Graph-based clustering under differential privacy

Rafaël Pinot, Anne Morvan, Florian Yger, Cedric Gouy-Pailler, Jamal Atif

Graph-based Clustering under Differential Privacy

Rafael Pinot *[1,2], Anne Morvan †[1,2], Florian Yger[2], Cédric Gouy-Pailler[1], and Jamal Atif[2]

[1]CEA, LIST, 91191 Gif-sur-Yvette, France
[2]Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, 75016 Paris, France

March 10, 2018

Abstract

In this paper, we present the first differentially private clustering method for arbitrary-shaped node clusters in a graph. This algorithm takes as input only an approximate Minimum Spanning Tree (MST) $\mathcal{T}$ released under weight differential privacy constraints from the graph. Then, the underlying nonconvex clustering partition is successfully recovered from cutting optimal cuts on $\mathcal{T}$. As opposed to existing methods, our algorithm is theoretically well-motivated. Experiments support our theoretical findings.

1 Introduction

Weighted graph data is known to be a useful representation data type in many fields, such as bioinformatics or analysis of social, computer and information networks. More generally, a graph can always be built based on the data dissimilarity where points of the dataset are the vertices and weighted edges express "distances" between those objects. For both cases, graph clustering is one of the key tools for understanding the underlying structure in the graph [Schaeffer, 2007]. These clusters can be seen as groups of nodes close in terms of some specific similarity.

Nevertheless, it is critical that the data representation used in machine learning applications protects the private characteristics contained into it. Let us consider an application where one wants to identify groups of similar web pages in the sense of traffic volume *i.e.* web pages with similar audience. In that case, the nodes stand for the websites. The link between two vertices represents the fact that some people consult them both. Edge weights are the number of common users and thus, carry sensitive information about individuals. During any graph data analysis, no private user surfing behavior should be breached *i.e.* browsing from one page to another should remain private. As a standard for data privacy preservation, differential privacy [Dwork et al., 2006b] has been designed: an algorithm is differentially private if, given two close databases, it produces statistically indistinguishable outputs. Since then, its definition has been extended to weighted graphs. Though, machine learning applications ensuring data privacy remain rare, in particular for clustering which encounters severe theoretical and practical limitations. Indeed, some clustering methods lack of theoretical support and most of them restrict the data distribution to convex-shaped clusters [Nissim et al., 2007, Blum et al., 2008, McSherry, 2009, Dwork, 2011] or unstructured data [Ho and Ruan, 2013, Chen et al., 2015]. Hence, the aim of this paper is to offer a theoretically motivated private graph clustering. Moreover, to the best of our knowledge, this is the first weight differentially-private clustering algorithm able to detect clusters with an arbitrary shape for weighted graph data.

*rafael.pinot@cea.fr
†anne.morvan@cea.fr. Partly supported by the *Direction Générale de l'Armement* (French Ministry of Defense).

Our method belongs to the family of Minimum Spanning Tree (MST)-based approaches. An MST represents a useful summary of the graph, and appears to be a natural object to describe it at a lower cost. For clustering purposes, it has the appealing property to help retrieving non-convex shapes [Zahn, 1971, Asano et al., 1988, Grygorash et al., 2006, Morvan et al., 2017]. Moreover, they appear to be well-suited for incorporating privacy constraints as will be formally proved in this work.

Contributions: Our contributions are threefold: 1) we provide the first theoretical justifications of MST-based clustering algorithms. 2) We endow DBMSTClu algorithm [Morvan et al., 2017], an MST-based clustering algorithm from the literature, with theoretical guarantees. 3) We introduce a differentially-private version of DBMSTClu and give several results on its privacy/utility tradeoff.

2 Preliminaries

2.1 Notations

Let $\mathcal{G} = (V, E, w)$ be a simple undirected weighted graph with a vertex set V, an edge set E, and a weight function $w := E \to \mathbb{R}$. One will respectively call the edge set and the node set of a graph $\mathcal{G}$ using the applications $E(\mathcal{G})$ and $V(\mathcal{G})$. Given a node set $S \subset V$, one denotes by $\mathcal{G}_{|S}$ the subgraph induced by S. We call $G = (V, E)$ the topology of the graph, and $\mathcal{W}_E$ denotes the set of all possible weight functions mapping E to weights in $\mathbb{R}$. For the remaining of this work, cursive letter are use to represent weighted graphs and straight letters refer to topological arguments. Since graphs are simple, the path $\mathcal{P}_{u-v}$ between two vertices u and v is characterized as the ordered sequence of vertices $\{u, \dots, v\}$. We also denote $V_{\mathcal{P}_{u-v}}$ the unordered set of such vertices. Besides, edges e_{ij} denote an edge between nodes i and j. Finally, for all positive integer K, $[K] := \{1, \dots, K\}$.

2.2 Differential privacy in graphs

As opposed to node-differential privacy [Kasiviswanathan et al., 2013] and edge-differential privacy [Hay et al., 2009], both based on the graph topology, the privacy framework considered here is weight-differential privacy where the graph topology $G = (V, E)$ is assumed to be public and the private information to protect is the weight function $w := E \to \mathbb{R}$. Under this model introduced by Sealfon [2016], two graphs are said to be neighbors if they have the same topology, and *close* weight functions. this framework allows one to release an almost minimum spanning tree with weight-approximation error of $O(|V| \log |E|)$ for fixed privacy parameters. Differential privacy is ensured in that case by using the Laplace mechanism on every edges weight to release a spanning tree based on a perturbed version of the weight function. The privacy of the spanning tree construction is thus provided by post-processing (cf. Th. 5). However, under a similar privacy setting, Pinot [2018] recently manages to produce the topology of a tree under differential privacy without relying on the post-processing of a more general mechanism such as the "Laplace mechanism". Their algorithm, called PAMST, privately releases the topology of an almost minimum spanning tree thanks to an iterative use of the "Exponential mechanism" instead. For fixed privacy parameters, the weight approximation error is $O\left(\frac{|V|^2}{|E|} \log |V|\right)$, which outperforms the former method from Sealfon [2016] on arbitrary weighted graphs under weak assumptions on the graph sparseness. Thus, we keep here privacy setting from Pinot [2018].

Definition 2.1 (Pinot [2018])**.** *For any edge set E, two weight functions $w, w' \in \mathcal{W}_E$ are neighboring, denoted $w \sim w'$, if $||w - w'||_\infty := \max_{e \in E} |w(e) - w'(e)| \leq \mu$.*

μ represents the sensitivity of the weight function and should be chosen according to the application and the range of this function. The neighborhood between such graphs is clarified in the following definition.

Definition 2.2. *Let $\mathcal{G} = (V, E, w)$ and $\mathcal{G}' = (V', E', w')$, two weighted graphs, $\mathcal{G}$ and $\mathcal{G}'$ are said to be neighbors if $V = V'$, $E = E'$ and $w \sim w'$.*

The so-called weight-differential privacy for graph algorithms is now formally defined.

