Studying the Utility Preservation in Social Network Anonymization via Persistent Homology

Tianchong Gao, Feng Li

Abstract—Following the trend of preserving privacy in online-social-network publishing, various anonymization mechanisms have been designed and applied. Differential privacy is an approach that guarantees the privacy level. Many existing mechanisms claim that they can also preserve the utility very well during anonymization. However, their utility analysis is always based on some specifically chosen metrics. While the existing metrics only partially present the graph utility, this paper aims to find a novel approach that describes the network in multiple scales. Persistent homology is a high-level metric, in that it reveals the parameterized topological features with various scales, and it is applicable for real-world applications. In this paper, four differential privacy mechanisms with different abstraction models are analyzed with traditional graph metrics and with persistent homology. The evaluation results demonstrate that all algorithms can partially or conditionally preserve certain graph utilities, but none of them are suitable for all metrics. Furthermore, none of the existing mechanisms fully preserves persistent homology, especially in high dimensions, which implies that the true graph utility is lost.

Index Terms—Social online networks; data publishing; utility and privacy; persistent homology; differential privacy

1 Introduction

Online Social Networks (OSNs) have exploded in popularity recently. The OSN providers release users' personal data to third parties for the purpose of feeding advertisements, recommending new friendships, and testing the effectiveness of applications. Since the data is sensitive to the users, anonymization mechanisms are applied to prevent the privacy leakage.

Differential privacy-based mechanisms are widely used because they provide a strong privacy guarantee [3, 5]. Different abstraction models are applied to capture the information of OSNs. Then noise is injected into the model according to the privacy level. *Sala et al.* employed the dK-2 series model, which transforms the OSN into pairs of degrees [15]. *Xiao et al.* proposed a similar model with the Hierarchical Random Graph (HRG) model in [18]. This model captures the probability of connectivity of nodes.

Different abstraction models have various advantages in preserving the network data. For instance, the dK-2 model keeps the information of the degree; then the published graph has a similar degree distribution. The HRG model keeps a cluster of nodes in the same branch on the tree when these nodes are closely linked in the original graph. Hence, the clustering information is partially preserved in the HRG model. In contrast, the dK-2 model only contains the linking information of two nodes, and contains little to no clustering data.

Do the existing anonymization mechanisms truly preserve the graph utility in their published graphs? Although these mechanisms claim to preserve the graph utility under some utility metrics, these claims are doubtful. First, the evaluation results in this paper demonstrate that none of the mechanisms perform well for all metrics. As mentioned above, the mechanisms based on the dK models may achieve unsatisfactory performance in preserving the clustering information. Second, each traditional utility metric cannot reveal the whole graph. For example, the information in the clustering coefficient is not covered in other metrics, like the shortest path length. Third, some traditional metrics, like the degree distribution, are not closely related with applications. While OSN data publishing aims to find a graph with similar properties to the original graph in real-world applications, preserving these metrics is not enough.

In this paper, in order to answer the question whether the existing anonymization mechanisms truly preserve the graph utility, we propose a novel angle, persistent homology, to analyze the anonymization mechanisms. Persistent homology comprehensively summarizes the OSNs and extracts the persistent structures. When we define the distance to be the number of hops of the shortest path between two users, it shows the difficulty of information transmission. Therefore, persistent homology is linked with applications such as friendship recommendation.

We evaluate the performance of the anonymization mechanisms under both traditional utility metrics and persistent homology barcodes. The traditional utility metrics contain both the graph utility metrics and an application utility metric. The three chosen graph utility metrics are the degree distribution, the clustering coefficient, and the shortest path length, while the chosen application metric is the influence maximization. The evaluation results under traditional metrics show that most mechanisms are suitable for preserving at least one metric; however, none of them can fully preserve all the metrics.

Under persistent homology, the evaluation results show that none of those mechanisms can preserve the barcode information. Because real-world OSNs have specific struc-

• *Tianchong Gao and Feng Li are with the school of Engineering and Technology, Indiana University - Purdue University Indianapolis, Indianapolis, IN 46202.*
E-mail: {tgao, fengli}@iupui.edu

This is the author's manuscript of the article published in final edited form as:

Gao, T., & Li, F. (2018). Studying the Utility Preservation in Social Network Anonymization via Persistent Homology. Computers & Security. https://doi.org/10.1016/j.cose.2018.04.003

HIGHLIGHTS

- A new topological feature, the persistent homology, is introduced into the utility analysis of the online social network.
- Existing differential privacy anonymization mechanisms are evaluated under both the traditional utility metrics and the persistent homology barcodes.
- The potential reasons of the persistent homology differences between the original OSN barcodes and the anonymized graph barcodes are studied.

ture information, anonymization mechanisms have difficulty duplicating the features of the original graph in the published graph. Specifically, we analyze the barcodes in the H_0, H_1, and H_2 domains. The analysis reveals that the original OSN graphs have stable structures. The existing OSN anonymization schemes overlooked the deep structure properties, and they are unable to simulate that through their design. Instead, the users in the anonymized graphs are more closely connected.

The major technique contributions are the following:

- We introduce a new topological feature, persistent homology, into the utility analysis of OSNs.
- We evaluate persistent homology in the published graphs of existing anonymization mechanisms.
- We study the potential reasons of persistent homology differences between the original OSN barcodes and the anonymized ones.

In this paper, we first propose the new metric, persistent homology, in Section 2. Then we introduce the differential privacy mechanisms and their abstraction models in Section 3. Afterwards the existing mechanisms are evaluated under different utility metrics in Section 4. Finally, we introduce some related research in Section 5 and discuss the conclusion and future work in Section 6.

2 Persistent Homology

In this paper, an OSN graph is modeled as an undirected graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. Each user is represented by a vertex in the graph and the relationships between users are the edges in the graph.

Persistent homology is proposed as a novel utility metric. Unlike the traditional metrics which describe the graph in specific angles, persistent homology gives multi-scales summarization of the graph. The barcodes are deployed to present persistent homology. They have two parts: the Vietoris-Rips simplicial complex records the structure change at different spatial resolutions in one dimension, and the Betti number records different dimensions.

2.1 Simplicial complex

Simplicial complex [16] is the basis of persistent homology. It contains points, line segments, triangles, and some high-dimensional components. A simplicial complex K is the set of simplices. It satisfies the following conditions:

1) Any face of a simplex from K is in K.
2) The intersection of any two simplices $\sigma_1, \sigma_2 \in K$ is either $\emptyset$ or a face of both σ_1 and σ_2.

A simplicial k-complex K has the property that the largest dimension of any simplex in K equals k. For instance, the 1-simplex is the line segment, the 2-simplex is the convex hull of the triangle, and the 3-simplex is the convex hull of the tetrahedron.

Fig. 1 shows an example with a 3-simplex, some 2-simplices, and some 1-simplices. The tetrahedron is the 3-simplex, where 3 means the simplex is in dimension 3. The face of the tetrahedron, e.g., the triangle {P, S, R}, is also a 2-simplex, and it is included in the simplicial complex. It is notable that there is no 2-simplex in the node set {T, S, U, W} because at least one pair of nodes has a distance greater than the desired threshold δ. Then {T, S, U, W} forms a 2-dimensional hole.

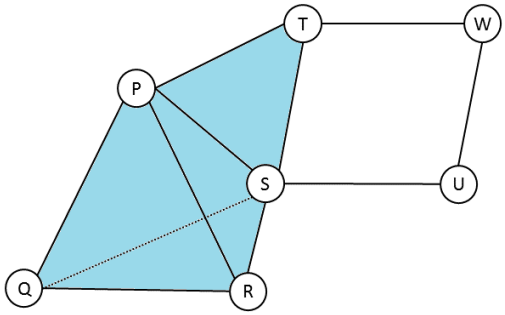


Fig. 1: Example of the simplicial complex

The threshold δ is applied in some abstract models like the Cech complex and the Vietoris-Rips complex [16]. In this study, we apply the Vietoris-Rips complex model. Given a distance parameter δ, this model determines the set of simplices $\{\sigma_1, \sigma_2, ..., \sigma_m, ...\}$ such that $d(v_i, v_j) \leqslant \delta$ for all node pairs (i, j) in any simplex σ_m.

2.2 Barcode

Persistent homology [6] is the homology of a filtration. In particular, changing the distance parameter δ results in an increasing sequence of Vietoris-Rips complexes:

$$K_0 \subseteq K_1 \subseteq ... \subseteq K_n = K,$$

While the single simplicial complex is related with δ, persistent homology collects the features in a wide range of distances, which gives a better view of the data space.

