Measuring Privacy and Utility in Privacy-Preserving Visualization

Aritra Dasgupta[1] Min Chen[2] Robert Kosara[1]

[1] University of North Carolina at Charlotte
[2] Oxford University

Abstract

In previous work, we proposed a technique for preserving the privacy of quasi-identifiers in sensitive data when visualized using parallel coordinates. This paper builds on that work by introducing a number of metrics that can be used to assess both the level of privacy and the amount of utility that can be gained from the resulting visualizations. We also generalize our approach beyond parallel coordinates to scatter plots and other visualization techniques.

Privacy preservation generally entails a trade-off between privacy and utility: the more the data is protected, the less useful the visualization. Using a visually-oriented approach, we can provide a higher amount of utility than directly applying data anonymization techniques used in data mining. To demonstrate this, we use the visual uncertainty framework for systematically defining metrics based on cluster artifacts and information theoretic principles. In a case study, we demonstrate the effectiveness of our technique as compared to standard data-based clustering in the context of privacy-preserving visualization.

1. Introduction

In visualization, one of the main challenges is to maximize data fidelity during the mapping between data space and the screen space of limited number of pixels. While visualizing sensitive data, however, the goal is to intentionally hide some information to prevent unauthorized disclosure. Since visualization entails inherent information loss and other types of uncertainty in the screen space, those can be exploited for privacy-preserving purposes. We recently introduced a technique for privacy-preserving data visualization (PPDV) using parallel coordinates [DK11], which is based on ideas from privacy-preserving data mining (PPDM). In contrast to the data-based approach to preserving privacy, a visual approach takes the properties of a visualization into account while manipulating the visual structures. A visualization model for privacy-preservation thus has to be treated differently from a data-based model for the same, and evaluation of PPDV techniques has to be based on the visual representation.

There are few established metrics for assessing the utility of information visualization techniques. Several researchers have argued for the need of measuring visual representations, as they form the interface between the data and the human mind [FLC*02]. In recent times, visual quality metrics for various techniques have been proposed for measuring effectiveness of representations. However, ambiguity in definition of these metrics and a lack of systematic comparison with respect to their usefulness make it difficult for visualization designers to objectively define metrics for assessing visualizations [BTK11].

While user studies are an accepted way of measuring the usefulness of visualization, they often do not provide clear insights into why some techniques work better, and do not generally allow the construction of optimization algorithms. In particular in the case of privacy, there are not only visual and utility criteria, but also the privacy aspects of the visualization that need to be taken into account. Although the latter can be validated in user interaction scenarios using studies, the issues of privacy and utility cannot be considered in isolation. To do this, we need to quantify the level of privacy that can be achieved by a particular visual representation and then assess the loss in utility those cause.

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

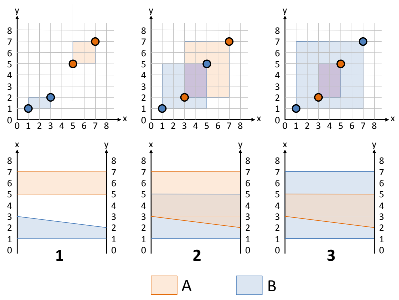


Figure 1: *Pixel-based binning and clustering in parallel coordinates and scatter plots. Illustrating ways for 2-anonymizing the four data points in a pixel grid. For very large number of points, estimation of privacy requires metrics.*

We use the model and taxonomy of visual uncertainty we introduced recently [DCK12] to select relevant metrics to define. This model is based on the uncertainty produced by the visualization process itself, rather than inherent in the data. By deconstructing a visualization into its smallest components, i.e., the visual structures, we are able to control the screen-space information. We address both types of uncertainty: *encoding uncertainty* that is concerned with the visual mapping process and *decoding uncertainty* that describes the perception and cognition of visual structures from a user's perspective. We propose a set of metrics that can help us quantify the different types of uncertainty and satisfy the requirements of a privacy-preserving model. The contributions of our work can be summarized as follows:

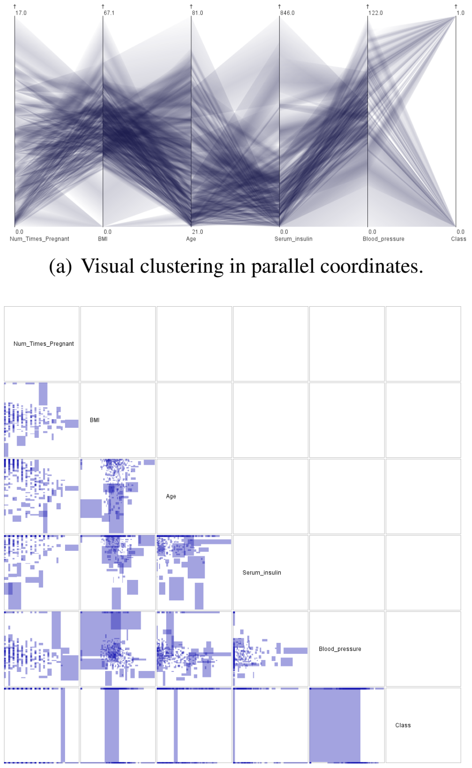
1. Identify causes and sources of visual uncertainty in a class of multivariate techniques that can be used for privacy-preservation of sensitive data.
2. Develop a set of metrics that quantify privacy and utility in term of encoding and decoding uncertainty.
3. Application of the metrics for comparison of data-based and visual approaches to privacy-preserving visualization.

2. Related Work

Since privacy preservation in the realm of information visualization is a relatively new concept, we discuss the relevant background in the area of data mining and provide context for the rest of the paper.

2.1. Criteria for Privacy

Among the different methods proposed in privacy-preserving data mining [AS00], we focus on the k-anonymity and l-diversity criteria for preserving privacy. The k-anonymity model [Swe02] is used extensively to prevent privacy breach by linking attributes that co-occur in private and public databases (known as *quasi-identifiers*) to externally available information. k-anonymity ensures records are aggregated in groups of k members so that they are indistinguishable with respect to the quasi-identifiers. This does not guard against the homogeneity attack due to the lack of diversity in sensitive attributes: if a quasi-identifer group only points to a single sensitive attribute, an attacker can breach privacy. To prevent this, Machanavajjhala et al. proposed l-diversity [MKGV07] which ensures each quasi-identifier group points to at least l different values for the sensitive attribute. In our previous work [DK11], we implemented a privacy-preserving visualization technique based on these two criteria, that is adaptive to user interactions. Here we not only evaluate the k-anonymity based privacy model with respect to different metrics, but also demonstrate that k-anonymity is not sufficient to guarantee privacy in the screen-space and therefore the metrics are necessary for quantifying the levels of privacy and utility.



(a) Visual clustering in parallel coordinates.

(b) Visual clustering in scatter plots.