Definition 2.3 (Sealfon [2016]). *For any graph topology $G = (V, E)$, let $\mathcal{A}$ be a randomized algorithm that takes as input a weight function $w \in \mathcal{W}_E$. $\mathcal{A}$ is called (ϵ, δ)-differentially private on $G = (V, E)$ if for all pairs of neighboring weight functions $w, w' \in \mathcal{W}_E$, and for all set of possible outputs S, one has*

$$\mathbb{P}\left[\mathcal{A}(w) \in S\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}(w') \in S\right] + \delta.$$

If $\mathcal{A}$ is (ϵ, δ)-differentially private on every graph topology in a class $\mathcal{C}$, it is said to be (ϵ, δ)-differentially private on $\mathcal{C}$.

One of the first, and most used differentially private mechanisms is the Laplace mechanism. It is based on the process of releasing a numerical query perturbed by a noise drawn from a centered Laplace distribution scaled to the sensitivity of the query. We present here its graph-based reformulation.

Definition 2.4 (reformulation Dwork et al. [2006b]). *Given some graph topology $G = (V, E)$, for any $f_G : \mathcal{W}_E \to \mathbb{R}^k$, the sensitivity of the function is defined as $\Delta f_G = \max_{w \sim w' \in \mathcal{W}_E} ||f_G(w) - f_G(w')||_1$.*

Definition 2.5 (reformulation Dwork et al. [2006b]). *Given some graph topology $G = (V, E)$, any function $f_G : \mathcal{W}_E \to \mathbb{R}^k$, any $\epsilon > 0$, and $w \in \mathcal{W}_E$, the graph-based Laplace mechanism is $\mathcal{M}_L(w, f_G, \epsilon) = f_G(w) + (Y_1, \ldots, Y_k)$ where Y_i are i.i.d. random variables drawn from $Lap(\Delta f_G/\epsilon)$, and $Lap(b)$ denotes the Laplace distribution with scale b $\left(\text{i.e probability density } \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)\right)$.*

Theorem 1 (Dwork et al. [2006b]). *The Laplace mechanism is ϵ-differentially private.*

We define hereafter the graph-based Exponential mechanism. In the sequel we refer to it simply as Exponential mechanism. The Exponential mechanism represents a way of privately answering arbitrary range queries. Given some range of possible responses to the query $\mathcal{R}$, it is defined according to a utility function $u_G := \mathcal{W}_E \times \mathcal{R} \to \mathbb{R}$, which aims at providing some total preorder on the range $\mathcal{R}$ according to the total order in $\mathbb{R}$. The sensitivity of this function is denoted $\Delta u_G := \max_{r \in \mathcal{R}} \max_{w \sim w' \in \mathcal{W}_E} |u_G(w, r) - u_G(w', r)|$.

Definition 2.6. *Given some graph topology $G = (V, E)$, some output range $\mathcal{R} \subset E$, some privacy parameter $\epsilon > 0$, some utility function $u_G := \mathcal{W}_E \times \mathcal{R} \to \mathbb{R}$, and some $w \in \mathcal{W}_E$ the graph-based Exponential mechanism $\mathcal{M}_{Exp}(G, w, u_G, \mathcal{R}, \epsilon)$ selects and outputs an element $r \in \mathcal{R}$ with probability proportional to $\exp\left(\frac{\epsilon u_G(w,r)}{2\Delta u_G}\right)$.*

The Exponential mechanism defines a distribution on a potentially complex and large range $\mathcal{R}$. As the following theorem states, sampling from such a distribution preserves ϵ-differential privacy.

Theorem 2 (reformulation McSherry and Talwar [2007]). *For any non-empty range $\mathcal{R}$, given some graph topology $G = (V, E)$, the graph-based Exponential mechanism preserves ϵ-differential privacy,* i.e *if $w \sim w' \in \mathcal{W}_E$,*

$$\begin{aligned}&\mathbb{P}\left[\mathcal{M}_{Exp}(G, w, u_G, \mathcal{R}, \epsilon) = r\right] \\ &\quad \leq e^{\epsilon}\mathbb{P}\left[\mathcal{M}_{Exp}(G, w', u_G, \mathcal{R}, \epsilon) = r\right].\end{aligned}$$

Further, Th 3 highlights the trade-off between privacy and accuracy for the Exponential mechanism when $0 < |\mathcal{R}| < +\infty$. Th 4 presents the ability of differential privacy to comply with composition while Th 5 introduces its post-processing property.

Theorem 3 (reformulation Dwork and Roth [2013]). *Given some graph topology $G = (V, E)$, some $w \in \mathcal{W}_E$, some output range $\mathcal{R}$, some privacy parameter $\epsilon > 0$, some utility function $u_G := \mathcal{W}_E \times \mathcal{R} \to \mathbb{R}$, and denoting $OPT_{u_G}(w) = \max_{r \in \mathcal{R}} u_G(w, r)$, one has $\forall\ t \in \mathbb{R}$,*

$$\begin{aligned}&u_G\left(G, w, \mathcal{M}_{Exp}(w, u_G, \mathcal{R}, \epsilon)\right) \\ &\quad \leq OPT_{u_G}(w) - \frac{2\Delta u_G}{\epsilon}\left(t + \ln|\mathcal{R}|\right)\end{aligned}$$

with probability at most $\exp(-t)$.

Theorem 4 (Dwork et al. [2006a]). *For any $\epsilon > 0$, $\delta \geq 0$ the adaptive composition of k (ϵ, δ)-differentially private mechanisms is $(k\epsilon, k\delta)$-differentially private.*

Theorem 5 (Post-Processing Dwork and Roth [2013]). *Let $\mathcal{A} : \mathcal{W}_E \to B$ be a randomized algorithm that is (ϵ, δ)-differentially private, and $h : B \to B'$ a deterministic mapping. Then $h \circ \mathcal{A}$ is (ϵ, δ)-differentially private.*

2.3 Differentially-private clustering

Differentially private clustering for unstructured datasets has been first discussed in Nissim et al. [2007]. This work introduced the first method for differentially private clustering based on the k-means algorithm. Since then most of the works of the field focused on adaptation of this method [Blum et al., 2008, McSherry, 2009, Dwork, 2011]. The main drawback of those works is that they are not able to deal with arbitrary shaped clusters. This issue has been recently investigated in Ho and Ruan [2013] and Chen et al. [2015]. They proposed two new methods to find arbitrary shaped clusters in unstructured datasets respectively based on density clustering and wavelet decomposition. Even though both of these works allow one to produce non-convex clusters, they only deal with unstructured datasets and thus are not applicable to node clustering in a graph. Our work focuses on node clustering in a graph under weight-differential privacy. Graph clustering has already been investigated in a topology-based privacy framework [Mülle et al., 2015, Nguyen et al., 2016], however, these works do not consider weight-differential privacy. Our work is, to the best of our knowledge, the first attempt to define node clustering in a graph under weight differential privacy.

