Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization

Aritra Dasgupta*, Robert Kosara†, Min Chen‡

Abstract

Minimization of disclosure risks is a key challenge in publicly available visualizations that can potentially reveal personal information. Such risks are inherently dependent on the amount of information that adversaries can gain by manipulating visual representations and by using their background knowledge. Conventional risk quantification models proposed in the field of privacy-preserving data mining suffer from a lack of transparency in letting data owners control privacy parameters and understand their implications for disclosure risks. To fill this gap, we propose a visual uncertainty model for letting data owners understand the relationships between privacy parameters and vulnerable visualization configurations. Our main contribution is a probabilistic analysis of the disclosure risks associated with vulnerabilities in privacy-preserving parallel coordinates and scatter plots. We quantify the relationship among attack scenarios, adversarial knowledge, and the inherent uncertainty in cluster-based visualizations that can act as defense mechanisms. We present examples and a case study to demonstrate the effectiveness of the model.

1 Introduction

We live in an era when the need to protect personal data against potential adversaries is one of the biggest socio-technical challenges. To develop defense mechanisms against attacks, data owners have to carefully evaluate disclosure risks involving publication or visualization of potentially sensitive data on online platforms. The implications of disclosure risk and adversary's background knowledge have been studied in depth [36] in the field of privacy-preserving data mining. However, privacy-preserving data visualization being a nascent field [27], no rigorous methodology exists for analyzing how visualizations can ensure disclosure risk is minimized.

The amount of information that can be inferred from a privacy-preserving visualization is not just a function of the underlying data model, but it also depends on the visual representation. For example, the k-anonymity model of privacy combines a minimum of k records belonging to the *quasi-identifier* group (attributes when combined can identify an individual, like age, sex, zip code, etc.) into one cluster for protecting disclosure of individual records. In privacy-preserving parallel coordinates and scatter plots, k-anonymized clusters are displayed as trapezoids or rectangles instead of lines or points (Figure 1). But precise borders of clusters divulge information about individual records. If adversaries already know that an individual's data exists in the visualization and know any single data point, they only have to make a limited number of guesses to know both coordinates of the record. Previous research had demonstrated that purely data-based metrics, such as k-anonymity are not sufficient for understanding the degree of anonymization in a visualization [9].

Visual uncertainty [8], which is the inherent uncertainty that stems from the visual mapping process between the data space and the screen space of limited resolution, can act as a defense mechanism for confusing adversaries and reducing disclosure risk probabilities. The merits of integrating uncertain data models and privacy models have been studied in the data mining community [1], but such analysis is absent in privacy-preserving visualization. To fill this gap, we build a visual uncertainty based model for analyzing disclosure risks from a visual perspective, and in a manner that is transparent enough for data owners to understand the dependency between privacy parameters and visualization configurations.

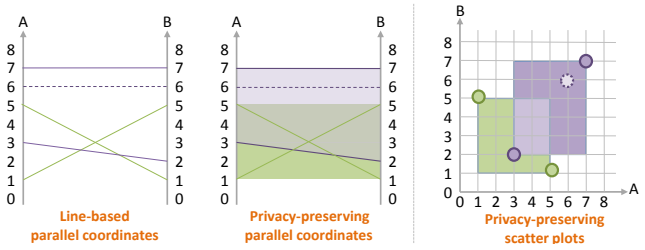


Figure 1: **Illustrating the k-anonymity model** of privacy preservation applied to cluster-based visualizations using pixel binning [11] that ensures at least k records belong to a group (where $k = 2$). Cluster edges represent data points and are more vulnerable to disclosure than the non-edge points, shown by dotted lines and points.

The contributions in this paper can be summarized as follows: 1) identify the *useful* sources of visual uncertainty that can act as defense mechanisms against adversarial attack scenarios, 2) develop a model for transparently evaluating disclosure risks associated with cluster-based visualizations, and 3) demonstrate how a data owner can control and configure the privacy parameters of a visualization, by accounting for disclosure risks in the presence or absence of adversarial background knowledge.

2 Related Work

In the field of privacy-preserving data mining [2], the goal of the k-anonymity model [37, 38] is to prevent re-identification by linking quasi-identifiers that co-exist in public and private databases. This is done by making k records identical to each other (Figure 1) and thus preventing identity disclosure as an adversary is unable to identify a particular individual. The k-anonymity model has its shortcomings for preventing attribute disclosure, where the identity of an individual is linked with sensitive values of an attribute (e.g., disease name as cancer). The field of privacy-preserving data visualization [7, 11, 39, 40] has adapted and extended models from the data mining community for striking a balance between privacy gain and loss of utility in anonymized visual representations. However, with the exception of the work by Chou et al. [6], which analyzes effects of perceptual masking of graph visualizations, relatively less attention has been paid to understanding the vulnerability and disclosure risks in a privacy-preserving visualization. This is important for understanding how such visualizations can be attacked and for data owners to get guidance on privacy parameters which guarantee a satisfactory level of non-disclosure guarantee.

Metrics like l-diversity [29] and t-closeness [25] have been proposed for analyzing attribute disclosure risks in privacy-preserving data mining. In this paper we restrict our scope to analyzing risks associated with identity disclosure, as that is a first step towards

*NJIT, E-Mail: aritra.dasgupta@njit.edu
†Tableau Research, E-Mail: rkosara@tableau.com
‡Oxford University, E-Mail: min.chen@eng.ox.ac.uk

ensuring disclosure risk in any form is minimized, and focus exclusively on the visualization adaptation of the *k*-anonymity model. In the context of privacy-preserving visualization [11, 13], *k*-anonymity is achieved by the *k*-member clustering algorithm [3] that has been adapted based on screen-space metrics [10, 12]. The analysis provided in the work reported here is based on the axis-pairwise clustering approach applied to both parallel coordinates and scatter plots.

In our previous work, we had examined a number of scenarios where visualization techniques might be used by adversaries to violate data privacy [13], implying that new methods need to be developed for minimizing disclosure risks, which is the focus of our work reported here. While analysis of disclosure risks [30, 36] has been studied in the context of databases, our focus is to exploit uncertainty in the screen-space to defend against attack scenarios. Uncertainty in the screen-space, or visual uncertainty [8] is a new perspective to deal with the uncertainty problem in visualization. So far, most existing work in visualization relates to data-space uncertainty (e.g., [21, 31, 35]) and uncertainty involving geometrical primitives, like iso-surface rendering [32]. The conceptualization of visual uncertainty takes communication of information into account and looks at both the encoding and decoding aspects of uncertainty on screen. The latter is similar in principle to the idea of uncertainty due to perception [33] and to the differentiation between input and output uncertainty [19]. Study of the sources of uncertainty can help data owners and visualization designers refine the visualization output for effective privacy preservation.

3 Visual Uncertainty Vs Disclosure Risk

Sources of visual uncertainty in cluster-based, privacy-preserving scatter plots and parallel coordinates can act as a defense mechanism against adversarial attacks. In this section, we describe the attack scenarios and our visual uncertainty model.

Attack Scenarios: Once data is anonymized, the main threat to privacy is the risk of re-identification [16, 24] of either the sensitive attributes or the individuals associated with them. When an adversary is able to learn about individual values of quasi-identifiers or sensitive attributes, this type of attack is categorized as attribute disclosure. In privacy-preserving visualization, where clusters have edges representing data points (Figure 1), those are particularly vulnerable to this type of attack. On the other hand, continuous cluster edges across different axes can lead to disclosure of multiple attributes and potentially reveal the identity of an individual, leading to identity disclosure. These disclosure risks are affected by how much an adversary knows about the data. Two kinds of re-identification scenarios [22] can be imagined based on the background knowledge of the adversary. The first one is called **prosecutor re-identification scenario**, where an intruder (e.g., a prosecutor) knows that a particular individual (e.g., a defendant) exists in an anonymized database and wishes to find out which record belongs to that individual. In the second one, known as the **journalist re-identification scenario**, an adversary tries to re-identify an arbitrary individual. The intruder does not care which individual is being re-identified, but is only interested in being able to claim that privacy breach is possible. We describe how these scenarios affect disclosure in privacy-preserving visualization in later sections.

