Nodes clustering in a graph under differential privacy constraints

A novel notion of differential privacy for structured datasets

Rafael PINOT *

Institut LIST, CEA, Université Paris-Saclay, F-91120, Palaiseau, France

- New theoretically motivated method for clustering under differential privacy constraints using a MST-based clustering algorithm.
- Some results on the robustness of such a clustering to sanitizing by randomization.
- Conditions on the edge weights in order to consider that the nodes form well separated clusters.

Differential privacy on structured datasets

Dwork and al. in [DMNS06]: *"The outcome of any analysis is essentially equally likely, independent of whether any individuals joins, or refrains from joining the database"*.

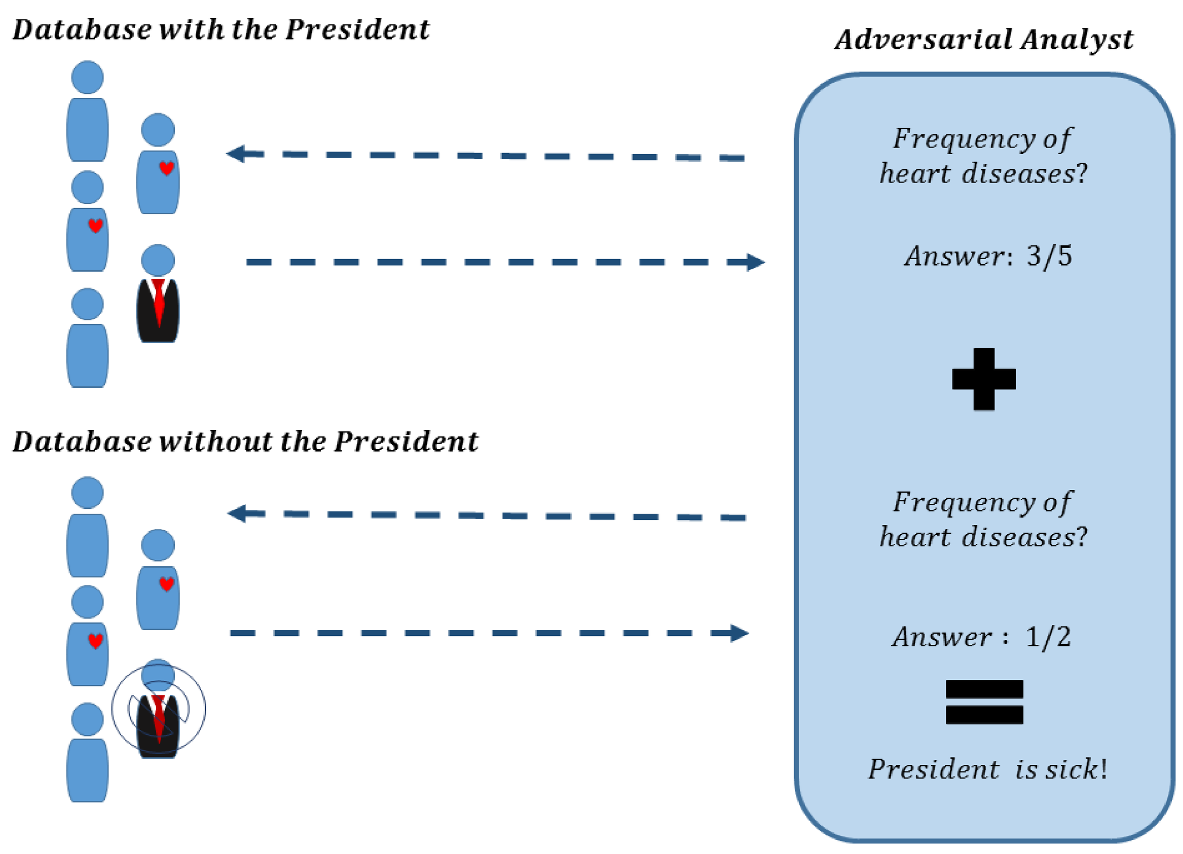

Figure 1: Inferring the French President medical condition asking the same query on two adversarial chosen databases

Definition 1 (Differential Privacy). *A randomized algorithm $\mathcal{A}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is called (ϵ, δ)-differentially private if for $S \subset Range(\mathcal{A})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:*

$$\mathbb{P}[\mathcal{A}(x) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(y) \in S] + \delta$$

Where the probability space is over the simplex of $\mathcal{A}$. Moreover, if $\delta = 0$, $\mathcal{A}$ is said to be ϵ-differentially private.