3 Differentially-private tree-based clustering

We aim at producing a private clustering method while providing bounds on the accuracy loss. Our method is an adaptation of an existing clustering algorithm DBMSTClu. However, to provide theoretical guarantees under differential privacy, one needs to rely on the same kind of guarantees in the non-private setting. Morvan et al. [2017] did not bring them in their initial work. Hence, our second contribution is to demonstrate the accuracy of this method, first in the non-private context.

In the following we present 1) the theoretical framework motivating MST-based clustering methods, 2) accuracy guarantees of DBMSTClu in the non-private setting, 3) PTClust our private clustering algorithm, 4) its accuracy under differential privacy constraints.

3.1 Theoretical framework for MST-based clustering methods

MST-based clustering methods, however efficient, lack of proper motivation. This Section closes this gap by providing a theoretical framework for MST-based clustering. In the sequel, notations from Section 2.1 are kept. The minimum path distance between two nodes in the graph is defined which enables to explicit our notion of Cluster.

Definition 3.1 (Minimum path distance). *Let be $\mathcal{G} = (V, E, w)$ and u, $v \in V$. The minimum path distance between u and v is*

$$d(u, v) = \min_{\mathcal{P}_{u-v}} \sum_{e \in V_{\mathcal{P}_{u-v}}} w(e)$$

with $\mathcal{P}_{u-v}$ a path from u to v in $\mathcal{G}$, and $V_{\mathcal{P}_{u-v}}$ the set of vertices contained in $\mathcal{P}_{u-v}$.

Definition 3.2 (Cluster). *Let be $\mathcal{G} = (V, E, w)$, $0 < w(e) \leq 1$ $\forall e \in E$ a graph, (V, d) a metric space based on the minimum path distance d defined on $\mathcal{G}$ and $D \subset V$ a node set. $C \subset D$ is a cluster iff. $|C| > 2$ and $\forall C_1, C_2$ s.t. $C = C_1 \cup C_2$ and $C_1 \cap C_2 = \emptyset$, one has:*

$$\operatorname*{argmin}_{z \in D \setminus C_1} \{ \min_{v \in C_1} d(z, v) \} \subset C_2$$

Assuming that a cluster is built of at least 3 points makes sense since singletons or groups of 2 nodes can be legitimately considered as noise. For simplicity of the proofs, the following theorems hold in the case where noise is neglected. However, they are still valid in the setting where noise is considered as singletons (with each singleton representing a generalized notion of cluster).

Theorem 6. *Let be $\mathcal{G} = (V, E, w)$ a graph and $\mathcal{T}$ a minimum spanning tree of $\mathcal{G}$. Let also be C a cluster in the sense of Def. 3.2 and two vertices $v_1, v_2 \in C$. Then, $V_{\mathcal{P}_{v_1-v_2}} \subset C$ with $\mathcal{P}_{v_1-v_2}$ a path from v_1 to v_2 in $\mathcal{G}$, and $V_{\mathcal{P}_{v_1-v_2}}$ the set of vertices contained in $\mathcal{P}_{v_1-v_2}$.*

Proof. Let be $v_1, v_2 \in C$. If v_1 and v_2 are neighbors, the result is trivial. Otherwise, as $\mathcal{T}$ is a tree, there exist a unique path within $\mathcal{T}$ between v_1 and v_2 denoted by $\mathcal{P}_{v_1-v_2} = \{v_1, \ldots, v_2\}$. Let now prove by *reductio ad absurdum* that $V_{\mathcal{P}_{v_1-v_2}} \subset C$. Suppose there is $h \in V_{\mathcal{P}_{v_1-v_2}}$ s.t. $h \notin C$. We will see that it leads to a contradiction. We set C_1 to be the largest connected component (regarding the number of vertices) of $\mathcal{T}$ s.t. $v_1 \in C_1$, and every nodes from C_1 are in C. Because of h's definition, $v_2 \notin C_1$. Let be $C_2 = C \backslash C_1$. $C_2 \neq \emptyset$ since $v_2 \in C_2$. Let be $z^* \in \underset{z \in V \backslash C_1}{\operatorname{argmin}} \{ \min_{v \in C_1} d(z, v) \}$ and $e^* = (z^*, v^*)$ an edge that reaches this minimum. Let us show that $z^* \notin C$. If $z^* \in C$, then two possibilities hold:

1. There is an edge $e_{z^*} \in \mathcal{T}$, s.t. $e_{z^*} = (z^*, z')$ with $z' \in C_1$. This is impossible, otherwise by definition of a connected component, $z^* \in C_1$. Contradiction.
2. For all $e_{z^*} = (z^*, z')$ s.t $z' \in C_1$, one has $e_{z^*} \notin \mathcal{T}$. In particular $e^* \notin \mathcal{T}$. Since h is the neighbor of C_1 in $\mathcal{G}$ there is also $e_h \in \mathcal{T}$, s.t. $e_h = (h, h')$ with $h' \in C_1$. Once again two possibilities hold:
   (a) $w(e_{z^*}) = \min_{z \in V \backslash C_1} \{ \min_{v \in C_1} d(z, v) \} < w(e_h)$. Then, if we replace e_h by e_{z^*} in $\mathcal{T}$, its total weight decreases. So $\mathcal{T}$ is not a minimum spanning tree. Contradiction.
   (b) $w(e_{z^*}) = w(e_h)$, therefore $h \in \underset{z \in V \backslash C_1}{\operatorname{argmin}} \{ \min_{v \in C_1} d(z, v) \}$. Since $h \notin C$, one gets that $\underset{z \in V \backslash C_1}{\operatorname{argmin}} \{ \min_{v \in C_1} d(z, v) \} \not\subset C_2$. Thus, C is not a cluster. Contradiction.

We proved that $z^* \notin C$. In particular, $z^* \notin C_2$. Then, $\underset{z \in V \backslash C_1}{\operatorname{argmin}} \{ \min_{v \in C_1} d(z, v) \} \not\subset C_2$. Thus, C is not a cluster. Contradiction. Finally $h \in C$ and $V_{\mathcal{P}_{v_1-v_2}} \subset C$. □

This theorem states that, given a graph $\mathcal{G}$, an MST $\mathcal{T}$, and any two nodes of C, every node in the path between them is in C. This means that a cluster can be characterized by a subtree of $\mathcal{T}$. It justifies the use of all MST-based methods for data clustering or node clustering in a graph. All the clustering algorithms based on successively cutting edges in an MST to obtain a subtree forest are meaningful in the sense of Th.6. In particular, this theorem holds for the use of DBMSTClu [Morvan et al., 2017] presented in Section 3.2.1.

