Tangles: from weak to strong clustering

or: Our adventure in machine learning

D Fiovoranti, S Klepper, L Rendsburg, U von Luxburg, E, K, T

Tangle definition

Given a set S of bipartitions (cuts), a **tangle** is a set τ which contains exactly one side of each bipartition such that

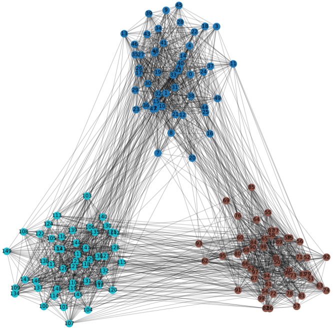
$$|A \cap B \cap C| \geq a \quad \forall A, B, C \in \tau.$$

a: **Agreement** parameter

Stochastic Block Model

k **blocks** of equal size $\frac{n}{k}$

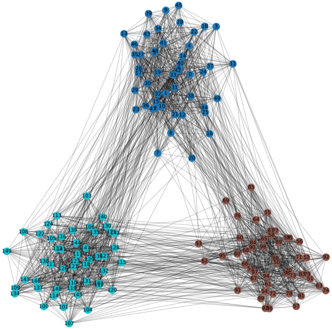
Edges within blocks with probability p, between blocks with probability $q < p$



Stochastic Block Model

k **blocks** of equal size $\frac{n}{k}$

Edges within blocks with probability p, between blocks with probability $q < p$



Consider all cuts up to order Ψ

$|A, A^{\complement}| := |E(A, A^{\complement})|$

When are the blocks (distinct) tangles?

When are there no other tangles?

SBM (Expectation Case)

2 blocks of equal size $\frac{n}{2}$

Edges within blocks with **weight** p, between blocks with weight q

All cuts up to order Ψ

$|A, A^{\complement}| := \sum_{a \in A, b \in A^{\complement}} \text{w(a,b)}$

SBM (Expectation Case)

2 blocks of equal size $\frac{n}{2}$

Edges within blocks with **weight** p, between blocks with weight q

All cuts up to order Ψ

$|A, A^{\complement}| := \sum_{a \in A, b \in A^{\complement}}$ w(a,b)

SBM (Expectation Case)

2 blocks of equal size $\frac{n}{2}$

Edges within blocks with **weight** p, between blocks with weight q

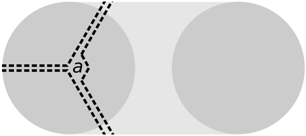
All cuts up to order Ψ

$|A, A^{\complement}| := \sum_{a\in A, b\in A^{\complement}}$ w(a,b)

SBM (Expectation Case)

2 blocks of equal size $\frac{n}{2}$

Edges within blocks with **weight** p, between blocks with weight q



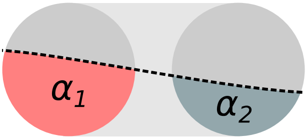
All cuts up to order Ψ

$|A, A^{\complement}| := \sum_{a \in A, b \in A^{\complement}} \mathrm{w(a,b)}$

SBM (Expectation Case)

2 blocks of equal size $\frac{n}{2}$

Edges within blocks with **weight** p, between blocks with weight q



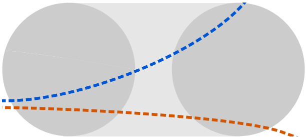
All cuts up to order Ψ
$|A, A^{\complement}| := \sum_{a \in A, b \in A^{\complement}}$ w(a,b)

$$\frac{n^2}{4}\left(p(\alpha_1 - \alpha_1^2 + \alpha_2 - \alpha_2^2) + q(\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2)\right)$$

SBM (Expectation Case)

2 blocks of equal size $\frac{n}{2}$

Edges within blocks with **weight** p, between blocks with weight q



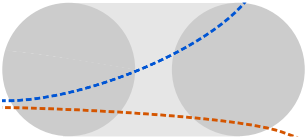
All cuts up to order Ψ

$|A, A^{\complement}| := \sum_{a \in A, b \in A^{\complement}} \mathrm{w(a,b)}$

SBM (Expectation Case)

