering S.

Traditionally, mindsets are found intuitively: they are first guessed, and only then established by quantitative evidence, perhaps even from a survey designed specifically to test them. Mindset tangles can do that, too. For example, we might feel that there is a 'socialist' way σ of answering our survey S. We could write these answers down without looking at the actual returns for S, just appealing to our intuitive notion of what a 'socialist' way to answer S would be. To test our intuition that this is indeed a typical mindset for the people in P we might then check whether, in the actual returns for S we received from the people in P, we can find a sizeable subset X of P as earlier for this particular $\tau = \sigma$.

But tangles can also do the converse: we can identify mindsets, as tangles, in the returns for S without having to guess their answers first:

> *Tangles offer a precise and quantitative way to test for suspected mindsets in a population and to discover unknown ones.*

For example, tangle analysis of political polls in the UK in the years before the Brexit referendum might have detected the existence of an unknown mindset of voters across the familiar political spectrum that helped establish the surprise majority for Brexit in 2016. And similarly in the US with the MAGA[5] mindset before 2016, or that of conservative Greens in Europe in the early 1970s. Tangles can identify previously unknown patterns of coherent views or behaviour.

1.3 Tangles in data science

One of the most basic, and at the same time most elusive, tasks in the analysis of big datasets is *clustering*: given a large set of points in some space, one seeks to identify within this set a small number of subsets, called 'clusters', of points that are in some sense similar. If we visualize similarity as distance, clusters will be sets of points that are, somehow, close to each other.

Figure 1.1 shows a simple example of points in the plane. In the picture on the left we can clearly see four clusters. Or can we? If a cluster is a set of points that are pairwise close, and the two points shown in green in the right half of the picture lie in the same cluster, should not the two red points – which are much closer – lie in a common cluster too?

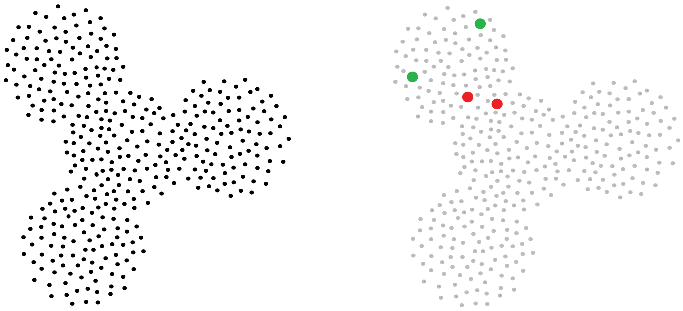

FIGURE 1.1. Four clusters?

For reasons such as this, and other more subtle ones, there is no universal notion of cluster in data science. In our example there are 'clearly' four clusters – but it is hard to come up with an abstract definition of 'cluster' that is satisfied by exactly four sets of points in Figure 1.1, let alone four sets resembling those that we intuitively see as clusters.

Tangles seek to describe clusters in an entirely different manner. Not by dividing the dataset into subsets in some clever new way, but without dividing it up at all: although there will be four tangles in our picture, these will not be defined as sets of points. In particular, questions such as whether the green points should end up in the same cluster but the red points, perhaps, should not, do not even arise.

By avoiding the issue of assigning points to clusters altogether, tangles can be precise without making arbitrary and unwarranted choices:

Tangles offer a precise, if indirect, way to identify fuzzy clusters.

Rather than looking for dense clouds of data points, tangles look for the converse: for obvious 'bottlenecks' at which the dataset naturally splits in two. We call ways of splitting our dataset into two disjoint subsets *partitions* of the set, and the two subsets the *sides* of the partition.

Figure 1.2 shows three partitions of our point set at bottlenecks.[6] Now, whatever formal definition of 'cluster' one might choose to work with, one thing will be clear: no bottleneck partition will divide any cluster roughly in half, since that should violate either the definition of a cluster or that of a bottleneck. For example, given one of the three bottlenecks in our picture, and one of the four obvious clusters, we might argue over a few points about whether they should count as belonging to that cluster or not, or on which side of the bottleneck they lie. But for almost all the points in our picture these questions will have a clear answer once we consider a fixed cluster and a fixed bottleneck, no matter how exactly these may be defined.

Put another way, whichever precise definition of cluster (and of bottleneck) someone chose to work with, each of our four intuitive clusters would lie *mostly* on the same side of any partition at a bottleneck. Let us then say that the cluster *orients* this partition towards the side on which most of it lies. Figure 1.2 shows how the central cluster, no matter how it was defined precisely, orients the partitions at the three bottlenecks in

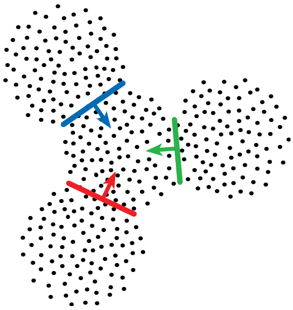

FIGURE 1.2. Orienting the bottlenecks consistently towards the central cluster.

this way. Each of the four clusters assigns its own set of arrows to these same three partitions, and the central cluster orients them all inwards.

Note that assignments of arrows to bottlenecks that come from one of the four clusters in this way are not arbitrary: the arrows are consistent in that they all point roughly the same way, towards that cluster.

The key idea behind tangles, now, is to keep for each cluster exactly this information – how it orients all the bottleneck partitions – and to forget everything else (such as which points might belong to it). More precisely, tangles will be *defined* as such abstract objects: as *consistent orientations of all the bottleneck partitions* in a data set.

In this way, tangles will extract from the various explicit ways of defining clusters as point sets something like their common essence. Tangles will be robust against small changes in the data, just as they are robust against small changes in any explicit definition of point clusters that we might use to specify them. But their definition as such will be perfectly precise, and involve no arbitrary choices of the kind one invariably has to make when one tries to define clusters as sets of points.

To make this approach work, of course, one has to define formally what the 'bottleneck partitions' of a given dataset are, and when an orientation of all these bottleneck partitions is deemed to be 'consistent'. In our example of Figure 1.2 we defined both these with reference to those four intuitive clusters. Indeed, as 'bottleneck partitions' we took those that split the set of points where this appeared narrow in the picture, which is just another way of saying that we took precisely those partitions that did not cut right through any of the four intuitive clusters; and we called a way of orienting these three partitions 'consistent' if the arrows indicating this pointed towards one of those intuitive clusters.

If we are serious about defining tangles as abstract objects, however, in a bid to bypass the difficulties inherent in trying to define clusters as point sets, then this would beg the question. *The challenge is to define both bottleneck partitions and consistency of their orientations without reference to any perceived cluster*, however vaguely defined. Only once we have achieved this can we 'define' clusters not explicitly as point sets but indirectly as tangles, as is our aim.[7]

To make this challenge a little clearer, let us look at a slightly modified example. Figure 1.3 again shows four clusters with three bottlenecks. This time, one of these has an elongated shape, like a handle. As before, there are clearly four clusters, yet there is no obvious way to define them directly as point sets.