3.2 Deterministic MST-based clustering

This Section introduces DBMSTClu [Morvan et al., 2017] that will be adapted to be differentially-private, and provide accuracy results on the recovery of the ground-truth clustering partition.

3.2.1 DBMSTClu algorithm

Let us consider $\mathcal{T}$ an MST of $\mathcal{G}$, as the unique input of the clustering algorithm DBMSTClu. The clustering partition results then from successive cuts on $\mathcal{T}$ so that a new cut in $\mathcal{T}$ splits a connected component into two new ones. Each final connected component, a subtree of $\mathcal{T}$, represents a cluster. Initially, $\mathcal{T}$ is one cluster containing all nodes. Then, at each iteration, an edge is cut if some criterion, called *Validity Index of a Clustering Partition* (DBCVI) is improved. This edge is greedily chosen to locally maximize the DBCVI at each step. When no improvement on DBCVI can be further made, the algorithm stops. The DBCVI is defined as the weighted average of all *cluster validity indices* which are based on two positive quantities, the *Dispersion* and the *Separation* of a cluster:

Definition 3.3 (Cluster Dispersion)**.** *The Dispersion of a cluster C_i (*DISP*) is defined as the maximum edge weight of C_i. If the cluster is a singleton (i.e. contains only one node), the associated Dispersion is set to* 0*. More formally:*

$$\forall i \in [K],\ \mathrm{DISP}(C_i) = \begin{cases} \max\limits_{j,\ e_j \in C_i} w_j & \text{if } |E(C_i)| \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Definition 3.4 (Cluster Separation)**.** *The Separation of a cluster C_i (*SEP*) is defined as the minimum distance between the nodes of C_i and the ones of all other clusters $C_j, j \neq i, 1 \leq i, j \leq K, K \neq 1$ where K is the total number of clusters. In practice, it corresponds to the minimum weight among all already cut edges from $\mathcal{T}$ comprising a node from C_i. If $K = 1$, the Separation is set to* 1*. More formally, with $incCuts(C_i)$ denoting cut edges incident to C_i,*

$$\forall i \in [K],\ \mathrm{SEP}(C_i) = \begin{cases} \min\limits_{j,\ e_j \in incCuts(C_i)} w_j & \text{if } K \neq 1 \\ 1 & \text{otherwise.} \end{cases}$$

Definition 3.5 (Validity Index of a Cluster)**.** *The Validity Index of a cluster C_i is defined as:*

$$V_C(C_i) = \frac{\mathrm{SEP}(C_i) - \mathrm{DISP}(C_i)}{\max(\mathrm{SEP}(C_i), \mathrm{DISP}(C_i))} \in [-1; 1]$$

Definition 3.6 (Validity Index of a Clustering Partition)**.** *The Density-Based Validity Index of a Clustering partition $\Pi = \{C_i\}, 1 \leq i \leq K$,* DBCVI($\Pi$) *is defined as the weighted average of the Validity Indices of all clusters in the partition where N is the number of vertices.*

$$\mathrm{DBCVI}(\Pi) = \sum_{i=1}^{K} \frac{|C_i|}{N} V_C(C_i) \in [-1, 1]$$

DBMSTClu is summarized in Algorithm 1: evaluateCut(.) computes the DBCVI when the cut in parameter is applied to $\mathcal{T}$. Initial DBCVI is set -1. Interested reader could refer to [Morvan et al., 2017] Section 4. for a complete insight on this notions.

3.2.2 DBMSTClu exact clustering recovery proof

In this section, we provide theoretical guarantees for the cluster recovery accuracy of DBMSTClu. Let us first begin by introducing some definitions.

Definition 3.7 (Cut)**.** *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K clusters, $\mathcal{T}$ an MST of $\mathcal{G}$. Let denote $(C_i^*)_{i \in [K]}$ the set of the clusters. Then, $Cut_{\mathcal{G}}(\mathcal{T}) := \{e_{kl} \in \mathcal{T} \mid k \in C_i^*,\ l \in C_j^*,\ i, j \in [K]^2,\ i \neq j\}$. In the sequel, for simplicity, we denote $e^{(ij)} \in Cut_{\mathcal{G}}(\mathcal{T})$ the edge between cluster C_i^* and C_j^*.*

$Cut_{\mathcal{G}}(\mathcal{T})$ is basically the set of effective cuts to perform on $\mathcal{T}$ in order to ensure the exact recovery of the clustering partition. More generally, trees on which $Cut_{\mathcal{G}}(.)$ enables to find the right partition are said to be a partitioning topology.

Definition 3.8 (Partitionning topology)**.** *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K clusters $C_1^*, \ldots, C_K^*$. A spanning tree $\mathcal{T}$ of $\mathcal{G}$ is said to have a partitioning topology if $\forall i, j \in [K],\ i \neq j,\ |\{e = (u, v) \in Cut_{\mathcal{G}}(\mathcal{T}) \mid u \in C_i^*, v \in C_j^*\}| = 1$.*

Def. 3.7 and 3.8 introduce a topological condition on the tree as input of the algorithm. Nevertheless, conditions on weights are necessary too. Hence, we define homogeneous separability which expresses the fact that within a cluster the edge weights are spread in a controlled manner.

Algorithm 1 DBMSTCLU($\mathcal{T}$)

Input: $\mathcal{T}$, the MST
$dbcvi \leftarrow -1.0$
$clusters \leftarrow \emptyset$
$cut\_list \leftarrow \{E(\mathcal{T})\}$
while $dbcvi < 1.0$ **do**
  $cut\_tp \leftarrow \emptyset$
  $dbcvi\_tp \leftarrow dbcvi$
  for each cut in cut_list **do**
    $newDbcvi \leftarrow \text{evaluateCut}(\mathcal{T}, cut)$
    if $newDbcvi \geq dbcvi\_tp$ **then**
      $cut\_tp \leftarrow cut$
      $dbcvi\_tp \leftarrow newDbcvi$
  if $cut\_tp \neq \emptyset$ **then**
    $clusters \leftarrow \text{cut}(clusters, cut\_tp)$
    $dbcvi \leftarrow dbcvi\_tp$
    $cut\_list \leftarrow cut\_list \backslash \{cut\_tp\}$
  else
    break
return $clusters$, $dbcvi$

Definition 3.9 (Homogeneous separability condition)**.** *Let us consider a graph $\mathcal{G} = (V, E, w)$, $s \in E$ and $\mathcal{T}$ a tree of $\mathcal{G}$. $\mathcal{T}$ is said to be homogeneously separable by s, if*

$$\alpha_{\mathcal{T}} \max_{e \in E(\mathcal{T})} w(e) < w(s) \text{ with } \alpha_{\mathcal{T}} = \frac{\max\limits_{e \in E(\mathcal{T})} w(e)}{\min\limits_{e \in E(\mathcal{T})} w(e)} \geq 1.$$

One will write for simplicity that $H_{\mathcal{T}}(s)$ is verified.