Definition 2 (ℓ_p sensitivity). *For any $p \in \mathbb{N}$, and $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the ℓ_p sensitivity of f is:*

$$\Delta_p f := \max_{x,y \in \mathbb{N}^{|\mathcal{X}|}, x \sim y} \|f(x) - f(y)\|_p$$

i.e the maximum ℓ_p-distance between the outcomes of any two neighboring databases.

Definition 3 (Laplace Mechanism). *Given any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, and $\epsilon > 0$, the Laplace mechanism is defined as*

$$\mathcal{M}_L(x, f, \epsilon) = f(x) + (Y_1, \ldots, Y_k)$$

where Y_i are i.i.d. random variables drawn from $Lap(\Delta_1 f/\epsilon)$.

Adam Sealfon in [Sea16]: In transport networks, genes correlation maps, world wide web monitoring, etc, the information is carried by the edge weights, this is why they must be protected.

Proposition 1 (e.g [DR13] Post-Processing). *Let $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ be a randomized algorithm that is (ϵ, δ)-differentially private, and $h : \mathcal{R} \to \mathcal{R}'$ a deterministic mapping. Then $h \circ \mathcal{A}$ is (ϵ, δ)-differentially private.*

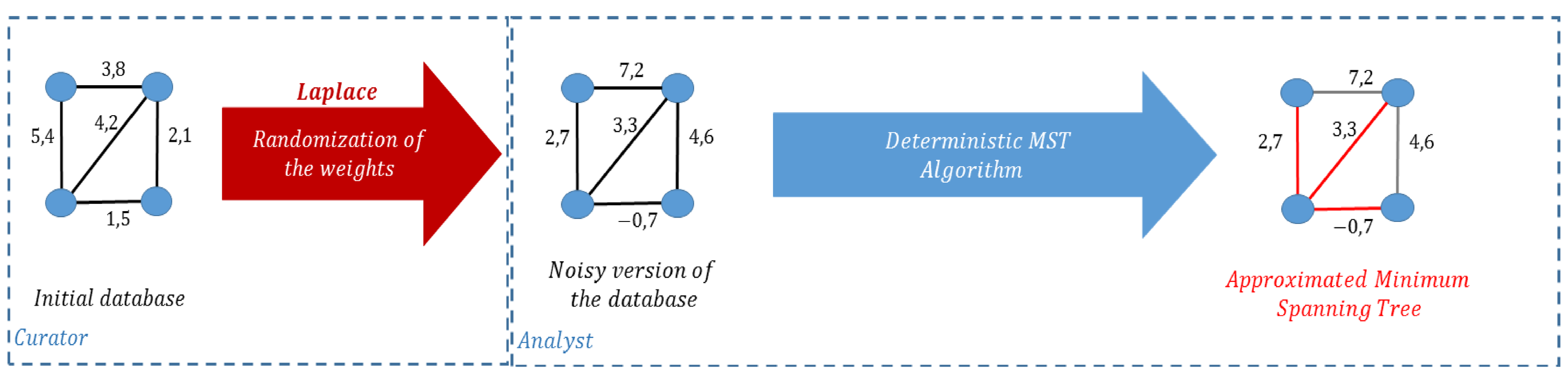

Figure 2: An almost minimum spanning tree under differential privacy conditions by post-processing a Laplace mechanism

New private MST-based nodes clustering

Xu and al. / Zhou and al. / Morvan and al. [YVD02, YOT11, MCGA17]: Minimum spanning tree based clustering algorithms help recognizing clusters with arbitrary shapes and thus can be used for wider applications than community detection.