2 blocks of equal size $\frac{n}{2}$

Edges within blocks with **weight** p, between blocks with weight q



All cuts up to order Ψ

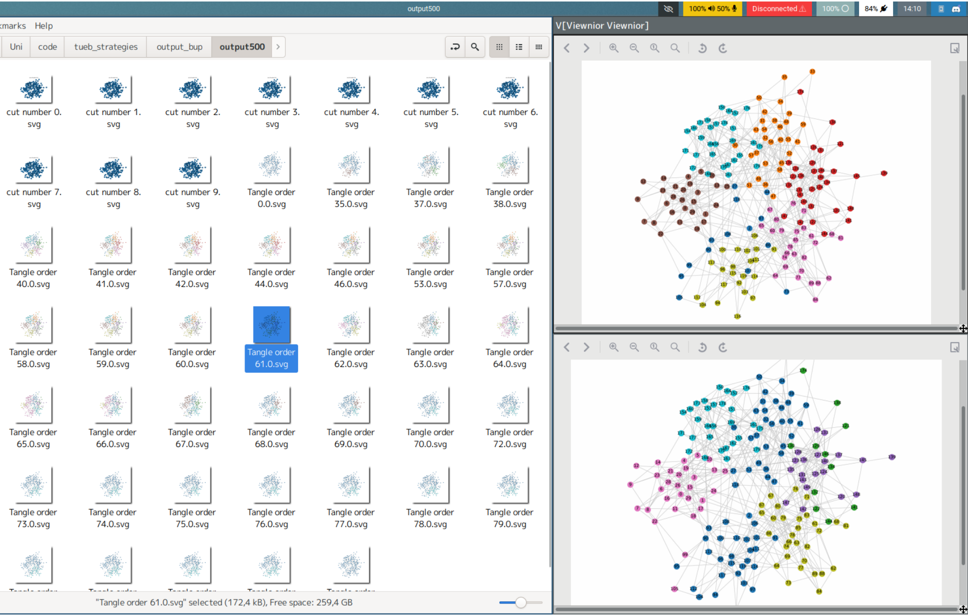
$|A, A^{\complement}| := \sum_{a \in A, b \in A^{\complement}} \mathrm{w(a,b)}$

$$\sum_{\substack{(i,j) \\ (\frac{i}{n}, \frac{j}{n}) \in A_\Psi}} \binom{n_1}{i} \cdot \binom{n_2}{j} \ll \sum_{\substack{(i,j) \\ (\frac{i}{\delta n}, \frac{j}{\delta n}) \in A_\Psi}} \binom{n_1}{i} \cdot \binom{n_2}{j}$$

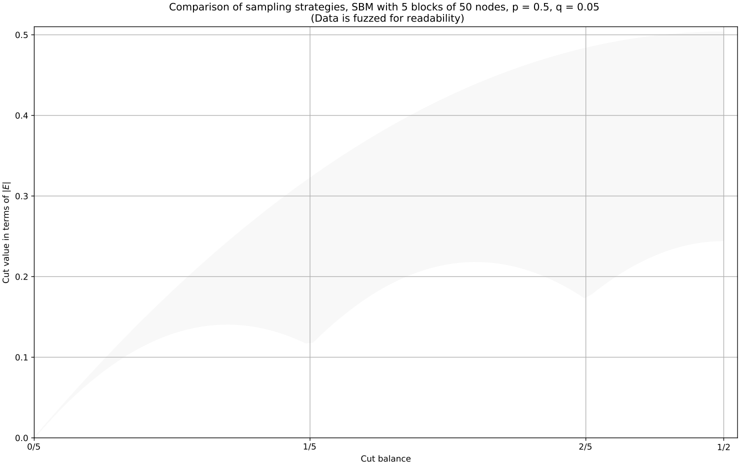
How do we sample good cuts?

How do we sample good cuts?

How do we evaluate?


output500
100% 50%
Disconnected
100%
84%
14:10
kmarks Help
Uni
code
tueb_strategies
output_bup
output500
cut number 0.svg
cut number 1.svg
cut number 2.svg
cut number 3.svg
cut number 4.svg
cut number 5.svg
cut number 6.svg
cut number 7.svg
cut number 8.svg
cut number 9.svg
Tangle order 0.0.svg
Tangle order 35.0.svg
Tangle order 37.0.svg
Tangle order 38.0.svg
Tangle order 40.0.svg
Tangle order 41.0.svg
Tangle order 42.0.svg
Tangle order 44.0.svg
Tangle order 46.0.svg
Tangle order 53.0.svg
Tangle order 57.0.svg
Tangle order 58.0.svg
Tangle order 59.0.svg
Tangle order 60.0.svg
Tangle order 61.0.svg
Tangle order 62.0.svg
Tangle order 63.0.svg
Tangle order 64.0.svg
Tangle order 65.0.svg
Tangle order 66.0.svg
Tangle order 67.0.svg
Tangle order 68.0.svg
Tangle order 69.0.svg
Tangle order 70.0.svg
Tangle order 72.0.svg
Tangle order 73.0.svg
Tangle order 74.0.svg
Tangle order 75.0.svg
Tangle order 76.0.svg
Tangle order 77.0.svg
Tangle order 78.0.svg
Tangle order 79.0.svg
"Tangle order 61.0.svg" selected (172,4 kB), Free space: 259,4 GB
V[Viewnior Viewnior]