Fig. 2 gives an example of the Vietoris-Rips complexes chain. We can define K_0 to be the set in the first plot, which containis four nodes. Similarly, K_1 is the second plot and K_2 is the third. K_1 has not only the four nodes, but also four edges, which means four more 1-simplices. K_2 has more triangles than K_1, i.e., 2-simplices. Then K_0, K_1, and K_2 form an increasing chain of Vietoris-Rips complexes.

The homology group is defined based on the boundary homomorphism. First, we need to define boundary. For example, K_2 in Fig. 2 is a tetrahedron, i.e., a 3-simplex, including a triangle {P, R, S} and an extra node Q. In this 3-simplex, the boundary is the set of 2-simplices, i.e., the four triangles on four faces. Second, having the boundary of chain σ_n, we can get the kernel homomorphism and image homomorphism [20]. Specifically, the kernel subspace of K_n is the vector space of n-cycles Z_n. The n-cycle is a chain in n-dimension with empty boundary. For example, the triangle $\{P, S, T\}$ in Fig. 1 is a 1-cycle, while the node group $\{P, S, T, W\}$ is not a 1-cycle when edge T–W is appended to the triangle. The image subspace of K_{n+1} is the vector space of n-boundaries B_n. The n-boundary of a chain is the sum of the boundaries of all simplices in the chain. For example, the 1-boundary in Fig. 1 is the tetragon $\{T, W, S, U\}$. Third, we get the n-th simplicial homology group H_n. We have $H_n := Z_n/B_n$ [21]. In Fig. 1, the hole inside the tetragon $\{T, W, S, U\}$ is a component in H_1. The rank of H_n is

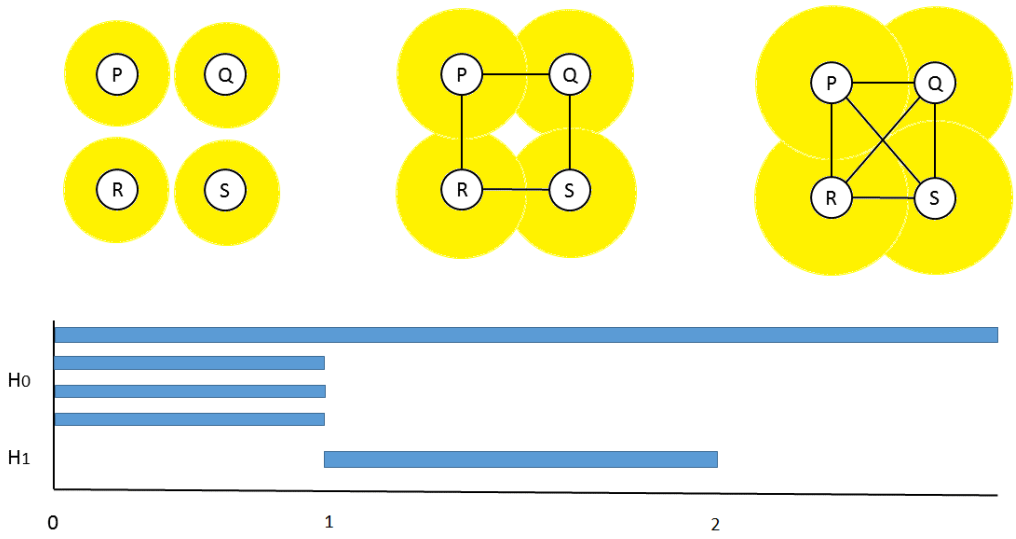
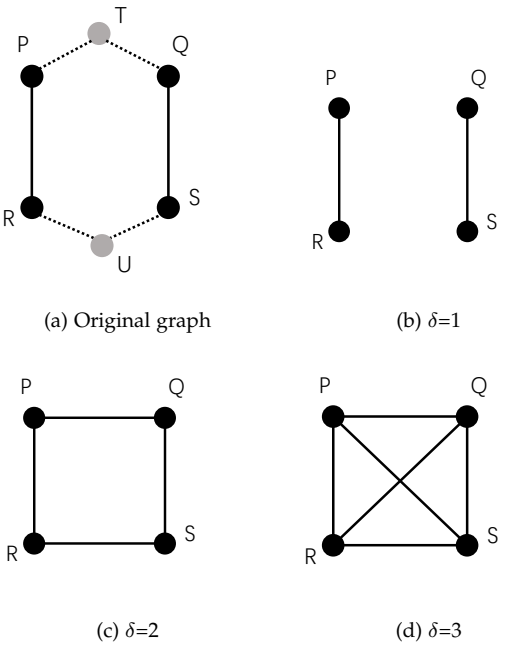


Fig. 2: Example of the barcode

denoted by the Betti number $Betti_n$. $Betti_n$ equals the number of $(n+1)$-dimensional holes. In particular, $Betti_0$ is the number of connected components, $Betti_1$ is the number of holes, and $Betti_2$ is the number of voids.

Applying the filtration gives the Betti intervals to describe the homology of H_n changes with δ. These intervals are called barcodes, where each one means a component or a hole in the corresponding dimension. The intervals show the birth time and death time of the components. In conclusion, the barcode collects the information of the existing periods of all components and holes when changing the distance δ.

Fig. 2 also shows a simple example of the barcode. The four nodes {P, Q, R, S} can form a square with a side of 1. We find that when $\delta < 1$, there are no edges in the graph. Each node is a component in H_0, so there are four bars in $[0, 1)$. When $\delta \geqslant 1$, the nodes are connected together to form a component, and this component exists until the end. Therefore, there is one bar of H_0 in $[1, \infty)$. When $\delta < 2$, the node pairs P–S and R–Q are not linked. Then the four nodes form a hole in 2-dimension, so there is one bar of H_1 in $[1, 2)$.



(a) Original graph (b) δ=1

(c) δ=2 (d) δ=3

Fig. 3: Example of the critical users and non-critical users in the graph

2.3 OSNs under persistent homology

The distance δ still needs to be defined such that persistent homology can be applied to the analysis of OSNs. Fundamentally, a basic OSN contains the users and their relationships, which are vertices and edges in the graph. And the anonymization mechanisms always focus on the topological perturbation, which also outputs graphs. Hence, the distance should capture the topology information of a graph. In this paper, the distance is defined to be the number of hops on the shortest path between the two vertices. In the recommendation application, this definition of distance is closely related to the difficulty of information transmission, which gives the analysis of barcodes practical meanings. Moreover, we can get the size of the components or holes through this definition (see Section 4.1.3).

Having the definition of graphs, the information in H_0 bars becomes a problem. Assuming the OSN is a connected graph, when $\delta \geqslant 1$, the whole graph becomes one component. It is similar to the bar $[0.5, \infty)$ in Fig. 2, that all connected graphs have a bar $[1, \infty)$. Assuming the OSN is disconnected, when δ increases, disconnected subgraphs are not able to connect. Then the H_0 bars can only show the number of disconnected subgraphs, which is trivial. To solve that problem, the nodes are randomly separated into two parts, critical users and non-critical users. Although the critical users are disconnected in the network, when δ increases, these critical users can be connected by non-critical users. Evaluation is performed on these critical users to show the

recommendation performance. When the distance metric of the whole network is generated, only the data of critical users are the input to obtain the barcodes.

Fig. 3 gives an example of the critical users and non-critical users. In the original graph, the four black nodes {P, Q, R, S} are the critical nodes while the two grey nodes {U, T} are non-critical users. The edges involving non-critical users are dotted. Because the persistent structure only cares about the critical users, the four nodes are apart when $\delta = 1$. However, in Fig. 3(c), these nodes are connected with the help of non-critical users and their friendships. These four nodes build an H_1 hole when $\delta = 2$, and the hole dies when $\delta = 3$.

3 Social Network Graph Anonymization

Anonymization mechanisms based on differential privacy are widely used in OSN data publishing, because they can achieve a strict guarantee of privacy. These mechanisms employ graph abstraction models to do the anonymization. In this section, we first introduce the definition of differential privacy and two kinds of abstraction models. Then we give a utility analysis with traditional metrics.

3.1 Differential privacy

Differential privacy is designed to protect the privacy between neighboring databases, which differ in only one element [5]. This means that the adversary cannot determine if one of the elements changes based on the releasing result. In the model of OSNs, the adversary cannot be sure if two users are linked in the original network.

