THE SHAPE OF EDGE DIFFERENTIAL PRIVACY

TOPOLOGICAL INFERENCE FROM DIFFERENTIALLY-PRIVATE RANDOM DOT-PRODUCT GRAPHS

SIDDHARTH VISHWANATH[†] AND JONATHAN HEHIR[†]

Extended Abstract

Graphs have become a mainstay in modeling several systems, with applications ranging from sociology and political science to public health and biochemistry. Random Dot Product Graphs (RDPGs) have emerged as a powerful paradigm for statistical analysis of graphs. These encompass a wide range of networks, including the stochastic (SBM) and its common extensions, such as the degree-corrected block model (Karrer and Newman, 2011) and mixed-membership SBM (Airoldi et al., 2008). RDPGs belong to the family of latent position models of Hoff et al. (2002), and typify the manifold hypothesis from statistical learning to the graphical setting: *high-dimensional data often lie near a lower-dimensional manifold*. Recent work in RDPGs has established that the spectral embedding of RDPGs are able to recover the low-dimensional structure in the latent space Athreya et al. (2017); Solanki et al. (2019); Rubin-Delanchy et al. (2017); Rubin-Delanchy (2020).

The defining quality of a graph is that it conveys relationships between vertices. In a wide variety of settings, the existence, lack of existence, or nature of such relationships may be sensitive. Preserving privacy in such settings can be achieved through the notion of edge differential privacy for graphs.

In this work, we consider a randomized-response mechanism called the edgeFlip, which releases a sanitized graph satisfying ϵ–edge differential privacy. We show that for a RDPG, the output of edgeFlip is also a RDPG. Then, using tools from the burgeoning area of Topological Data Analysis (TDA), we show that if the structure underlying a RDPG in the latent space is supported on a lower-dimensional manifold $\mathcal{M}$, then the ϵ–edge differentially private synthetic graph obtained via edgeFlip is also supported on a manifold $\mathcal{M}'$ with identical topological features (to be made precise later). Additionally, for the privacy budget $\epsilon = 0$, the manifold $\mathcal{M}'$ retracts to a single point, making the RDPG equivalent to an Erdős-Rényi graph. In essence, for $\epsilon > 0$, the privacy mechanism warps the original manifold $\mathcal{M}$ in a way such that the subtle topological features are still preserved. Furthermore, we assess the quality of the spectral embedding of the RDPG using persistence diagrams. Asymptotically, we can show that even though the limiting persitence diagram obtained via edgeFlip is different from that of the original graph, the shift-invariant bottleneck distance (a variant of the bottleneck distance which identifies the same input metric space measured in two different units) between the two limiting persitence diagrams converges to zero. We illustrate the advantage of employing edgeFlip as opposed to other alternatives. Lastly, we highlight the benefit of the topological perspective by employing ToMaTO—a topologically-motivated clustering algorithm Chazal et al. (2013)—as an alternative to the k-Means algorithm for spectral clustering. To the best of our knowledge, our work is the first to examine the structure of a differential-privacy mechanism through the lens of algebraic topology and statistical inference.

Key words and phrases: Edge Differential Privacy; Persistent Homology; Random Dot-product Graphs.

[†] The Pennsylvania State University. suv87@psu.edu, jsh435@psu.edu .

Background

Random Dot-product Graphs. Statistical inference from graphs (equivalently networks)—given by $G = (V, E)$, and consisting of $V = \{v_1, v_2, \ldots, v_n\}$ distinct vertices (equivalently nodes) and $E \subseteq V \times V$ edges—ordinarily begins by embedding the vertices of the graph as points in a space; most commonly via spectral embedding of the adjacency matrix A, or the Laplacian L associated with the graph G. A rigorous justification for spectral embedding from a statistical standpoint stems from recent work establishing that the embedding asymptotically recovers meaningful latent information for the class of *random dot-product graphs.* (Lei, 2018; Rubin-Delanchy et al., 2017; Rubin-Delanchy, 2020; Solanki et al., 2019). We refer the reader to Athreya et al. (2017) for a recent survey of results. The defining characteristic for RDPGs is that the likelihood of a connection between two vertices v_i, v_j is characterized by the dot-product $\langle \boldsymbol{x}(v_i), \boldsymbol{x}(v_j)\rangle_2$ of their respective latent positions $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$. Given G, a RDPG with latent positions $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and adjacency matrix A, the *adjacency spectral embedding* of G into $\mathbb{R}^m$ for $0 < m \leq d$ is given as follows. Consider the spectral decomposition $A = P\Lambda P^{\top} + Q\Omega Q^{\top}$, where Λ is the $m \times m$ diagonal matrix comprising of the m–largest eigenvalues of A (by absolute value). The adjacency spectral embedding is the collection of points $\{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_n\} \subset \mathbb{R}^m$ obtained from the columns of the matrix $\left[\hat{\boldsymbol{x}}_1 | \ldots | \hat{\boldsymbol{x}}_n\right] = P|\Lambda|^{\frac{1}{2}}$.