Cut finding strategies



Karger’s algorithm

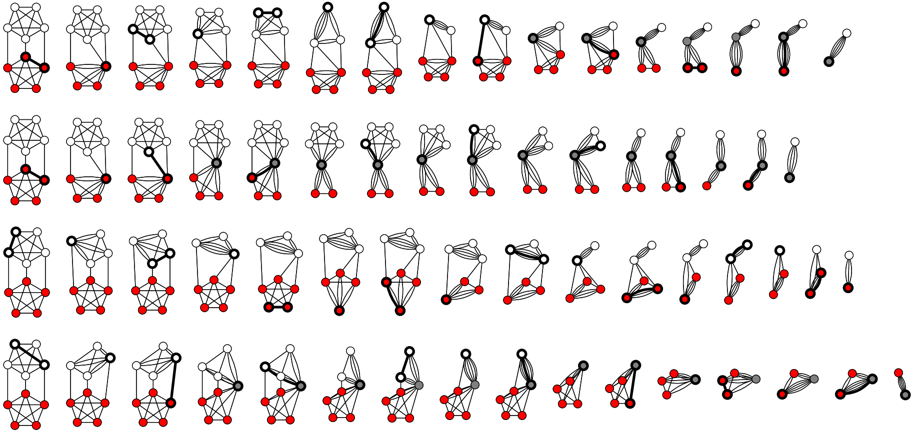
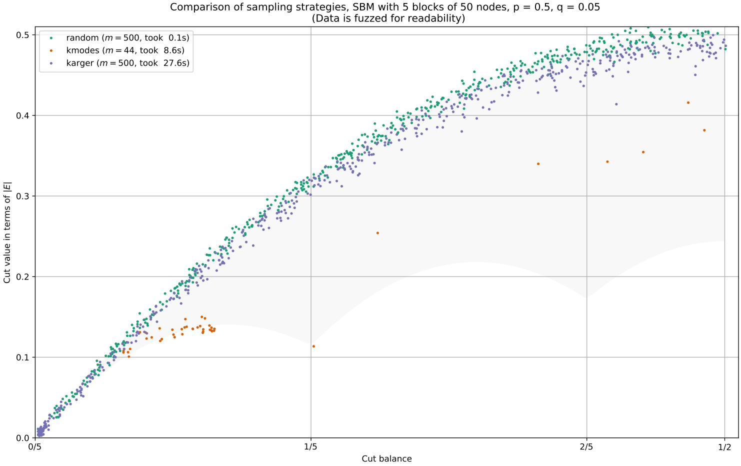
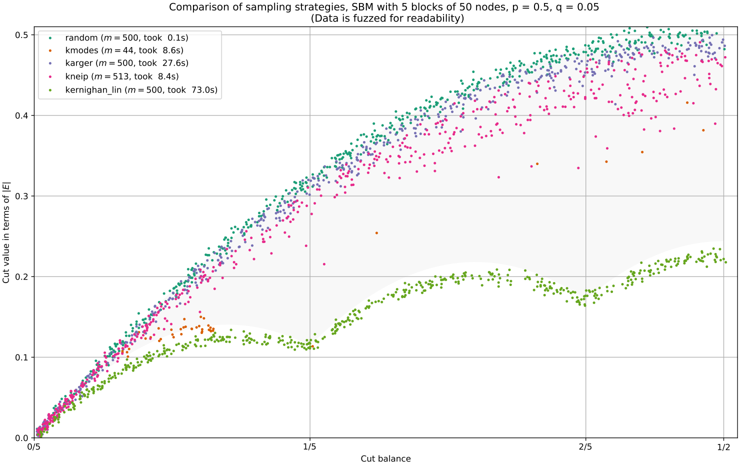