Definition 1 (NEIGHBOR DATABASE). *Given a database D_1, its neighbor database D_2 differs from D_1 in at most one element.*

In this paper, the neighbor database/graph refers to an OSN with one edge added or deleted.

Definition 2 (SENSITIVITY). *The sensitivity ($\triangle f$) of a function f is the maximum distance of any two neighbor databases in the ℓ_1 norm.*

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\| \tag{1}$$

Definition 3 (ϵ-DIFFERENTIAL PRIVACY). *A randomized algorithm $\mathcal{A}$ achieves ϵ-differential privacy if for all neighbor datasets D_1 and D_2, and all $S \subseteq Range(\mathcal{A})$*

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D_2) \in S] \tag{2}$$

Equation (2) calculates the probability that two neighbor databases have the same result under the same algorithm. Based on this definition, researchers designed some graph abstraction models to calculate the numerical results and then add sufficient noise. To publish the anonymized graph, these abstraction models should be able to rebuild the graph from the perturbed numerical results. Several well known models, like the dK model and the HRG model, were designed to achieve differential privacy.

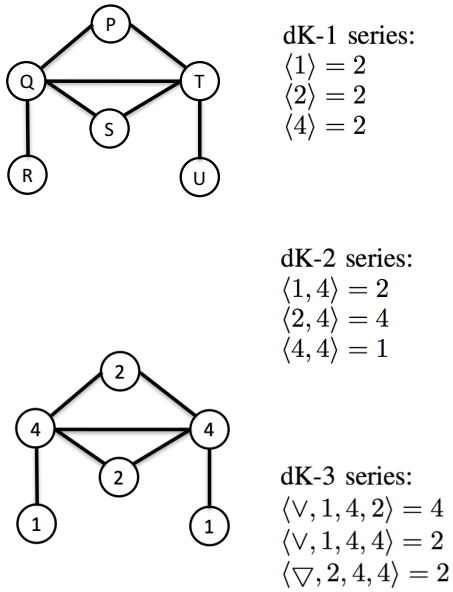


Fig. 4: The dK graph model

3.2 The dK model

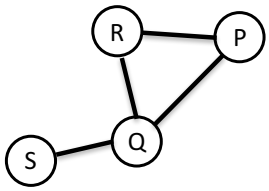
The dK-N model captures the degree distribution of connected components of size N in a target graph [11]. For example, dK-1 counts the number of nodes in each degree value. So dK-1 is also known as the node-degree distribution. The dK-2 model, also called the joint-degree distribution, captures the number of edges in each combination of two degree values, corresponding to the two nodes linked by that edge. The dK-3 model gets the number of 3-node subgraphs with different combinations of node degrees. There are two kinds of 3-node subgraphs with different structures: wedges and triangles.

Fig. 4 shows an example of the dK graph model. In particular, the dK-1 series $\langle 1 \rangle = 2$ means there are two nodes with degree 1. The dK-2 series $\langle 1, 4 \rangle = 2$ means there are two pairs of nodes with degrees 1 and 4. In the dK-3 series, the symbol $\vee$ shows the wedge structure, while the symbol $\bigtriangledown$ shows the triangle structure. After deriving the dK series, the differential privacy anonymization mechanism requires us to add noise to the query result. For instance, let $\langle 1 \rangle = 1$ replace $\langle 1 \rangle = 2$. Then the graph is regenerated based on the perturbed dK series.

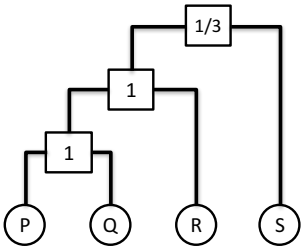
When the dK-1 model directly captures the degree-distribution information, the dK-2 and other dK models also store that information. In the example shown in Fig. 4, there are four dK-2 series containing degree 2. Since each degree-2 node builds two edges, there should be two nodes of degree 2 in the graph. Hence, the degree perturbation in the dK anonymized graphs mostly comes from the privacy request, and the dK models fully preserve the degree information.

3.3 The HRG model

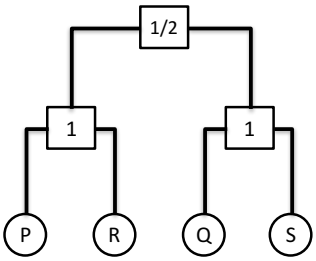
The HRG model captures the connection probability of nodes. Specifically, an HRG model is a dendrogram $\mathcal{T}$, which is a rooted binary tree with $|V|$ leaf nodes corresponding to $|V|$ vertices in the graph G. Each node on the



(a) Original graph



(b) Dendrogram $\mathcal{T}_1$



(c) Dendrogram $\mathcal{T}_2$

Fig. 5: The HRG model

tree except the leaf node has a number on it, which shows the probability of connection between its left part and right part.

Let L_r and R_r denote the left and right subtrees rooted at r, respectively. n_{Lr} and n_{Rr} are the numbers of leaf nodes in L_r and R_r. Let E_r be the total number of edges between the two groups of nodes L_r and R_r. Then, the posterior probability for the subtrees rooted at r is $p_r = E_r/(n_{Lr}n_{Rr})$. The posterior probability of the whole HRG model $\mathcal{T}$ to represent G is given by

$$p(\mathcal{T}) = \prod_{r\in\mathcal{T}} p_r^{E_r}(1-p_r)^{n_{Lr}n_{Rr}-E_r} \quad (3)$$

Fig. 5 shows an example of two possible dendrograms of the original graph. The p_r in each root node is first calculated. For instance, in the dendrogram $\mathcal{T}_2$, the root node of subtrees $\{P, R\}$ and $\{Q, S\}$ have a probability $1/2$. Because there are two edges between the two sets of nodes, we have $E_r = 2$ so $p_r = 2/(2*2) = 1/2$. Then we get the posterior probability of the two HRGs. $p(\mathcal{T}_1) = (1/3)(2/3)^2 \approx 0.148$ while $p(\mathcal{T}_2) = (1/2)^2(1/2)^2 \approx 0.006$. $p(\mathcal{T}_1)$ is greater than $p(\mathcal{T}_2)$, so $\mathcal{T}_1$ has more probability to represent the graph.

After collecting a group of dendrograms and their probabilities, the differential privacy anonymization mechanism requires us to resample one dendrogram with a noisy probability distribution [18]. The final graph is regenerated based on that dendrogram. Based on the HRG model, the

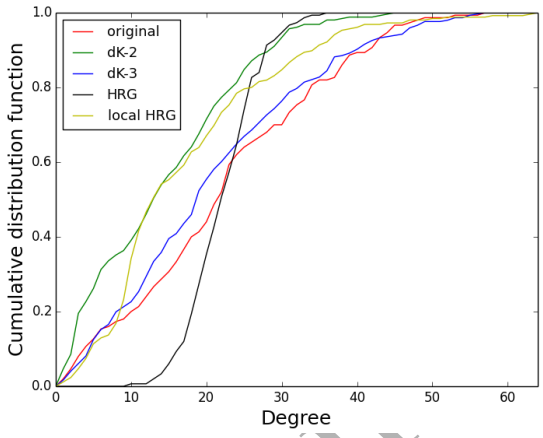


Fig. 6: Degree distribution of the Facebook dataset

local HRG means cutting the original graph into subgraphs, applying the HRG anonymization mechanism, and pasting them together.

The HRG model, including the local HRG model, builds dendrograms by grouping nodes together. When the models choose a cluster and transfer that cluster into a subtree in the dendrogram, there is a high probability that the regenerated graph has a similar cluster. When some nodes in the subtree do not belong to the cluster, the dendrogram also captures that information and stores it in the posterior probability p_r.

3.4 Graph utility analysis

In this paper, four different graph models, including the dK-2 model, the dK-3 model, the HRG model, and the local HRG model, are compared with each other. The graph utility is analyzed with four utility metrics: the degree distribution, the clustering coefficient, the shortest path length and the influence maximization.

The degree of a node in a network is the number of edges the node has to other nodes. The dK models directly store degree information, but the HRG model does not. Although the connection probability of nodes contains some degree information, some information is lost because of the model itself. Because the dK models directly store degree information, these models may, by definition, have better performance than the HRG model.

The clustering coefficient is a measure of how nodes in a graph tend to cluster together. The HRG model directly stores the clustering information, but the dK models do not. In the dK models, the nodes information is separately stored in groups of size N. The dK models can not fully preserve the clustering information when N is smaller than the size of clusters.