Visual Uncertainty as Protection Against Disclosure: We follow the classification of causes and effects of visual uncertainty described by Dasgupta et al. [8]. Figure 2 describes how sources of visual uncertainty can act as a defense mechanism and how adversaries can try to exploit vulnerabilities, causing unintended disclosure of attribute values or records. We describe these sources below:

Cluster configuration: With or without any background knowledge, an adversary can analyze internal configurations of clusters, such as: determining the number of data points in a cluster, finding connections among them, determining if a specific record is in a particular cluster, finding coordinates of a two-dimensional record,

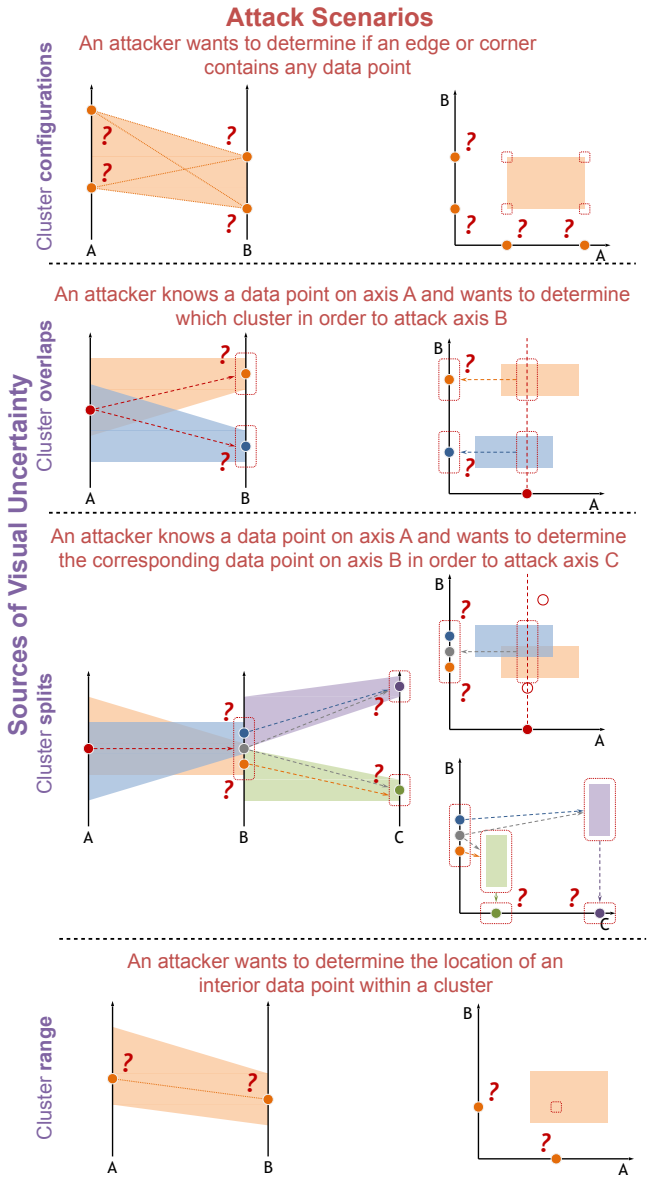


Figure 2: **Illustrating the relationship between adversarial attack scenarios and sources of visual uncertainty** in cluster-based parallel coordinates and scatter plots. Uncertainty due to cluster ranges and configurations helps in privacy-preservation when an adversary attempts to gain knowledge about the data at a lower level of granularity than what is shown, in case of **journalist re-identification scenario**. Uncertainty due to overlaps and split helps in privacy-preservation when an adversary attempts to determine the cluster membership of a known data point in case of the **prosecutor re-identification scenario**. These scenarios may not follow a particular sequence and one scenario can lead to another.

and so on. It is not difficult to combine such basic attacks with acquired knowledge to enable complex attack actions, for instance, given a disclosed value of a record (i.e., an end point on an axis), find all other values of this record (i.e., all end points of this line; or given a disclosed line in a cluster, find all other lines in the cluster.

Cluster overlaps: If an adversary knows that a certain data point or a record exists in the database in case of a prosecutor re-identification scenario, overlaps can make it difficult for them to identify which cluster that entity belongs to. In absence of such a background

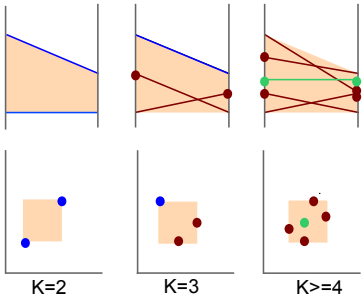


Figure 3: **Illustrating different cluster configurations**. From left to right, a cluster can be defined by: all edge elements, a combination of edge and pivot elements, and also no edge element when $k >= 4$. Edge elements are shown in blue, pivot elements in red, and free elements in green.

knowledge, overlaps can also confuse the adversary if a data point at all exists or not. Both of these cases thus lead to identity uncertainty.

Cluster splits: Once adversaries find out a point on an axis, they would want to find the point on the adjacent axis. Due to axis-pairwise clustering [11], clusters appear to be non-contiguous and it can be difficult to trace the path of the record across different axis-pairs (Section 5). However the splitting points can also reveal locations of data points and lead to attribute disclosure.

Cluster range: Cluster range causes precision and granularity uncertainty which have to be overcome to gain knowledge about number of records per cluster (which is at least k) and lack of spatial accuracy which has to be overcome to know the exact coordinates of a point. Exact coordinates of points can also be revealed by cluster splits between adjacent axes, for which traceability uncertainty needs to be overcome.

4 Uncertainty Due to Cluster Configurations

A two-dimensional cluster configuration (Figure 3) in scatter plots and parallel coordinates is defined by the pixel coordinates of the data points within the cluster. A pixel coordinate in scatter plots is represented by a point, while in parallel coordinates, it is a line, by the point-line duality principle [20]. For the sake of clarity, we refer to the points/lines as elements within the cluster. The shape of a cluster in parallel coordinates can be a quadrilateral or a triangle and the same in scatter plots is either a rectangle or a line. In case of triangular clusters, the uncertainty is very low, since coordinates of the two borders or end-points are always known. Thus $k = 2$ has no anonymization effect in case of triangular clusters. In case of numerical dimensions, the probability of occurrence of quadrilateral clusters is much higher than triangular or linear ones. In this section, we study the orientation of the elements within clusters for different values of k, how they cause different types of uncertainty and their relation to privacy for different values of k.

4.1 Different elements in a cluster

Location of the elements within the cluster determines the uncertainty associated with them. We define the different elements within the cluster as follows:

Edge elements: In parallel coordinates these are the lines connecting two pairs of corner points: they can be either the borders or the diagonals. In scatter plots these are the corner points. These are marked in blue in Figure 3. These are most vulnerable to disclosure as one of the coordinates of the corner points is always known to the adversary.

Pivot elements: In parallel coordinates these are the lines that connect corner points to free points. In scatter plots these are free points located on the edges. These are marked in red in Figure 3.

