SUPPLEMENTARY MATERIAL

6 PROOFS REGARDING THE ACCURACY OF DBMSTCLU

6.1 PROOF OF THEOREM 3.2

This theorem relies on the following lemma:

Lemma 6.1. *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K clusters $C_1^*, \ldots, C_K^*$ and $\mathcal{T}$ an MST of $\mathcal{G}$. If for all $i \in [K]$, C_i^* is weakly homogeneous, then $\operatorname*{argmax}_{e\in\mathcal{T}} w(e) \subset Cut_{\mathcal{G}}(\mathcal{T})$ i.e. the heaviest edges in $\mathcal{T}$ are in $Cut_{\mathcal{G}}(\mathcal{T})$.*

Proof. Let us consider C_i^* a cluster of $\mathcal{G}$. As C_i^* is weakly homogeneous, $\forall j \in [K]$ s.t. $e^{(ij)} \in Cut_{\mathcal{G}}(\mathcal{T})$, $\max_{e\in\mathcal{T}_{|C_i^*}} w(e) < w(e^{(ij)})$. Hence, $\operatorname*{argmax}_{e\in E(\mathcal{T})} w(e) \subset Cut_{\mathcal{G}}(\mathcal{T})$. □

Theorem. *3.2 Let us consider a graph $\mathcal{G} = (V, E, w)$ with K homogeneous clusters $C_1^*, \ldots, C_K^*$ and $\mathcal{T}$ an MST of $\mathcal{G}$. Let now assume that at step $k < K-1$,* DBMSTCLU *built $k+1$ subtrees $\mathcal{C}_1, \ldots, \mathcal{C}_{k+1}$ by cutting $e_1,\ e_2,\ \ldots,\ e_k \in E$.*

Then, $Cut_k := Cut_{\mathcal{G}}(\mathcal{T}) \setminus \{e_1,\ e_2,\ \ldots,\ e_k\} \neq \emptyset \implies \mathrm{DBCVI}_{k+1} \geq DBCVI_k$, i.e. if there are still edges in Cut_k, the algorithm will continue to perform some cut.

Proof. Let note DBCVI at step k, $DBCVI_k = \sum_{i=1}^{k+1} \frac{|\mathcal{C}_i|}{N} V_C(\mathcal{C}_i)$. Let assume that $Cut_k \neq \emptyset$. Therefore, there is $e^* \in Cut_k$ and $i \in \{1,\ \ldots,\ k+1\}$ s.t. $e^* \in E(\mathcal{C}_i)$. Since $e^* \in Cut_{\mathcal{G}}(\mathcal{T})$, using Lem. 6.1, one can always take $e^* \in \operatorname*{argmax}_{e\in E(\mathcal{C}_i)} w(e)$. Then, if we denote $\mathcal{C}_i^1$, $\mathcal{C}_i^2$ the two subtrees of $\mathcal{C}_i$ induced by the cut of e^* (see Fig. 1 for an illustration) and $DBCVI_{k+1}(e^*)$ the associated DBCVI value,

$$\begin{aligned}\Delta &= DBCVI_{k+1}(e^*) - DBCVI_k \\ &= \frac{|\mathcal{C}_i^1|}{N} \underbrace{\left(\frac{\mathrm{SEP}(\mathcal{C}_i^1) - \mathrm{DISP}(\mathcal{C}_i^1)}{\max(\mathrm{SEP}(\mathcal{C}_i^1), \mathrm{DISP}(\mathcal{C}_i^1))}\right)}_{V_C(\mathcal{C}_i^1)} + \frac{|\mathcal{C}_i^2|}{N} \underbrace{\left(\frac{\mathrm{SEP}(\mathcal{C}_i^2) - \mathrm{DISP}(\mathcal{C}_i^2)}{\max(\mathrm{SEP}(\mathcal{C}_i^2), \mathrm{DISP}(\mathcal{C}_i^2))}\right)}_{V_C(\mathcal{C}_i^2)} - \frac{|\mathcal{C}_i|}{N} \underbrace{\left(\frac{\mathrm{SEP}(\mathcal{C}_i) - \mathrm{DISP}(\mathcal{C}_i)}{\max(\mathrm{SEP}(\mathcal{C}_i), \mathrm{DISP}(\mathcal{C}_i))}\right)}_{V_C(\mathcal{C}_i)}.\end{aligned}$$

There are two possible cases:

1. $V_C(\mathcal{C}_i) \leq 0$, then $\mathrm{SEP}(\mathcal{C}_i) \leq \mathrm{DISP}(\mathcal{C}_i) = w(e^*)$. As for $l \in \{1, 2\}$, $\mathrm{SEP}(\mathcal{C}_i^l) \geq \mathrm{SEP}(\mathcal{C}_i)$ and $\mathrm{DISP}(\mathcal{C}_i^l) \leq \mathrm{DISP}(\mathcal{C}_\rangle)$ because $e^* \in \operatorname*{argmax}_{e\in E(\mathcal{C}_\rangle)} w(e)$, then, for $l \in \{1, 2\}$,

   $$\frac{\mathrm{SEP}(\mathcal{C}_i^l) - \mathrm{DISP}(\mathcal{C}_i^l)}{\max(\mathrm{SEP}(\mathcal{C}_i^l), \mathrm{DISP}(\mathcal{C}_i^l))} \geq \frac{\mathrm{SEP}(\mathcal{C}_l) - \mathrm{DISP}(\mathcal{C}_i)}{\max(\mathrm{SEP}(\mathcal{C}_i), \mathrm{DISP}(\mathcal{C}_i))} = \frac{\mathrm{SEP}(\mathcal{C}_i)}{w(e)} - 1$$

   and $\Delta \geq 0$.