Image by Thore Husfeldt for Wikimedia Commons, Creative Commons BY-SA

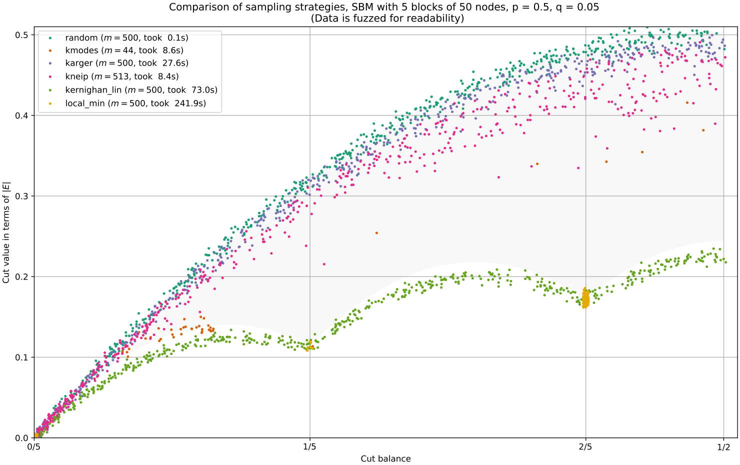
Cut finding strategies



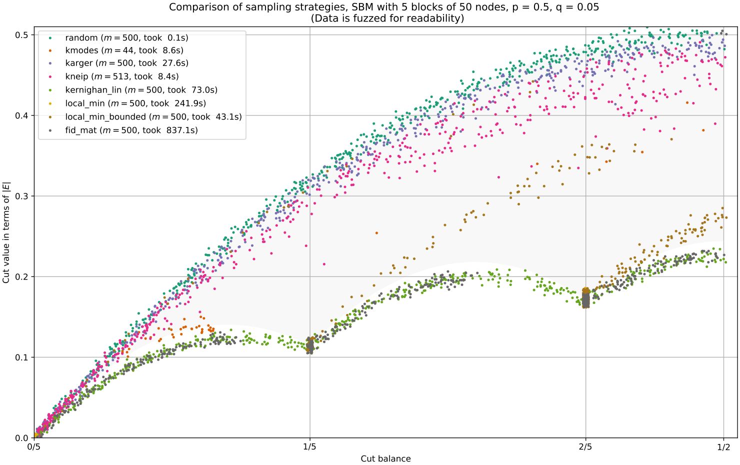
Cut finding strategies



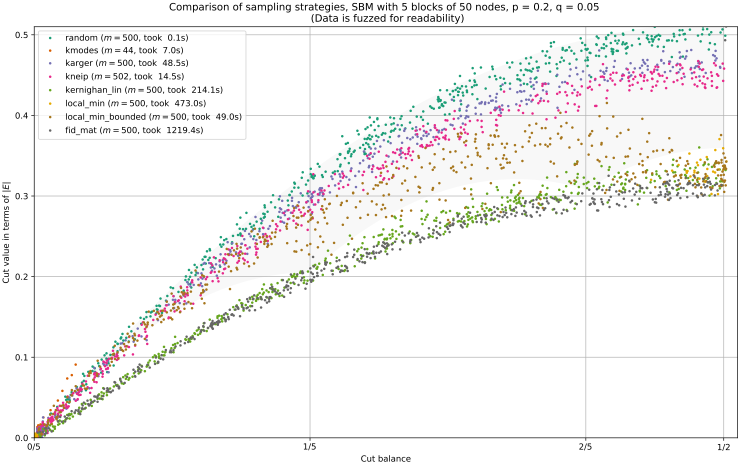
Cut finding strategies



Cut finding strategies



Cut finding strategies



The mindset model

A ‘typical’ pattern of answering a questionaire.

The mindset model

k mindsets, m questions, n people

Step 1: Sample k template vectors $\mu_1, \ldots, \mu_k \in \{0, 1\}^m$ **(mindsets)**

Step 2: For each μ_i, a set of $\frac{n}{k}$ people answers as μ_i does, but deviates on each question independently with probability $p < \frac{1}{2}$

The mindset model

k mindsets, m questions, n people

Step 1: Sample k template vectors $\mu_1, \ldots, \mu_k \in \{0,1\}^m$ **(mindsets)**

Step 2: For each μ_i, a set of $\frac{n}{k}$ people answers as μ_i does, but deviates on each question independently with probability $p < \frac{1}{2}$

Cuts are induced by questions.