Figure 2: *Privacy-preserving clustering in parallel coordinates and scatter plots for k=3 in case of the* Diabetes *dataset.*

2.2. Privacy and Utility With Respect to Visual Uncertainty

While the trade-off between privacy and utility is very much an open research area in PPDM, several metrics have been proposed to quantify privacy and utility individually. A comprehensive survey have been done by Bertino et al. [BLJ08] where the different metrics have been categorized and described. Entropy as a privacy metric was first proposed by Aggrawal et al. [AA01] and was developed further by others [BFP05, Bez08]. Utility has been measured in terms of data quality and clustering quality and also with respect to preservation of patterns with respect to specific data mining techniques. In our work, although we do not directly apply any of these metrics, we adopt some of the approaches for privacy-preservation in the screen-space. For example, we use entropy to measure the effect of overlaps and also clustering quality with respect to the centroids of the polygons. Most importantly, we systematically define metrics according to the visual uncertainty taxonomy that we had proposed in our previous work.

Uncertainty in the screen-space, or visual uncertainty [DCK12] 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., [PWL97, Joh04]) and uncertainty involving geometrical primitives, like isosurface rendering [RLBS03]. 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 [RNC*95] and to the differentiation between input and output uncertainty [HLS*12]. Study of the sources and causes of uncertainty enables the visualization designers to analyze and refine the visualization output for effective privacy-preservation in an interactive environment.

2.3. Information-Theoretic Measures

Several authors have advocated the use of information-theoretic measures to quantify the information-content of a visualization. Purchase et al. argue about conceptualizing the visualization pipeline as a lossy information channel and mentions that information-theoretic measures can be used to measure the loss [PAJKW08]. Rundensteiner et al. propose some measures of data quality and abstraction quality to make the connection between data and the screen spaces [RWX*07]. Yang-Peláez and Flowers have demonstrated how information content in visualization can be quantified without taking semantics into account [YPF00]. More recently, Chen and Jänicke have shown how different information-theoretic concepts like entropy and mutual information can be used at different stages of the visualization pipeline [CJ10]. We demonstrate the use of some of those metrics for quantification of privacy and different constraints that guide the interaction.

3. Conceptualizing Privacy-Preserving Visualization

The goal of a privacy model in visualization is to protect the sensitive values of individual records, so that with or without interaction, a user without appropriate access rights is not able to read data at a precision that is higher than allowed by the data owner. For this, we need to transform the records in the data space to an anonymized form which masks their real values, yet preserves the overall patterns. Based on the visual variables [CM97] being used, values can be masked based on color (pixel-oriented techniques), shape (glyphs), position (scatter plots, parallel coordinates, line charts, etc.).

At the core of the k-anonymity privacy model is the indistinguishability of k data items with respect to each other. In screen space, data is represented by visual variables. Anonymity in the data space means hiding the values, but the same in the screen space can be achieved through manipulation of these different visual variables. Depending on the key component of a visual representation, appropriate visual variables are manipulated to achieve anonymity. In previous work, we have proposed privacy-preserving parallel coordinates. By the line-point duality principle [ID90], lines in parallel coordinates have a one-to-one mapping with points in scatter plots. Exploiting this mapping, we extended the privacy-preserving model to be applied to scatter plots. The only change is in the cluster shapes, quadrilaterals or triangles in case of parallel coordinates, while rectangles/squares or lines in case of scatter plots. The concepts that we discuss in this paper are applicable directly to both scatter plots and parallel coordinates and can be generalized for other position-based visual representations like line charts.

3.1. Pixel Binning

In case of position-based representations like scatter plots and parallel coordinates binning is based on pixel coordinates (Figure 1). In case of other representations like glyphs, binning can be based on shape, color, etc. Let D be a tabular dataset, which has n_r m-dimensional records: $R_1, R_2, \ldots, R_{n_r}$ where $R_i = \langle x_{i,1}, x_{i,2}, \ldots, x_{i,m} \rangle$. Let each axis be of h pixels high, and S be the set of valid pixel coordinates, i.e., $S = \{0, 1, \ldots, h-1\}$. Let $X_j, j = 1, 2, \ldots, m$ denote the set of all values on the j^{th} axis, i.e., $X_j = \{x_{1,j}, x_{2,j}, \ldots, x_{n_r,j}\}$. Each data value $x_{i,j}$ is thus projected onto a screen pixel coordinate on the j^{th} axis by $f_j : X_j \rightarrow S$.

3.2. Clustering

Since the minimum number of original records per cluster for all clusters is guaranteed to be at least k, this privacy-preservation principle is referred to as k-anonymity. We use the k-members algorithm proposed by Byun et. al. [BKBL07] to cluster records in screen-space. Clustering can be done in two ways: *Data-based clustering*: clustering multiple dimensions at a time, using data-space properties and *Visual clustering*: axis-pairwise clustering using screen-space metrics. In our previous work [DK11] we have established the benefits of visual clustering, as it preserves the structures in the screen space and the user can understand the trends and relationships much more effectively. In this

	Cause/Effect of Uncertainty	Measurable Quantities	Measured Criteria
Encoding	Precision Granularity	Binning, Cluster range	Privacy Privacy
Decoding	Spatial Accuracy Identity Traceability Pattern Complexity	Cluster Range Cluster Overlaps Cluster Splits Semantic Structures	Privacy Privacy, Utility Privacy, Utility Utility

Table 1: *Connecting sources and effects of visual uncertainty to measurable quantities and their relationship to privacy and/or utility. This helps in systematically defining metrics for a privacy-preserving visualization.*

work, we apply the visual uncertainty metrics for measuring the privacy and utility of the two types of clustering.

Advantages of visual clustering over data clustering are two-fold: the selection of seed clusters is guided by properties of the pixel bins, that guarantees optimum cluster sizes and the fact, that, individual axis pairs are clustered rather than all the dimensions at once, thus preserving local structures. Properties of pixel bins like over plotting and convergence/divergence provide an implicit aggregation for the data points. The seed points for clusters are selected from the bins with higher over plotting or convergence/divergence. Since these artifacts are taken into account in visual clustering, the quality of clustering is better, both from privacy and utility points-of-view, that we prove with our metrics.

4. Privacy Model Based on Visual Uncertainty

In this section we conceptualize the relationship between privacy and visual uncertainty in the context of position-based representations like scatter plots, parallel coordinates, etc. We refer to malicious users with intention of privacy breach as *attackers*, and outline our assumptions regarding their background knowledge about the data in course of our description of the metrics.

4.1. Applying the Visual Uncertainty Taxonomy

Visual uncertainty can be decomposed into a set of encoding and decoding uncertainties, according to the visual uncertainty taxonomy [DCK12]. This taxonomic approach offers an opportunity to identify the causes, effects and sources of uncertainty that can be related to privacy and utility of a visualization. In Table 1 we tie these different elements together. In course of our ensuing discussion we refer to the various levels of the taxonomy tree proposed in that work.

As discussed in Section 3, binning and clustering are the basic elements of our privacy model. For encoding uncertainty, one option can be hiding sensitive data values which would lead to completeness uncertainty at the data mapping stage. However, since the goal is to minimize information loss, we do not consider that option. Instead, we focus on quantifying uncertainty introduced at the visual mapping stage, since privacy-preservation is based on screen-space properties. Encoding uncertainty in the form of precision and granularity are introduced due to binning and clustering. These are causes of intended uncertainty, affect the static visual representation of the clusters and are not influenced by interaction.