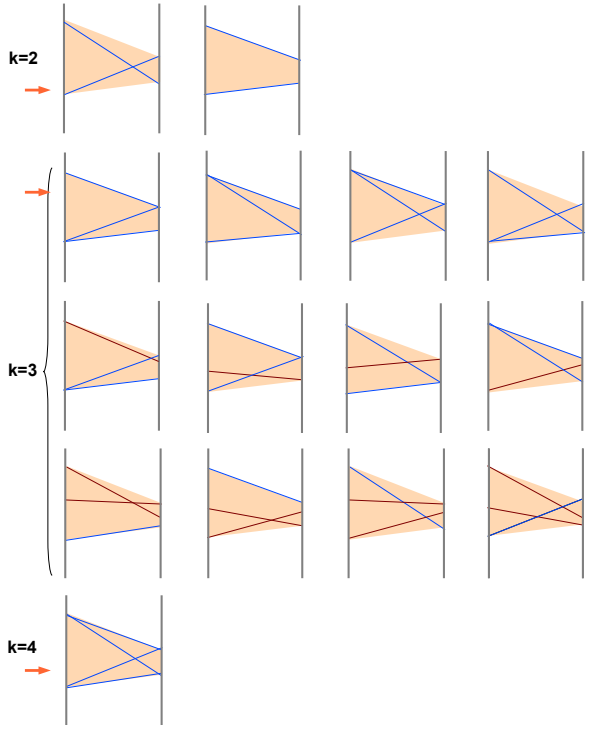


Figure 4: **Base configurations** in a) parallel coordinates on which the configurations for higher k are based upon. Arrowheads denote edge-only configurations, while others are mixed edge configurations.

These are less vulnerable than the edge elements, as only one of the coordinates is revealed by the visualization.

Free elements: In parallel coordinates, these are the lines that connect a pair of free points. In scatter plots, these are the non-corner coordinates. These are marked in green in Figure 3. These have the highest degree of privacy as they can be located anywhere within the cluster and both coordinates are difficult to guess. In triangular or linear clusters, free elements can be located only on one of the axes, therefore the uncertainty is much lower as one of the coordinates is always known. Depending on the value of k, a configuration can be defined by only edge elements, a combination of edge and pivot elements, or only pivot elements. These are also shown in Figure 3. In the following sections we define and quantify how different configurations can be formed by the cluster elements.

4.2 Base Configurations

To define a cluster configuration minimally, the edges or the corners have to be defined first. We term those cluster configurations as base configurations, which are concerned with the orientation of edge elements, and from which others can be derived. Base configurations (n_{b_k}) for $k = 2$, $k = 3$ and $k = 4$ are shown in Figure 4. When the number of edge elements is equal to k, we call the configuration a edge-only configuration. Otherwise, it is either a mixed edge configuration, made of edge elements and pivot elements or for $k >= 4$ there can be pivot edge configurations made of edge and pivot elements. For triangular clusters, there can be only one edge-only configuration. The following analysis therefore, applies to quadrilateral clusters.

Edge-only configuration: Configurations that are formed by only edge elements. Let the number of possible edge-only configurations for a certain k be n_{el_k}. For $k = 2$, $n_{el_k} = 2$. For $k = 3$, there can be four additional configurations as shown in Figure 4 and For $k = 4$ there is one added configuration as all the edge elements can form edges. For $k > 4$, there cannot be any new edge added, so $n_{el_k} = 7$.

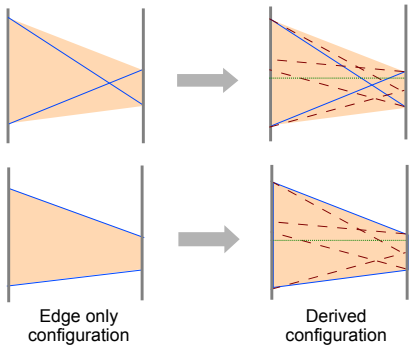


Figure 5: **Illustrating derived configurations** based on edge-only configurations for $k = 2$. The dotted red lines represent the possible pivot edges that can be added and the dotted green line denotes a free edge that can be added, giving a total of five degrees of freedom for a derived configuration.

Mixed edge configuration: In a mixed edge configuration, two pivot elements can define an edge/corner while the other edge/corner consists of an edge element. Since a minimum of three elements are needed, for $k = 2$ there is no mixed edge configuration. For $k = 3$, there can be four different possibilities for a one-edge configuration and four more, for a two-edge configuration, giving a total of 8 mixed-edge configurations (n_{em_k}). These are shown in Figure 4.

Pivot edge configurations: Configurations where no edge/corner is formed by an edge element but only by the pivot elements are referred to as pivot configurations (n_{ep_k}). These configurations have a higher level of privacy as in absence of real edges, there can be many different possibilities for connection among corner and free points. Pivot configurations are only possible for $k >= 4$ and their structure depend on the number and distribution of free points.

Let the number of free points be f, the number of data points within a cluster be k, and the number of corner points on each side be 2. The maximum number of free points possible for any k is given by $f_{max} = 2(k-2) = 2k-4$. Figure 6 illustrates how pivot configurations are built from free points in parallel coordinates. Let p and q be the free points. If they have to define the corner points, then they can connect with either the top corner point or the bottom one on the other axis: lines from p and q either intersect or do not. The same argument applies to r and s and this is how the false edges denoted by the dotted lines are formed. This denotes an ordered selection of free points on each axis, the order (pq in the left image and qp in the right image) being the direction. A minimum of 4 free points are needed to define a pivot configuration and from the formula of f_{max}, we can deduce that pivot configurations are not possible for $k < 4$. The number of possible pivot configurations when $k = 4$ is thus given by the number of possible selection of two points for each axis, that is $2! * 2! = 4$. For higher k, this is similar to a combinatorial problem when we have to select n different things, taken r at a time and the order matters. Here n refers to the free points, that is f and r refers to the available locations, i.e., 2. To define a pivot configuration, the minimum number of free points on one axis is 2 and so the maximum number of free points on the other axis is $f_{max} - 2$. The total number of pivot configurations is thus given by:

$$n_{ep_k} = \sum_{i=2}^{f_{max}-2} P(i,2) * P(f_{max} - 2 - i, 2) \quad k > 4 \tag{1}$$

The total number of base configurations is given by the sum of the edge-only, mixed and pivot edge configurations.

$$n_{b_k} = n_{el_k} + n_{em_k} + n_{ep_k}$$

4.3 Derived Configurations

Derived configurations are those that can be constructed from the base configurations for $k = 2, 3, 4$ as the added elements become pivot elements or free elements. The pivot elements have four degrees of freedom: they are attached to any one of the four corners. For the free element there is an added degree of freedom, giving a total of five degrees of freedom for a derived configuration. For example, if $k = 5$, a configuration can be built with an edge-only configuration with two edge elements (these are the two possible edge-only configurations for $k = 2$). Each of the additional three elements have 5 degrees of freedom each, and therefore total number of possible configurations with two edge elements is $2 * 3 * 5 = 30$. The formation of a derived configuration from a edge-only configuration for $k = 2$ is shown in Figure 5.

Now let us generalize the formula for any k. If the number of edge elements is i, the number of non-edge elements is $k - i$. From the discussion above, the number of derived configurations (n_{d_k}) is given by:

$$n_{d_k} = \sum_{i=2}^{k_{max}} n_{b_i} * (k-i) * 5 \quad n_{b_i} = n_{el_i} + n_{em_i} + n_{ep_i} \tag{2}$$

where $k_{max} = k$ if $k < 4$ and $k_{max} = 4$ if $k => 4$. The factor of $(k-i)*5)$ is a multiplicative factor that gives the configurations for added pivot elements and free elements, without adding edge-elements.