The mindset model

k mindsets, m questions, n people

Step 1: Sample k template vectors $\mu_1, \ldots, \mu_k \in \{0, 1\}^m$ **(mindsets)**

Step 2: For each μ_i, a set of $\frac{n}{k}$ people answers as μ_i does, but deviates on each question independently with probability $p < \frac{1}{2}$

Cuts are induced by questions.

When *are* the mindsets tangles?

When are there no other tangles?

Stochastics

Everything's just Bernoulli random variables.

Binomial distributions are well understood.

Stochastics

If $1 - 3p > ka/n$ then with probability at least $1 - km\exp(-2n(\frac{ka}{n} - 1 + 3p)^2\frac{1}{9k})$ **every mindset is a tangle**.

Stochastics

If $1 - 3p > ka/n$ then with probability at least $1 - km\exp(-2n(\frac{ka}{n} - 1 + 3p)^2\frac{1}{9k})$ **every mindset is a tangle**.

If $p \leq a/n$ then with probability at least $1 - mk\exp(-\frac{2n}{k}(p - \frac{ka}{n})^2)$ **every triple with large intersection comes from a mindset**.

Stochastics

If $1 - 3p > ka/n$ then with probability at least $1 - km\exp(-2n(\frac{ka}{n} - 1 + 3p)^2\frac{1}{9k})$ **every mindset is a tangle**.

If $p \leq a/n$ then with probability at least $1 - mk\exp(-\frac{2n}{k}(p - \frac{ka}{n})^2)$ **every triple with large intersection comes from a mindset**.

But how do we turn this into 'Every tangle is a mindset'?

The problem

Suppose we have these mindsets:

$$
\begin{aligned}
&(1, 1, 1,\ 0, 0, 0,\ 0, 0, 0,\ 0, 0, 0)\\
&(0, 0, 0,\ 1, 1, 1,\ 0, 0, 0,\ 0, 0, 0)\\
&(0, 0, 0,\ 0, 0, 0,\ 1, 1, 1,\ 0, 0, 0)\\
&(0, 0, 0,\ 0, 0, 0,\ 0, 0, 0,\ 1, 1, 1)
\end{aligned}
$$

The problem

Suppose we have these mindsets:

$$(1, 1, 1,\ 0, 0, 0,\ 0, 0, 0,\ 0, 0, 0)$$
$$(0, 0, 0,\ 1, 1, 1,\ 0, 0, 0,\ 0, 0, 0)$$
$$(0, 0, 0,\ 0, 0, 0,\ 1, 1, 1,\ 0, 0, 0)$$
$$(0, 0, 0,\ 0, 0, 0,\ 0, 0, 0,\ 1, 1, 1)$$

Then we also get a tangle for

$$(0, 0, 0,\ 0, 0, 0,\ 0, 0, 0,\ 0, 0, 0)$$

The problem

Suppose we have these mindsets:

$$(1,1,1,\ 0,0,0,\ 0,0,0,\ 0,0,0)$$
$$(0,0,0,\ 1,1,1,\ 0,0,0,\ 0,0,0)$$
$$(0,0,0,\ 0,0,0,\ 1,1,1,\ 0,0,0)$$
$$(0,0,0,\ 0,0,0,\ 0,0,0,\ 1,1,1)$$

Then we also get a tangle for

$$(0,0,0,\ 0,0,0,\ 0,0,0,\ 0,0,0)$$

Assumption. If $\tau \in \{0,1\}^m$ satisfies that for all $x, y, z \leq m$ there exists a mindset μ such that $\tau(x) = \mu_i(x)$ as well as $\tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, then τ is a mindset, i.e. $\tau = \mu_j$ for some j.

How often is this satisified?

Assumption. If $\tau \in \{0,1\}^m$ satisfies that for all $x, y, z \leq m$ there exists a mindset μ such that $\tau(x) = \mu_i(x)$ as well as $\tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, then τ is a mindset, i.e. $\tau = \mu_j$ for some j.

How often is this satisified?

Assumption. If $\tau \in \{0,1\}^m$ satisfies that for all $x, y, z \leq m$ there exists a mindset μ such that $\tau(x) = \mu_i(x)$ as well as $\tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, then τ is a mindset, i.e. $\tau = \mu_j$ for some j.

Easily holds if every partition of the mindsets is induced by a question.

How often is this satisified?

Assumption. If $\tau \in \{0,1\}^m$ satisfies that for all $x, y, z \leq m$ there exists a mindset μ such that $\tau(x) = \mu_i(x)$ as well as $\tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, then τ is a mindset, i.e. $\tau = \mu_j$ for some j.