2. $V_C(\mathcal{C}_i) \geq 0$, then $\mathrm{SEP}(\mathcal{C}_i) \geq \mathrm{DISP}(\mathcal{C}_i) = w(e^*)$ i.e. $\max(\mathrm{SEP}(\mathcal{C}_i), \mathrm{DISP}(\mathcal{C}_i)) = \mathrm{SEP}(\mathcal{C}_i)$, for $l \in \{1, 2\}$, $\mathrm{DISP}(\mathcal{C}_i^l) \leq \mathrm{DISP}(\mathcal{C}_i)$ i.e. $\mathrm{DISP}(\mathcal{C}_i^l) \leq w(e^*)$, $\mathrm{SEP}(\mathcal{C}_i^l) = w(e^*)$ hence $\mathrm{SEP}(\mathcal{C}_i^l) \geq \mathrm{DISP}(\mathcal{C}_i^l)$. Thus, $V_C(\mathcal{C}_i) = 1 - \frac{\mathrm{DISP}(\mathcal{C}_i)}{\mathrm{SEP}(\mathcal{C}_i)}$ and for $l \in \{1, 2\}$, $V_C(\mathcal{C}_i^l) = 1 - \frac{\mathrm{DISP}(\mathcal{C}_i^l)}{\mathrm{SEP}(\mathcal{C}_i^l)}$. Then, for $l \in \{1, 2\}$, $V_C(\mathcal{C}_i^l) \geq V_C(\mathcal{C}_i)$ and $\Delta \geq 0$.

For both cases, $\Delta = DBCVI_{k+1}(e^*) - DBCVI_k \geq 0$. Hence, at least the cut of e^* improves the current DBCVI, so the algorithm will perform a cut at this stage.

□

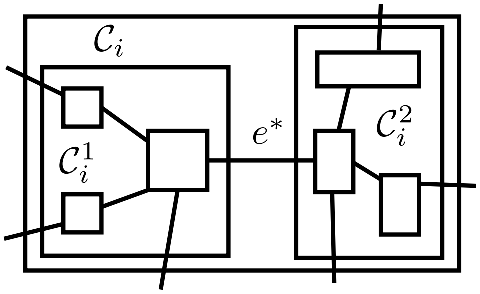


Figure 1: Illustration for Th. 3.2's proof.

6.2 PROOF OF THEOREM 3.3

Theorem. *3.3 Let us consider a graph $\mathcal{G} = (V, E, w)$ with K homogeneous clusters $C_1^*, \ldots, C_K^*$ and $\mathcal{T}$ an MST of $\mathcal{G}$.*

Let now assume that at step $k < K-1$, DBMSTCLU *built $k+1$ subtrees $\mathcal{C}_1, \ldots, \mathcal{C}_{k+1}$ by cutting $e_1,\ e_2,\ \ldots,\ e_k \in E$. We still denote $Cut_k := Cut_{\mathcal{G}}(\mathcal{T}) \backslash \{e_1,\ e_2,\ \ldots,\ e_k\}$.*

Then, $Cut_k \neq \emptyset \implies \underset{e \in \mathcal{T} \backslash \{e_1,\ e_2,\ \ldots,\ e_k\}}{\text{argmax}} DBCVI_{k+1}(e) \subset Cut_k$ i.e. the edge that the algorithm cuts at step $k+1$ is in Cut_k.

Proof. It is sufficient to show that, at step k, if there exists an edge e^* whose cut builds two clusters, then e^* maximizes DBCVI among all possible cuts in the union of itself and both resulting clusters. Indeed, showing this for two clusters, one can easily generalize to the whole graph as a combination of couples of clusters (see Fig. 3 for an illustration): if for each couple, the best local solution is in Cut_k, then the best general solution is necessary in Cut_k.

Let us consider at step k of the algorithm two clusters C_1^* and C_2^* such that e^* the edge separating them in $\mathcal{T}$ is in Cut_k (see Fig. 2 for an illustration). For readability we denote $\mathcal{T}_{|C_1^*} = \mathcal{C}_1^*$ and $\mathcal{T}_{|C_2^*} = \mathcal{C}_2^*$ Let us proof that for all $\tilde{e} \in \mathcal{T}_{|C_1^* \cup C_2^*}$, one has: $DBCVI_{k+1}(e^*) > DBCVI_{k+1}(\tilde{e})$. W.l.o.g. let assume $\tilde{e} \in \mathcal{C}_1^*$ and let denote $\mathcal{C}_{1,1}^*$ and $\mathcal{C}_{1,2}^*$ the resulting subtrees from the cut of $\tilde{e}$. We still denote $DBCVI_{k+1}(e)$ the value of the DBCVI at step $k+1$ for the cut of e.

