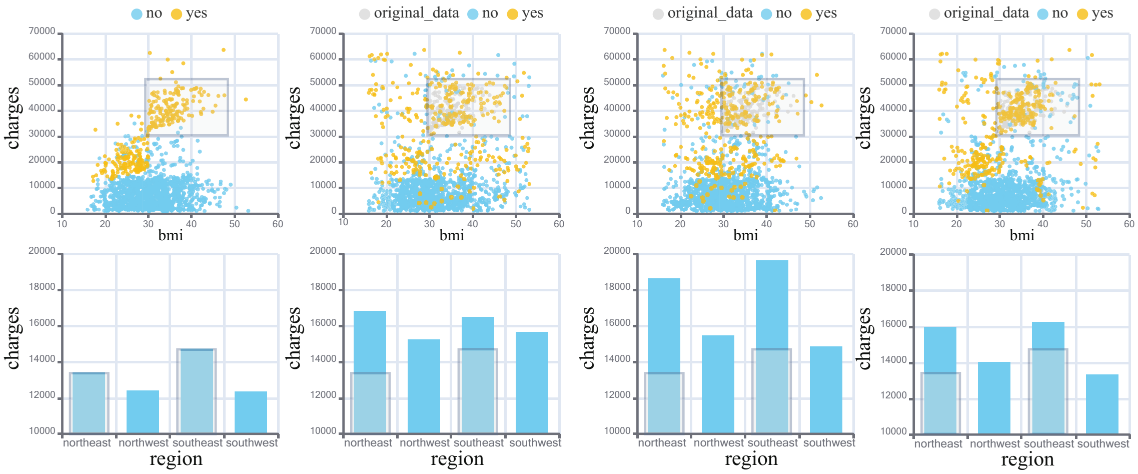
DPVisCreator: Incorporating Pattern Constraints to Privacy-preserving Visualizations via Differential Privacy

Jiehui Zhou, Xumeng Wang, Jason K. Wong, Huanliang Wang, Zhongwei Wang, Xiaoyu Yang, Xiaoran Yan, Haozhe Feng, Huamin Qu, Haochao Ying, and Wei Chen



(a) Original (b) PrivBayes (c) PriVis (w_cluster=0.5, w_order=0.5) (d) PriVis (w_cluster=1,w_order=0.1)

Fig. 1. Comparison of a series of privacy-preserving visualization charts under the same privacy budget. (a) The original data. (b) The PrivBayes [76] results, where the cluster in the gray boxes are barely retained, and the ranking of the selected bars changes. (c, d) The *PriVis* results with different pattern constraints, showing that adjusting the importance weights can help maintain different patterns.

Abstract— Data privacy is an essential issue in publishing data visualizations. However, it is challenging to represent multiple data patterns in privacy-preserving visualizations. The prior approaches target specific chart types or perform an anonymization model uniformly without considering the importance of data patterns in visualizations. In this paper, we propose a visual analytics approach that facilitates data custodians to generate multiple private charts while maintaining user-preferred patterns. To this end, we introduce pattern constraints to model users' preferences over data patterns in the dataset and incorporate them into the proposed Bayesian network-based Differential Privacy (DP) model *PriVis*. A prototype system, *DPVisCreator*, is developed to assist data custodians in implementing our approach. The effectiveness of our approach is demonstrated with quantitative evaluation of pattern utility under the different levels of privacy protection, case studies, and semi-structured expert interviews.

Index Terms—Privacy-preserving visualization, visual analytics, differential privacy, tabular data

1 Introduction

Data visualization is widely applied to support the exploration and understanding of domain data patterns [34, 53, 54]. For instance, visualizations help medical practitioners analyze the development of patients' diseases by intuitively showing their health status throughout time [56]. However, when the analyzed data contains sensitive information, unrestricted and unscrutinized representations of such data can result in privacy leakage. A large number of data breaches [19, 62, 67] have caused severe financial losses and reputation crises. Disclosing sensitive disease information may also provoke discrimination against patients and carry legal consequences [57]. Many countries and jurisdictions have tightened privacy regulations to protect their citizens' data security and privacy rights. For example, the EU enacted the GDPR to strengthen individuals' control over personal data. In the UK, health data must be processed and analyzed in the Trusted Research Environment (TRE) to ensure that no personal information is revealed in any analytical results. Therefore, visualization publishing has an urgent requirement of privacy preservation.