The components of decoding uncertainty affect how an attacker is able to gain information by using interaction. Cluster overlaps cause identity uncertainty and and splits cause traceability uncertainty (in parallel coordinates). These are related to both privacy and utility. For example, if an attacker knows the existence of a data value on one of the dimensions, and tries to guess the values for the other dimensions, then cluster overlaps help in creating identity uncertainty and making privacy breach difficult by hiding the cluster membership. However, too many overlaps create clutter and therefore make effective perception of patterns more difficult. Thus decoding uncertainty includes both intended and unintended forms of uncertainty. Metric-based analysis of visual uncertainty helps quantify these different forms and design a privacy-preserving visualization that balances these trade-offs.

4.2. Connecting Privacy, Utility, and Visual Uncertainty

Information is a measure of the decrease of uncertainty for the receiver of a message [Sha48]. If visualization is viewed as a communication channel from the data space to the perceptual and cognitive mental space of the user [PAJKW08], it is important to trace the uncertainty along different stages of the pipeline, so that the information communicated to the user can be optimized. In case of privacy-preserving visualizations, some forms of uncertainty would be intended, to prevent disclosure of sensitive information. In general, increasing the amount of *visual uncertainty* in a visualization will increase *privacy* of the visualization while decreasing its *utility*. In other words, privacy and utility are functions of visual uncertainty.

Let $u_1, u_2, \dots, u_k \in [0,1]$ be the quantities corresponding to a set of measurable uncertainties in a visualization, with 0 being most certain and 1 being most uncertain. As certainty is usually treated as probability [Hal03], here we consider an uncertainty measure u_i corresponds to a probability measure p_i as a dual, i.e., $u_i = 1 - p_i$. In this work we propose different measures u_i that address the different causes and effects of uncertainty as categorized by the taxonomy of visual uncertainty [DCK12]. Here we use the term “measure” in a broad sense, including both computational measurement and human-centered quantitative evaluation. Let us consider two approximated measurements for privacy (m_p) and utility (m_t) of the visualization. In general, m_p and each of u_i are positively (or non-negatively) correlated, while m_t and u_i are negatively (or non-positively) correlated. However, the integration of different u_i is much more complex in our context. For example, given two uncertainty measurements, u_a and u_b, for representing the visual uncertainty of two different features a and b in a visualization,

- a privacy measure, $m_p = \min(u_a, u_b)$ may be used if a privacy concern can be compromised by ascertaining either a or b;
- $m_p = u_a \cdot u_b$ may be used if a privacy concern can only be compromised by ascertaining both a and b jointly, and u_a and u_b are independent;
- $m_p = u_a + u_b$ may be used if a privacy concern is accumulative over different observations and u_a and u_b are mutually exclusive.

We can make similar observations about utility measure m_t. However, one cannot assume a uniform pair-wise relationship across different uncertainty measures to be discussed in this work. This will become clear after detailing different uncertainty in the following sections. In this section, we give two generalized formulae for m_p and m_t respectively.

$$m_p = 1 - \sqrt[\beta_p]{\sum_1^k \left(\alpha_{p,i}(1-u_i)^{\beta_p}\right)} \quad (1)$$

$$m_t = \sqrt[\beta_t]{\sum_1^k \left(\alpha_{t,i}(1-u_i)^{\beta_t}\right)} \quad (2)$$

where $\beta_p > 0$, $\alpha_{p,i} \geq 0$, and $\sum_i \alpha_{p,i} = 1$. $\alpha_{t,i}$ and β_t are constrained in the same manner. Although m_p and m_t are within the range of [0, 1] and encode a fair amount of probabilistic information, we should not treat them as probability measures. In fact, it is more appropriate to consider both as distance measures. For example, by setting $\beta_p \to \infty$, the right side of m_p becomes a min function for all u_i. However, as these measures are not always applied to the same privacy concern, we cannot really say if a specific part of a visualization is compromised (e.g., $\exists i, u_i = 0$), all other uncertainties will also be compromised (i.e., $\forall j \neq i, u_j = 0$). After weighting various pros and cons, we decide to set β_p to 1 in this work, which is the simple city block distance to emphasize the difference of these measures. Similarly, by treating each measure as a separate event, the simple city block distance offers an adequate approximation to the m_t in eq.(1).

In practice, privacy and utility may also be affected by factors other than visual uncertainty, such as the environment where the visualization is used. The above measurement, m_p and m_t should be used only for comparing visualization with different forms of anonymization while those other factors remain unchanged.

4.3. Choice of Metrics

For some forms of visual uncertainty, like that due to pattern complexity, screen-space metrics already exist, like Scagnostics for scatter plots [WAG05], Pargnostics for parallel coordinates [DK10], etc. However, metrics for other types of uncertainty are missing in the current literature. Some of the metrics that we propose are applicable beyond the context of privacy, where the issue of disparity between the large number of data points and limited number of pixels arise.

Encoding uncertainty serves as the initial defensive mechanism against attackers with no background knowledge about the data. Loss of precision due to binning and high-level of granularity due to clustering make it difficult for an attacker to guess the exact value and number of data points within a cluster. To quantify these, we have developed the *cluster range metric* and the *cluster summary error metric*. The cluster range metric also captures the decoding uncertainty involving spatial accuracy for guessing the location of the data points within a cluster. Encoding uncertainty cannot be reduced by using interaction.

When an attacker has some background knowledge about the data, the different components of decoding uncertainty help in confusing the attacker. When an attacker knows about the existence about a particular data-point in the database, the process of privacy breach starts by associating a data point with a cluster. Identity uncertainty due to cluster overlaps make that association difficult. Overlaps also lead to clutter, where identifying paths of clusters itself is difficult. We quantify the privacy aspect of identity uncertainty through the *overlap entropy* metric and the utility aspect dealing with clutter, through the *overlap clutter* metric. For line-based parallel coordinates, traceability of lines across multiple dimensions is an advantage over scatter plots. However, in privacy-preserving parallel coordinates, due to axis pair-wise clustering, clusters appear to split across axis. This leads to uncertainty due to lack of traceability, that we capture through our *average split count* metric. For controlling the uncertainty due to pattern complexity we use the Pargnostics metrics and also compute the mutual information between adjacent dimensions. These are essential for choosing the best adjacency configurations for parallel coordinates and also point out the pairs of axes with high utility in scatter plots.

5. Metrics

In the following section we introduce the metrics for measuring the different types of uncertainty. We describe the type of uncertainty measured by each metric and quantitatively describe each with illustrations and examples from the *Diabetes* dataset that are described in detail in Section 6.

