# Master Thesis Alexander Conzelmann

Ulrike von Luxburg / Solveig Klepper, David Künstle

November 11, 2021

## 1 Tangles on triplet data

A triplet is a tuple of three elements for which we can only access the relationships between the elements. The approach is comparison based and usually does not include quantitative information. For example consider a tuple (a, b, c), the triplet information could tell us if a is closer to b or to c but no absolut distances. To better understand triplet data we can use machine learning to process and aggregate the triplet information. With tangles we consider an algorithm to cluster a set of datapoints using only their triplet information without any embedding as a preprocessing. In first toy examples this approach looks promising. In real world datasets there are obstacles to overcome. Usually the data is very sparse and we only have triplet information of a small subset of the triplets, as well as noise.
In this work the focus is on experimenting with different approaches to overcome these obstacles and evaluation empirically and investigate from a theoretical perspective the different properties and biases of this approach.

**Getting started:**

- As a first step you should get familiar with the algorithm and the implementation:
  - To do so we suggest you experiment on a simple toy example:
    - Generate a set $N$ of $n$ datapoints from a mixture of gaussians. Each mixture component represents one ground truth cluster.
    - Based on a distance function you want to consider a set of triplets $(a, b, c) \in N \times N \times N$, $a \neq b \neq c$. For every triplet you observe the information if $dist(a, b) \leq dist(a, c)$.
    - Based on this triplet information you build a questionnaire. For every possible pair of the datapoints $\{b, c\} \in N \times N, b \neq c$ represents a question. Every point $a \in N$ answers the corresponding question $dist(a, b) \leq dist(a, c)$. The result is a questionnaire with $n$ 'participants' and $\binom{n}{2}$ 'questions'. Each question now represents one bipartition of the data and we can use the Hamming similarity to assign a cost.
  - How stable is the performance considering hyper parameter choices as the agreement parameter (and parameters of the data generation process, e.g. dimension, distance of means, number of components, noise (in terms of variance of the gaussians as well as 'wrong' answers to the questionnaire))?
  - To evaluate your results you want to choose different metrics for the performance of the algorithm. As a first step we suggest to stick to the hard clustering output of the tangle algorithm and standard clustering metrics.
- As a next step we want to move towards real world applications looking at sparse datasets. (1) Given your derived questionnaire, we only get $d$ percent of the answers (density $d$). (2) We want to sample efficiently, so we can choose $d$ percent of pairs ('questions') for which we know the 'answers' for all datapoints. Can we still cluster the dataset in these scenarios? There are two possible directions:
  - Heuristics to preprocessing the input data, e.g. fill the missing information with random samples (randomly select an answert) or neutral answers (randomly choose 0.5 for every missing entry).
    You maybe want to read into some other approaches to deal with missing data in other settings and try to adapt these to tangles.
  - Adaption of the tangle algorithm (evaluate at which steps in the pipeline we need the data and how can we adapt the algorithm to apply to questionnaires with missing data?

  Implement and evaluate different options. (Visualize your experiments.)

**To move on:** There are several interesting questions which we could consider from here on. The idea is to choose one or two and investigate them in greater detail. For all of them we suggest to start with experiments to get a better understandig and then try to analyse the behaviour theoretically.

- Investigate the tangles algorithm on real world triplet data.
  - How well do tangles perform on triplet data with noise and missing triplets.
  - What does a "question" mean in this setting and how to interpret the results? This question is especially interesting for the soft clustering case!
  - The insights might suggest better ways to preprocess triplets to bipartitons.
  - Consider their theoretical implications and think about the bias they might induce.
- How well does the triplet-tangle clustering work in comparison to other triplet-based clustering algorithms and standard clustering methods, applied on the ordinal embedding of the triplets?
- How many triplets do we need to recover the "ground truth" for clustering?
  There are theoretical results on embedding triplets. Maybe we can extend them to clustering with the tangle algorithm.
- Which inductive bias does the tangle algorithm induce (also considering the hyper parameters)? What do we implicitly assume about the underlaying data?
- If tangles perform well on triplet data, can we use triplets as a "cut finding strategy"? In practice we have a triplet dataset. That is we inherently only consider information of the form $dist(a, b) < dist(a, c)$. For tangle sone of the major questions is how to generate a representative set of initial cuts (if not given). Can we is use data (like the GMM or a graph) to generate (all or a 'good' subset of) triplets and the corresponding questionnaire and use this information to cluster? Again whats the bias we introduce?