However, it is challenging to generate a group of privacy-preserving visualizations that represent multiple data patterns to support comprehensive data understanding. Privacy-preserving visualizations can be achieved through screen-space sanitization and data-space sanitiza-

- *J. Zhou, H. Wang, Z. Wang, X. Yang, H. Feng, and W. Chen are with the State Key Lab of CAD&CG, Zhejiang University. W. Chen is also with the Laboratory of Art and Archaeology Image (Zhejiang University), Ministry of Education, China. E-mail: {zhoujiehui, 22051090, wzw09, 22051142, fenghz, chenvis}@zju.edu.cn.*
- *X. Wang is with TMCC, CS, Nankai University. E-mail: wangxumeng@nankai.edu.cn.*
- *J. Wong and H. Qu are with the Hong Kong University of Science and Technology. Email: {kkwongar, huamin}@cse.ust.hk.*
- *X. Yan is with the Zhejiang Lab. E-mail: xiaoran.a.yan@gmail.com.*
- *H. Ying is with the School of Public Health, Zhejiang University. He is also with the Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province. E-mail: haochaoying@zju.edu.cn.*
- *Haochao Ying and Wei Chen are the corresponding authors.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

tion [23]. Screen-space sanitization creates visual clutter by overlapping [20] and splitting [21] elements to obfuscate the visual effects of sensitive information. However, privacy preservation from the perception perspective is disputable without extensive evaluation. Existing data-space sanitization [13, 15, 74] applies automatic models to detect privacy issues and handle them by manipulating the data directly. Recent advanced techniques such as Differential Privacy (DP) [26] are theoretically proven to avoid privacy disclosure. However, these approaches are not designed for visualization; They perform privacy-preserving processing at each data point uniformly without considering the importance of the data patterns when visualized. Therefore, important patterns may be highly distorted or even destroyed. Visualizing these results would decrease the usability of the charts and increase the risk of communicating incorrect information.

To address this challenge, we develop a novel visual analytics approach that enables data custodians to generate privacy-preserving visualizations that reserve preferred patterns. We introduce *pattern constraints* to model users' preferences over different patterns in a sensitive dataset. The multiple pattern constraints can be represented as a distribution of importance over the entire dataset. We further propose *PriVis*, a privacy-preserving visualization generation model, which extends the Bayesian network-based DP data generation method to consider pattern constraints in the process of structure learning on the data. We present *DPVisCreator*, an interactive prototype system that allows users to interactively explore data patterns in sensitive datasets and specify the pattern constraints. It has a set of coordinated views to help users understand the impacts of different pattern constraints on the privacy-preserving model and the final generated visualizations. It supports efficient privacy-preserving process configuration and intuitive validation of results through user-friendly interactions. *DPVisCreator* is capable of meeting data custodians' needs to publish different types of charts that preserve both privacy and desired patterns. Our contributions are as follows:

- A privacy-preserving visualization generation model, *PriVis*, to maintain multiple data patterns by extending a DP data publishing model based on Bayesian networks with pattern constraints.
- An interactive prototype system, *DPVisCreator*, to specify pattern constraints, understand the privacy-preserving process, and compare various privacy protection schemes.

2 Background

We present an example scenario to illustrate the motivation and application of privacy-preserving visualization. Then, we introduce the basic concepts of DP used in this paper. Finally, we describe evaluation metrics for the utility of the visualization.

2.1 Example Scenario in Healthcare

Hospitals hold many electronic health records (EHRs) containing patients' private information, such as those in Table 1. In this scenario, there are three stakeholders: *data owner*, *data custodian*, and *data consumer*. The data owner is the hospital, which owns the patients' EHRs. The data custodian is a clinical data analyst with access to the raw data and is tasked with publishing data through visualizations. The data custodian needs to have some basic knowledge of privacy. The data consumer receives the published visualizations for information.

Table 1. An example sensitive EHRs dataset.

Name	Age	Cholesterol	Heart attack
Liam Clark	23	6.2	Yes
Walter Woods	52	2.5	No
Rebecca Fraser	30	3.8	No
Caitlyn Gonzales	42	4.2	Yes
Laura Hernandez	69	3.1	Yes

The data custodian wants to demonstrate that people over 60 are at a much greater risk of heart disease than others. A bar chart can visually summarize the number of heart attack patients in each age group, but it may also disclose private information about individuals' illnesses. For example, if malicious attackers know that Liam is the only patient in the 20-25 age range who had been to the hospital, they could infer that Liam had a heart attack based on the distribution pattern in the bar chart. Leaking such personal information may lead Liam to suffer from targeted advertisements and high premiums for health insurance. Therefore, data custodians face a dilemma between publishing significant data facts and protecting patients' privacy. On the one hand, it is necessary to ensure that the charts do not reveal private information about individuals. On the other hand, privacy-preserving charts should be able to preserve important data patterns and communicate insights to data consumers.