The total number of possible cluster configurations is given by:

$$n_{c_k} = n_{b_k} + n_{d_k}$$

4.4 Useful Uncertainty

In this section we quantify granularity uncertainty caused by cluster configurations and also look at the potential reduction in uncertainty about a configuration from visual artifacts. The most fundamental metric for analyzing the disclosure risk within a cluster is the number of data entities in the cluster. $k > 1$. One could assume that all records have equal probability to be identified, the risk for a specific record L_i to be identified is thus:

$$P(L_i) = \frac{1}{k}, \quad i = 1, 2, \ldots, k$$

In practice, this almost equates to a scenario where k records are put into a bag, and an adversary can pick one record out of the bag randomly. This may be an over-simplification in analyzing the risks of cluster-based parallel coordinates and scatter plots, because, to know the connection among the elements within a cluster, or in other words, all the two-dimensional coordinates, an adversary has to guess the correct configuration, i.e., the orientation of the records within the cluster. The probability of a correct guess is given by the following equation:

$$P(c_k) = \frac{1}{n_{c_k}} \tag{3}$$

4.5 Disclosure Risk of Cluster Configurations

Uncertainty about a cluster configuration can be reduced from the knowledge of free points. Here we quantify the disclosure risk of particular cluster configurations due to knowledge about free points. If the adversary knows that the number of free points on one of the axes (the j^{th} axis in the following equation) is zero, then there have to be real edge elements that define the edges of the cluster. This would imply that the configuration is edge-only as no pivot configurations or mixed edge configuration are possible. If c_k^j is a cluster on the j^{th} axis, then it follows:

$$P(c_k^j | f^j = 0) = \frac{1}{n_{el_k}} \quad since \;\; n_{ep_k}, n_{em_k} = 0 \tag{4}$$

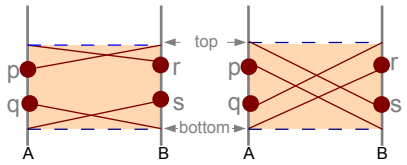


Figure 6: **Illustrating how free points define false edges**. The dotted lines represent false edges. For each pair of free points, pq or rs, there are two options: either they cross or do not cross. There are thus two ways for each free point to be associated with a corner point: top or bottom. This denotes an ordered selection of points for defining a false edge.

From our computation we found that for increasing k, the possible number of edge-only configurations remains constant at *seven* and thus poses a higher risk than the other two types of configuration.

If an adversary knows that the number of free points is non-zero but less than *two* on any axis, then there cannot be a pivot configuration, because a minimum of *two* free points are needed to define both edges of a cluster, as discussed earlier in Section 4.2. Thus the number of possible configurations is given by the following equation:

$$P(c_k^j | f^j < 2) = \frac{1}{n_{el_k} + n_{em_k}} \quad since \;\; n_{ep_k} = 0 \tag{5}$$

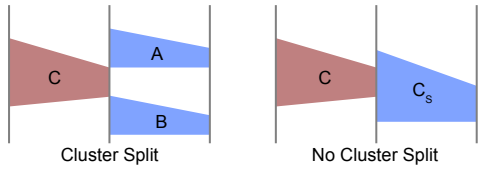
The increase in number of pivot configurations is much steeper than that of mixed-edge configurations as k goes higher than 6. In case of pivot configurations, although an adversary can tentatively guess a configuration from the distribution of free points, the connections among those points and the precise location of those would still be unknown in most cases. Even if the location of the free points were revealed by the visualization itself, the configurations do not reveal information about the connection among the free points. So, the uncertainty reduced due to analysis of cluster configurations, would be mostly restricted to knowledge about the corner elements.

5 Uncertainty Due to Cluster Overlaps

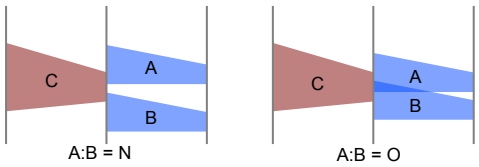
Overlapping pixel ranges of clusters are an artifact of their overlapping data ranges. Some of these overlaps are less risky and some are more risky from a privacy breach point-of-view. In parallel coordinates there is the additional case of overlap between clusters across adjacent axis pairs, that leads to uncertainty in tracing the records within the clusters, across multiple axes. In this section we study the role of cluster overlaps in potential privacy breach scenarios.

Number of Splits: The number of splits is bounded by the number of records in each cluster, i.e., if a cluster contains k records, then there can be a maximum of k splits. If there are fewer than k splits, then there is a higher probability of uncertainty in tracing which records belong to which split clusters. On the other hand, if there are exactly k split clusters, then each split cluster contains a record each from the originating cluster and there is no uncertainty in guessing the distribution of records.

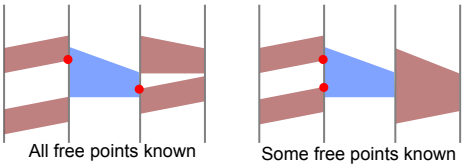
Knowledge about configuration: With user selection of clusters in parallel coordinates or scatter plots, clusters that contain the same record for all axes are visible and therefore the split configurations on both sides of a two-dimensional cluster can be known. Two such configurations are shown in Figure 7(c). For the left image, free points on both sides are known. Since the number of free points on both side is one, it follows from our discussion in Section 4, that the cluster cannot have a pivot edge configuration and thus, has at least one real edge line. In the second case, any free points that might exist are not revealed. So, the configuration of a cluster is not known exactly and one has to work through a large number of combinations for guessing the exact location of the end-points of the records.



(a) A cluster may or may not split, depending on the interval relationship between clusters across adjacent axes.



(b) Relationship between split cluster determines if we know the free points.



(c) *left:* Free points known, not a pivot configuration, *right:* Free points not known, might be edge or pivot configuration.

Figure 7: **Illustrating how the different types of cluster overlaps and splits reveal information about the clusters.** *N* denotes no overlap between clusters A and B, while *O* denotes an overlap between clusters A and B.

5.1 Useful Uncertainty Due to Overlaps

In this section, we formally model the uncertainty caused by overlaps based on two types of uncertainty: identity uncertainty and traceability uncertainty; and also discuss any uncertainty reduced due to the overlaps.

Identity Uncertainty: Between adjacent axes, cluster overlaps can lead to identity uncertainty about the existence of a data point or about the cluster membership of a data point, based on the adversary's background knowledge. In that case, if we assume an adversary knows the existence of a record, the probability of inference would be reduced, proportional to the number of overlapped clusters. If the adversary does not know if the data point exists or not, the uncertainty would further depend on the overlap type. If the point under consideration lies at the meeting point of lines (case E), then the probability is not affected as the precise location is revealed. However, if the point lies in an overlapped region (case O), then the probability of inference is further reduced, proportional to the area of overlap. The overlap entropy metric [9] that quantifies the uncertainty due to overlaps can be used to measure this.

Traceability: Traceability uncertainty is about the confusion in knowing the distribution of the records in the split/continuing clusters, once the configuration of a cluster is known. This depends on the number of splits. If the number of splits is equal to the number of records within the cluster, then the split clusters share one record each with the originating cluster. This is a less uncertain case than when the number of splits is less than the number of records in each cluster. Then the distribution of records in the continuing clusters is not immediately known.