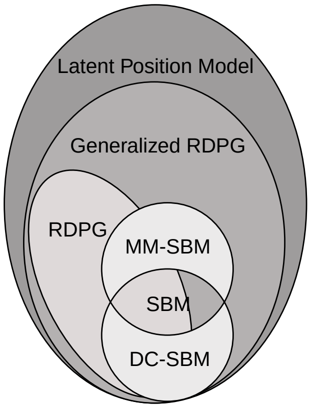


FIG. Graphical Models

Definition 1 (RDPG and Generalized RDPG)**.** *Let $\mathbb{P}$ be a probability measure on $\mathbb{R}^d$ and $p, q \in \mathbb{Z}_+$ such that (i) $p + q = d$, (ii) for $\boldsymbol{X} \sim \mathbb{P}$ the second-moment matric $\Delta_{\mathbb{P}} \doteq \mathbb{E}\left(\langle \boldsymbol{X}, \boldsymbol{X}\rangle_2\right)$ has rank d, and (iii) for $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \sim \mathbb{P}$ it holds that $\langle \boldsymbol{X},\ \mathbb{I}_{p,q}\boldsymbol{Y}\rangle \in [0, 1]$ a.s., where $\mathbb{I}_{p,q} = \mathsf{Diag}\left(\mathbb{1}_p{}^{\top}, -\mathbb{1}_q{}^{\top}\right)$ is the indefinite identity matrix with signature (p, q).*

Then $G \sim \mathsf{gRDPG}(\mathbb{P}, p, q)$ is said to be a generalized random dot-product graph with signature (p, q) and base measure $\mathbb{P}$, if for $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ iid from $\mathbb{P}$ and $i, j < n$ the adjacency matrix

$$A_{ij} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim \text{Bernoulli}\left(\left\langle \boldsymbol{X}_i, \mathbb{I}_{p,q}\boldsymbol{X}_j \right\rangle_2\right).$$

Furthermore, when $p = d$ and $q = 0$, then $G \sim \mathsf{RDPG}(\mathbb{P})$ is a random dot-product graph.

It is easy to see that RDPGs are identifiable up to arbitrary transformations of the orthogonal group $\mathcal{O}(d)$; similarly, gRDPGs are identifiable up to transformations of the indefinite orthogonal group $\mathcal{O}(p, q)$. This is referred to as the inherent non-identifability of RDPGs. Notwithstanding, they still encompass a large class of commonly used models. Of special interest in this work are cases when $\mathbb{P}$ is supported on a low-dimensional structure $\mathcal{M} \subset \mathbb{R}^d$. For example, if $\mathbb{P} = \delta_{\boldsymbol{x}}$ is a dirac mass at point $\boldsymbol{x} \in \mathbb{R}^d$ with $\pi \doteq \|\boldsymbol{x}\|_2 \in (0, 1)$, then $\mathsf{RDPG}(\mathbb{P}) = \mathsf{gRDPG}(\mathbb{P}, p, q)$ is an Erdős-Rényi graph with parameter π. Similarly, if $\mathbb{P} = \sum_k \alpha_k \delta_{\boldsymbol{x}_k}$ is a mixture of k dirac masses, then the resulting $\mathsf{RDPG}(G)$ will be an SBM. The sociability network from Caron and Fox (2017) is another example of a gRDPG where the underlying structure is a 1-dimensional manifold.

Differentially-Private Synthetic Graphs. Differential privacy for graphs typically constitue one of two cases: (i) Node differential privacy, and (ii) Edge differential privacy. They differ in how one interprets the notion of "neighboring" graphs. Node differential privacy is a stronger notion of privacy, where the identity of the nodes in the network is protected. In contrast, edge differential privacy is used to protect the worst-case disclosure risk of interactions (represented by the edges) between the components (represented by the nodes), when the identity of the nodes is known *a priori*. Indeed, satisfying node differential privacy implies edge differential privacy. In many applications, however, node differential privacy offers such strong privacy protection that it precludes meaningful analysis (Qin et al., 2017).