The shortest path length measures the average length from one node to every other node. Theoretically, there is no direct shortest path length information in any of the four models. The models can only preserve this information indirectly. For example, the dK model builds chains of nodes, so that two nodes in the same chain can preserve some path information. In the HRG model and the local HRG model, the linking information of nodes and clusters

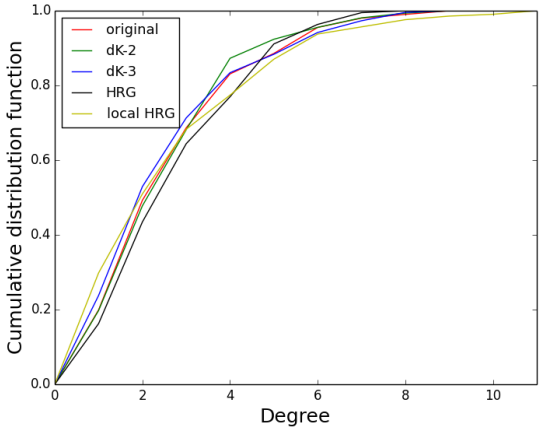


Fig. 7: Degree distribution of the ca-HepPh dataset

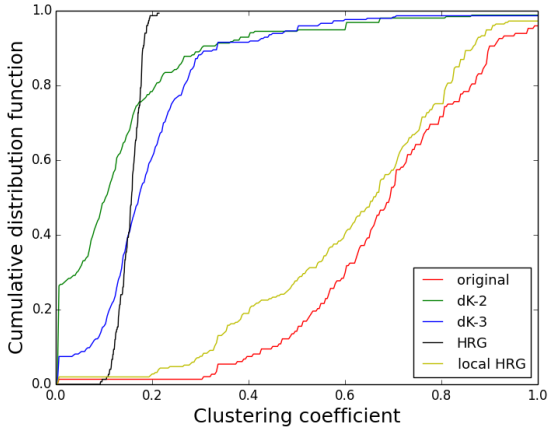


Fig. 8: Clustering coefficient of Facebook

allows us to derive the shortest path length. The shortest path length between two clusters is the minimum value of the shortest path length between any two nodes in the clusters.

In influence maximization, the greedy algorithm is applied to choose a set of seeds [8]. Then the algorithm based on the independent cascade model is evaluated to find the percentage of influenced users [14]. Similar to the shortest path length, there is no direct information of influence maximization theoretically stored in any of the models. Both the clusters and the shortest paths in the network may have impact on the propagation of information.

The analysis shows that existing mechanisms are able to preserve several particular utility metrics, which are limited by their models. The success in preserving one metric, like the dK-2 model preserving the degree distribution, does not equal success in preserving graph utility. If researchers do not want to evaluate multiple kinds of utility metrics, finding a comprehensive utility metric like persistent homology is significant.

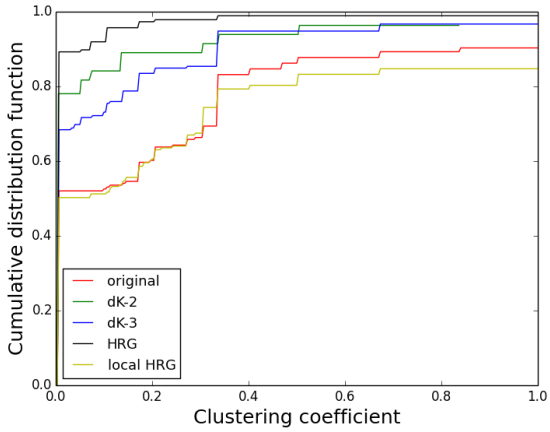


Fig. 9: Clustering coefficient of the ca-HepPh dataset

4 Evaluation and Analysis

Anonymization mechanisms based on these four models are used to anonymize the same graphs. Three datasets are analyzed, including the Facebook social network, the ca-HepPh collaboration network, and the Enron email network [10]. The privacy parameter, ϵ, is set to 5 for all mechanisms, which means these mechanisms will effect the same ability to preserve privacy. Then, performance is analyzed under various utility metrics, including the traditional metrics and the high-level metric, persistent homology.

4.1 Traditional utility metrics

4.1.1 Degree distribution

Fig. 6 shows the degree distribution of the Facebook dataset. The original graph has an average degree of 22.57, while the anonymized graph has the average degree of 14.52, 21.11, 22.57, and 17.45 in the dK-2 model, the dK-3 model, the HRG model, and the local HRG model anonymized graph. The standard deviations of the degree are 13.06, 10.08, 13.16, 4.66, and 11.61 in the original graph and the dK-2, dK-3, HRG, and local HRG results. We can find that the dK-3 model can preserve the degree distribution well, while the other three models only partially preserve the degree information.

Fig. 7 shows the degree distribution under the ca-HepPh dataset. The average degree is 2.98, 2.91, 2.89, 3.12, and 3.01 in the original graph and the dK-2, dK-3, HRG, and local HRG results. The standard deviations are 1.76, 1.64, 1.82, 1.64, and 2.11, respectively. The results show that all four models preserve the degree distribution.

Only the dK-3 model perfectly preserves the degree information in both of these datasets. The dK-2 model fails becase it does not contain as much information as the dK-3 model. Consequently, it is hard to use the dK-2 series to regenerate a graph. The HRG model and the local HRG model fail because the utility of degree is lost when generating the models (which is analyzed in Section 3.4).

4.1.2 Clustering coefficient

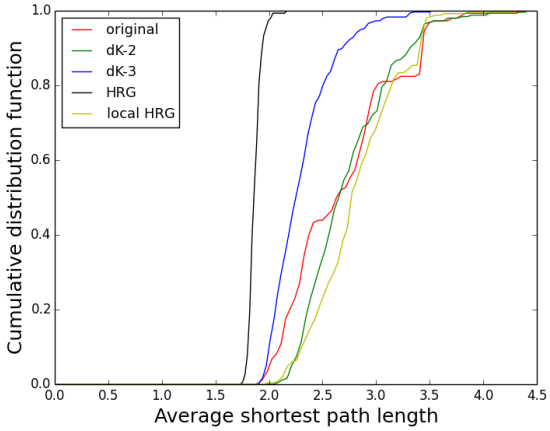


Fig. 10: Shortest path length of the Facebook dataset

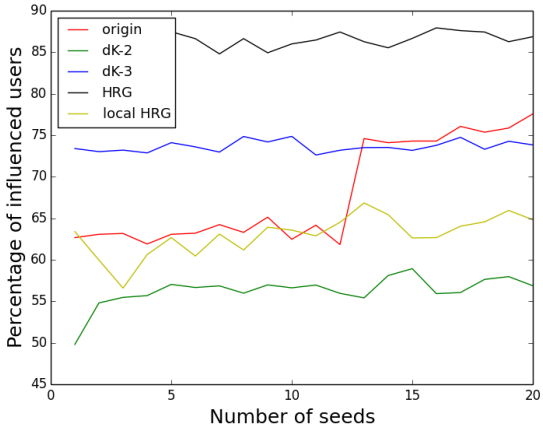


Fig. 12: Percentage of influenced users in Facebook

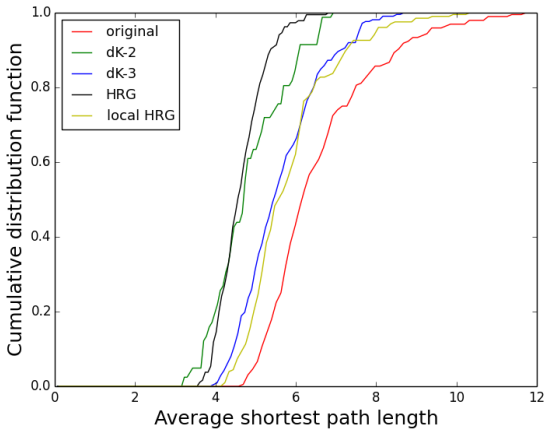


Fig. 11: Shortest path length of the ca-HepPh dataset

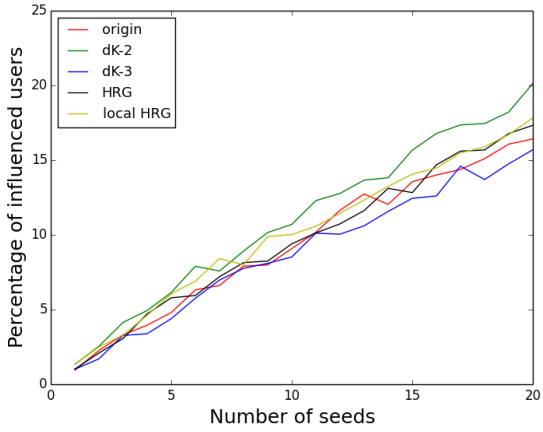


Fig. 13: Percentage of influenced users in ca-HepPh

Fig. 8 shows the clustering coefficient distribution of the Facebook dataset. The average clustering coefficient is 0.68, 0.14, 0.20, 0.15, and 0.62 in the original graph and the dK-2, dK-3, HRG, and local HRG results. The standard deviations are 0.19, 0.18, 0.15, 0.02, and 0.21, respectively. The results show that only the local HRG model can preserve the clustering coefficient information well, while other models break the clustering coefficient information.