Let t be the number of splits where $t < k$. The number of records in each split cluster can range from 1 to $k - t + 1$. The problem

of finding the number of possible distributions of records in the split clusters then reduces to finding the number of t-subsets of k elements. Let us assume the worst-case scenario of the adversary knowing the records within a cluster and then trying to find out their distribution in the split clusters. This is analogous to the problem of distributing k distinguishable objects (the records) in t non-empty indistinguishable boxes. This is given by Stirling number of the second kind [34], $S(k,t)$ which is given by the following standard formula:

$$S(k,t) = \frac{1}{t!}\sum_{i=1}^{t} -1^{i} C(t,i)(t-i)^{k} \tag{6}$$

where $C(t,i)$ denotes the combination of t things, taken i at a time. A two-dimensional table of values for the above formula relating k to t is readily available in the discrete mathematics literature [28]. The number of possibilities when the number of splits approaches k increases drastically and thus, if the adversary does not have any further background knowledge about the data, it would be very difficult to know the distribution of records in the split clusters.

5.2 Disclosure Risk Due to Cluster Splits

Since cluster borders are formed by data-points, elements located at the edge or corner of a cluster have high vulnerability. On the other hand, the cluster coordinates that are not on the corner have a higher level of privacy. We term these as *free points*. Knowing about location and number of free points can lead to attribute disclosure and also enable adversaries to know about cluster configurations which is described in Section 4. Here we focus on disclosure risk of free points due to cluster splits. In case of an overlap (O) between the originating cluster and split cluster, free points are revealed, while in case of an edge-meeting (E), they are not. Let f_j denote the number of free points on the j^{th} axis. The ability of an adversary to know the exact number of free points depends on the overlap relation between the originating cluster and the split clusters. If t_j is the number of split clusters on the j^{th}, C^j is the originating cluster, and C^i_j is a corresponding split cluster, then the number of free points known is given by:

$$f_j = \sum_{i=1}^{t} \begin{cases} 1 & \text{if } C_j : C^i_j = O \\ 0 & \text{if } C_j : C^i_j = E \end{cases}$$

For a given dimension, we calculate the net risk of known free points on the j^{th} axis by the following formula, where f^i_j denotes the number of split clusters on the j^{th} axis for the i^{th} originating clusters, n being the number of originating clusters:

$$R(f_j) = \frac{1}{n}\sum_{i=1}^{n} \frac{f^i_j}{t^i_j} \quad 0 < R(f_j) < 1 \tag{7}$$

In case of all free points known for every cluster, $R(f_j) = 1$. The lesser the value of $R(f_j)$, the lower is the risk for the j^{th} dimension.

6 Uncertainty Due to Cluster Range

Cluster configurations are not affected by pixel resolution. For a detailed analysis of the attack scenarios, we have to take the pixel range of clusters into account and analyze the varying k and varying cluster range affect disclosure risk. When an adversary has background knowledge about some attribute values (end points in parallel coordinates) and/or is interested in discovering attribute values, the plan of attack would be based on working through a number of combinatorial cases for overcoming the uncertainty caused by lack of spatial accuracy owing to the pixel ranges of clusters.

In this section we study the disclosure risk attached with knowledge of end points. The methods for determining the amount of uncertainty or vulnerability about end points can also be extended to scatter plots, except that the notion of end points is transformed to coordinates of sample points in a scatter plots. Since for a triangle-shaped cluster one end is always known, we focus our analysis on two-dimensional quadrilateral-shaped clusters between adjacent axes. Figure 8 shows two example scenarios where an adversary would try to breach privacy: in relation to two attributes (a) and (b), the adversaries may have the knowledge that a specific line must be in the cluster, but not much more. Hence the adversaries have to make wild guesses about possible values on axes (a) and (b). These guesses translate visually to the guesses of both green end points on the two axes. Second, the adversaries may have already discovered an attribute value on axis (a) about this target line, and want to find out attribute value on axis (b). In other words, they need to guess where the green end point is.

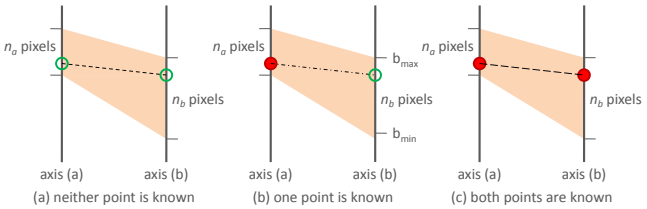


Figure 8: **Three basic scenarios about adversary's knowledge about the end points of a line**. This relates to spatial accuracy in guessing the exact locations of the points on the axes.

In the following subsections, we first examine scenario 2, and provide a method for quantifying the uncertainty in relation to the number of lines in the cluster, k, and the resolution of the visualization. Building on the analysis for 2, we examine the scenario 1. We assume that the adversary has some kind of background knowledge using which they can confirm the existence of a line, given two correct attribute values (i.e., both end points).

6.1 Disclosure Risk of One End-point of a Specific Line

Considering Figure 8(b), let k be the number of lines in this cluster. Without losing generality, we assume that the adversary has discovered the value of attribute (a) associated with a specific line L_i. If the adversary does not gain any knowledge from the visualization about the possible location of the other end of the cluster, then the probability of a correct guess depends on a number of combinatorial cases. We can compute the number of valid combinations of k lines that pass through some of the n_b pixels, that is the cluster range on axis (b) as:

$$G(k,n_b) = \begin{cases} 1 & n_b = 1 \\ n_b^k - 2(n_b-1)^k + (n_b-2)^k & n_b > 1 \end{cases} \tag{8}$$

where the first term in the case of $n_b > 1$ is the number of all possible combinations, the second term defines the number of invalid cases where no line passes through either of the corner point, b_{min} and b_{max}, and the third term defines the number of cases where no line passes through either b_{min} nor b_{max}.

Consider two numbers that define the number of combinations of k lines with at least one line passing b_{min} and b_{max} respectively. The two numbers can be defined by using the same recurrence function:

$$\begin{aligned} &B(0,n_b) = 0 \\ &B(1,n_b) = 1 \\ &B(2,n_b) = 2n_b - 1 \\ &\quad \cdots \\ &B(k,n_b) = n_b^{k-1} + (n_b-1)B(k-1,n_b) \end{aligned} \tag{9}$$

Therefore, for the target line L_i to pass through either b_{min} or b_{max}, the number of combinations will be $B(k-1,n_b)$, since there must

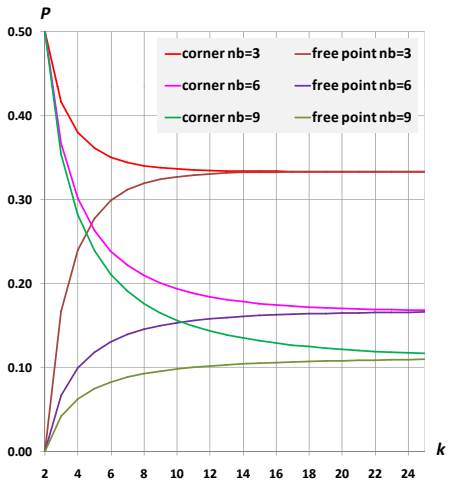


Figure 9: **Vulnerability of an unknown point**. Given the adversary knows one end point of a line, the plot shows the certainty (or vulnerability) of the other end-point as a function of k and n_b where n_b is the cluster range on axis b. Whether or not the end-point is a corner point poses different risk factor.

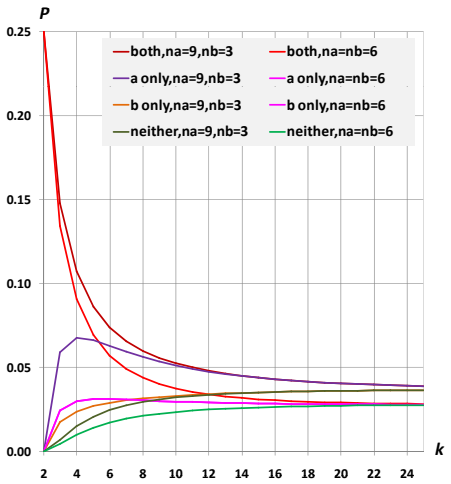


Figure 10: **Vulnerability of two unknown points**. Given the adversary only knows the containment of a line in a cluster but not the end points, the plot shows the certainty (or vulnerability) of both points with two different combinations of n_a and n_b.

be at least one other line that passes through the other corner point. Meanwhile, for L_i to pass through a free point, $b_{min} < b_x < b_{max}$, the number of valid combinations is simply $G(k-1,n_b)$.