	$Y_A > Y_B$	$Y_A < Y_B$	$Y_A m Y_B$ $Y_B m Y_A$	$Y_A s Y_B$ $Y_B s Y_A$	$Y_A f Y_B$ $Y_B f Y_A$	$Y_A = Y_B$	$Y_A o Y_B$ $Y_B o Y_A$	$Y_A d Y_B$ $Y_B d Y_A$
$X_A > X_B$	N	B	E	OB	OB	OB	OB	OB
$X_A < X_B$	B	N	E	OB	OB	OB	OB	OB
$X_A = X_B$	OB	OB	OB	OB	OB	OB	OB	OB
$X_A m X_B$ $X_B m X_A$	E	EB	EB	OB	OB	OB	OB	OB
$X_A s X_B$ $X_B s X_A$	OB	OB	OB	OB	OB	OB	OB	OB
$X_A f X_B /$ $X_B f X_A$	OB	OB	OB	OB	OB	OB	OB	OB
$X_A o X_B$ $X_B o X_A$	OB	OB	OB	OB	OB	OB	OB	OB
$X_A d X_B$ $X_B d X_A$	OB	OB	OB	OB	OB	OB	OB	OB

Table 2: *Cluster overlaps depending on the relationship of the clusters (A and B) on the axes (X and Y)in parallel coordinates.* ***N****: No overlap either on the axes or between the axes;* ***B****: Overlap only between axes;* ***OB****: Overlap between as well as on the axes;* ***E****: Meeting at the edge,****EB****: Meeting on the axes and overlap between axes.*

5.1. Cluster Summary Error

The further an actual record is perceived to be located from its actual position, the more difficult it would be to precisely guess the value of a record. In case of pixel-based representation, binning already introduces loss in precision due to quantization error. A cluster-based representation accentuates the error: the further away a record is from the cluster centroid, the more difficult it will be for knowing the exact value. We measure the summary error for privacy-preserving clustering as the Manhattan distance between the its actual pixel coordinate and the pixel coordinate of the cluster centroid. This is similar to the class consistency measure [SNLH09] for selecting optimal views of the data using scatter plots. Although in case of cluster-based representation this is not directly reflected in what the user sees on screen, the metric gives a quantitative measure of precision uncertainty.

Consider a cluster C_t consists of n_l records. It intersects with an axis, spanning over several pixel bins, $a_t, a_t + 1, \ldots, b_t - 1, b_t$ where $0 \le a_t \le b_t \le h - 1$, where h is the total number of pixels of this axis. The centroid of the intersected section of the cluster is thus:

$$\eta_t = \frac{a_t + b_t}{2}$$

The error of this intersection ε_t can be defined in terms of the Manhattan distance of the pixel coordinates of the data points within a cluster from its centroid, as follows:

$$\varepsilon_t = \frac{1}{n_l h} \sum_{i=1}^{n_l} |s_i - \eta_t| \tag{3}$$

The normalized average error over all clusters is given by:

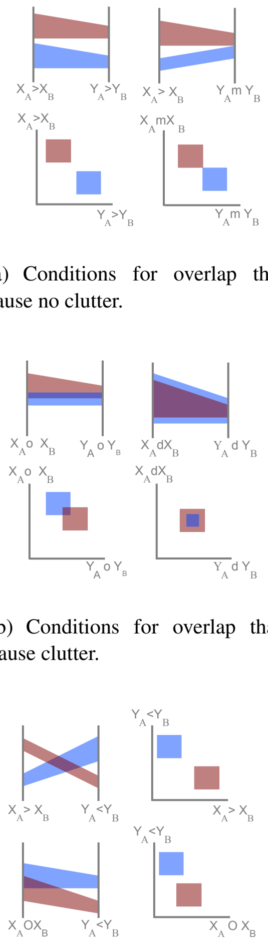
$$\varepsilon = \frac{1}{n_c} \sum_{i=1}^{n_c} \varepsilon_t \tag{4}$$

where s_i is the actual mapped pixel coordinate of record R_i on that axis.

We compare the cases in Figure 1 for the cluster summary error metric. For x-axis, we have: $\varepsilon_{A1} = 0.056$, $\varepsilon_{B1} = 0.056$; $\varepsilon_{A2} = 0.222$, $\varepsilon_{B2} = 0.222$; and $\varepsilon_{A3} = 0.111$, $\varepsilon_{B3} = 0.333$. For y-axis, we have: $\varepsilon_{A1} = 0.056$, $\varepsilon_{B1} = 0.056$; $\varepsilon_{A2} = 0.278$, $\varepsilon_{B2} = 0.222$; and $\varepsilon_{A3} = 0.167$, $\varepsilon_{B3} = 0.333$. Therefore configuration 1 is less private than configurations 2 and 3 on both axes. The average cluster summary error is much higher in case of data-based clustering. This coincides with higher cluster ranges which leads to clutter and reduces the visual quality. Information loss in terms of precision of data values is also much higher in this case.

5.2. Cluster Range

Cluster ranges on the axes mask the precise location of the data points. A cluster in the data-space is perceived in terms of the number of record it contains. In the screen space, it is perceived in terms of the number of pixel bins covered by the cluster on the axes, which we define as cluster range. When the analysts has no background knowledge about the data and tries to randomly guess if data points exist or not, cluster ranges lead to granularity uncertainty and that uncertainty due to lack of spatial accuracy. The larger the range, the less accurate will be estimation of the value of any record within this range, and at the same time, the more likely will it cause overlapping among clusters. Though a cluster range can be perceived as both a privacy and utility metric, since its primary role is masking data values and the uncertainty is intended, we consider cluster range as a privacy metric. Unlike cluster summary error, cluster range is independent



(a) Conditions for overlap that cause no clutter.

(b) Conditions for overlap that cause clutter.

(c) Conditions for overlap that cause clutter in parallel coordinates but not in scatter plots.

Figure 3: *Illustrating difference in effects of cluster overlaps for scatter plots and parallel coordinates. The red cluster is represented by A and the blue cluster by B.*

of the number of records in the cluster or their individual values.

Consider a cluster C_t. Its intersection with the axis spans between pixel coordinates a_t and b_t, where The normalized range of this cluster is thus $(b_t - a_t)/(h-1)$. We can define an axis-based metric as the average range of all clusters intersecting with the axis as:

$$\gamma = \frac{1}{n_c(h-1)} \sum_{t=1}^{n_c} (b_t - a_t) \tag{5}$$

where n_c is the total number of clusters. We compare the cases in Figure 1 for the cluster range metric. For x-axis, we have: $\gamma_{A1} = 0.125$, $\gamma_{B1} = 0.125$; $\gamma_{A2} = 0.5$, $\gamma_{B2} = 0.5$; and $\gamma_{A3} = 0.25$, $\gamma_{B3} = 0.75$. For y-axis, we have: $\gamma_{A1} = 0.125$, $\gamma_{B1} = 0.125$; $\gamma_{A2} = 0.625$, $\gamma_{B2} = 0.500$; and $\gamma_{A3} = 0.375$, $\gamma_{B3} = 0.750$. Configurations 2 and 3 are therefore more private than 1.

5.3. Overlap Clutter

Cluster overlaps lead to identity and traceability uncertainty in perceiving the path of the clusters, therefore leading to clutter. In line-based parallel coordinates, the vertical distance between the start and end points of a line on adjacent axes can be treated as intervals [All83] to determine when lines cross [DK10]. In the case of cluster-based parallel coordinates and scatter plots, we treat the cluster ranges on each axis as intervals for detecting cluster overlaps. Allen's interval algebra defines 13 possible cases between two intervals, X and Y: X before Y, X starts Y, X ends Y, X meets Y, X during Y, X overlaps Y, and X equals Y. All but the last condition also have a symmetrical case. Given the 13 cases between the two clusters on each axis, we have to investigate $13 \times 13 = 169$ possible cases.