$$\begin{aligned}
\Delta &:= DBCVI_{k+1}(e^*) - DBCVI_{k+1}(\tilde{e}) \\
&= \underbrace{\frac{|\mathcal{C}_1^*|}{N}\left(\frac{\text{SEP}(\mathcal{C}_1^*) - \text{DISP}(\mathcal{C}_1^*)}{\max(\text{SEP}(\mathcal{C}_1^*), \text{DISP}(\mathcal{C}_1^*))}\right) + \frac{|\mathcal{C}_2^*|}{N}\left(\frac{\text{SEP}(\mathcal{C}_2^*) - \text{DISP}(\mathcal{C}_2^*)}{\max(\text{SEP}(\mathcal{C}_2^*), \text{DISP}(\mathcal{C}_2^*))}\right)}_{A} \\
&- \underbrace{\left(\frac{|\mathcal{C}_{1,1}^*|}{N}\left(\frac{\text{SEP}(\mathcal{C}_{1,1}^*) - \text{DISP}(\mathcal{C}_{1,1}^*)}{\max(\text{SEP}(\mathcal{C}_{1,1}^*), \text{DISP}(\mathcal{C}_{1,1}^*))}\right) + \frac{|\mathcal{C}_{1,2}^*|}{N}\left(\frac{\text{SEP}(\mathcal{C}_{1,2}^*) - \text{DISP}(\mathcal{C}_{1,2}^*)}{\max(\text{SEP}(\mathcal{C}_{1,2}^*), \text{DISP}(\mathcal{C}_{1,2}^*))}\right)\right)}_{B}
\end{aligned}$$

By weak homogeneity of C_1^* and C_2^*, $A = \frac{|\mathcal{C}_1^*|}{N}\left(1 - \frac{\text{DISP}(\mathcal{C}_1^*)}{\text{SEP}(\mathcal{C}_1^*)}\right) + \frac{|\mathcal{C}_2^*|}{N}\left(1 - \frac{\text{DISP}(\mathcal{C}_2^*)}{\text{SEP}(\mathcal{C}_2^*)}\right) > 0$

$$B = \underbrace{\frac{|\mathcal{C}_{1,1}^*|}{N}\left(\frac{\text{SEP}(\mathcal{C}_{1,1}^*) - \text{DISP}(\mathcal{C}_{1,1}^*)}{\max(\text{SEP}(\mathcal{C}_{1,1}^*), \text{DISP}(\mathcal{C}_{1,1}^*))}\right)}_{B_1} + \underbrace{\frac{|\mathcal{C}_{1,2}^*|}{N}\left(\frac{\text{SEP}(\mathcal{C}_{1,2}^*) - \text{DISP}(\mathcal{C}_{1,2}^*)}{\max(\text{SEP}(\mathcal{C}_{1,2}^*), \text{DISP}(\mathcal{C}_{1,2}^*))}\right)}_{B_2}$$

By Lem. 6.1, $e^* \in \underset{e \in E(\mathcal{T}_{|C_1^* \cup C_2^*})}{\text{argmax}} w(e)$ so $\text{DISP}(\mathcal{C}_{1,2}^*) = w(e^*)$.

Since $e^* \in Cut_{\mathcal{G}}(\mathcal{T})$, one has $w(e^*) \geq \max(\text{SEP}(\mathcal{C}_1^*), \text{SEP}(\mathcal{C}_2^*))$. Moreover, as $\mathcal{C}_2^*$ is a subtree of $\mathcal{C}_{1,2}^*$, then $\text{SEP}(\mathcal{C}_{1,2}^*) \leq \text{SEP}(\mathcal{C}_2^*)$. Thus, $w(e^*) \geq \text{SEP}(\mathcal{C}_{1,2}^*)$. Finally, $B_2 = \frac{|\mathcal{C}_{1,2}^*|}{N}\left(\frac{\text{SEP}(\mathcal{C}_{1,2}^*)}{\text{DISP}(\mathcal{C}_{1,2}^*)} - 1\right) \leq 0$.

Besides, $w(\tilde{e}) \leq \text{SEP}(\mathcal{C}_1^*) \implies \text{SEP}(\mathcal{C}_{1,1}^*) = w(\tilde{e}) \leq \max_{e \in E(\mathcal{C}_1^*)} w(e)$ and $\text{DISP}(\mathcal{C}_{1,1}^*) = \max_{e \in E(\mathcal{C}_{1,1}^*)} w(e) \geq \min_{e \in E(\mathcal{C}_1^*)} w(e)$. Then, two possibilities hold:

1. $B_1 < 0 \implies B < 0 < A$.
2. $B_1 \geq 0$, thus one has $B_1 = \frac{|\mathcal{C}_{1,1}^*|}{N}\left(1 - \frac{\text{DISP}(\mathcal{C}_{1,1}^*)}{\text{SEP}(\mathcal{C}_{1,1}^*)}\right) \leq \frac{|\mathcal{C}_{1,1}^*|}{N}\left(1 - \frac{\min_{e \in \mathcal{C}_1^*} w(e)}{\max_{e \in \mathcal{C}_1^*} w(e)}\right)$. Under weak homogeneity condition, there is: $\frac{\text{DISP}(C_1^*)}{\text{SEP}(C_1^*)} < \frac{\min_{e \in \mathcal{C}_1^*} w(e)}{\max_{e \in \mathcal{C}_1^*} w(e)}$. Thus,

$$\begin{aligned} B_1 &< \frac{|C_{1,1}^*|}{N}\left(1 - \frac{\text{DISP}(\mathcal{C}_1^*)}{\text{SEP}(\mathcal{C}_1^*)}\right) \\ &< \frac{|\mathcal{C}_1^*|}{N}\left(1 - \frac{\text{DISP}(\mathcal{C}_1^*)}{\text{SEP}(\mathcal{C}_1^*)}\right) \text{ because } \mathcal{C}_{1,1}^* \text{ is a subtree of } \mathcal{C}_1^* \\ &< A \end{aligned}$$

So, $B_1 + B_2 = B < A = DBCVI_{k+1}(e^*)$.