Let $L_i \multimap \rho$ denote that the line L_i passes a point on axis (b), where ρ can be one of the values, $b_{min}, b_{min}+1, \ldots, b_{max}-1, b_{max}$. The vulnerability or certainty of this specific line that is to be guessed depends on its position on axis (b). For $k > 1, n_b > 2$, we have:

$$P(L_i \multimap \rho) = \begin{cases} \dfrac{B(k-1,n_b)}{G(k,n_b)} & \text{if } \rho = b_{min} \text{ or } \rho = b_{max} \\ \dfrac{G(k-1,n_b)}{G(k,n_b)} & \text{if } b_{min} < \rho < b_{max} \end{cases} \tag{10}$$

As shown in Figure 9, when k is relatively small, there is a significant difference between lines that pass through corner points (b_{min} or b_{max}, and lines that do not. Such a gap closes when k increases (i.e., there are more lines) as there can be a large number of pivot configurations that are possible due to the addition of pivot lines. In general, the higher value n_b is, the lower the certainty is, except that when $k = 2$, the resolution does not affect the uncertainty.

6.2 Disclosure Risk of Both End-points of a Specific Line

Building on the above analysis, we now consider scenario 1 in Figure 8(a). Assume that the adversaries know the fact that a specific line is in a cluster, but do not know the value of either attribute. As the probabilistic distributions on the two axes are independent, we can derive the joint distribution from the distributions on individual axes.

Let $L_i \multimap\!\!\circ\, (\rho_a, \rho_b)$ denote the line L_i passes two points on axes (a) and (b) respectively, where ρ_a can be one of the integer values between a_{min} and a_{max}. For $k > 1, n_a > 2, n_b > 2$, the vulnerability or certainty of this specific line is to be guessed is:

$$P(L_i \multimap\!\!\circ\, (\rho_a,\rho_b)) = \begin{cases} \dfrac{B(k-1,n_a)B(k-1,n_b)}{G(k,n_a)G(k,n_b)} & \rho_a \text{ and } \rho_b \text{ are corners} \\ \dfrac{B(k-1,n_a)G(k-1,n_b)}{G(k,n_a)G(k,n_b)} & \text{only } \rho_a \text{ is a corner} \\ \dfrac{G(k-1,n_a)B(k-1,n_b)}{G(k,n_a)G(k,n_b)} & \text{only } \rho_b \text{ is a corner} \\ \dfrac{G(k-1,n_a)G(k-1,n_b)}{G(k,n_a)G(k,n_b)} & \text{neither is a corner} \end{cases} \tag{11}$$

Figure 10 shows the vulnerability in the case of both end points are unknown. It shows two situations when $n_a = 9, n_b = 3$ and when $n_a = n_b = 6$ respectively. It is useful to note that the first situation is slightly more vulnerable than the second, though the sum of the pixel resolutions on both axes are the same. This indicates that the lower pixel resolution on the right axis incurs more risks. When the sum of the pixel resolutions on the left and right axes are the same, a cluster patch of a trapezoid shape is more risky than a parallelogram. In comparison with Fig. 9, the risk is noticeably lower.

7 Case Study

We use the German Credit dataset [17] to illustrate some real-world examples about attack scenarios. We demonstrate that k-anonymity is a necessary but not a sufficient condition for privacy-preservation in the screen space; we need additional metrics to guide the appearance of clusters and subsequent user interaction. The dataset has 1000 instances which classify bank account holders into credit classes *Good* or *Bad*. Each data object is described by 20 attributes that include 13 categorical and 7 numerical attributes. Since the privacy model is mostly applicable to numerical attributes, we leave out the other categorical attributes, and use the following attributes: *duration of loan*, *creditamount*, and *age*. The quantification of different sources of visual uncertainty can be used for iterative refinement of the design of privacy-preserving parallel coordinates and scatter plots. In this section we perform a probabilistic analysis of the disclosure risks when an adversary uses background knowledge and interaction for breaching privacy.

7.1 Journalist Attack scenario

In an attempt to randomly breach privacy in case of the journalist attack scenario, adversaries will try to reorder the axes in such a way that the axis with the most vulnerable clusters is adjacent to the sensitive attribute dimension. From our model, that axis is the one that i) produces the most number of edge-only configurations when put adjacent with a sensitive attribute and ii) lowers the disclosure risk of free points of the clusters. In light of these two attack goals we examine the possible defense mechanisms.

Analyzing Disclosure Risk of Cluster Configurations: In a journalist attack scenario, the next step after exploiting adjacency configurations is to analyze cluster configurations that are most vulnerable. These are the edge-only configurations. How does the probability of edge-only configurations vary with respect to k? In Figure 11a the number of edge only configurations, computed using Equation

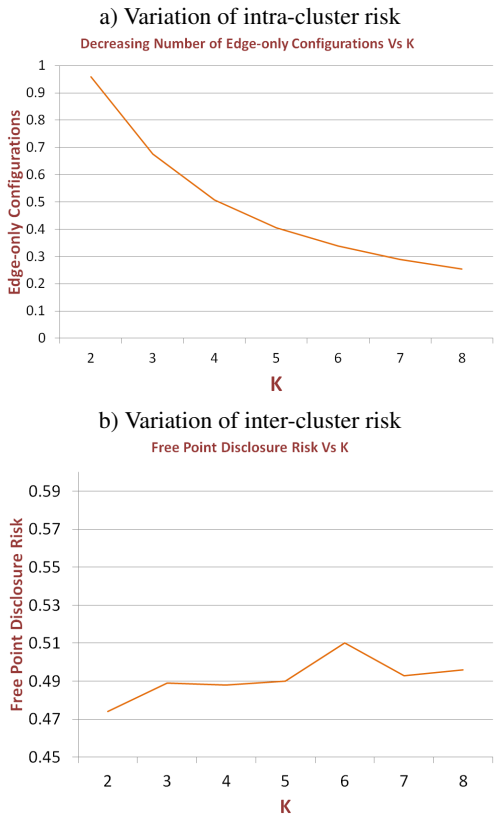


Figure 11: **Number of edge-only configurations decreases with increasing** k, while the disclosure risk for free points can increase with increasing k due to the uncertainty involving cluster splits.

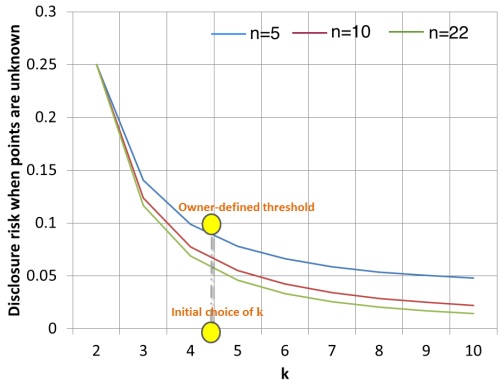
4 is plotted for increasing k. The decreasing number of edge-only configurations means that the probability of the number of pivot configurations increases with an increasing k. This correlation with k apparently signifies more privacy as increasing probability of pivot configurations makes it difficult to guess the edge records. However, in Figure 11b we see that the probability of knowing the location of free points, computed by Equation 2, can increase with increasing k.