For parallel coordinates there are four possible pairwise relationship between cluster ranges: no overlap (N), meeting at the edges (E); overlap on and between axes (OB) and meeting at the edges and overlap between axes (EB). In fact, all cases where overlap happens on the axes also means that there is overlap between the axes. For scatter plots on the other hand, there is no distinction between an overlap *on* the axis and that *between* the axes, because the coordinate positions can be anywhere between the axes. The possible relationships between cluster ranges in scatter plots, are therefore, N, E, and O. A simple enumeration of the $13 * 13$ conditions enables us to draw a distinction among these overlap cases. The different possibilities for parallel coordinates are shown in Table 2 and those for scatter plots are shown in Table 3. The symmetrical conditions are shown together except for the greater than ($>$) and less than ($<$) condition as there is a distinction between parallel coordinates and scatter plots in this case.

Based on these conditions we can derive three cases that are relevant for clutter: A) overlap conditions that do not lead to clutter in either parallel coordinates or scatter plots (Figure 3(a)), B) overlap conditions that lead to clutter in both parallel coordinates or scatter plots (Figure 3(b)), and C) overlap conditions that lead to clutter in parallel coordinates but not in scatter plots (Figure 3(c)). For A, the conditions are either 'before/after' or 'meets' on both axes. For B, the conditions are 'overlaps', 'starts/finishes' and 'during on both axes.

One distinction between parallel coordinates and scatter plots is when there is a 'before/after' condition on one axis and a different condition on the other. This is reflected in a much higher number of N in Table 3 for scatter plots than parallel coordinates. In these cases clusters overlap between

	$Y_A > Y_B$	$Y_A < Y_B$	$Y_A m Y_B$ $Y_B m Y_A$	$Y_A s Y_B$ $Y_B s Y_A$	$Y_A f Y_B$ $Y_B f Y_A$	$Y_A = Y_B$	$Y_A o Y_B$ $Y_B o Y_A$	$Y_A d Y_B$ $Y_B d Y_A$
$X_A > X_B$	N	N	N	N	N	N	N	N
$X_A < X_B$	N	N	N	N	N	N	N	N
$X_A = X_B$	N	N	O	O	O	O	O	O
$X_A m X_B$ $X_B m X_A$	N	N	E	O	O	O	O	O
$X_A s X_B$ $X_B s X_A$	N	N	O	O	O	O	O	O
$X_A f X_B$ $X_B f X_A$	N	N	O	O	O	O	O	O
$X_A o X_B$ $X_B o X_A$	N	N	O	O	O	O	O	O
$X_A d X_B$ $X_B d X_A$	N	N	O	O	O	O	O	O

Table 3: *Cluster overlaps depending on the relationship of the clusters (A and B) on the axes (X and Y)in scatter plots. The notations are the same as in parallel coordinates except for the case B as there is no distinction between overlap between and on the axes in scatter plots.*

axes in parallel coordinates but there is no perceptual overlap in scatter plots. These conditions are the basis for our overlap clutter metric.

For the clusters, the upper bound for the number of overlaps is $\frac{n_c(n_c-1)}{2}$. Therefore we compute clutter in parallel coordinates (C_P) as:

$$C_P = \frac{2n_o}{n_c(n_c-1)} \tag{6}$$

where n_o is the total number of overlaps in parallel coordinates or scatter plots. For scatter plots we denote clutter by C_S. Since n_o is much lower in scatter plots than parallel coordinates, i.e., identity and traceability uncertainty are lower in scatter plots than parallel coordinates, $C_P > C_S$ irrespective of higher or lower k.

5.4. Overlap Entropy

The previous metrics do not take any possible background knowledge of an attacker into account. If an attacker knows which cluster a data point belongs to, then the privacy breach becomes easier, than the case when there is identity uncertainty regarding associating a data point with a cluster. Overlapping cluster ranges on the axes lead to uncertainty because of difficulty in knowing which cluster a pixel bin belongs to and consequently, tracking the clusters across different axis-pairs. When certain data values are known to attackers, overlaps help in creating uncertainty about cluster membership of a data point as illustrated in Figure 4(a). We use an information theoretic measure in the form of Shannon's entropy to quantify the uncertainty in tracking precise membership of a bin in cluster. This is similar to the privacy metric based on entropy suggested by Agrawal et al. [AA01] and Bertino et al. [BLJ08].

Consider n_c clusters on an axis in a privacy preserving parallel coordinates visualization. The axis has h pixel bins. Each bin may intersect with zero, one or several clusters, while each cluster may span over one or more bins. As identifying an empty bin is trivial, the uncertainty is thus associated with those bins that intersect with one or more clusters. Assume that the attacker has no a priori knowledge about any cluster, so the probability of making a correct guess of the association between a bin and a cluster is independent and identically-distributed.

Let α_i be the number of clusters intersect with bin β_i, where $0 \le i \le h-1$. Given a cluster, C_t, the probability mass function for identifying this cluster at bin β_i is thus

$$P(t@i) = \begin{cases} 0 & \text{if } C_t \text{ does not intersect with } \beta_i \\ 1/\alpha_i & \text{if } C_t \text{ intersects with } \beta_i \end{cases}$$

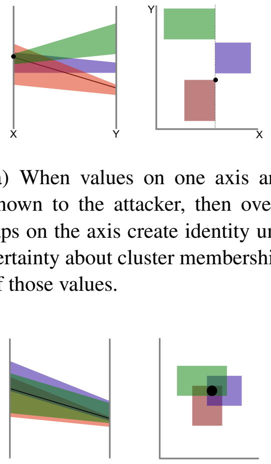
The entropy in relation to cluster C_t is thus the following sum computed over all non-zero $Pt@i$.

$$H_t = -\sum_{i=0}^{h-1} P(t@i)\ln P(t@i)$$

We can compute an information-theoretic measure of uncertainty of the axis as

$$\Phi = \frac{\sum_{t=1}^{n_c} H_t}{n_c H_{max}} \tag{7}$$

H_{max} is the maximum entropy value for H_t, which is associated with a situation where every cluster spans over every pixel bin, that is, $P(t@i) = 1/n_c$ for every cluster and every bin. The lower H_t is, the lower information it contains about C_t, and thus higher privacy. From empirical results, we have observed that the absolute value of entropy increases with increasing k. Since with increasing k there are more overlaps, we have more uncertainty in the screen-space. When



(a) When values on one axis are known to the attacker, then overlaps on the axis create identity uncertainty about cluster membership of those values.

(b) When values on both axes are known to the attacker, then overlap on both axes create identity uncertainty. Large number of such overlaps also reduces the mutual information between adjacent axes.

Figure 4: *Illustrating how overlaps on one axis and that on both axes lead to uncertainty.*

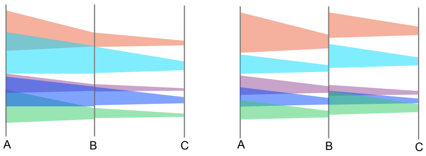


Figure 5: *In case of multi-dimensional clustering,on the left, there is no traceability uncertainty on brushing, as clusters are continuous. In case of axis pairwise clustering, on the right, traceability uncertainty of an axis pair depends on the average number of split cluster on the adjacent axis.*

we compare with data-based clustering, the entropy value is lower as compared to visual clustering, because of less number of overlaps among clusters (Figure 6b).