Definition 3.10 (Weak homogeneity condition of a Cluster)**.** *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K clusters $C_1^*, \ldots, C_K^*$. A given cluster C_i^*, $i \in [K]$, C_i^* is weakly homogeneous if: for all $\mathcal{T}$ an MST of $\mathcal{G}$, and $\forall j \in [K]$, $j \neq i$, s.t. $e^{(ij)} \in Cut_{\mathcal{G}}(\mathcal{T})$, $H_{\mathcal{T}_{|C_i^*}}(e^{(ij)})$ is verified. For simplicity, one denote $\underline{\alpha}_i = \max\limits_{\mathcal{T} \text{ MST of } \mathcal{G}} \alpha_{\mathcal{T}_{|C_i^*}}$*

Definition 3.11 (Strong homogeneity condition of a Cluster)**.** *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K clusters $C_1^*, \ldots, C_K^*$. A given cluster C_i^*, $i \in [K]$, C_i^* is strongly homogeneous if: for all $\mathcal{T}$ a spanning tree (ST) of $\mathcal{G}$, and $\forall j \in [K]$, $j \neq i$, s.t. $e^{(ij)} \in Cut_{\mathcal{G}}(\mathcal{T})$, $H_{\mathcal{T}_{|C_i^*}}(e^{(ij)})$ is verified. For simplicity, one denote $\bar{\alpha}_i = \max\limits_{\mathcal{T} \text{ ST of } \mathcal{G}} \alpha_{\mathcal{T}_{|C_i^*}}$*

We show that the weak homogeneity condition is implied by the strong homogeneity condition.

Proposition 3.1. *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K clusters $C_1^*, \ldots, C_K^*$. If a given cluster C_i^*, $i \in [K]$ is strongly homogeneous, then, it is weakly homogeneous.*

Proof. If $\mathcal{T}$ a spanning tree of $\mathcal{G}$, and $\forall j \in [K]$, $j \neq i$, s.t. $e^{(ij)} \in Cut_{\mathcal{G}}(\mathcal{T})$, $H_{\mathcal{T}_{|C_i^*}}(e^{(ij)})$ is verified, then in particular, it is true for any MST. □

Strong homogeneity condition appears to be naturally more constraining on the edge weights than the weak one. The accuracy of DBMSTCLU is proved under the weak homogeneity condition, while the accuracy of its differentially-private version is only given under the the strong homogeneity condition.

Theorem 7. *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, $Cut_k := Cut_{\mathcal{G}}(\mathcal{T}) \setminus \{e_1,\ e_2,\ \ldots,\ e_k\} \neq \emptyset \implies \text{DBCVI}_{k+1} \geq DBCVI_k$, i.e. if there are still edges in Cut_k, the algorithm will continue to perform some cut.

Proof. See supplementary material. □

Theorem 8. *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 $\mathcal{G}$.*

Assume now 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 $Cut_k := Cut_{\mathcal{G}}(\mathcal{T}) \backslash \{e_1,\ e_2,\ \ldots,\ e_k\}$.

If $Cut_k \neq \emptyset$ then $\underset{e \in \mathcal{T} \backslash \{e_1,\ e_2,\ \ldots,\ e_k\}}{\operatorname{argmax}} DBCVI_{k+1}(e) \subset Cut_k$ i.e. the cut edge at step $k+1$ is in Cut_k.

Proof. See supplementary material. □

Theorem 9. *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.

Proof. See supplementary material. □

Corollary 3.1. *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 $\mathcal{G}$. $DBMSTClu(\mathcal{T})$ stops after $K-1$ iterations and the K subtrees produced match exactly the clusters i.e. under homogeneity condition, the algorithm finds automatically the underlying clustering partition.*

Proof. Th. 7 and 9 ensure that under homogeneity condition on all clusters, the algorithm performs the $K-1$ distinct cuts within $Cut_{\mathcal{G}}(\mathcal{T})$ and stops afterwards. By definition of $Cut_{\mathcal{G}}(\mathcal{T})$, it means the DBMSTClu correctly builds the K clusters. □

3.3 Private MST-based clustering

This section presents our new node clustering algorithm PTClust for weight differential privacy. It relies on a mixed adaptation of PAMST algorithm [Pinot, 2018] for recovering a differentially-private MST of a graph and DBMSTClu.

3.3.1 PAMST algorithm

Given a simple-undirected-weighted graph $\mathcal{G} = (V, E, w)$, PAMST outputs an almost minimal weight spanning tree topology under differential privacy constraints. It relies on a Prim-like MST algorithm, and an iterative use of the graph based Exponential mechanism. PAMST takes as an input a weighted graph, and a utility function. It outputs the topology of a spanning tree which weight is almost minimal. Algorithm 3 presents this new method, using the following utility function:

$$\begin{array}{rccc} u_G & : & \mathcal{W}_E \times \mathcal{R} & \to & \mathbb{R} \\ & & (w, r) & \mapsto & -|w(r) - \min\limits_{r' \in \mathcal{R}} w(r')|. \end{array}$$

PAMST starts by choosing an arbitrary node to construct iteratively the tree topology. At every iteration, it uses the Exponential mechanism to find the next edge to be added to the current tree topology while keeping the weights private. This algorithm is the state of the art to find a spanning tree topology under differential privacy. For readability, let us introduce some additional notations. Let S be a set of nodes

from G, and $\mathcal{R}_S$ the set of edges that are incident to one and only one node in S (also denoted xor-incident). For any edge r in such a set, the incident node to r that is not in S is denoted $r_{\rightarrow}$. Finally, the restriction of the weight function to an edge set $\mathcal{R}$ is denoted $w_{|\mathcal{R}}$.

Algorithm 2 PAMST(G, u_G, w, ϵ)

Input: $\mathcal{G} = (V, E, w)$ a weighted graph (separately the topology G and the weight function w), ϵ a degree of privacy and u_G utility function.
Pick $v \in V$ at random
$S_V \leftarrow \{v\}$
$S_E \leftarrow \emptyset$
while $S_V \neq V$ **do**
  $r = \mathcal{M}_{Exp}(\mathcal{G}, w, u_G, \mathcal{R}_{S_V}, \frac{\epsilon}{|V|-1})$
  $S_V \leftarrow S_V \cup \{r_{\rightarrow}\}$
  $S_E \leftarrow S_E \cup \{r\}$
return S_E

Theorem 10 states that using PAMST to get an almost minimal spanning tree topology preserves weight-differential privacy.