Fig. 9 shows the clustering coefficient distribution of the ca-HepPh dataset. The average clustering coefficient is 0.22, 0.07, 0.10, 0.02, and 0.25 in the original graph, the dK-2, dK-3, HRG, and local HRG results, respectively. The standard deviations are 0.31, 0.18, 0.21, 0.11, 0.35, respectively. Similarly, only the local HRG model preserves the clustering information.

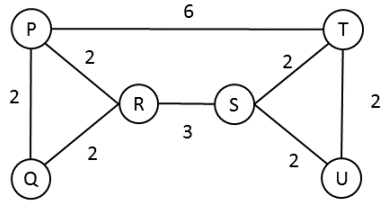
Besides the local HRG model, the utility with clustering coefficient is lost in the other three models. The experiment of the HRG model shows that it does not preserve the grouping information, though it theoretically has the ability. The potential reason is that a normal OSN has a large number of dendrograms and each dendrogram only has a small posterior probability of representing the graph.

In the HRG model, the number of possible dendrograms is $|\mathcal{T}|=(2|V|-3)!!$ for a graph with $|V|$ vertices, where !! is the semi-factorial symbol [4]. The HRG model is able to capture the clustering information and stores this information into some dendrograms. However, if the total probability of these useful dendrograms is relatively small, then it is hard for the final graph to preserve that information. Compared with the HRG model, the local HRG model splits the graph into small subgraphs, which significantly reduces the number of dendrograms. Hence, the local HRG is the only model to preserve the clustering information.

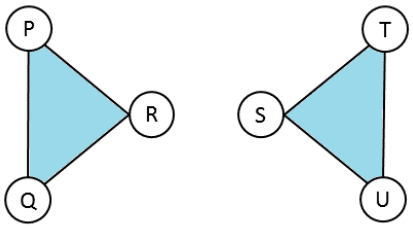
4.1.3 *Shortest path length*

Fig. 10 shows the average shortest path length of each node in the Facebook subgraph. The overall average shortest path length is 2.67, 2.73, 2.30, 1.87, and 2.81, corresponding to the original graph and the dK-2, dK-3, HRG, and local HRG results. The standard deviations are 0.50, 0.41, 0.29, 0.06, and 0.39, respectively. The results show that the local HRG model and the dK-2 model can partially preserve some shortest path length information.

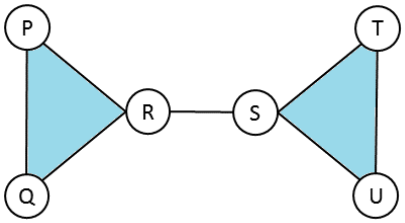
Fig. 11 shows the shortest path length in the ca-HepPh dataset. The overall average shortest path length is 6.55,



(a) Original graph with distance information



(b) δ=2



(c) δ=3

Fig. 14: Persistent homology example of H_0

4.81, 5.61, 4.62, and 5.81, corresponding to the original graph and the dK-2, dK-3, HRG, and local HRG results. The standard deviations are 1.38, 0.92, 1.00, 0.58, and 1.10, respectively. Here, none of the four models preserves the shortest path length.

Similar to the analysis, the experiment results reveal that all models cannot fully preserve the information. Because the models do not directly store the data, the anonymized graphs introduce errors that are independent of the privacy requirement. Consequently, the utility of the graphs is lost.

4.1.4 Influence maximization

In the evaluation, various sizes of the seed sets are chosen to test the number of influenced users. The propagation rate of information is set to 0.1. Fig. 12 shows the percentage of influenced users in the Facebook dataset. Compared with the original data, the root-mean-square errors (RMSEs) are 12.98, 8.17, 18.90, and 7.10 in the dK-2, dK-3, HRG, and local HRG data. None of the four models preserves the influence maximization information, since we can find a shape increase when there are 13 seeds in the original Facebook graph but it not exists in the anonymized graphs.

Fig. 12 shows the percentage of influenced users in the ca-HepPh dataset. Compared with the original data, the RMSEs are 1.87, 0.91, 0.70, and 0.92 in the dK-2, dK-3, HRG, and local HRG data. The anonymized graphs of the ca-HepPH all preserve some of the influence maximization information of the original graph.

4.1.5 Conclusion

The anonymized results are evaluated under three traditional graph utility metrics (the degree distribution, the clustering coefficient, and the shortest path length) and one application utility metric (the influence maximization). Generally, the dK-based models can preserve more degree distribution information, while the HRG-based models can preserve more clustering information. The dK-3 model always outperforms the dK-2 model, and the local HRG model outperforms the HRG model. Although all models preserve some of the information, none of them are suitable for preserving all four kinds of metrics.

In previous research, utility metrics are always carefully chosen and the adverse metrics are sometimes ignored. Because each traditional metric can only describe the graph in some aspects, the true utility of the anonymized graph is questionable. Hence, this paper deploys persistent homology to evaluate the anonymization mechanisms.

4.2 High-level utility metric

In this section, the barcodes of the original graph and the anonymized graphs are generated. Fig. 15 shows the barcodes of the Facebook dataset, while Fig. 16 shows the barcodes of the ca-HepPh dataset. In each graph, 60 users are randomly chosen to be the critical users. The results show that most components or holes are in the H_0 and H_1 dimensions. However, some anonymized graphs contain 3-dimensional holes. Then the barcodes show persistent homology in H_0, H_1, and H_2 dimensions. In the following subsections, we evaluate the ability of preserving persistent homology evaluated among different dimensions and examine why persistent homology information is lost.

4.2.1 H_0 dimension

Longest bar. The barcodes of the H_0 dimension shows the number of components existing in a distance interval. The longest bar from the beginning to the end exists in the barcode of every connected graph. When the distance δ is small, there are some disconnected components. When δ is long enough, the graph becomes a connected graph, and the disconnected components are combined into one. Hence, all barcodes have a bar from 0 to the end of the x-axis.

The length of the x-axis is chosen to be the length of longest shortest path in the graph, which ensures that all components and holes are included in the barcode. In OSNs, the longest distance shows the maximum number of hops of information transmission. When the distance is small, users are closely connected and the information is rapidly transmitted. The original Facebook graph has the length of 6, while the anonymized results are from 3 to 5. The original ca-HepPh graph has the length of 18, while the anonymized results are from 8 to 14.

Second longest bar. Besides the longest bar, the second one is $[0, 3)$ in the original Facebook network. This bar shows that when δ is no less than 3, the graph becomes one united component. The length of the second longest bar is a critical number, which shows the number of hops connecting the graph together.