Our contribution:
- Reformulation and proof of theoretical motivation for MST-based clustering
- MST-based clustering algorithm using a differentially private almost minimum spanning tree.

Definition 4 (Separability condition of a cluster). *Let $G = (V, E, w)$ a simple weighted graph, (V, d) a metric space defined in G according to the minimal-weighted path between nodes, and $D \subset V$ a dataset. $C \subset D$ is called a cluster if and only if for any partition $C = C_1 \cup C_2$ one has*

$$\arg\min_{v \in D - C_1} \{\min\{d(v, c) | c \in C_1\}\} \in C_2.$$

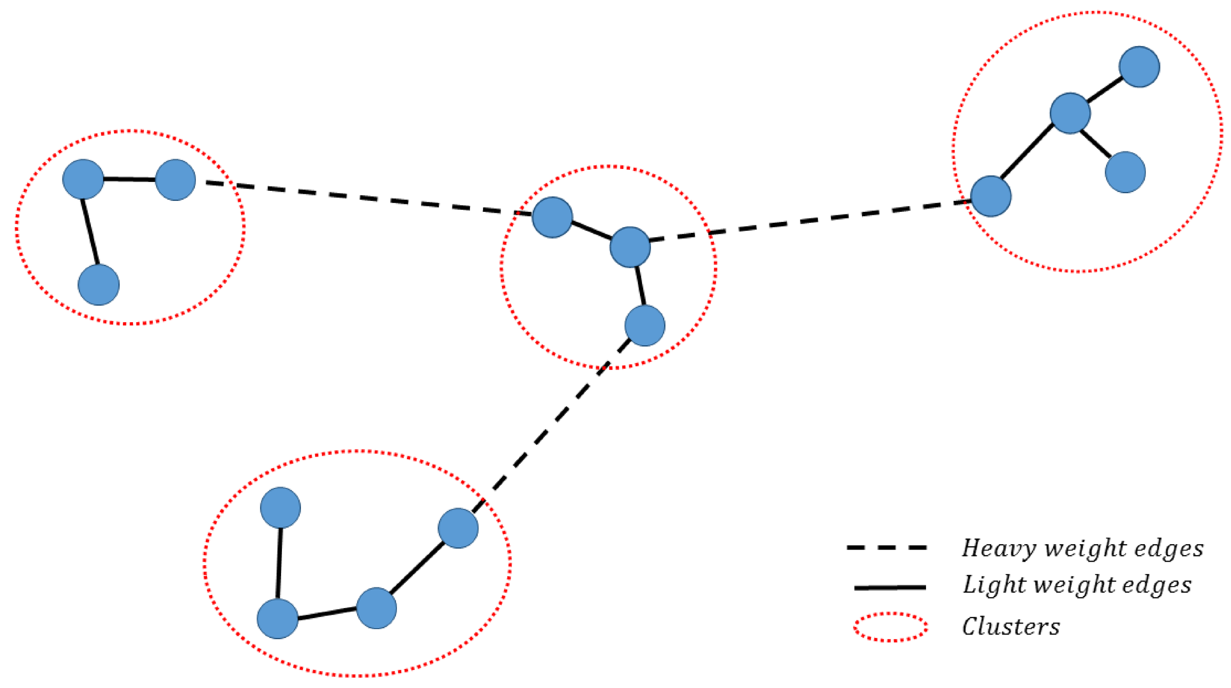

Figure 3: Illustration of a minimum spanning tree based clustering

Proposition 2 (Reformulation of [YVD02]). *If one takes two points c_1, c_2 of a cluster C, then all data points in the tree path connecting c_1 and c_2 in the MST must be in C.*

Proposition 3. *Let $G = (V, E, w)$ be an undirected weighted graph, if one constructs $K > 1$ sets of edges $(E_1, \ldots, E_K)$ such that:*

$$i, j \in [K], i < j \implies \forall e \in E_i, e' \in E_j, w(e) < w(e') \tag{1}$$