5.5. Mutual Information

When an attacker knows both coordinates of a two-dimensional data point, then uncertainty due to overlaps is only caused when overlaps are on both axes. As shown in Figure 4(b), these types of overlaps reduce the mutual information between the two adjacent axes. We consider the mutual information as an utility metric and this metric is important for handling interaction scenarios like reordering axes. The mutual information, a measure of the reduction in uncertainty of one variable due to the knowledge of the other, needs to be maximized for utility purposes. The general formula for the mutual information between two random variables X and Y is

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\ln\frac{P(x,y)}{P(x)P(y)} \tag{8}$$

Where $P(x,y)$ is the joint probability of x and y, and $P(x)$ and $P(y)$ are the marginal probabilities. For a record with values $x_{i,j}$ and $x_{i,j+1}$ on adjacent axes on a parallel coordinates plot, the joint probability is equal to the uncertainty of that record's exact location, which is determined by the number of clusters that contain it. Consider two axes, x and y. Given a specific cluster C_t, we can compute the probability mass functions $P(t@x_i)$ and $P(t@y_j)$ as in the previous section. The joint probability mass function can be defined as:

$$P(t@x_i,t@y_j) = \begin{cases} 0 & \text{if condition (i)} \\ \frac{1}{\alpha_{x,i}\alpha_{y,j}} & \text{if condition (ii)} \end{cases}$$

where condition (i) is when C_t does not intersect with the i^{th} bin on x-axis or the j^{th} bin on y-axis; and condition (ii) is when C_t intersects with both the i^{th} bin on x-axis and the j^{th} bin on y-axis. We can thus compute the mutual information for cluster C_t between the two axes using $I(X;Y)$. The maximum mutual information is when all clusters intersect with the two axes at the exactly same bins. As expected, mutual information decreases for increasing k, but not monotonically for every k, as shown in Figure 6a).

5.6. Average Split Count

The average number of split cluster per axis pair is an indication of the traceability uncertainty. When an attacker selects a cluster of interest, the larger the number of cluster splits on the adjacent axes, the more difficult will it be to trace the cluster that contains the same record as the selected cluster. This form of uncertainty helps in meeting the l-diversity criteria [MKGV07], which ensures sufficient diversity between a quasi-identifier axis and a sensitive attribute axis, so that a cluster cannot be associated with exactly one sensitive value. In our previous work on privacy-preserving parallel coordinates, we had shown how cluster splits help us achieve l-diversity in the interactive scenario [DK11]. For computing the average split count, we have to consider two axis pairs together as shown in Figure 5. For each cluster, we compute the number of splits on adjacent axes. The average number of spits per cluster in axis pair, indicates the level of traceability uncertainty where T is given by the following equation:

$$T = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{1}{Split(C_i)} \tag{9}$$

where $Split(C_i)$ counts the number of split clusters for the i_{th}

cluster on adjacent axis. The lowest traceability uncertainty is when $T = 1$, that is, there is a one-one association between clusters on adjacent axes. In case of multidimensional clustering T is always equal to 1. We consider T as a utility metric and $1-T$ as a privacy metric. In this respect multidimensional clustering has higher utility than visual clustering. However, the privacy is lower because each cluster can be associated with exactly one cluster on the adjacent axis. In case of l-diversity this becomes a problem and can lead to disclosure of sensitive attributes.

5.7. Measures based on Pargnostics

Visual structures in a visualization represent semantic patterns within the data. Transformation of the representation distorts those structures as in privacy-preserving clustering. The different visual artifacts, like parallel lines, converging/diverging lines and line-crossings between adjacent axes; represent the trends and relationship between adjacent data dimensions. In previous work we had developed a set of metrics that quantify these different structural properties [DK10]. In case of cluster-based visualization, it is useful to see how the structures get preserved or distorted with comparison to line-based parallel coordinates. Since most cluster boundaries represent connection between actual data points, we treat the cluster boundaries as lines and apply the Pargnostics metrics on this lines. Even in some cases where the cluster boundaries do not represent actual data points, their orientation between an axis-pair leads to the overall perception of the dominant visual structure there.

Cluster Parallelism. To describe parallelism, we compute a vertical distance histogram between any two cluster boundaries on adjacent axes. Then we look at the distribution of the distance values and estimate the interquartile range. Narrower range implies higher parallelism. We normalize the distances between 0 and 1, by dividing by the highest possible distance. With large cluster ranges cluster parallelism ($Par_{cluster}$) gets distorted.

Cluster Convergence/Divergence. In the original parallel coordinates, lines converging to or diverging from a few points on the adjacent axis form a frequently occurring pattern. We exploit these properties for seeding our clusters. The points with most convergence/divergence (which of them is the dominant structure) are our starting points for clustering. Similar to Pargnostics, we use the two-dimensional axis histogram to calculate the amount of convergence/divergence between adjacent axes and normalize the values with the maximum value of convergence/divergence.

To compute utility in terms of pattern preservation for an axis-pair, we compute the ratio of cluster-based parallelism ($Par_{cluster}$) and line-based parallelism (Par_{lines}), and the ratio of cluster based convergence/divergence ($CD_{cluster}$) and line-based convergence/divergence (CD_{lines}) and compute the average pattern-preservation for an axis-pair (V) as follows:

$$V = \frac{1}{2}\left(\frac{Par_{cluster}}{Par_{lines}} + \frac{CD_{cluster}}{CD_{lines}}\right) \qquad (10)$$

5.8. Additivity of Uncertainty Metrics

In the preceding sections, we described a number of uncertainty and certainty measures, including *cluster summary error* (uncertainty, ε), *cluster range* (uncertainty, γ), *overlap clutter* (certainty, C), *overlap entropy* (uncertainty, Φ), *mutual information* (certainty, I), *average split count* (certainty, T), and *pargnostics* (certainty, V). We consider that V, I and C do not contribute much towards any privacy concerns. We can remove their effects in formulating the privacy measure m_p by setting the corresponding $\alpha_{p,i}$ to zero. For the remaining measures, we simply treat them to have an equal contribution by setting $\alpha_{p,i}$ to the same value, i.e., $1/4$ in this case. We can easily observe that these measures are not applied to the same privacy concern, we cannot really say if ε is compromised, γ, Φ and $1-T$ are also compromised. This justifies our choice to use the Manhattan distance by setting β_p to 1 in Equation 1. Similarly we also remove the effects of ε, γ, Φ in computing m_t. For the remaining measures, we simply treat them to have an equal contribution by setting $\alpha_{t,i}$ to the same value, i.e., $1/4$ in this case.

6. Case Study

We use the *Diabetes* dataset [FA10] to illustrate the application of our metrics. This dataset has 768 records and consists of 6 dimensions: *number of times pregnant*, *blood pressure*, *serum insulin level*, *body mass index (BMI)*, *age*, and the binary attribute *class*. The sensitive dimension is the *class* attribute, all others are considered to make up the quasi-identifier attribute.