Theorem 10. *Let $G = (V, E)$ be the topology of a simple-undirected graph, then $\forall \epsilon > 0$,* PAMST$(G, u_G, \bullet, \epsilon)$ *is ϵ- differentially private on G.*

3.3.2 Differentially private clustering

The overall goal of this Section is to show that one can obtain a differentially private clustering algorithm by combining PAMST and DBMSTClu algorithms. However, PAMST does not output a weighted tree which is inappropriate for clustering purposes. To overcome this, one cloud rely on a sanitizing mechanism such as the Laplace mechanism. Moreover, since DBMSTClu only takes weights from (0,1], two normalizing parameters τ and p are introduced, respectively to ensure lower and upper bounds to the weights that fit within DBMSTClu needs. This sanitizing mechanism is called the Weight-Release mechanism. Coupled with PAMST, it will allows us to produce a weighted spanning tree with differential privacy, that will be exploited in our private graph clustering.

Definition 3.12 (Weight-Release mechanism)**.** *Let $\mathcal{G} = (G, w)$ be a weighted graph, $\epsilon > 0$ a privacy parameter, s a scaling parameter, $\tau \geq 0$, and $p \geq 1$ two normalization parameters. The Weight-Release mechanism is defined as*

$$\mathcal{M}_{w.r}(G, w, s, \tau, p) = \left(G, w' = \frac{w + (Y_1, ..., Y_{|E|}) + \tau}{p}\right)$$

where Y_i are i.i.d. random variables drawn from Lap$(0, s)$. With $w + (Y_1, ..., Y_{|E|})$ meaning that if one gives an arbitrary order to the edges $E = (e_i)_{i \in [|E|]}$, one has $\forall i \in [|E|]$, $w'(e_i) = w(e_i) + Y_i$.

The following theorem presents the privacy guarantees of the Weight-Release mechanism.

Theorem 11. *Let $G = (V, E)$ be the topology of a simple-undirected graph, $\tau \geq 0$, $p \geq 1$, then $\forall \epsilon > 0$, $\mathcal{M}_{w.r}\left(G, \bullet, \frac{\mu}{\epsilon}, \tau, p\right)$ is ϵ- differentially private on G.*

Proof. Given $\tau \geq 0$, $p \geq 1$, and $\epsilon > 0$, the Weight release mechanism scaled to $\frac{\mu}{\epsilon}$ can be break down into a Laplace mechanism and a post-processing consisting in adding τ to every edge and dividing them by p. Using Theorems 1 and 5, one gets the expected result. □

So far we have presented DBMSTClu and PAMST algorithms, and the Weight-Release mechanism. Let us now introduce how to compose those blocks to obtain a Private node clustering in a graph, called

Algorithm 3 PTCLUST$(G, w, u_G, \epsilon, \tau, p)$

1: **Input:** $\mathcal{G} = (V, E, w)$ a weighted graph (separately the topology G and the weight function w), ϵ a degree of privacy and u_G utility function.
2: $T = \text{PAMST}(G, w, u_G, \epsilon/2)$
3: $\mathcal{T}' = \mathcal{M}_{w.r}(T, w_{|E(T)}, \frac{2\mu}{\epsilon}, \tau, p)$
4: **return** DBMSTCLU$(\mathcal{T}')$

PTCLUST. The algorithm takes as an input a weighted graph (dissociated topology and weight function), a utility function, a privacy degree and two normalization parameters. It outputs a clustering partition. To do so, a spanning tree topology is produced using PAMST. Afterward a randomized and normalized version of the associated weight function is released using the Weight-release mechanism. Finally the obtained weighted tree is given as an input to DBMSTCLU that performs a clustering partition. The following theorem ensures that our method preserves ϵ-differential privacy.

Theorem 12. *Let $G = (V, E)$ be the topology of a simple-undirected graph, $\tau \geq 0$, and $p \geq 1$, then $\forall \epsilon > 0$,* PTCLUST$(G, \bullet, u_G, \epsilon, \tau, p)$ *is ϵ-differentially private on G.*

Proof. Using Theorem 10 one has that T is produced with $\epsilon/2$-differential privacy, and using Theorem 11 one has that w' is obtained with $\epsilon/2$-differential privacy as well. Therefore using Theorem 4, $\mathcal{T}'$ is released with ϵ-differential privacy. Using the post-processing property (Theorem 5) one gets the expected result. □

3.4 Differential privacy trade-off of clustering

The results stated in this section present the security/accuracy trade-off of our new method in the differentially-private framework. PTCLUST relies on two differentially private mechanisms, namely PASMT and the Weight-Release mechanism. Evaluating the accuracy of this method amounts to check whether using these methods for ensuring privacy does not deteriorate the final clustering partition. The accuracy is preserved if PAMST outputs the same topology as the MST-based clustering, and if the Weight-Release mechanism preserves enough the weight function. According to Def. 3.8, if a tree has a partitioning topology, then it fits the tree-based clustering. The following theorem states that with high probability PAMST outputs a tree with a partitioning topology.

Theorem 13. *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K strongly homogeneous clusters $C_1^*, \ldots, C_K^*$ and $T = \text{PAMST}(\mathcal{G}, u_\mathcal{G}, w, \epsilon)$, $\epsilon > 0$. T has a partitioning topology with probability at least*

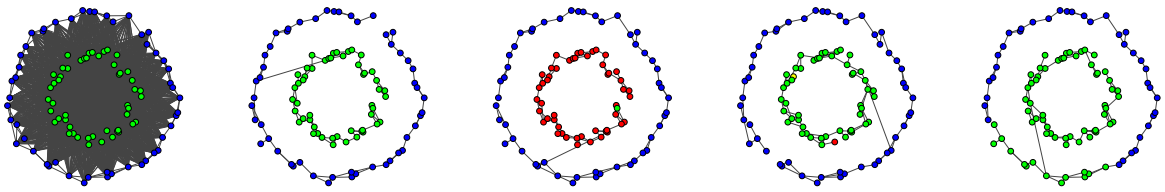
$$1 - \sum_{i=1}^{K} (|C_i^*| - 1) \exp\left(-\frac{A}{2\Delta u_\mathcal{G}(|V| - 1)}\right)$$

with $A = \epsilon \left(\bar{\alpha}_i \max_{e \in E\left(\mathcal{G}_{|C_i^*}\right)}(w(e)) - \min_{e \in E\left(\mathcal{G}_{|C_i^*}\right)}(w(e)) \right) + \ln |E|$.

Proof. See supplementary material. □

The following theorem states that given a tree $\mathcal{T}$ under the strong homogeneity condition, if the subtree associated to a cluster respects Def. 3.9, then it still holds after applying the Weight-Release mechanism to this tree.

Theorem 14. *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K strongly homogeneous clusters $C_1^*, \ldots, C_K^*$ and $T = PAMST(\mathcal{G}, u_\mathcal{G}, w, \epsilon)$, $\mathcal{T} = (T, w_{|T})$ and $\mathcal{T}' = \mathcal{M}_{w.r}(T, w_{|T}, s, \tau, p)$ with $s << p, \tau$. Given some*



(a) Homogeneous graph (b) DBMSTClu (c) PTClust, $\epsilon = 1.0$ (d) PTClust, $\epsilon = 0.7$ (e) PTClust, $\epsilon = 0.5$

Figure 1: Circles experiments for $n = 100$. PTClust parameters: $w_{min} = 0.1$, $w_{max} = 0.3$, $\mu = 0.1$.