Then Eq. 1 holds on $G' = \mathcal{M}_{GbL}(G, \epsilon)$ with probability greater than

$$1 - \sum_{i \in [K-1]} \exp(-t_{i,i+1}\epsilon)\left(\frac{1}{2} + \frac{t_{i,i+1}\epsilon}{4}\right)(E_i + E_{i+1} - 1) \text{ With } t_{i,i+1} = \min_{e \in E_i, e' \in E_{i+1}} \{w(e') - w(e)\}.$$

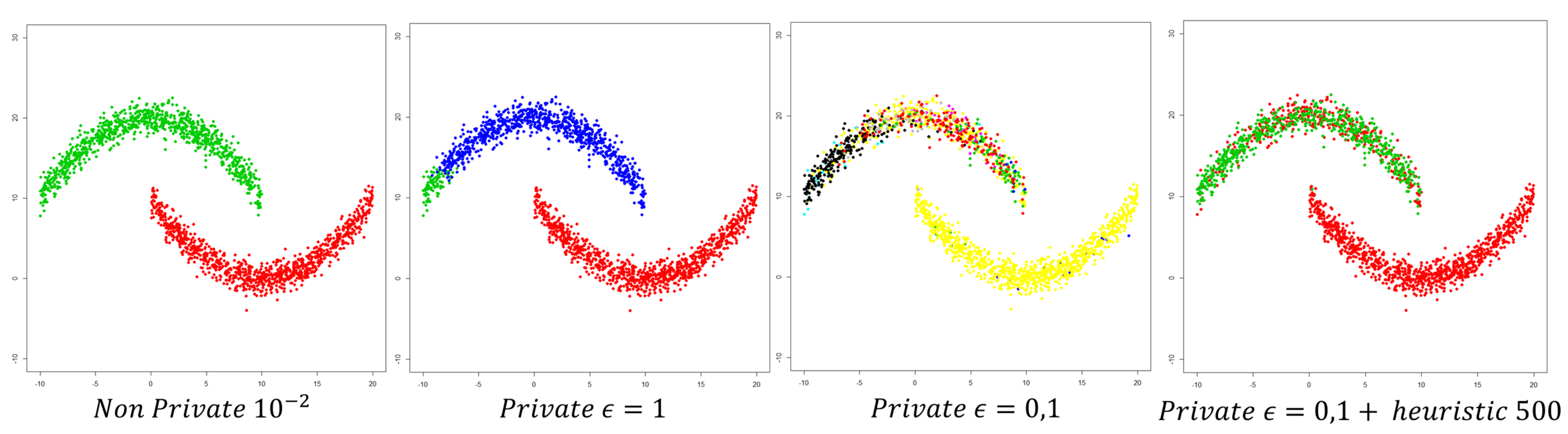

Figure 4: Experiment investigating the robustness of MSDR to privacy mechanisms and simple heuristic

Future Work

- Find new ways for releasing an almost minimum spanning tree under differential privacy constraints.
- Investigate gene clustering using genes map interaction graphs.

Acknowledgements

*With the collaboration of **A. MORVAN** (CEA/ U. Paris Dauphine), and under supervision of **F. YGER** (U. Paris Dauphine), **C. GOUY-PAILLER** (CEA) and **J. ATIF** (U. Paris Dauphine).

References

[DMNS06] Dwork, McSherry, Nissim, and Smith. Calibrating noise to sensitivity in private data analysis. pages 265–284, 2006.

[DR13] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. Now Publishers, 2013.

[MCGA17] A. Morvan, K. Choromanski, C. Gouy-Pailler, and J. Atif. Graph sketching-based Massive Data Clustering. *ArXiv e-prints*, March 2017.

[Sea16] 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.

[YOT11] Y.Zhou, O.Grygorash, and T.F.Hain. Clustering with minimum spanning tree. *International Journal on Artificial Intelligence Tools*, 20(01):139–177, feb 2011.

[YVD02] Y.Xu, V.Olman, and D.Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, apr 2002.

CONTACT|rafael.pinot@cea.fr