Easily holds if every partition of the mindsets is induced by a question.

This is bound to happen as $m \to \infty$.

How often is this satisified?

Assumption. If $\tau \in \{0,1\}^m$ satisfies that for all $x, y, z \leq m$ there exists a mindset μ such that $\tau(x) = \mu_i(x)$ as well as $\tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, then τ is a mindset, i.e. $\tau = \mu_j$ for some j.

Easily holds if every partition of the mindsets is induced by a question.

This is bound to happen as $m \to \infty$.

Caveat: This requires m to be exponential in k.

How often is this satisified?

Assumption. If $\tau \in \{0,1\}^m$ satisfies that for all $x, y, z \leq m$ there exists a mindset μ such that $\tau(x) = \mu_i(x)$ as well as $\tau(y) = \mu_i(y)$ and $\tau(z) = \mu_i(z)$, then τ is a mindset, i.e. $\tau = \mu_j$ for some j.

Easily holds if every partition of the mindsets is induced by a question.

This is bound to happen as $m \to \infty$.

Caveat: This requires m to be exponential in k.

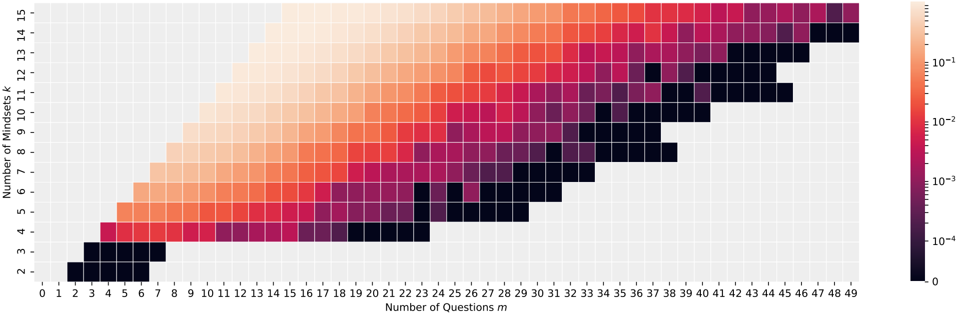
Theorem. Asympotically, m has to be exponential in k, or else the assumption fails with high probability.

How often is it *really* satisified?

Realistically $k \leq 15$.

How often is it *really* satisified?

Realistically $k \leq 15$.



Back to experiments

How do we evaluate the quality of our clustering numerically?

Back to experiments

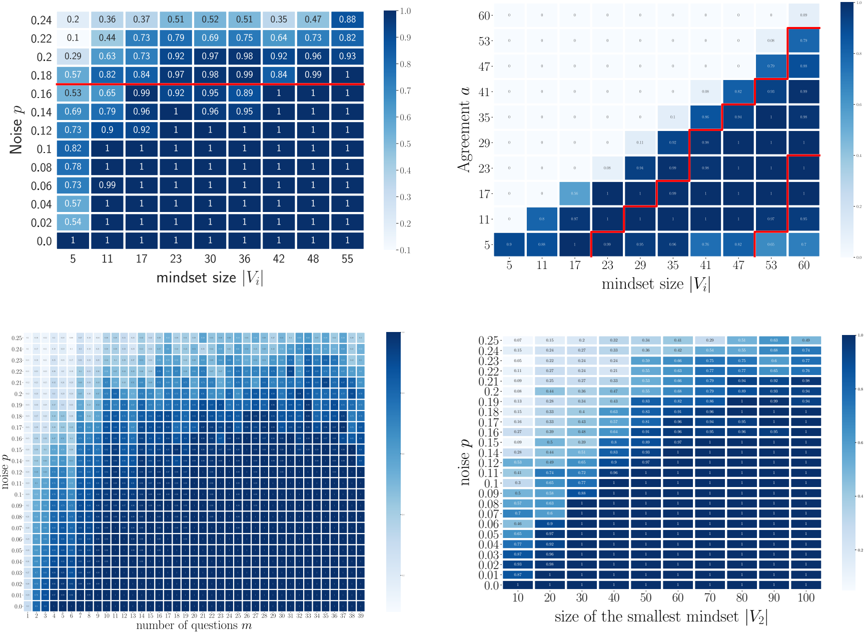
How do we evaluate the quality of our clustering numerically?