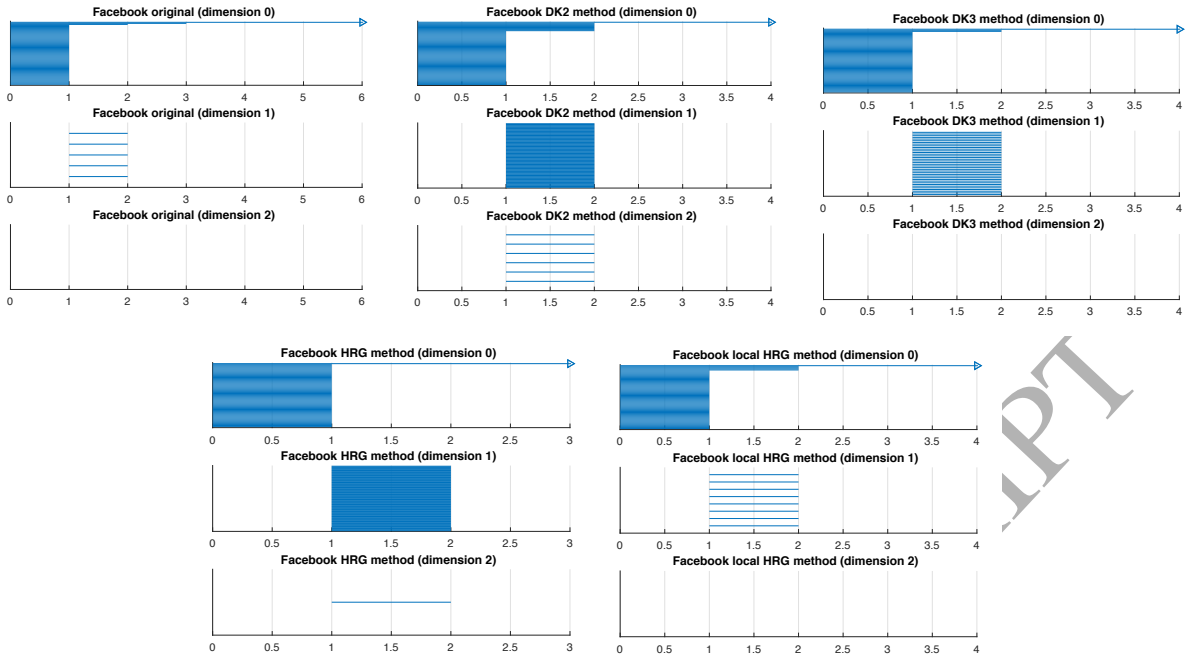


Fig. 15: Barcode of the Facebook dataset

Fig. 14 shows a simple example, in which the barcode is similar to the original Facebook graph. Some distance information is shown in Fig. 14(a), e.g., the node pair P–T has the longest distance, which equals 6. In Fig. 14(b), when δ=2, the simplices are disconnected. In Fig. 14(c), when δ=3, the whole graph is a connected component. In Fig. 15, we find that only the barcode of the original graph has the second longest bar, length 3, the barcodes of the dK-2, dK-3, and the local HRG results have length 2; and the barcode of the HRG result only has length 1.

The second longest bar length is meaningful in real-world applications. Assuming OSNs, for instance, transmit recommendation information, the second longest bar length represents the ability to cover the critical users. When the length is 1, the network has maximum transmission ability. If the critical users are guaranteed to recommend the product to their neighbors when they receive that information, the group of all critical users is covered. When the length is 2, the group of critical users is not fully covered when the non-critical users refuse to broadcast the information. However, the group can be fully covered when the non-critical users adjacent to the critical users can recommend the product. Like the example in Fig. 14(b), {P, Q, R} becomes connected with the help of some median non-critical users. When the length is 3, the transmission becomes more difficult. This requires all non-critical users who are within range of 2 hops of critical users to broadcast the recommendation information.

As described above, different second longest bar lengths show different ability of to transmit information. When the second longest bar length is 3 in the original Facebook graph, the anonymized results are 1 and 2. When the second longest bar length is 5 in the original ca-HepPh graph, the dK-3 anonymized result is 6, the dK-2 anonymized result is 1, the HRG result is 3, and the local HRG result is 4.

H_0 distribution. The H_0 distribution has meaning similar to the second longest bar length. One bar shows the number of hops to connect a subgraph together. In the Facebook dataset, the number of the $[0, 2)$ bars is 8 in the dK-2 anonymized results, while it is just 1 in the original graph. Other anonymized results have a very similar H_0 distribution. In the ca-HepPh dataset, the dK-2, HRG and local HRG results have more short bars than the original graph. The dK-3 result has more long bars than the original graph. As demonstrated above, more long bars means more difficulty transmitting information in OSNs.

The H_0 barcode shows that most anonymized OSNs are more closely connected than the original graph. Compared with the original results, all anonymized graphs have shorter longest bars and second longest bars. Except the dK-2 result of Facebook and the dK-3 result of ca-HepPh, all anonymized graphs have more short bars than the original results. In the application like recommendation, the anonymized graphs transmit information more easily than the original graphs.

4.2.2 H_1 dimension

In the barcode, an H_1 hole is a 2-dimensional hole surrounded by a 1-dimensional cycle. Fig. 17 shows a pentagon hole corresponding to the bar $[1, 2)$. Fig. 18 shows a hexagon hole corresponding to the bar $[1, 2)$ in H_1 and $[2, 3)$ in H_2. The first subgraph of each graph shows the distance information. According to the definition, the distance is the number of hops of the shortest path between two users. For instance, in Fig. 17, the distance of P–S is 2 if P–Q and Q–S are two pairs of adjacent nodes.

In the two examples, when $\delta = 1$, the holes are formed. In Fig. 18(c) we find that when $\delta = 2$, the hexagon is fully

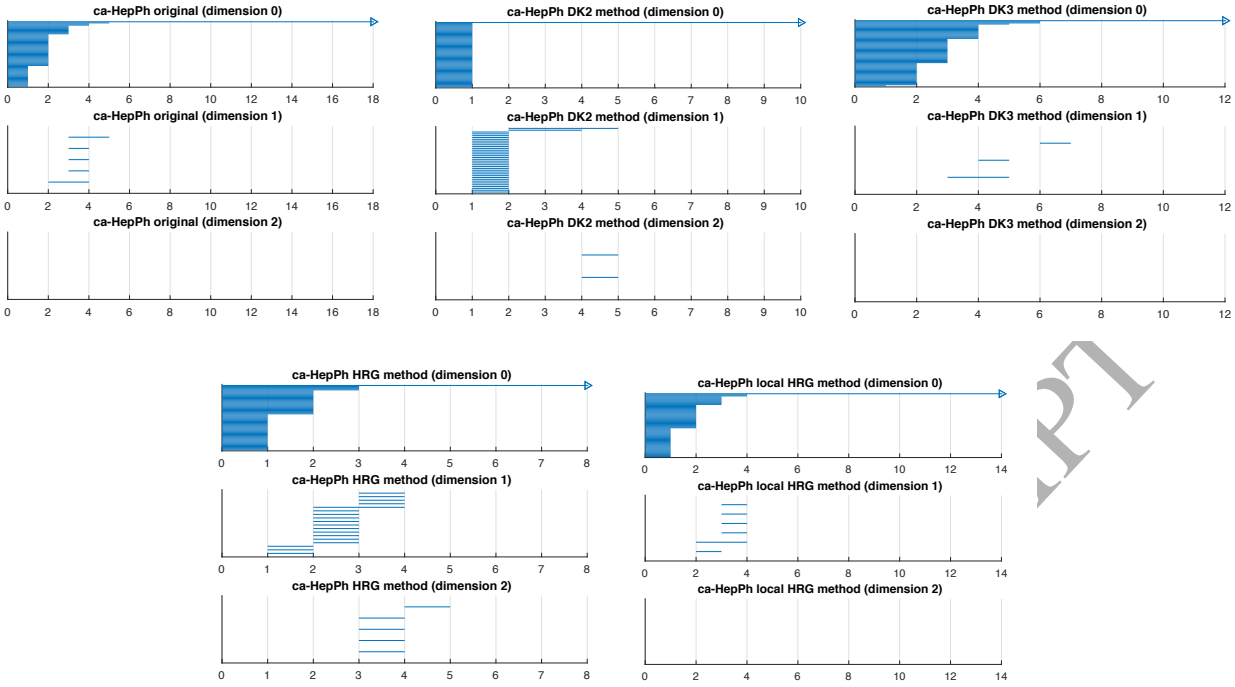


Fig. 16: Barcode of the ca-HepPh dataset

filled by triangles. Because triangles are not viewed as holes in persistent homology, there is no H_1 hole after $\delta = 2$. Although some polygons cannot be directly cut into triangles, they are filled by the combination of parts of triangles. For example, the quadrilateral $\{U, P, S, R\}$ is covered by the triangles $\{U, Q, S\}$, $\{U, P, Q\}$, and $\{S, R, Q\}$. The last two triangles are not limited in the quadrilateral. However, the H_1 bar only cares about the holes in dimension 2, i.e., the occupancy of triangles. If the whole area is occupied by triangles, the H_1 bar dies.

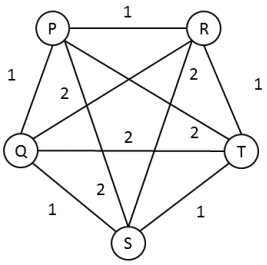
The two examples show that the birth time and the death time of the H_1 bar are related to the size of the hole. Specifically, having lengths of edges along the cycle, the birth time of a hole equals the largest value of these lengths. For instance, the birth time is 1 in Fig. 17. The death time is the longest distance, which makes the polygon become filled with triangles. If the birth time is $\delta = 1$ and the death time is $\delta = n$ in H_1, the hole is a polygon with $(3n - 2)$ sides, $(3n - 1)$ sides, or $(3n)$ sides. Fig. 19, 20, and 21 give the examples of polygons with 7, 8, and 9 sides. In these three polygons, edge lengths are all 1. The examples show that when $\delta = 2$, there is at least one area not filled by triangles in each polygon. When $\delta = 3$, all the polygons are filled by triangles. Hence, these polygons all have H_1 bar $[1, 3)$.