Definition 2 (ϵ–edge differential privacy)**.** *Let $\epsilon > 0$ and $\mathcal{G}^n = \{(V, E) : |V| = n\}$ denote the set of all graphs with n vertices. A randomized mechanism $\mathcal{A} : \mathcal{G}^n \to \mathcal{G}^n$ satisfies ϵ–edge differential privacy if for all graphs $G_1 \sim G_2$ differing in a single edge, and for all sets of graphs $S \subseteq \mathcal{G}^n$,*

$$\mathbb{P}\left(\mathcal{A}(G_1) \in S\right) \le e^{\epsilon}\,\mathbb{P}\left(\mathcal{A}(G_2) \in S\right)$$

The privacy mechanism we focus on in this work is the symmetric edge-flip mechanism edgeFlip, as described in Karwa et al. (2017); Qin et al. (2017); Imola et al. (2020).

Definition 3 (edgeFlip)**.** *Let $\mathbb{U}^n$ denote the class of symmetric, holllow, binary $n \times n$ matrices. Given a graph G, edgeFlip is the randomized mechanism $\mathcal{A}_\epsilon : \mathbb{U}^n \to \mathbb{U}^n$ given by*

$$\mathcal{A}_\epsilon(A)_{ij}|A_{ij} = \begin{cases} Z_{ij}|A_{ij} & i \le j \\ Z_{ji}|A_{ji} & i > j, \end{cases}$$

where, for $\pi(\epsilon) \doteq \frac{1}{1+\exp(\epsilon)}$ the upper-triangular random variables $\{Z_{ij}\}_{i \le j}$ are given by

$$Z_{ij}|A_{ij} = \begin{cases} 1 - A_{ij} & w.p.\ \pi(\epsilon) \\ A_{ij} & w.p.\ 1 - \pi(\epsilon). \end{cases}$$

Karwa et al. (2017) show that edgeFlip satisfies ϵ–edge differential privacy. In addition, edgeFlip is simple to implement and flexible. Karwa et al. (2017) discusses a central differential privacy setting for edgeFlip where the synthetic graph is released by a trusted curator, while Qin et al. (2017) discusses a local differential privacy implementation. In this work, we show that for the class of RDPGs, edgeFlip preserves the local geometric and global topological features which underlie the graph.

Topological Data Analysis. TDA is a framework that provides mathematical, statistical and algorithmic tools to extract geometric and topological structures in complex data. Informally speaking, given a collection of points $\mathbb{X}_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$, persistent homology summarizes the multiscale features underlying the data concisely as a persistence diagram $\mathsf{Dgm}(\mathbb{X}_n)$. The basic process is as follows. At each resolution $r > 0$, an object, called the simplicial complex K_r, is constructed to encode the geometric and topological information underlying the data. For example, the number of connected components, loops, holes, etc. Persistent homology keeps track of the evolution and changes in homology at different resolutions – for example, as r increases a new topological features are born at $r = b$, and subsequently vanishes at resolution $r = d > b$. The pair of birth and death times associated with the collection topological features, $\{(b, d) : 0 \le b < d < \infty\} =: \mathsf{Dgm}(\mathbb{X}_n)$, is called a *persitence diagram*. By examining the sample points under a spectrum of resolutions, persistent homology sheds light on the local geometric *and* global topological features which underlie the sample points. The space of persistence diagrams is endowed with a collection of Wasserstein metrics $\{W_p(\cdot, \cdot)\}_{p \ge 1}$, and the special case of $W_\infty(\cdot, \cdot)$ is referred to as the *bottleneck distance*. An important property of persistence diagrams is that, although they encode subtle features underlying the data, they are invariant to $\mathcal{O}(p, q)$ transformations. This makes persistence diagrams particularly useful for analyzing the latent structure underlying RDPGs – which have rich geometric information but are limited by their inherent non-identifiability. We refer the reader to Edelsbrunner and Harer (2010) for a comprehensive introduction to TDA, and Chazal and Michel (2017) for a concise overview.

Main Results

Suppose $\mathbb{P}$ is a distribution supported on a low-dimesnional structure $\mathcal{M} \in \mathbb{R}^d$ satisfying the conditions of Def. 1. If $G \sim \mathsf{RDPG}(\mathbb{P})$, we begin by showing that for edgeFlip, $\mathcal{A}_\epsilon(G)$ is also an RDPG.