Turn it into a hard clustering. Count the number of wrongly separated pairs. Adjust for expectation. ($\leadsto$ Adjusted Rand Index)

Dimensions

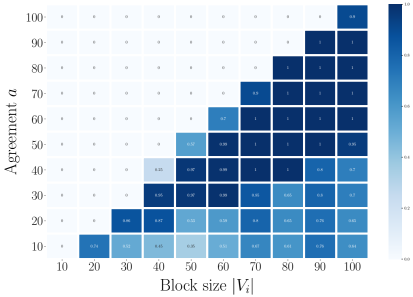
k: number of mindsets
m: number of questions
n: number of people
p: noise probability
a: tangle agreement
additional noise questions

A 6-dimensional space that needs to be explored!



Noise p
mindset size $|V_i|$
Agreement a
mindset size $|V_i|$
noise p
number of questions m
noise p
size of the smallest mindset $|V_2|$

The same goes for the SBM

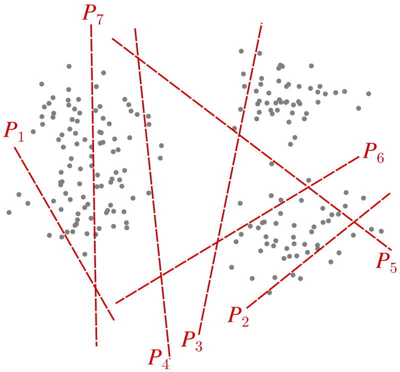


Visualizing tangles

Suppose our data points are embedded in the plane

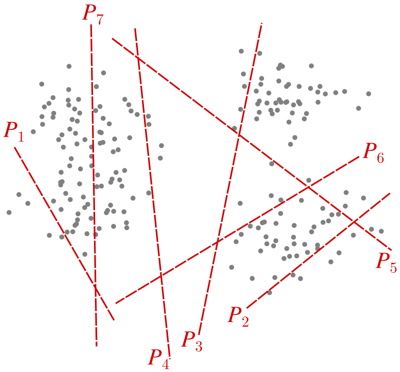
Visualizing tangles

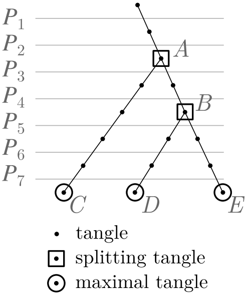
Suppose our data points are embedded in the plane



Visualizing tangles

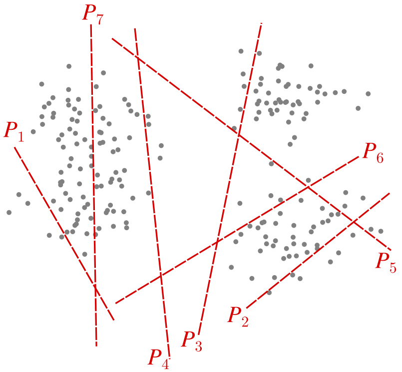
Suppose our data points are embedded in the plane