6.1. Privacy

Privacy is measured in terms of both encoding and decoding uncertainty as outlined in Table 1. On the encoding side, cluster range (γ) and cluster summary error (η) are higher in case of data-based clustering than visual clustering. This implies that precision and granularity uncertainty are higher in case of data-based clustering, signifying higher privacy. However, as shown in Figure 6b, decoding uncertainty due to overlaps, as measured by entropy, is much higher in case of visual clustering. Here, we observe that the uncertainty measured in terms of the overlap entropy of the axes (computed based on equation 7), increases for increasing k, which signifies higher uncertainty for guessing exactly which cluster a record belongs to. This metric not only measures the entropy in the static image, i.e, when highlighting/brushing

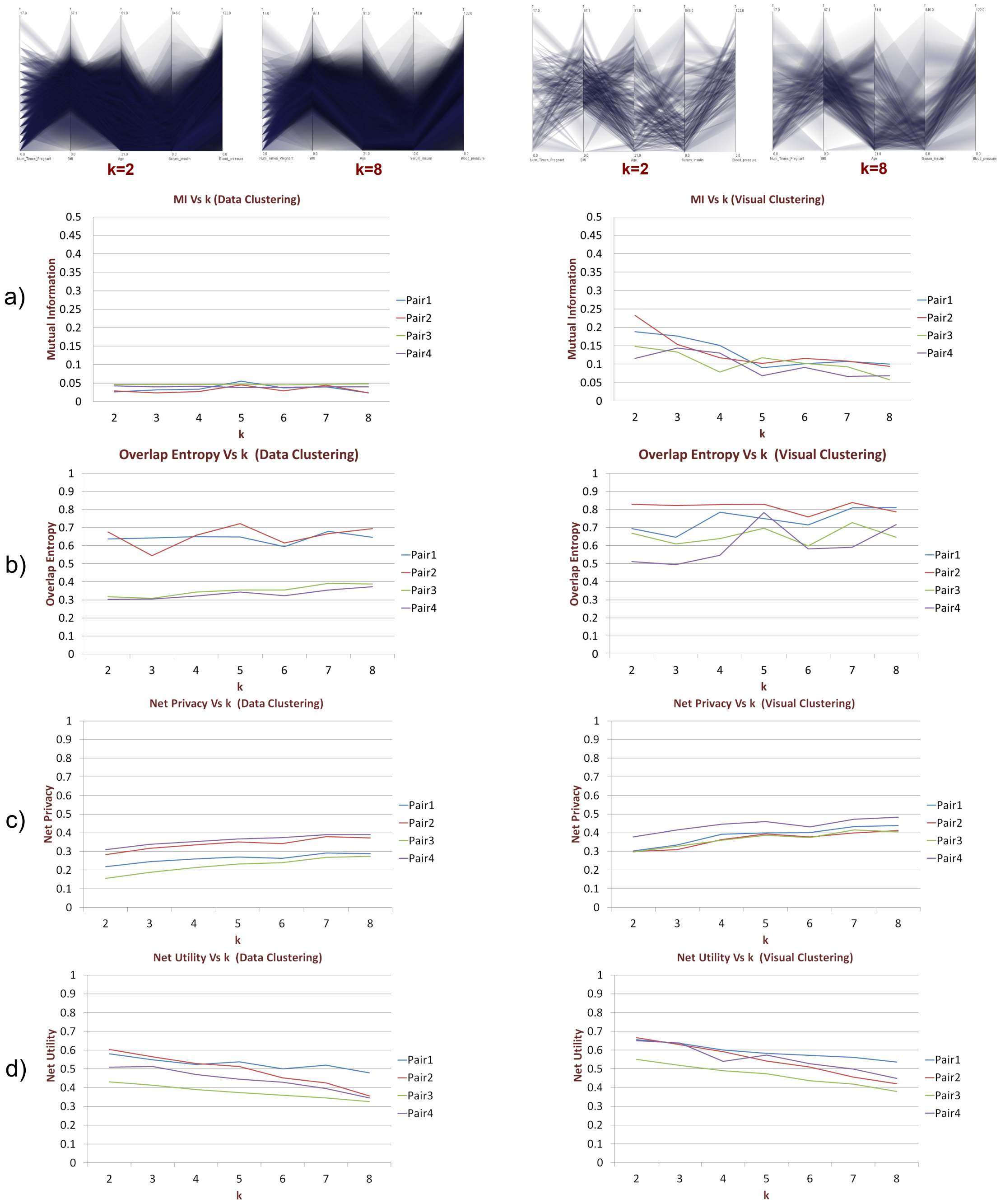


Figure 6: *Comparison of privacy and utility metrics for four different axis pairs in case of data clustering and visual clustering.*

is not available, but also covers cases where a user can select certain clusters by interaction. In our technique, when several clusters overlap on a pixel bin, we only highlight the smallest cluster. The attacker would thus not be certain whether a record which he/she is trying to guess the value of, belongs to that particular cluster.

In our experiments we have also found that the number of clusters, whose boundaries exactly coincide (the 'meets' condition in Allen's interval algebra), also decreases with increasing k, because with increasing k, there is a greater loss in precision of the values on the axis. This is beneficial from a privacy point-of-view because for smaller k, cluster boundaries meeting precisely on a pixel can potentially cause disclosure of the the data-points on these boundaries [DK11].

Traceability uncertainty is also much higher in case of visual clustering due to the higher average split count than in case of data-based clustering, where there are no cluster splits and the clusters have a one-to-one correspondence with continuing clusters on adjacent axes. As we discuss in Section 5.6, this creates the lack of l-diversity problem. The effect of higher cluster range and cluster summary error in case of data-based clustering is offset by the higher overlap entropy and higher average split count in case of visual clustering. This is reflected in the graphs for net privacy (m_p) in Figure 6c, which shows that privacy achieved in case of visual clustering is higher on average, than data-based clustering.

6.2. Utility

Utility is expressed in terms of the different metrics for decoding uncertainty, mainly clutter and pattern complexity. High mutual information between adjacent dimensions maximizes utility. Figure 6a) shows the variation of mutual information for increasing k for four different axis pairs of the *Diabetes* dataset. The difference of mutual information between the two types of clustering is very pronounced due to the large cluster ranges in case of data-based clustering, which ensures that a two-dimensional data point is overlapped by multiple clusters on both axes in most cases.

With increasing k, it is expected that patterns in the visualization will get distorted and will be more difficult to discern. We are able to have better utility in terms of screen-space clarity in case of visual clustering because of less pattern complexity. For increasing k the effect of parallelism (computed based on our discussion in Section 5.7) gets reduced. However, the decrease happens in quite small increments and therefore does not degrade the visualization much. In our experiments we have observed that for converging-diverging structures, there is no significant variation with changing k. With very small k, like $k = 2$, the patterns are almost same as in raw parallel coordinates. With increasing k, the number of converging/diverging lines do not increase or decrease significantly. This is because we use convergence-divergence as a criterion for seeding the clusters and there is not much change in choice of seeds with changing k.

Due to higher traceability uncertainty in case of visual clustering, the utility is reduced. However, the effect of the other uncertainty components is much more significant when m_t is computed. The comparison for net utility is shown in Figure 6d, where we can observe that higher net utility in case of visual clustering than data-based clustering.