Proposition 1 (Closure of edgeFlip)**.** *If $G \sim \mathsf{RDPG}(\mathbb{P})$ with $\mathrm{supp}(\mathbb{P}) = \mathcal{M} \in \mathbb{R}^d$. Then for $\epsilon > 0$ $\mathcal{A}_\epsilon(G) \sim \mathsf{RDPG}(\mathbb{Q})$ with $\mathrm{supp}(\mathbb{Q}) = \mathcal{M}'$, where $\mathcal{M}' \subset \mathbb{R}^{d+1}$ is the image of the map*

$$\phi : \boldsymbol{x} \mapsto \sqrt{1 - 2\pi(\epsilon)} \cdot \boldsymbol{x} \oplus \sqrt{\pi(\epsilon)},$$

and $\mathbb{Q} = \phi_\sharp \mathbb{P}$ is the pushforward of $\mathbb{P}$ through ϕ.

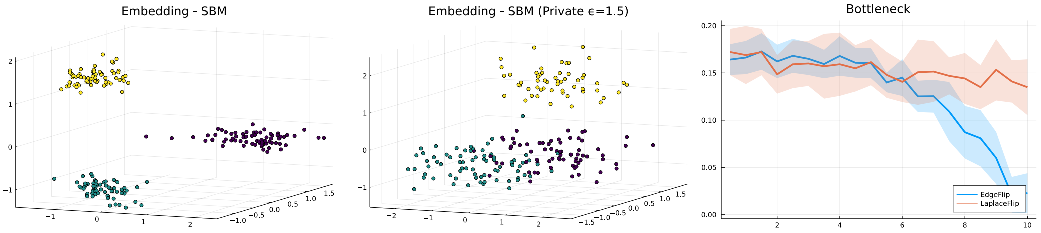


Figure 1. (Left) The effect of edgeFlip brings the clusters closer together as per Corollary 1. (Right) edgeFlip outperforms LaplaceFlip, which is another ϵ–edge differentially private mechanism. Simulation results from some other mechanisms tested are omitted for clarity.

The following corollary is obtained as a consequence of Proposition 1. The first part establishes that $\mathcal{M}' \subset \mathbb{R}^d$ contracts towards the center as the privacy budget ϵ decreases. This phenomenon is illustraed in Figure 1. The second part establishes that, under the edgeFlip mechanism, the resulting $\mathcal{M}'$ is topologically equivalent to $\mathcal{M}$ in a strong sense. Third, when $\epsilon = 0$, the resulting RDPG for $\mathcal{A}_\epsilon(G)$ is an Erdős-Rényi graph with parameter $\frac{1}{2}$.

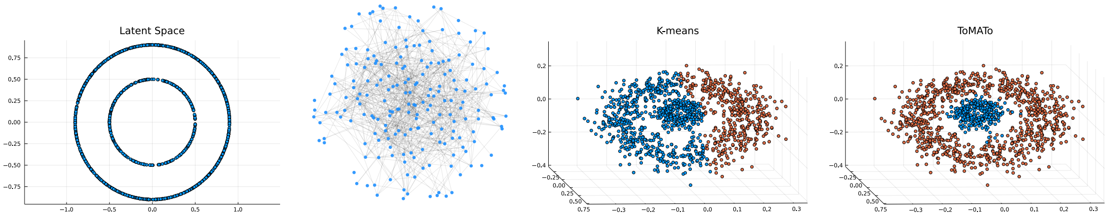
Corollary 1. *Under the same conditions as Proposition 1:* (i) *If G is a RDPG then $\mathcal{M}$ is bounded, and for $\epsilon_1 < \epsilon_2$, $\operatorname{diam}\mathcal{M}'_1 < \operatorname{diam}\mathcal{M}'_2$.* (ii) *When $\epsilon > 0$, $\mathcal{M}'$ is diffeomorphic to $\mathcal{M}$.* (iii) *When $\epsilon = 0$, $\mathcal{M}' \equiv \boldsymbol{p} \in \mathbb{R}^{d+1}$, where $\|\boldsymbol{p}\| = \frac{1}{2}$ and $\boldsymbol{p} \perp \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{M}$.*

Although Proposition 1 and Corollary 1 are explicitly stated for RDPGs, a similar holds for gRDPGs. However, when the signature (p, q) is non-trivial, the analogue of Proposition 1 relies on recognizing the indefinite inner product introduced by $\mathbb{I}_{p,q}$ as an inner-product in a finite dimensional Kreĭn space. We omit the details for brevity, and refer the reader to Lei (2018) for an excellent introduction.