Since $B < A$, $\Delta > 0$ and e^* maximizes DBCVI among all possible cuts in the union of itself and both resulting clusters. Q.E.D. □

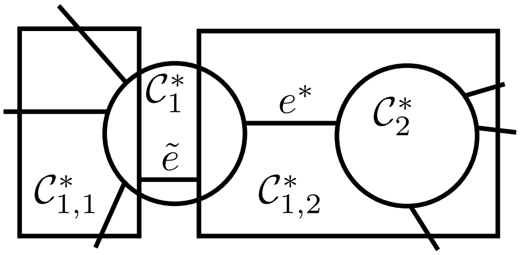


Figure 2: Illustration for Th. 3.3's proof.

6.3 PROOF OF THEOREM 3.4

Theorem. *3.4 Let us consider a graph $\mathcal{G} = (V, E, w)$ with K weakly homogeneous clusters $C_1^*, \ldots, C_K^*$ and $\mathcal{T}$ an MST of $\mathcal{G}$. Let now assume that at step $K-1$,* DBMSTCLU *built K subtrees $\mathcal{C}_1, \ldots, \mathcal{C}_K$ by cutting $e_1,\ e_2,\ \ldots,\ e_{K-1} \in E$. We still denote $Cut_{K-1} := Cut_{\mathcal{G}}(\mathcal{T}) \backslash \{e_1,\ e_2,\ \ldots,\ e_{K-1}\}$.*

Then, for all $e \in \mathcal{T} \backslash \{e_1,\ e_2,\ \ldots,\ e_{K-1}\}$, $DBCVI_K(e) < DBCVI_{K-1}$ i.e. the algorithm stops: no edge gets cut during step K.

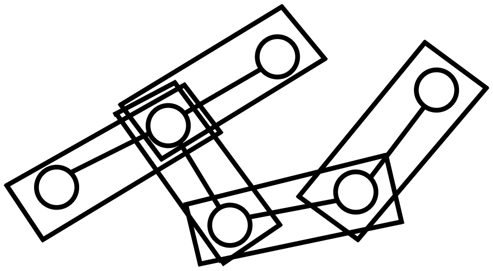


Figure 3: Illustration for Th. 3.3's proof. Each circle corresponds to a cluster. The six clusters are handled within five couples of clusters.

Proof. According to Th. 3.2 and Th. 3.3, for all $k < K$, if $Cut_k \neq \emptyset$, the algorithm performs some cut from $Cut_{\mathcal{G}}(\mathcal{T})$. We still denote for all $j \in [K]$ $\mathcal{C}_j^* = \mathcal{T}_{|C_j^*}$.Since $|Cut_{\mathcal{G}}(\mathcal{T})| = K-1$, the $K-1$ first steps produce $K-1$ cuts from $Cut_{\mathcal{G}}(\mathcal{T})$. Therefore, $\text{DBCVI}_{K-1} = \sum\limits_{j\in[K-1]} \frac{|\mathcal{C}_j^*|}{N} V_C(\mathcal{C}_j^*)$.

Let be e the (expected) edge cut at step K, splitting the tree $\mathcal{C}_i^*$ into $\mathcal{C}_{i,1}^*$ and $\mathcal{C}_{i,2}^*$.

$$\begin{aligned}
&\Delta = \text{DBCVI}_{K-1} - \text{DBCVI}_K \\
&= \frac{|\mathcal{C}_i^*|}{N} V_C(\mathcal{C}_i^*) - \frac{|\mathcal{C}_{i,1}^*|}{N} V_C(\mathcal{C}_{i,1}^*) - \frac{|\mathcal{C}_{i,2}^*|}{N} V_C(\mathcal{C}_{i,2}^*) \\
&= \frac{|\mathcal{C}_i^*|}{N} \frac{\text{SEP}(\mathcal{C}_i^*) - \text{DISP}(\mathcal{C}_i^*)}{\max(\text{SEP}(\mathcal{C}_i^*), \text{DISP}(\mathcal{C}_i^*))} - \frac{|\mathcal{C}_{i,1}^*|}{N} \frac{\text{SEP}(\mathcal{C}_{i,1}^*) - \text{DISP}(\mathcal{C}_{i,1}^*)}{\max(\text{SEP}(\mathcal{C}_{i,1}^*), \text{DISP}(\mathcal{C}_{i,1}^*))} - \frac{|\mathcal{C}_{i,2}^*|}{N} \frac{\text{SEP}(\mathcal{C}_{i,2}^*) - \text{DISP}(\mathcal{C}_{i,2}^*)}{\max(\text{SEP}(\mathcal{C}_{i,2}^*), \text{DISP}(\mathcal{C}_{i,2}^*))}
\end{aligned}$$