To analyze the effect of holes, we can imagine the breaking of holes. Take Fig. 17 as an example, if the node pairs P–S and R–S have direct links in the graph, then the holes are broken, and it becomes three triangles. The edges of the triangles all have the distance of 1. Because of the definition of simplicial complex, these triangles are not holes and consequently not H_1 barcodes. Whether the pentagon or the three triangles, the five nodes are connected. This means that if these users are guaranteed to transmit information, all of them are covered. However, comparing the two structures, we can see that the triangular structure is more stable. If two edges in the pentagon are deleted, the component becomes disconnected and some users are not covered. By contrast, when at most one edge is deleted from each of the three triangles, the structure is still connected. Hence, the stable OSN has less H_1 bars.

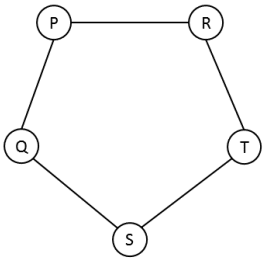
Fig. 15 shows that the original Facebook graph only has 5 bars in H_1. However, all anonymized graphs have more H_1 bars. The dK-2, dK-3, HRG, and local HRG results have 82, 38, 50, and 8 H_1 bars, respectively. In Fig. 16, the original graph, the dK-2, dK-3, HRG, and local HRG results have 5, 36, 3, 18, and 6 bars, respectively. Furthermore, although the H_1 bars in the dK-3 result of ca-HepPh graph is less than the bars in the original graph, the anonymized result is not more stable because it has a bar $[6, 7)$. According to the analysis above, this bar is a tetragon or a pentagon with edge distance of 6. The edge distance of 6 means that at least one pair of critical users needs 4 non-critical users to connect. This structure becomes weak when some users refuse to transmit the information. Hence, the bars with large birth time or death time have a huge impact on the stability of the network. In conclusion, all anonymized graphs are not as stable as the original graph, because they have more 2-dimensional holes or larger holes.

4.2.3 H_2 dimension

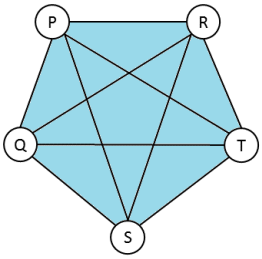
Similar to the H_1 holes, the H_2 bars represent the 3-dimensional voids bounded by 2-dimensional surfaces. In OSNs, the 2-dimensional surfaces are the clusters of users. According to the definition of simplicial complex, the cluster has to have at least three users to form a triangle. Then the H_2 bars show the stability of structures between clusters.



(a) Original graph with distance information



(b) δ=1



(c) δ=2

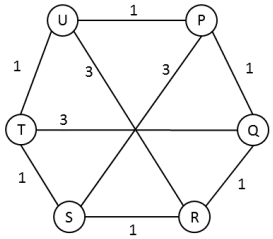
Fig. 17: Example of hole with bar $[1, 2)$ in H_1

In Fig. 18, there is an H_2 bar $[2, 3)$. Although it is difficult to analyze the hole in high dimension, i.e., H_2, we can combine H_1 and H_2 together to find some insights. In Fig. 18, the six nodes that form the hole have the maximum pairwise distance of 3, and the holes, in H_1 and H_2, die when $\delta = 3$. This implies that the death time, of holes in all dimensions, is the longest pairwise distance of the links formed by nodes in the hole. If the birth time is $\delta = 1$ and the death time is $\delta = n$ in H_1 and H_2, the hole is a $(2n)$-sided polygon or a $(2n + 1)$-sided polygon.

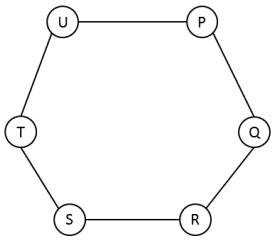
Fig. 15 and 16 show that the original graphs do not have H_2 bars. The H_2 bars exist in the dK-2 and the HRG anonymized graphs. The H_2 bars show that in the anonymized graph of the two models, some clusters are not strongly connected. If some edges are broken, these clusters will become disconnected with each other. In the recommendation application, if some linking users refuse to recommend the product, some clusters of users may not obtain the information. The anonymized network of the two models get bad performance.

4.2.4 *Impact of different parameters*

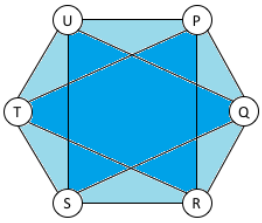
In order to evaluate the performance of the anonymization methods under different privacy criteria, we set ϵ to different values and generate the respective barcodes. In Fig.



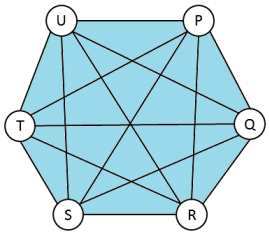
(a) Original graph with distance information



(b) δ=1

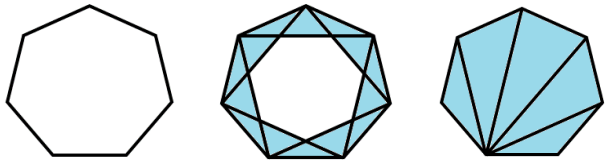


(c) δ=2



(d) δ=3

Fig. 18: Example of hole with bar $[1, 2)$ in H_1

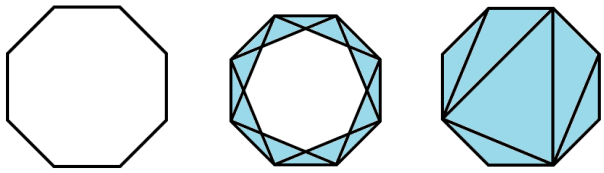


(a) $\delta = 1$ (b) $\delta = 2$ (c) $\delta = 3$

Fig. 19: Example of 7-sided polygons

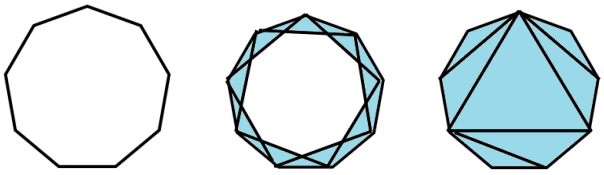
22, we present the results of utilizing the dK-3 model in the Enron email dataset, which preserves some persistent homology information. We get the same results in other models and other datasets.

Fig. 22 shows that when ϵ is smaller, i.e., the privacy level is stricter, there is less persistent homology information in the anonymized graph. For example, there are no H_1 and H_2 bars in the original graph. However, there are 4 H_2 bars



(a) $\delta = 1$ (b) $\delta = 2$ (c) $\delta = 3$

Fig. 20: Example of 8-sided polygons



(a) $\delta = 1$ (b) $\delta = 2$ (c) $\delta = 3$

Fig. 21: Example of 9-sided polygons

when $\epsilon = 0.5$ and 2 H_2 bars when $\epsilon = 1$. The presence of the H_1 and H_2 holes means that the anonymized graphs are not strongly connected. When $\epsilon \geqslant 5$, the H_1 or H_2 bars are eliminated, which means the corresponding holes do not exist in the anonymized graph.

The length of the longest H_0 bar is another important factor in the anonymized graph. The figure shows that even when $\epsilon = 50$, which is a very loose privacy criteria, the longest H_0 bar is much longer than the one in the original graph. As analyzed before, the longest distance in the original graph is 4, but it is 11 to 35 in the anonymized graphs. Furthermore, when the privacy requirement gets looser, the anonymized graph does not preserve more precise information of the longest bar length. It shows that the anonymization model cannot perfectly preserve the persistent homology information, rather than that the privacy noise causes the data distortion.

4.2.5 Conclusion

Comparing the four models, we found that the local HRG model is the best to preserve persistent homology information. However, some information, like the longest bar length, the second longest bar length and the H_0, H_1 distributions are not well preserved. The other models significantly change persistent homology, especially in adding H_1 and H_2 bars.