6.3. Devising an effective k

We have shown that all the various uncertainty measures do not increase or decrease monotonically with k. Especially for overlap entropy and mutual information, there is a high degree of variability across different k (Figures 6a and 6b). A similar pattern is also reflected in the graphs for m_p and m_t in Figures 6c and 6d. This implies that higher k does not necessarily signify higher privacy and/or lower utility in the screen-space. This is an important difference from privacy preservation in the data space, where variation of k is directly proportional to the privacy achieved or utility lost. Metric-based analysis of visual uncertainty, therefore, will enable visualization designers to choose the effective k based on the requirements for privacy and utility.

6.4. Comparing parallel coordinates and scatter plots

Encoding uncertainty is identical in parallel coordinates and scatter plots as we get the same values for the metrics. On the decoding side, however, clutter and traceability produce different results. Although clutter is less in scatter plots than parallel coordinates, as described in Section 5.3, it cannot be readily concluded that in terms of perception , the former is better. This is because the degree of distortion of the visual structures is higher in scatter plots than parallel coordinates. In scatter plots, points (zero-dimensional entities) are transformed to rectangles (two-dimensional entities). In case of parallel coordinates, lines (one-dimensional entities) are transformed into polygons (two-dimensional entities). Therefore the structural properties are much less distorted. This can be observed in Figure 2, as the degree of linear relationship in case of parallel coordinates is much better perceptible than scatter plots.

7. Conclusions and Future Work

In the work reported here we have presented a privacy-preservation model for visualization based on scatter plots and parallel coordinates and proposed a set of metrics that measure the privacy and utility as functions of visual uncertainty. We have compared the data-based approach to ours with respect to these metrics and have proved the effectiveness of the latter in terms of utility and privacy. Some of the proposed metrics are also applicable beyond the confines of a privacy-preserving application, especially in case of cluster-based visualization [NH06]. The systematic quantification of visual uncertainty also shows a new approach to evaluating visualizations, so that they can can be iteratively refined based on various constraints and requirements.

As a next step, we will design an optimization function that balances the privacy and utility based on the metrics and guides the configuration of the display accordingly. We also want to apply the information-theoretic measures to compare the level of privacy and utility that can be achieved in a visualization to that in PPDM and point out the advantages and disadvantages of each approach.

References

[AA01] Agrawal D., Aggarwal C.: On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the ACM Symposium on Principles of database systems* (2001), ACM, pp. 247–255. 3, 8

[All83] Allen J.: Maintaining knowledge about temporal intervals. *Communications of the ACM 26* (1983), 832–843. 7

[AS00] Agrawal R., Srikant R.: Privacy-preserving data mining. *ACM Sigmod Record 29*, 2 (2000), 439–450. 2

[Bez08] Bezzi M.: An entropy based method for measuring anonymity. In *Third International Conference on Security and Privacy in Communications Networks* (2008), IEEE, pp. 28–32. 3

[BFP05] Bertino E., Fovino I. N., Provenza L. P.: A Framework for Evaluating Privacy Preserving Data Mining Algorithms*. *Data Mining and Knowledge Discovery 11*, 2 (2005), 121–154. 3

[BKBL07] Byun J., Kamra A., Bertino E., Li N.: Efficient k-anonymization using clustering techniques. In *Proceedings Database Systems for Advanced Applications* (2007), Springer, pp. 188–200. 3

[BLJ08] Bertino E., Lin D., Jiang W.: A survey of quantification of privacy preserving data mining algorithms. *Privacy-Preserving Data Mining* (2008), 183–205. 2, 8

[BTK11] Bertini E., Tatu A., Keim D.: Quality metrics in high-dimensional data visualization: an overview and systematization. *Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2203–2212. 1

[CJ10] Chen M., Jänicke H.: An information-theoretic framework for visualization. *Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1206–1215. 3

[CM97] Card S., Mackinlay J.: The structure of the information visualization design space. In *Proceedings, Symposium on Information Visualization,* (1997), IEEE, pp. 92–99. 3

[DCK12] Dasgupta A., Chen M., Kosara R.: Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum 31*, 3pt2 (2012), 1015–1024. 2, 3, 4

[DK10] Dasgupta A., Kosara R.: Pargnostics: Screen-space metrics for parallel coordinates. *Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1017–26. 5, 7, 10

[DK11] Dasgupta A., Kosara R.: Adaptive privacy-preservation using parallel coordinates. *Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2241–2248. 1, 2, 3, 9, 12

[FA10] Frank A., Asuncion A.: UCI machine learning repository. http://archive.ics.uci.edu/ml, 2010. 10

[FLC*02] Freitas C., Luzzardi P., Cava R., Winckler M., Pimenta M., Nedel L.: On evaluating information visualization techniques. In *Proceedings, Advanced Visual Interfaces* (2002), ACM, pp. 373–374. 1

[Hal03] Halpern J. Y.: *Reasoning about Uncertainty*. The MIT Press, 2003. 4

[HLS*12] Holzhüter C., Lex A., Schmalstieg D., Schulz H.-J., Schumann H., Streit M.: Visualizing uncertainty in biological expression data. In *Proceedings Visualization and Data Analysis* (2012). 3

[ID90] Inselberg A., Dimsdale B.: Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization* (1990), IEEE CS Press, pp. 361–378. 3

[Joh04] Johnson C.: Top scientific visualization research problems. *Computer graphics and applications, IEEE 24*, 4 (2004), 13–17. 3

[MKGV07] Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M.: l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data 1*, 1 (2007), 3. 2, 9

[NH06] Novotny M., Hauser H.: Outlier-preserving focus+context visualization in parallel coordinates. *Transactions on Visualization and Computer Graphics 12*, 5 (2006), 893–900. 13

[PAJKW08] Purchase H., Andrienko N., Jankun-Kelly T., Ward M.: Theoretical foundations of information visualization. In *Information Visualization: Human-Centered Issues and Perspectives*. Springer, 2008, pp. 46–64. 3, 4

[PWL97] Pang A., Wittenbrink C., Lodha S.: Approaches to uncertainty visualization. *The Visual Computer 13*, 8 (1997), 370–390. 3

[RLBS03] Rhodes P., Laramee R., Bergeron R., Sparr T.: Uncertainty visualization methods in isosurface rendering. In *Eurographics* (2003), pp. 83–88. 3

[RNC*95] Russell S., Norvig P., Canny J., Malik J., Edwards D.: *Artificial intelligence: a modern approach*. Prentice hall, 1995. 3

[RWX*07] Rundensteiner E. A., Ward M. O., Xie Z., Cui Q., Wad C. V., Yang D., Huang S.: Xmdvtool: Quality-aware interactive data exploration. In *SIGMOD Conference* (2007), pp. 1109–1112. 3

[Sha48] Shannon C. E.: A mathematical theory of communication. *The Bell System Technical Journal 27* (1948), 379–423. 4

[SNLH09] Sips M., Neubert B., Lewis J., Hanrahan P.: Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 831–838. 6

[Swe02] Sweeney L.: k-anonymity: A model for protecting privacy. *IEEE Security And Privacy 10*, 5 (2002), 1–14. 2

[WAG05] Wilkinson L., Anand A., Grossman R.: Graph-theoretic scagnostics. In *Proceedings Information Visualization* (2005), IEEE CS Press, pp. 157–164. 5

[YPF00] Yang-Peláez J., Flowers W. C.: Information content measures of visual displays. In *Proceedings of the IEEE Symposium on Information Vizualization 2000* (2000), IEEE Computer Society, pp. 99–103. 3