Since C_i^* is a weakly homogeneous cluster, therefore $\text{SEP}(\mathcal{C}_i^*) \geq \text{DISP}(\mathcal{C}_i^*)$. Then, minimal value of Δ, Δ_{min} is reached when $\text{SEP}(\mathcal{C}_{i,1}^*) \geq \text{DISP}(\mathcal{C}_{i,1}^*)$, $\text{SEP}(\mathcal{C}_{i,2}^*) \geq \text{DISP}(\mathcal{C}_{i,2}^*)$, $\text{SEP}(\mathcal{C}_{i,1}^*) = \text{SEP}(\mathcal{C}_{i,2}^*) = \min\limits_{e'\in E(\mathcal{C}_i^*)} w(e')$, $\text{DISP}(\mathcal{C}_{i,1}^*) = \text{DISP}(\mathcal{C}_{i,2}^*) = \max\limits_{e'\in E(\mathcal{C}_i^*)} w(e')$. Then,

$$\begin{aligned}
N \times \Delta_{min} &= |\mathcal{C}_i^*| \left(1 - \frac{\text{DISP}(\mathcal{C}_i^*)}{\text{SEP}(\mathcal{C}_i^*)}\right) - |\mathcal{C}_{i,1}^*| \left(1 - \frac{\text{DISP}(\mathcal{C}_{i,1}^*)}{\text{SEP}(\mathcal{C}_{i,1}^*)}\right) - |\mathcal{C}_{i,2}^*| \left(1 - \frac{\text{DISP}(\mathcal{C}_{i,2}^*)}{\text{SEP}(\mathcal{C}_{i,2}^*)}\right) \\
&= |\mathcal{C}_i^*| \left(1 - \frac{\text{DISP}(\mathcal{C}_i^*)}{\text{SEP}(\mathcal{C}_i^*)}\right) - |\mathcal{C}_{i,1}^*| \left(1 - \frac{\max\limits_{e'\in E(\mathcal{C}_i^*)} w(e')}{\min\limits_{e'\in E(\mathcal{C}_i^*)} w(e')}\right) - |\mathcal{C}_{i,2}^*| \left(1 - \frac{\max\limits_{e'\in E(\mathcal{C}_i^*)} w(e')}{\min\limits_{e'\in E(\mathcal{C}_i^*)} w(e')}\right) \\
&= |\mathcal{C}_i^*| \left(-\frac{\text{DISP}(\mathcal{C}_i^*)}{\text{SEP}(\mathcal{C}_i^*)} + \frac{\max\limits_{e'\in E(\mathcal{C}_i^*)} w(e')}{\min\limits_{e'\in E(\mathcal{C}_i^*)} w(e')}\right)
\end{aligned}$$

By weak homogeneity condition on C_i^*, $\frac{\text{DISP}(\mathcal{C}_i^*)}{\text{SEP}(\mathcal{C}_i^*)} < \frac{\min\limits_{e'\in E(\mathcal{C}_i^*)} w(e'))}{\max\limits_{e'\in E(\mathcal{C}_i^*)} w(e')} \leq \frac{\max\limits_{e'\in E(\mathcal{C}_i^*)} w(e'))}{\underset{e'\in E(\mathcal{C}_i^*}{)} \min w(e')}$. Therefore, $\Delta_{min} > 0$ and $\Delta > 0$. □

7 PROOFS REGARDING THE ACCURACY OF PTCLUST

7.1 PROOF OF THEOREM 3.8

Theorem. *3.8 Let us consider a graph $\mathcal{G} = (V, E, w)$ with K strongly homogeneous clusters $C_1^*, \dots, C_K^*$ and $\mathcal{T} = \text{PAMST}(\mathcal{G}, u_\mathcal{G}, w, \epsilon)$, $\epsilon > 0$. $\mathcal{T}$ has a partitioning topology with probability at least*

$$1 - \sum_{i=1}^{K} (|C_i^*| - 1) e^{-\frac{\epsilon}{2\Delta u_\mathcal{G}(|V|-1)}(\bar{\alpha}_i \max\limits_{e \in E(\mathcal{G}_{|C_i^*})}(w(e)) - \min\limits_{e \in E(\mathcal{G}_{|C_i^*})}(w(e))) + \ln(|E|)}$$

Proof. Let $\mathcal{T} = \text{PAMST}(\mathcal{G}, u_\mathcal{G}, w, \epsilon)$, $\{\mathcal{R}_1, ..., \mathcal{R}_{|V|-1}\}$ denotes the ranges used in the successive calls of the Exponential mechanism in $\text{PAMST}(\mathcal{G}, u_\mathcal{G}, w, \epsilon)$, $r_k = \mathcal{M}_{Exp}(\mathcal{G}, w, u_\mathcal{G}, \mathcal{R}_k, \underbrace{\frac{\epsilon}{|V| - 1}}_{\epsilon'})$, and Steps$(C_i^*)$ the set of steps k of the algorithm were $\mathcal{R}_k$ contains at least one edges from $\mathcal{G}_{|C_i^*}$. Finally for readability we denote $u_k = u_\mathcal{G}(w, r_k)$

$$\begin{aligned}
&\mathbb{P}[\mathcal{T} \text{ has a partitioning topology}] \\
=&\mathbb{P}[\underbrace{\forall i, j \in [K],\ i \neq j,\ |\{(u, v) \in E(\mathcal{T}),\ u \in C_i^*,\ v \in C_j^*\}| = 1}_{A}] = 1 - \mathbb{P}[\neg A]
\end{aligned}$$

If we denote $B =$ "$\forall i \in [K],\ \forall k > 1 \in$ Steps(C_i^*), if $r_{k-1} \in E(\mathcal{G}_{|C_i^*})$ then $r_k \in E(\mathcal{G}_{|C_i^*})$" One easily has: $B \implies A$, therefore $\mathbb{P}[\neg A] \leq \mathbb{P}(\neg B)$. Moreover, by using the privacy/accuracy trade-off of the exponential mechanism, one has

$$\forall t \in \mathbb{R}, \forall i \in [K], \forall k \in \text{Steps}(C_i^*)\ \mathbb{P}\left[\underbrace{u_k \leq -\frac{2\Delta u_\mathcal{G}}{\epsilon'}(t + \ln|\mathcal{R}_k|)}_{A_k(t)}\right] \leq \exp(-t).$$