While this can be counter-intuitive, given our model and metrics, it is not difficult to reason about this apparent anomaly. One reason for the revelation of free points with increasing k is the increasing number of pivot configurations (as computed by Equation 3). Another factor is the number of overlaps cases as we had described in Section 5: the condition for overlap for meeting of edges (condition E) is exceeded by the same for overlap of edges (condition O), with greater k. These two factors lead to a higher probability of free points being known with increasing k. With free points being known, this can lead to disclosure of other data points within the cluster. This signifies higher k is not sufficient and necessitates additional measures, that would objectively compute the probability of disclosure given such uncertain patterns with higher k-anonymity. For this we will use the parameterized measures (the parameter being cluster range) derived in Section 6.

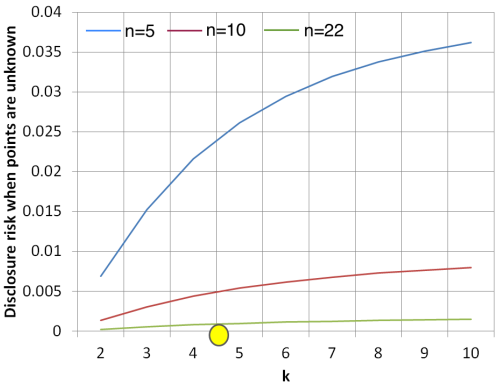
7.2 Prosecutor Attack Scenario

A journalist attack scenario that leads to disclosure of attributes, can lead to a prosecutor attack scenario. The latter can also be initiated by an adversary's background knowledge about the data.

Assumption about background knowledge: In many real world datasets, certain types of knowledge can be assumed. For example, in a disease dataset, with the sensitive value being breast cancer, the adversary already knows about the most probable gender of the patient, and can use that knowledge for breaching privacy further.



(a) Variation of disclosure risk of edge coordinates of the sensitive clusters, when one of the coordinates of a record is known.



(b) Variation of disclosure risk of free coordinates of the sensitive clusters, when one of the coordinates of a record is known.

Figure 12: **Disclosure risk when an adversary knows that an individual belongs to a group** but *exact points are unknown*. By analyzing disclosure risks with respect to assumptions about the adversary's background knowledge, data owners can use thresholds for disclosure probability and choose a k accordingly. In this case, a data owner sets the threshold for disclosure probability to be less than 0.1 and therefore the initial choice of k is 4.

In this dataset, we make the assumption that the adversary knows very young or very old people are more likely to have bad credit history. We examine the clusters that belong to people of these age groups (where we assume *old* age to be above 60 and *young* to be below 20), having sensitive value, for two cases: case A, where the adversaries know a person belongs to this group but does not know specific data points and case B, when the adversaries know a person belongs to this group and knows one of the data points, for example they know the age but not their loan or credit amount.

Identifying Sensitive Clusters: Cluster range is one of the input parameters for computing the disclosure risks quantified in Equations 8 – 12. For this dataset we compute the interquartile range of the cluster ranges on the axis for the sensitive clusters. Then based on equations 8, 9, 10, and 11; we perform a detailed analysis for deciding the k that can be used for an axis pair or for the dataset. In this case the average interquartile range for the chosen set of dimensions is 5 to 22, with the median being 10. Since disclosure risk is a function of both k and cluster range, we examine the properties of the graphs where disclosure risk values are plotted against different k, for the cluster range values of 5, 10 and 22. It should be noted that by selecting a smaller subset of probable sensitive clusters, we are assuming a worst-case scenario for disclosure: the fact that an adversary has been able to break through the inter-cluster

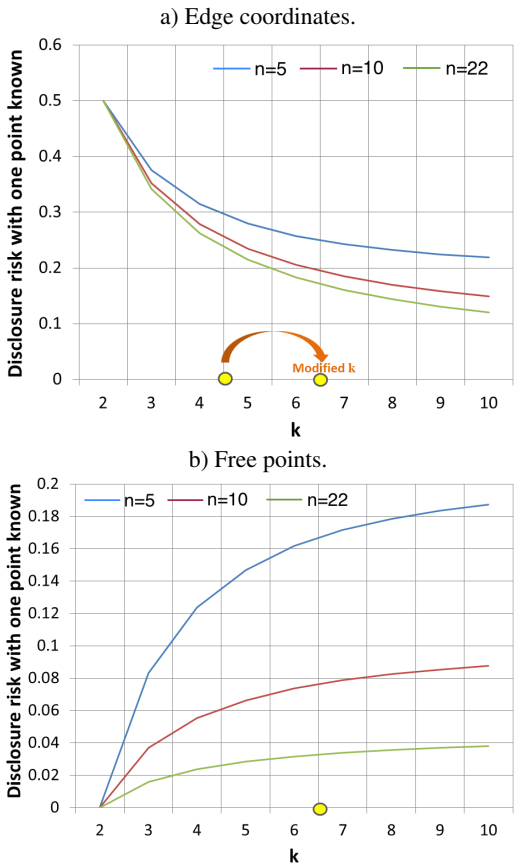


Figure 13: **Disclosure risk with background knowledge**, when an adversary knows that an individual belongs to a group but *one of the points is known*. In this case a data owner changes his/her choice of k from 4 to 6, so that the disclosure probability stays below the modified threshold of 0.2.

uncertainties due to overlaps and splits and are faced with the challenge of negating intra-cluster uncertainty due to cluster range and configurations.

Analyzing disclosure risk of sensitive clusters: In Figure 12a, the disclosure risk of edge points vs k, for case A when none of the coordinates is known by the adversary, is plotted using the first condition in Equation 10. By looking at the graph, the data owner or the visualization designer sets a threshold of disclosure probability to 0.1, as he/she desires that the disclosure risk of edge points to be below this probability. It can be observed in the graph that $k = 4$ gives that desired probability for this dataset. In Figure 12b, the disclosure of non-edge or free points is plotted for varying k using the second condition in Equation 10. While the disclosure risk for those points increase with increasing k as we had observed in Figure 11*a*, the disclosure probability is below 0.1 when $k = 4$.

Next the data owner would want to examine case B, that is when the adversary knows that an individual belongs to a group and also one of the data points. The relation between k and disclosure risk is plotted using the first condition in Equation 11. As we can observe from the graph, the initial threshold for disclosure risk of 0.1 will necessitate a very high k, but that will degrade the utility to a large extent as we had shown in our earlier work [9].

To address this the data owner would ideally decide to increase the risk threshold to about 0.2 which allows for modifying k to 6. Since the disclosure risks for lower cluster ranges are a bit higher, the data owner can choose to constrain brushing by not displaying those clusters on interaction. To verify if $k = 6$ poses a risk for the free points in case B, we plot disclosure risk against k for the edge coordinates and free points (Figure 13). We can observe that the disclosure risk here is below 0.2 and therefore the final choice of k is 6. While we illustrated probabilistic analysis of a few disclosure risk scenarios, there can be others as perceived by the data owners. Using a systematic and step-by-step analysis as detailed above, they can choose an appropriate k and control interaction based on reordering and brushing, by evaluating the different configurations of the visualization using our metrics.

8 Discussion

In this section, we reflect on the limitations, applicability, and generalization of our proposed model.

Limitations of our model: While we have modeled some aspects of the background knowledge that an adversary might possess, there are other aspects that can still be incorporated in the analysis, like multi-dimensional background knowledge and other types of knowledge based on attribute types. The fact that background knowledge is subjective and therefore hard to model has been widely acknowledged in the privacy-preserving data mining literature [15]. There has been some recent work related to modeling of background knowledge in the context of PPDM [26] and we would like to integrate our visualization model with the data-based model.