In the H_0 analysis, the anonymized graphs are more closely connected than the original graphs. However, the H_1 analysis and H_2 analysis show that these connections are based on some specific linking users. Hence, the anonymized graphs are not stable. By contrast, although the original graphs cannot rapidly transmit information like the anonymized graphs, their structures are more stable. The real-world dataset contains a compact structure that existing anonymization mechanisms find difficult to duplicate.

5 Related Research

Persistent homology is a description of topology features [21]. It was applied in analyzing persistent air passengers' networks [13], obtaining the distance of lower bounds between networks [7], and scheduling robot paths in uncertain environments [1]. The idea of barcode was proposed in [6] but it is novel in security analysis. In [17], the authors made an attempt to employ zigzag persistent homology [2] to achieve K-anonymity [19].

The simplest OSN anonymization mechanism is naive ID removal [12]. It does not change the topology of the social networks, and proved to be vulnerable to de-anonymization attacks [9]. K-anonymity based mechanisms are designed to anonymize relational data [19, 22]. Although they are suitable for some specific structural semantics, e.g., the degree distribution, attackers can use other structure information to conquer the anonymized data. Fortunately, advanced mechanisms like differential privacy-based methods are proposed to solve the vulnerability [3, 5].

6 Conclusion and Future Work

Existing anonymization mechanisms claim to achieve a balance between utility and privacy. In this paper, we demonstrate that when the privacy level is not extremely high (ϵ=5), the utility of the anonymized graphs are significantly violated. From a comprehensive point of view, with persistent homology, the trade-off between the utility and the privacy is broken. The violation of the utility is mainly related to the models, but not the privacy request. The performance of existing mechanisms on real-world datasets becomes untrustworthy.

In the future, we aim to design a new anonymization mechanism to preserve persistent homology. As shown in the analysis, persistent homology is related to the components and holes in the network. Hence, our work will start with the basic structural components. Compared to existing work, which preserves degree information or clustering information, persistent homology is believed to be more significant because it can be deployed in recommendation applications.

References

[1] Subhrajit Bhattacharya, Robert Ghrist, and Vijay Kumar. Persistent homology for path planning in uncertain environments. *IEEE Transactions on Robotics*, 31(3):578–590, 2015.

[2] Gunnar Carlsson, Vin De Silva, and Dmitriy Morozov. Zigzag persistent homology and real-valued functions. In *Proceedings of the 25th Annual Symposium on Computational Geometry*, pages 247–256. ACM, 2009.

[3] Rui Chen, Gergely Acs, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, pages 638–649. ACM, 2012.

[4] Aaron Clauset, Cristopher Moore, and Mark Newman. Structural inference of hierarchies in networks. *Statistical Network Analysis: Models, Issues, and New Directions*, pages 1–13, 2007.

[5] Cynthia Dwork. Differential privacy. In *Encyclopedia of Cryptography and Security*, pages 338–340. Springer, 2011.

[6] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[7] Weiyu Huang and Alejandro Ribeiro. Persistent homology lower bounds on network distances. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4845–4849. IEEE, 2016.

[8] Sergei Ivanov and Panagiotis Karras. Harvester: Influence optimization in symmetric interaction networks. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 61–70. IEEE, 2016.

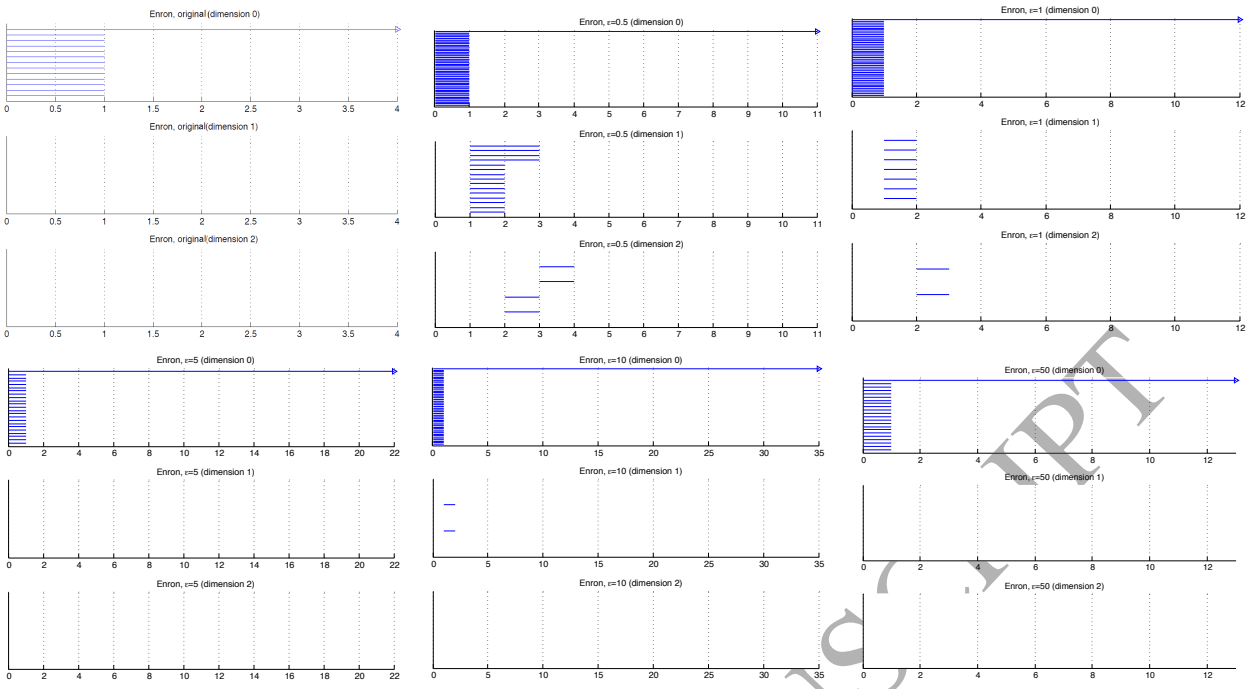


Fig. 22: Barcode of the Enron dataset with different ϵ

[9] Shouling Ji, Weiqing Li, Prateek Mittal, Xin Hu, and Raheem Beyah. Secgraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization. In *Proc. of USENIX Security Symposium*, 2015.

[10] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[11] Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. Systematic topology analysis and generation using degree correlations. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 135–146. ACM, 2006.

[12] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium*, pages 173–187. IEEE, 2009.

[13] Giovanni Petri, Martina Scolamiero, Irene Donato, and Francesco Vaccarino. Topological strata of weighted complex networks. *PloS One*, 8(6):e66506, 2013.

[14] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In *Knowledge-based intelligent information and engineering systems*, pages 67–75. Springer, 2008.

[15] Alessandra Sala, Xiaohan Zhao, Christo Wilson, Haitao Zheng, and Ben Y Zhao. Sharing graphs using differentially private graph models. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement*, pages 81–98. ACM, 2011.

[16] Edwin H Spanier. *Algebraic topology*, volume 55. Springer Science & Business Media, 1994.

[17] Alberto Speranzon and Shaunak D Bopardikar. An algebraic topological perspective to privacy. In *American Control Conference (ACC), 2016*, pages 2086–2091. IEEE, 2016.

[18] Qian Xiao, Rui Chen, and Kian-Lee Tan. Differentially private network data release via structural inference. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 911–920. ACM, 2014.

[19] Bin Zhou and Jian Pei. Preserving privacy in social networks against neighborhood attacks. In *IEEE 24th International Conference on Data Engineering, 2008*, pages 506–515. IEEE, 2008.

[20] Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, pages 1953–1959, 2013.

[21] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

[22] Lei Zou, Lei Chen, and M Tamer Özsu. K-automorphism: A general framework for privacy preserving network publication. *Proceedings of the VLDB Endowment*, 2(1):946–957, 2009.

Tianchong Gao is a PhD student with the Department of Electrical and Computer Engineering of Indiana University-Purdue University Indianapolis. His co-advisors are Dr. Feng Li and Dr. Xiaojun Lin. He has worked on problems in security, privacy, and social networks. His research vision is to explore privacy issues in computing and networking.

Feng Li is an Associate Professor of Computer and Information Technology at Indiana University-Purdue University Indianapolis (IUPUI). He received his PhD in Computer Science from Florida Atlantic University in Aug. 2009. His PhD advisor is Dr. Jie Wu. He joined the Department of Computer and Information Technology at IUPUI in Aug. 2009. His research interests include the areas of cybersecurity and trust issues, cloud, and mobile computing. He has published more than 50 papers in top conferences including INFOCOM and ICDCS.