Moreover one can major $\mathbb{P}[\neg B]$ as follows

$$\mathbb{P}\left[\exists i \in [K], \exists k \in \text{Steps}(C_i^*) \text{ s.t } r_{k-1} \in E(\mathcal{G}_{|C_i^*}) \text{ and } r_k \notin E(\mathcal{G}_{|C_i^*})\right]$$

By using the union bound, one gets

$$\leq \sum_{i \in [K]} \mathbb{P}\left[\exists k \in \text{Steps}(C_i^*) \text{ s.t } r_{k-1} \in E(\mathcal{G}_{|C_i^*}) \text{ and } r_k \notin E(\mathcal{G}_{|C_i^*})\right]$$

Using the strong homogeneity of the clusters, one has

$$\begin{aligned}
&= \sum_{i \in [K]} \mathbb{P}\left[\exists k \in \text{Steps}(C_i^*) \text{ s.t } u_k \leq -|\bar{\alpha}_i \max_{e \in E(\mathcal{G}_{|C_i^*})} w(e) - \min_{r \in \mathcal{R}_k} w(r)|\right] \\
&\leq \sum_{i \in [K]} \mathbb{P}\left[\exists k \in \text{Steps}(C_i^*) \text{ s.t } u_k \leq -|\bar{\alpha}_i \max_{e \in E(\mathcal{G}_{|C_i^*})} w(e) - \min_{e \in E(\mathcal{G}_{|C_i^*})} w(e)|\right]
\end{aligned}$$

By setting $t_{k,i} = \frac{\epsilon'}{2\Delta u_\mathcal{G}}(\bar{\alpha}_i \max\limits_{e \in E(\mathcal{G}_{|C_i^*})}(w(e)) - \min\limits_{e \in E(\mathcal{G}_{|C_i^*})}(w(e))) + \ln(|\mathcal{R}_k|)$ one gets

$$= \sum_{i \in [K]} \mathbb{P}\left[\exists k \in \text{Steps}(C_i^*) \text{ s.t } A_k(t_{k,i})\right]$$

Since for all $i \in [K]$, and $k \in \text{Steps}(C_i^*)$, $|\mathcal{R}_k| \leq |E|$, and using a union bound, one gets

$$\begin{aligned}
&\leq \sum_{i \in [K]} \sum_{k \in \text{Steps}(C_i^*)} \mathbb{P}\ [A_k(t_{k,i})] \leq \sum_{i \in [K]} \sum_{k \in \text{Steps}(C_i^*)} \exp\ (-t_{i,k}) \\
&\leq \sum_{i=1}^{K} (|C_i^*| - 1) \exp \left(-\frac{\epsilon}{2\Delta u_{\mathcal{G}}(|V| - 1)} \left(\bar{\alpha}_i \max_{e \in E(\mathcal{G}_{|C_i^*})} w(e) - \min_{e \in E(\mathcal{G}_{|C_i^*})} w(e) \right) + \ln(|E|) \right)
\end{aligned}$$

□

7.2 PROOF OF THEOREM 3.9

Let recall the theorem from S. Kotz on the Laplace distribution and generalizations (2001):

Theorem 7.1. *Let $n \in \mathbb{N}$, $(X_i)_{i \in [n]} \underset{iid}{\sim} Lap(\theta, s)$, denoting $X_{r:n}$ the order statistic of rank r one has for all $k \in \mathbb{N}$,*

$$\mathbb{E}\left[(X_{r:n} - \theta)^k\right] = s^k \underbrace{\frac{n!\Gamma(k+1)}{(r-1)!(n-r)!} \left((-1)^k \sum_{j=0}^{n-r} a_{j,r,k} + \sum_{j=0}^{r-1} b_{j,r,k} \right)}_{\alpha(n,k)}$$

Theorem. *3.9 Let us consider a graph $\mathcal{G} = (V, E, w)$ with K strongly homogeneous clusters $C_1^*, \dots, C_K^*$ and $T = \text{PAMST}(\mathcal{G}, u_{\mathcal{G}}, w, \epsilon)$, and $\mathcal{T}' = \mathcal{M}_{w.r}(T, w_{|T}, s, \tau, p)$ with $s << p, \tau$. Given some cluster C_i^*, and $j \neq i$ s.t $e^{(ij)} \in Cut_{\mathcal{G}}(\mathcal{T})$, if $H_{\mathcal{T}_{|C_i^*}}(e^{(ij)})$ is verified, then $H_{\mathcal{T'}_{|C_i^*}}(e^{(ij)})$ is verified with probability at least*

$$1 - \frac{\Lambda_1 + (\theta_{(ij)}^2 + \delta)\Lambda_2 - (\Lambda_3^2 + \theta_{(ij)}^2 \Lambda_4^2)}{\Lambda_1 + (\theta_{(ij)}^2 + \delta)\Lambda_2 + 2\Lambda_3\Lambda_4}$$

with the following notations:

- $\delta = \frac{s}{p}$, $\theta_{\min} = \frac{\min\limits_{e \in E(\mathcal{T})} w(e) + \tau}{p}$
- $\theta_{\max} = \frac{\max\limits_{e \in E(\mathcal{T})} w(e) + \tau}{p}$,$\theta_{(ij)} = \frac{w(e^{(ij)} + \tau}{p}$
- $\Lambda_1 = 24\delta^4 n\alpha(n,4) + 12\theta_{\max}\delta^3 n\alpha(n,3) + 12\theta_{\max}^2\delta^2 n\alpha(n,2) + 4\theta_{\max}^3\delta n\alpha(n,1) + \theta_{\max}^4$
- $\Lambda_2 = 2\delta^2 n\alpha(1,2) + 2\theta_{\min}\delta n\alpha(1,1) + \theta_{\min}^2$
- $\Lambda_3 = 2\delta^2 n\alpha(n,2) + 2\theta_{\max}\delta n\alpha(n,1) + \theta_{\max}^2$
- $\Lambda_4 = \delta n\alpha(1,1) + \theta_{\min}$

Proof. Let $\tau > 0$ and $p > 1$, according to the weight-release mechanism, all the randomized edge weights $w'(e)$ with $e \in E(\mathcal{T}')$ are sampled from independents Laplace distributions $Lap(\frac{w(e)+\tau}{p}, \frac{s}{p})$. Given some cluster C_i^*, and $j \neq i$ s.t $e^{(ij)} \in Cut_{\mathcal{G}}(\mathcal{T})$, $H_{\mathcal{T}_{|C_i^*}}(e^{(ij)})$ is verified. Finding the probability that $H_{\mathcal{T'}_{|C_i^*}}(e^{(ij)})$ is equivalent to find the probability $\mathbb{P}\left[\frac{(\max\limits_{e \in E(C_i^*)} X_e)^2}{\min\limits_{e \in E(C_i^*)} X_e} < X^{out}\right]$ with $X_e \underset{indep}{\sim} Lap(\frac{w(e)+\tau}{p}, \frac{s}{p})$ and $X^{out} \sim Lap(\frac{w(e^{(ij)})+\tau}{p}, \frac{s}{p})$. Denoting with $Y_i \underset{iid}{\sim} Lap\left(\theta_{\max}, \delta\right), Z_i \underset{iid}{\sim} Lap\left(\theta_{\min}, \delta\right)$ and $X^{out} \sim Lap(\theta_{(ij)}, \delta)$, one can lower bounded this probability by $\mathbb{P}\left[\frac{(\max\limits_{i\in[|C_i^*|-1]} Y_i)^2}{\min\limits_{i\in[|C_i^*|-1]} Z_i} < X^{out}\right]$. Choosing τ big enough s.t $\min\limits_{i\in[|C_i^*|-1]} Z_i < 0$ is negligible, one has

$$\begin{aligned}
&\mathbb{P}\left[\frac{(\max\limits_{i\in[|C_i^*|-1]} Y_i)^2}{\min\limits_{i\in[|C_i^*|-1]} Z_i} < X^{out}\right] \\
=&\mathbb{P}\left[\underbrace{(\max_{i\in[|C_i^*|-1]} Y_i)^2 - \min_{i\in[|C_i^*|-1]} Z_i \times X^{out}}_{\varphi} < 0\right].
\end{aligned}$$

Moreover since τ, $p \gg s$, one has $\mathbb{E}(\varphi) \leq 0$. Therefore,

$$\begin{aligned}
\mathbb{P}\left[\varphi < 0\right] =&\mathbb{P}\left[\varphi - \mathbb{E}(\varphi) < \underbrace{-\mathbb{E}(\varphi)}_{\geq 0}\right] \\
=&1 - \mathbb{P}\left[\varphi - \mathbb{E}(\varphi) > -\mathbb{E}(\varphi)\right]
\end{aligned}$$

Using the one-sided Chebytchev inequality, one gets

$$\geq 1 - \frac{\mathbb{V}(\varphi)}{\mathbb{V}(\varphi) + \mathbb{E}(\varphi)^2} = 1 - \frac{\mathbb{V}(\varphi)}{\mathbb{E}(\varphi^2)}$$

By giving an analytic form to $\mathbb{E}(\varphi)$ and $\mathbb{V}(\varphi)$ by using Theorem 7.1 one gets the expected result. □

8 FURTHER EXPERIMENTS

Both in private and non-private settings, another successful experiment has been conducted on the real NYC "Taxi & Limousine Commission Trip Record"[1] dataset taking the yellow and green taxis.

8.1 EXPERIMENTAL SETUP

From the dataset is built a graph, taking locations for vertices, edges for trips, and setting the weights by the number of trips between locations. Some preprocessing is made on the graph. Edges with strictly less than 140 trips are removed in order to sparsify the graph. Self-loops are removed since they are not supported by the clustering algorithm. Two points belonging to the same airport are also merged because the goal is to distinguish airport places from Manhattan and it was considered as noise. Finally, only the biggest connected component is kept since it is more relevant to perform the clustering on it. As a result, the considered connected graph contains $N = 162$ nodes and $m = 236$ edges. For our clustering algorithm, the weights on the edges should represent a dissimilarity between two vertices. So the following softmax-like transformation taken from the R package[2], representing a dissimilarity function, is made on the weights of the graph:

$$\forall i \in [m],\ w_i \leftarrow f(w_i - w_{min}) \in (0,1] \tag{1}$$

where $w_{min} = 140$ is the minimum weight in the graph and f is defined such that for all weight w:

$$f(w) = \frac{1}{1 + \exp\left(\frac{w-\mu}{\sigma \times \pi/2}\right)} \tag{2}$$

where μ is the average weight of the edges before, namely the average number of trips on the edges between the vertices and σ the standard deviation.