Applicability in open data: Modern open data portals publish both data and visualizations for helping public and domain experts find relevant data for their analysis. Researchers have demonstrated that open data could be vulnerable to unintended disclosure through adversarial attacks [14]. To mitigate such risks, data owners can leverage our visual uncertainty model for evaluating visualizations before publishing them via data-sharing interfaces [18].

Generalizability of our approach: The basic premise of this work is to analyze how the low-level aspects of visual encoding using shape and size of clusters and location of records can lead to information disclosure. In this specific application scenario of privacy-preservation our goal is to minimize such disclosure by controlled visual uncertainty. We believe our work has important implications beyond the realm of privacy preservation: a generalized model of visual uncertainty can enable visualization designers build displays that would be optimized for maximizing insight from an interactive visualization system. Some recent developments [4, 5, 23] have indicated that it may be feasible to formulate a generalized model for visual uncertainty that takes into account the background knowledge of users. Such a model would enable visualization designers to optimize interactive visualization systems in a systematic and rigorous manner.

9 Conclusions and Future Work

In this work, we have presented a detailed analysis of the relationship among visual uncertainty, attack scenarios and their associated disclosure risks, and privacy-preserving visualization in the context of scatter plots and parallel coordinates. We have quantified the different types of visual uncertainty and illustrated how they help in creating an additional layer of defense besides k-anonymity to prevent disclosure of attributes. We are currently working on two aspects that would be natural extensions of this work. We are working on a client-server based privacy-preserving visualization system where the disclosure risk metrics along with previously proposed metrics for privacy and utility will be integrated for optimizing the rendering of the clusters. data owners can control the visualization from the server-side using a risk-based evaluation based on our metrics. The system will be integrated with open data portals such that data owners and stakeholders can get immediate feedback about potential disclosure risks while providing publicly accessible visualization of the data. This will be a step towards more transparent privacy-preservation integrated with accessible data-sharing practices.

References

[1] C. Aggarwal. On unifying privacy and uncertain data models. In *ICDE*, pages 386–395. IEEE, 2008.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. *ACM Sigmod Record*, 29(2):439–450, 2000.

[3] J. Byun, A. Kamra, E. Bertino, and N. Li. Efficient k-anonymization using clustering techniques. *Advances in Databases: Concepts, Systems and Applications*, pages 188–200, 2007.

[4] M. Chen and D. S. Ebert. An ontological framework for supporting the design and evaluation of visual analytics systems. In *Computer Graphics Forum*, volume 38, pages 131–144. Wiley Online Library, 2019.

[5] M. Chen and A. Golan. What may visualization processes optimize? *IEEE transactions on visualization and computer graphics*, 22(12):2619–2632, 2015.

[6] J.-K. Chou, C. Bryan, J. Li, and K.-L. Ma. An empirical study on perceptually masking privacy in graph visualizations. In *2018 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 1–8. IEEE, 2018.

[7] J.-K. Chou, Y. Wang, and K.-L. Ma. Privacy preserving visualization: A study on event sequence data. In *Computer Graphics Forum*, volume 38, pages 340–355. Wiley Online Library, 2019.

[8] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum*, 31(3pt2):1015–1024, 2012.

[9] A. Dasgupta, M. Chen, and R. Kosara. Measuring privacy and utility in privacy-preserving visualization. In *Computer Graphics Forum*, volume 32, pages 35–47. Wiley Online Library, 2013.

[10] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Trans. on Visualization and Computer Graphics*, 16(6):1017–26, 2010.

[11] A. Dasgupta and R. Kosara. Adaptive privacy-preserving visualization using parallel coordinates. *IEEE Trans. on Visualization and Computer Graphics*, 17(12):2241–2248, 2011.

[12] A. Dasgupta and R. Kosara. Privacy-preserving data visualization using parallel coordinates. In *Visualization and Data Analysis 2011*, volume 7868, page 78680O. International Society for Optics and Photonics, 2011.

[13] A. Dasgupta, E. Maguire, A.-R. Alfie, and M. Chen. Opportunities and challenges for privacy-preserving visualization of electronic health record data. In *Proceedings of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records*, 2014.

[14] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva. Anonymizing nyc taxi data: Does it matter? In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 140–148. IEEE, 2016.

[15] W. Du, Z. Teng, and Z. Zhu. Privacy-maxent: Integrating background knowledge in privacy quantification. In *Proceedings, SIGMOD International Conf. on Management of Data*, pages 459–472, 2008.

[16] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Assn.*, 81(393):pp. 10–18, 1986.

[17] A. Frank and A. Asuncion. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2010.

[18] M. Gaboardi, J. Honaker, G. King, J. Murtagh, K. Nissim, J. Ullman, and S. Vadhan. Psi ({\Psi}): a private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016.

[19] C. Holzhüter, A. Lex, D. Schmalstieg, H.-J. Schulz, H. Schumann, and M. Streit. Visualizing uncertainty in biological expression data. In *Proceedings Visualization and Data Analysis*, 2012.

[20] A. Inselberg. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer, 2009.

[21] C. Johnson. Top scientific visualization research problems. *Computer graphics and applications, IEEE*, 24(4):13–17, 2004.

[22] F. D. K. El Emam. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15:627–637, 2008.

[23] N. Kijmongkolchai, A. Abdul-Rahman, and M. Chen. Empirically measuring soft knowledge in visualization. In *Computer Graphics Forum*, volume 36, pages 73–85. Wiley Online Library, 2017.

[24] D. Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9:313–331, 1993.

[25] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE International Conference on Data Engineering*, pages 106–115, 2007.

[26] T. Li, N. Li, and J. Zhang. Modeling and integrating background knowledge in data anonymization. In *In Proceedings, International Conference on Data Engineering*, pages 6–17. IEEE, 2009.

[27] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[28] D. Loeb. A generalization of the stirling numbers. *Discrete mathematics*, 103(3):259–269, 1992.

[29] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):3, 2007.

[30] G. Paass. Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics*, pages 487–500, 1988.

[31] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[32] P. Rhodes, R. Laramee, R. Bergeron, and T. Sparr. Uncertainty visualization methods in isosurface rendering. In *Eurographics*, pages 83–88, 2003.

[33] S. Russell, P. Norvig, J. Canny, J. Malik, and D. Edwards. *Artificial intelligence: a modern approach*. Prentice hall, 1995.

[34] H. Sharp Jr. Cardinality of finite topologies. *Journal of Combinatorial Theory*, 5(1):82–86, 1968.

[35] M. Skeels, B. Lee, G. Smith, and G. Robertson. Revealing uncertainty for information visualization. *Information Visualization*, 9(1):70–81, 2009.

[36] C. Skinner and M. Elliot. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 64(4):855–867, 2002.

[37] L. Sweeney. k-anonymity: A model for protecting privacy. *IEEE Security And Privacy*, 10(5):1–14, 2002.

[38] S. F. V. Ciriani, S. De Capitani di Vimercati and P. Samarati. k-anonymous data mining: A survey. In *Privacy-Preserving Data Mining: Models and Algorithms*, pages 105–136. Springer-Verlag, 2007.

[39] X. Wang, W. Chen, J.-K. Chou, C. Bryan, H. Guan, W. Chen, R. Pan, and K.-L. Ma. Graphprotector: A visual interface for employing and assessing multiple privacy preserving graph algorithms. *IEEE transactions on visualization and computer graphics*, 25(1):193–203, 2018.

[40] X. Wang, J.-K. Chou, W. Chen, H. Guan, W. Chen, T. Lao, and K.-L. Ma. A utility-aware visual approach for anonymizing multi-attribute tabular data. *IEEE transactions on visualization and computer graphics*, 24(1):351–360, 2017.