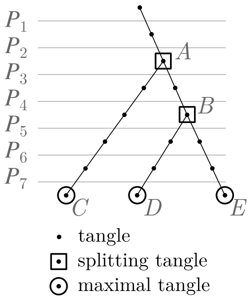


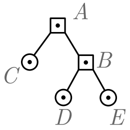


Visualizing tangles

Suppose our data points are embedded in the plane

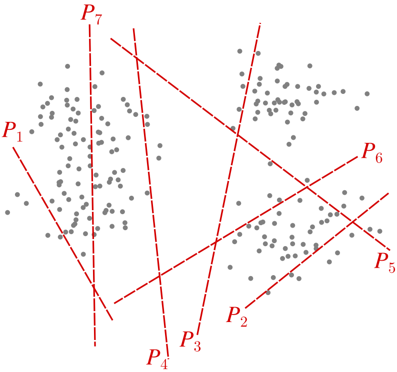


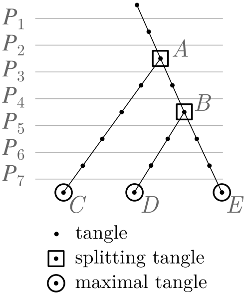


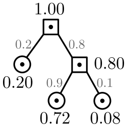


Visualizing tangles

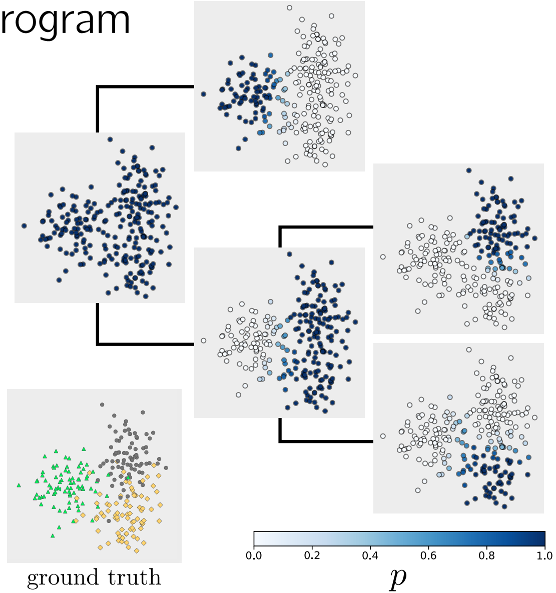
Suppose our data points are embedded in the plane



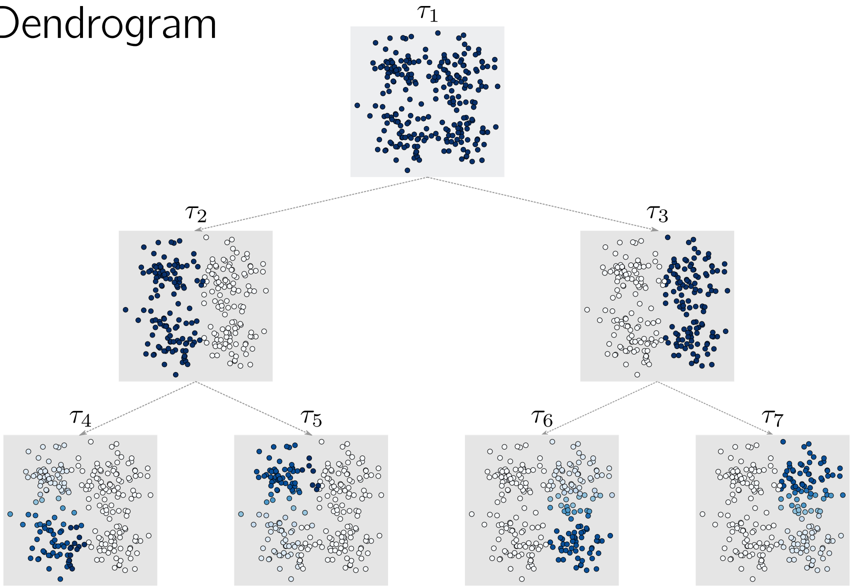


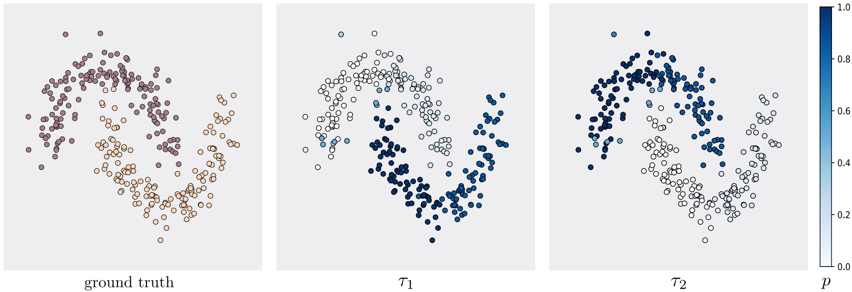


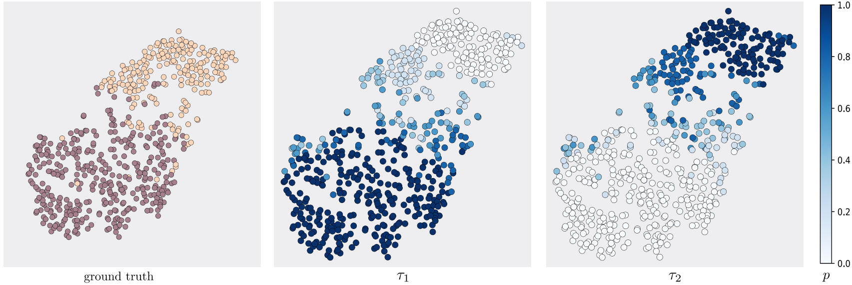
Dendrogram



Dendrogram




ground truth
τ_1
τ_2
p
1.0
0.8
0.6
0.4
0.2
0.0


ground truth
τ_1
τ_2
p
1.0
0.8
0.6
0.4
0.2
0.0

Thank you!