[1] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

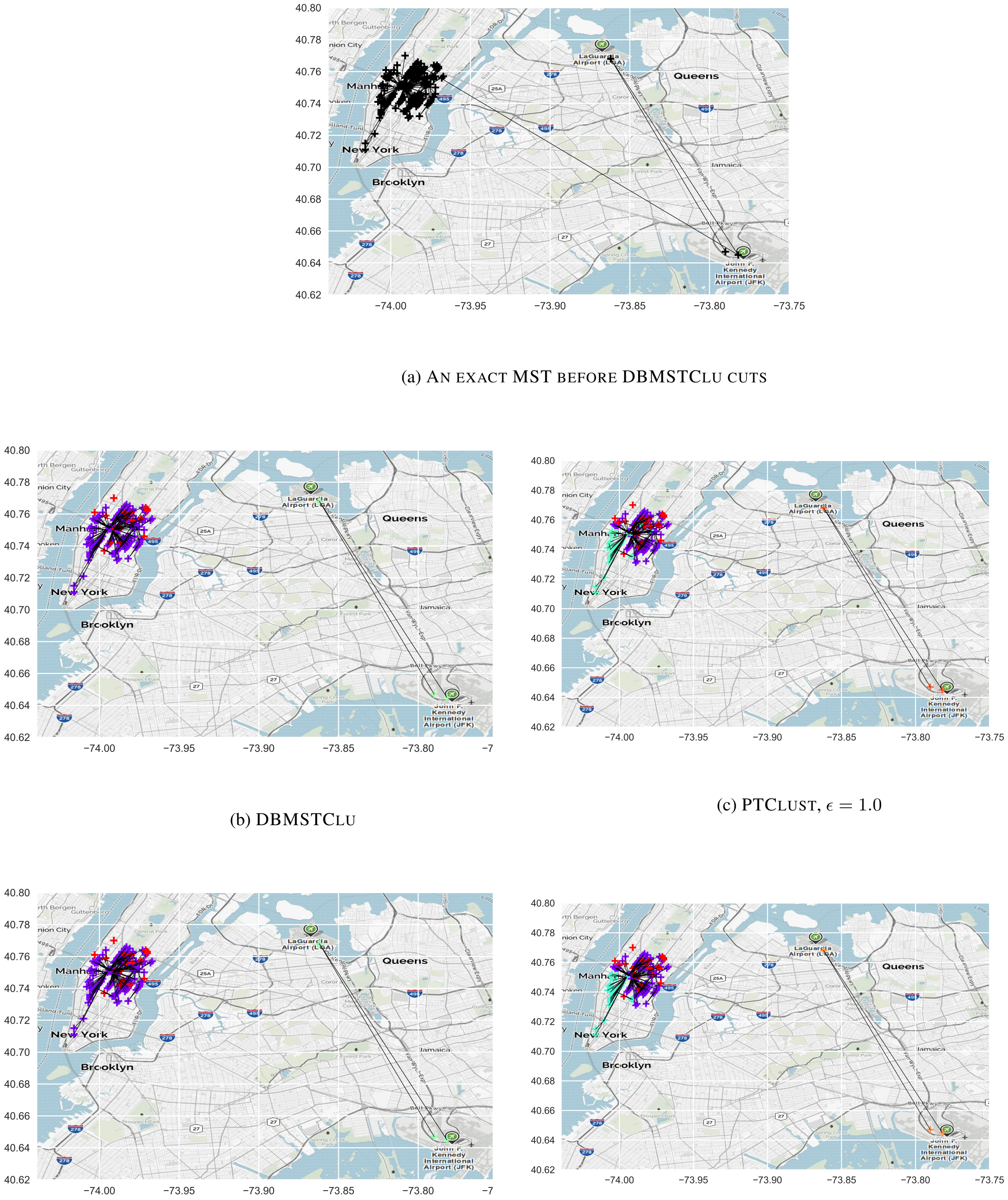
[2] https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SoftMax

In this framework, the graph topology is public, and the private information is carried by the weights. In fact, what an individual would want to be kept private is rather if he/she participated or not to a trip (thus keeping his/her location/moves private).

8.2 RESULTS

Figure 4a shows an exact MST that has been given as input to DBMSTCLU algorithm. The results of the latter are demonstrated in Figure 4b. Finally, Figures 4c, 4d and 4e exhibit the visualization results for PTCLUST algorithm with $\epsilon \in \{1.0, 0.7, 0.5\}$.

From only the number of trips made by yellow and green taxis between two locations (but not the GPS coordinates), the private and non-private clustering algorithms (respectively PTCLUST and DBMSTCLU) separate Manhattans island from the two neighboring airports. Even if the softmax-like function helps a lot, it is really important to emphasize that geographical information have never been used to obtain these clustering partitions, which is an unusual, but meaningful way of considering this kind of datasets.



(a) AN EXACT MST BEFORE DBMSTCLU CUTS

(b) DBMSTCLU

(c) PTCLUST, $\epsilon = 1.0$

(d) PTCLUST, $\epsilon = 0.7$

(e) PTCLUST, $\epsilon = 0.5$

Figure 4: NYC taxis experiments.