Let $\mathbb{X}_n = \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_n\}$ denote the spectral embedding of G and $\mathbb{Y}_n = \{\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_n\}$ the embedding for $\mathcal{A}_\epsilon(G)$. The next establishes that even though $\mathsf{Dgm}(\mathbb{X}_n)$ and $\mathsf{Dgm}(\mathbb{Y}_n)$ do not converge to the same population limit; if we consider the *shift-invariant bottleneck distance* $W_\infty^{SI}(\cdot, \cdot)$ — a variant of the bottleneck distance introduced in Sheehy et al. (2018) — then $\mathsf{Dgm}(\mathbb{X}_n)$ and $\mathsf{Dgm}(\mathbb{Y}_n)$ converge to the same population equivalence class.

Proposition 2. *If $\mathbb{P}$ satisfies the (a, b)–standard condition for $\alpha > 1, \beta > 0$ (c.f. Chazal et al., 2015), and $G \sim \mathsf{RDPG}(\mathbb{P})$, then the following hold:* (i) $W_\infty(\mathsf{Dgm}(\mathbb{X}_n), \mathsf{Dgm}(\mathcal{M}')) \xrightarrow{p} 0$ *as* $n \to 0$. *The convergence rate follows from Solanki et al. (2019), and is identical to the non-private case albeit with sub-optimal constants owing to privacy. It follows that $W_\infty(\mathsf{Dgm}(\mathbb{X}_n), \mathsf{Dgm}(\mathbb{Y}_n)) \not\to 0$.* (ii) *Furthermore, for shift-invariant bottleneck distance,* $W_\infty^{SI}(\mathsf{Dgm}(\mathbb{X}_n), \mathsf{Dgm}(\mathbb{Y}_n)) \xrightarrow{p} 0$ *as* $n \to 0$.

Lastly, as noted in Rubin-Delanchy et al. (2017), several theoretical arguments can be made against the use of the k-Means algorithm for spectral clustering. Building on the topological perspective developed here, we propose using the ToMATo algorithm (Chazal et al., 2013) for spectral clustering. As illustrated in Figure 2, a clustering algorithm more suited to the data and the privacy mechanism leads to arguably better results. As future work, we hope to explore similar connections to graphons.



(A) Latent Space $\mathcal{M}$ (B) $G \sim \mathsf{RDPG}(\mathrm{unif}(\mathcal{M}))$ (C) k-Means clustering (D) ToMaTO Clustering

Figure 2. Illustration of topology-aware spectral clustering.

References

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 2008.

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

F. Caron and E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(5):1295, 2017.

F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017.

F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):1–38, 2013.

F. Chazal, M. Glisse, C. Labruère, and B. Michel. Convergence Rates for Persistence Diagram Estimation in Topological Data Analysis. *Journal of Machine Learning Research*, 16(110):3603–3635, 2015. ISSN 1533-7928. URL https://jmlr.org/papers/v16/chazal15a.html.

H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

J. Imola, T. Murakami, and K. Chaudhuri. Locally differentially private analysis of graph statistics. *arXiv preprint arXiv:2010.08688*, 2020.

B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

V. Karwa, P. N. Krivitsky, and A. B. Slavković. Sharing social network data: differentially private estimation of exponential family random-graph models. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 66(3):481–500, 2017.

J. Lei. Network Representation Using Graph Root Distributions. *arXiv*, Feb 2018. URL https://arxiv.org/abs/1802.09684v2.

Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren. Generating synthetic decentralized social graphs with local differential privacy. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 425–438, 2017.

P. Rubin-Delanchy. Manifold structure in graph embeddings. *Advances in Neural Information Processing Systems*, 33, 2020.

P. Rubin-Delanchy, J. Cape, M. Tang, and C. E. Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*, 2017.

D. Sheehy, O. Kisielius, and N. J. Cavanna. Computing the shift-invariant bottleneck distance for persistence diagrams. In *CCCG*, pages 78–84, 2018.

V. Solanki, P. Rubin-Delanchy, and I. Gallagher. Persistent homology of graph embeddings. *arXiv preprint arXiv:1912.10238*, 2019.