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*

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

with the following notations :

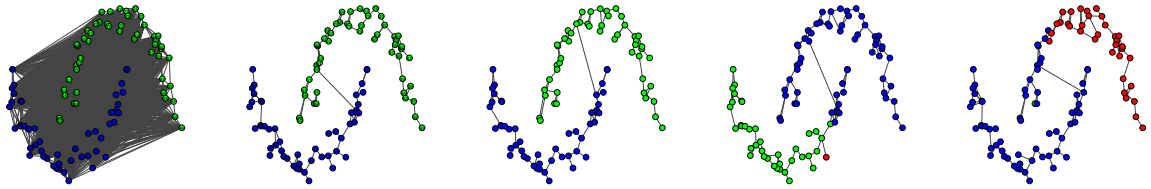
- $\varphi = \underset{j\in[|C_i^*|-1]}{(\max} Y_j)^2 - \underset{j\in[|C_i^*|-1]}{\min} Z_j \times X^{out}$
- $Y_j \underset{iid}{\sim} Lap\left(\frac{\max_{e\in E(\mathcal{T})} w(e)+\tau}{p}, \frac{s}{p}\right)$
- $Z_j \underset{iid}{\sim} Lap\left(\frac{\min_{e\in E(\mathcal{T})} w(e)+\tau}{p}, \frac{s}{p}\right)$
- $X^{out} \sim Lap\left(\frac{w(e^{(ij)})+\tau}{p}, \frac{s}{p}\right)$,

Proof. See supplementary material. □

Note that Theorem 14 is stated in a simplified version. A more complete version (specifying an analytic version of $\mathbb{V}(\varphi)$ and $\mathbb{E}(\varphi)$) is given in the supplementary material.

4 Experiments

So far we have exhibited the trade-off between clustering accuracy and privacy and we experimentally illustrate it with some qualitative results. Let us discuss hereafter the quantitative performances of our algorithm. We have performed experiments on two classical synthetic graph datasets for clustering with nonconvex shapes: two concentric circles and two moons, both in their noisy versions. For the sake of readability and for visualization purposes, both graph datasets are embedded into a two dimensional Euclidean space. Each dataset contains 100 data nodes that are represented by a point of two coordinates. Both graphs have been built with respect to the strong homogeneity condition: edge weights within clusters are between $w_{min} = 0.1$ and $w_{max} = 0.3$ while edges between clusters have a weight strictly above $w_{max}^2/w_{min} = 0.9$. In practice, the complete graph has trimmed from its irrelevant edges (*i.e.* not respecting the strong homogeneity condition). Hence, those graphs are not necessarily Euclidean



(a) Homogeneous graph (b) DBMSTClu (c) PTClust, $\epsilon = 1.0$ (d) PTClust, $\epsilon = 0.7$ (e) PTClust, $\epsilon = 0.5$

Figure 2: Moons experiments for $n = 100$. PTClust parameters: $w_{min} = 0.1$, $w_{max} = 0.3$, $\mu = 0.1$.

since close nodes in the visual representation may not be connected in the graph. Finally, weights are normalized between 0 and 1.

Figures 1 and 2 (best viewed in color) show for each dataset (a) the original homogeneous graph $\mathcal{G}$ built by respecting the homogeneity condition, (b) the clustering partition[1] of DBMSTClu with the used underlying MST, the clustering partitions for PTClust with $\mu = 0.1$ obtained respectively with different privacy degrees[2] : $\epsilon = 0.5$ (c), $\epsilon = 0.7$ (c) and $\epsilon = 1.0$ (e). The utility function $u_{\mathcal{G}}$ corresponds to the graph weight. Each experiment is carried out independently and the tree topology obtained by PAMST will eventually be different. This explains why the edge between clusters may not be the same when the experiment is repeated with a different level of privacy. However, this will marginally affect the overall quality of the clustering.

As expected, DBMSTClu recovers automatically the right partition and the results are shown here for comparison with PTClust. For PTClust, the true MST is replaced with a private approximate MST obtained for suitable τ and p ensuring final weights between 0 and 1.

When the privacy degree is moderate ($\epsilon \in \{1.0, 0.7\}$), it appears that the clustering result is slightly affected. More precisely, in Figures 1c and 1d the two main clusters are recovered while one point is isolated as a singleton. This is due to the randomization involved in determining the edge weights for the topology returned by PAMST. In Figure 2c, the clustering is identical to the one from DBMSTClu in Figure 2b. In Figure 1d, the clustering is very similar to the DBMSTClu one, with the exception of an isolated singleton. However, as expected from our theoretical results, when ϵ is decreasing, the clustering quality deteriorates, as DBMSTClu is sensitive to severe changes in the MST (cf. Figure 1e, 2e).

5 Conclusion

In this paper, we introduced PTClust, a novel graph clustering algorithm able to recover arbitrarily-shaped clusters while preserving differential privacy on the weights of the graph. It is based on the release of a private approximate minimum spanning tree of the graph of the dataset, by performing suitable cuts to reveal the clusters. To the best of our knowledge, this is the first differential private graph-based clustering algorithm adapted to nonconvex clusters. The theoretical analysis exhibited a trade-off between the degree of privacy and the accuracy of the clustering result. This work suits to applications where privacy is a critical issue and it could pave the way to metagenomics and genes classification using

[1]For the sake of clarity, the edges in those Figures are represented based on the original weights and not on the privately released weights.

[2]Note that, although the range of ϵ is in $\mathbb{R}_+^\star$, it is usually chosen in practice in $(0, 1]$ [Dwork and Roth, 2013, Chap 1&2].

individual gene maps while protecting patient privacy. Future work will be devoted to deeply investigate these applications.

References

T. Asano, B. Bhattacharya, M. Keil, and F. Yao. Clustering algorithms based on minimum and maximum spanning trees. In *Proceedings of the Fourth Annual Symposium on Computational Geometry*, SCG '88, pages 252–257, New York, NY, USA, 1988. ACM.

A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 609–618, New York, NY, USA, 2008. ACM.

L. Chen, T. Yu, and R. Chirkova. Wavecluster with differential privacy. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1011–1020, New York, NY, USA, 2015. ACM.

C. Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95, Jan. 2011.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407, 2013.

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Eurocrypt*, volume 4004, pages 486–503. Springer, 2006a.

C. Dwork, F. McSherry, K. Nissim, and A.Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006b.

O. Grygorash, Y. Zhou, and Z. Jorgensen. Minimum spanning tree based clustering algorithms. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, pages 73–81, Nov 2006.

M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 169–178, Dec 2009.

S.-S. Ho and S. Ruan. Preserving privacy for interesting location pattern mining from trajectory data. *Trans. Data Privacy*, 6(1):87–106, Apr. 2013.

S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In *Proceedings of the 10th Theory of Cryptography Conference on Theory of Cryptography*, TCC'13, pages 457–476, Berlin, Heidelberg, 2013. Springer-Verlag.

F. McSherry. Privacy integrated queries. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. Association for Computing Machinery, Inc., June 2009.

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Providence, RI, October 2007. IEEE.

A. Morvan, K. Choromanski, C. Gouy-Pailler, and J. Atif. Graph sketching-based massive data clustering. *SIAM Data Mining 2018 (to appear)*, 2017.

Y. Mülle, C. Clifton, and K. Böhm. Privacy-integrated graph clustering through differential privacy. In *EDBT/ICDT Workshops*, 2015.

H. H. Nguyen, A. Imine, and M. Rusinowitch. Detecting communities under differential privacy. In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, WPES '16, pages 83–93, New York, NY, USA, 2016. ACM.

K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing - STOC*. ACM Press, 2007.

R. Pinot. Minimum spanning tree release under differential privacy constraints. *ArXiv e-prints*, Jan. 2018.

S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27 – 64, 2007.

A. Sealfon. Shortest paths and distances with differential privacy. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS*. ACM Press, 2016.

C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.*, 20(1):68–86, Jan. 1971.

SUPPLEMENTARY MATERIAL

6 Proof regarding the accuracy of DBMSTClu

6.1 Proof of Theorem 7

This theorem relies on the following lemma:

Lemma 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. 7 *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, $Cut_k := Cut_{\mathcal{G}}(\mathcal{T}) \setminus \{e_1,\ e_2,\ \ldots,\ e_k\} \neq \emptyset \implies \text{DBCVI}_{k+1} \geq DBCVI_k$, 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. 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. 3 for an illustration) and $DBCVI_{k+1}(e^*)$ the associated DBCVI value,

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

There are two possible cases:

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

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

   and $\Delta \geq 0$.

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

□

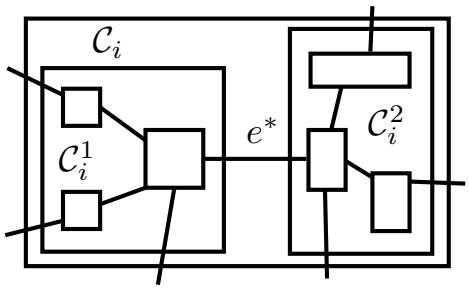


Figure 3: Illustration for Th. 7's proof.

6.2 Proof of Theorem 8

Theorem. ***8*** *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, $Cut_k \neq \emptyset \implies \underset{e \in \mathcal{T} \backslash \{e_1,\ e_2,\ \ldots,\ e_k\}}{\operatorname{argmax}} DBCVI_{k+1}(e) \subset Cut_k$ 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. 5 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. 4 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.

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

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

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

By Lem. 1, $e^* \in \underset{e \in E(\mathcal{T}_{|\mathcal{C}^*_1 \cup \mathcal{C}^*_2})}{\text{argmax}} w(e)$ 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. □

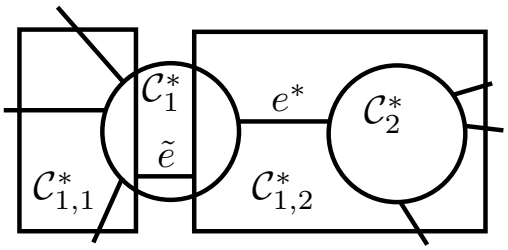


Figure 4: Illustration for Th. 8's proof.

6.3 Proof of Theorem 9

Theorem. ***9*** *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 $Cut_{K-1} := Cut_{\mathcal{G}}(\mathcal{T}) \backslash \{e_1,\ e_2,\ \ldots,\ e_{K-1}\}$.*

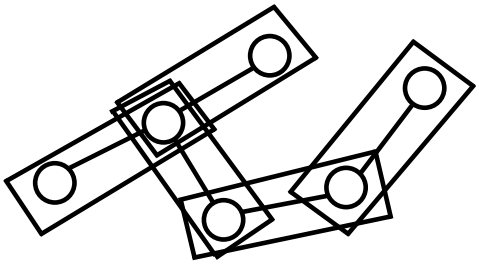


Figure 5: Illustration for Th. 8's proof. Each circle corresponds to a cluster. The six clusters are handled within five couples of clusters.

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*.*

Proof. According to Th. 7 and Th. 8, 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, $\text{DBCVI}_{K-1} = \sum_{j \in [K-1]} \frac{|\mathcal{C}_j^*|}{N} V_C(\mathcal{C}_j^*)$.

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 $\text{SEP}(\mathcal{C}_{i,1}^*) \geq \text{DISP}(\mathcal{C}_{i,1}^*)$, $\text{SEP}(\mathcal{C}_{i,2}^*) \geq \text{DISP}(\mathcal{C}_{i,2}^*)$, $\text{SEP}(\mathcal{C}_{i,1}^*) = \text{SEP}(\mathcal{C}_{i,2}^*) = \min_{e' \in E(\mathcal{C}_i^*)} w(e')$, $\text{DISP}(\mathcal{C}_{i,1}^*) = \text{DISP}(\mathcal{C}_{i,2}^*) = \max_{e' \in E(\mathcal{C}_i^*)} w(e')$. Then,

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

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

7 Proofs regarding the accuracy of PTClust

7.1 Proof of Theorem 13

Theorem. ***13*** *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K strongly homogeneous clusters $C_1^*, \ldots, C_K^*$ 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 $\text{Steps}(C_i^*)$ 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 $B =$ "$\forall i \in [K],\ \forall k > 1 \in \text{Steps}(C_i^*)$, 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

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

□

7.2 Proof of Theorem 14

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

Theorem 15. *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 $k \in \mathbb{N}$,*

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

Theorem. ***14*** *Let us consider a graph $\mathcal{G} = (V, E, w)$ with K strongly homogeneous clusters $C_1^*, \ldots, C_K^*$ 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 verified 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 $Y_i \underset{iid}{\sim} Lap(\theta_{\max}, \delta)$, $Z_i \underset{iid}{\sim} Lap(\theta_{\min}, \delta)$ and $X^{out} \sim$

$Lap(\theta_{(ij)}, \delta)$, one can lower bounded this probability by $\mathbb{P}\left[\frac{(\max_{i\in[|C_i^*|-1]} Y_i)^2}{\min_{i\in[|C_i^*|-1]} Z_i} < X^{out}\right]$. Choosing τ big enough s.t $\min_{i\in[|C_i^*|-1]} Z_i < 0$ is negligible, one has

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

Moreover since τ, $p >> s$, 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 15 one gets the expected result. □