2.2 Preliminary Privacy Preservation

We focus on the widely used *tabular data*, which stores data records with a set of attributes as rows in a multidimensional table. Privacy preservation for tabular data is mainly achieved by syntactic anonymity and DP. Syntactic anonymity (e.g., k-anonymity [64], l-diversity [44], and t-closeness [42]) constructs equivalence groups to prevent attackers from distinguishing individuals. Unfortunately, equivalence groups could be cracked by background knowledge [17].

Compared with syntactic anonymity approaches, DP is gradually being applied to more real-world scenarios because of its robust mathematical guarantees [26]. DP is defined based on two neighboring datasets D_1 and D_2, which differ by adding or removing a record. Given a user-defined privacy budget ε ($\varepsilon > 0$), a randomized algorithm G satisfies **ε-differential privacy** (ε-DP), if and only if the following equation holds for any possible output O:

$$\Pr(G(D_1) = O) \leq e^{\varepsilon} \cdot \Pr(G(D_2) = O). \quad (1)$$

Existing studies have proved that the Laplace and exponential mechanisms are feasible randomized algorithms. For a function f with numerical output, the **Laplace mechanism** [26] constructs a corresponding G_f by adding noise sampled from a Laplace distribution $Lap(\frac{\Delta f}{\varepsilon})$, where Δf denotes the l_1-sensitivity of f, that is, the maximum difference between the output of two neighboring datasets. The **exponential mechanism** [48] applies to function f with categorical output, which allows privately selecting the "best" element from a set. Assume that $q(D)$ is a designed score function that gives each discrete value a probability; the algorithm G_f provides ε-DP if it approximately maximizes the score by returning values from the discrete domain with the probability proportional to $exp(\frac{\varepsilon q(D)}{2\Delta q})$, where Δq is the l_1-sensitivity of q.

2.3 Utility Metrics of Private Visualization

The dilemma faced by data custodians in Sect. 2.1 motivated us to analyze the utility of privacy-preserving visualizations on data pattern maintenance, which is important for pattern-centric data analysis tasks [59], such as identifying anomalies. Seeking automatic utility evaluation, Bertini et al. [8] and Behrisch et al. [5] studied visual quality metrics and categorized them into image-related metrics, perception-related metrics, and data-related metrics.

Image and perception-related metrics mainly aim to evaluate the effectiveness of a visual representation. Treating visualizations as images, image-related metrics inspect the effectiveness through techniques from computer vision and digital image processing [49, 50]. Related metrics can detect distribution patterns [61] and evaluate visual clutters [22, 28]. Perception-related metrics simulate the results of perception experiments [16]. Those metrics can measure visualizations from the perspectives of memorability [10], aesthetics [33], and engagement [60]. Data-related metrics focus on the statistical characteristics of the data visualized in charts. Johansson et al. [38] propose user-defined quality metrics to reorder parallel coordinate plots. Scagnostics [72] describes various measures of interest for pairs of variables based on their appearance on a scatterplot. Since the proposed approach manipulates the data space with DP techniques, we evaluate the data pattern retention with data-related metrics.

3 Related Work

In this section, we summarize related work in privacy-preserving data visualization and interactive interfaces for privacy protection.

3.1 Privacy-Preserving Data Visualization

Privacy-preserving data visualization [9] aims to make visualizations publicly available without privacy violation. Two approaches are widely used: screen-space and data-space sanitization [23]. Screen-space sanitization leverages visual uncertainty such as a blur or overlap to protect privacy [15]. Novel visual designs [2,23] and evaluation metrics [12,21] have been proposed to prevent privacy breaches. However, there is currently no rigorous privacy definition available for screen-space sanitization [75]; thus, its usage is still disputable without extensive evaluation.

Data-space sanitization mainly uses privacy protection algorithms in the data-space and visualizes the resulting data. For example, Lin et al. [43] used k-anonymity and numeric variance algorithms to preserve privacy in taxpayers' profiles and transaction records. Chou et al. [14, 15] applied syntactic anonymity techniques on event sequences and then extended it to social networks. They reduced data-level privacy leaks by modifying graph data such that the nodes and edges satisfy k-anonymity and l-diversity [13]. Oksanen et al. [55] introduced the privacy-preserving heat map from mobile sports tracking application data to protect the location privacy of individuals.

A more robust and popular approach in data-space sanitization is to use DP because of its resistance to several privacy attacks, such as re-identification [35], reconstruction attacks [24], and differencing attacks [27]. In order to protect individual privacy in geolocation data visualization, Ren et al. [36] proposed IDP-kmeans, which introduces DP in the k-means algorithm to balance privacy disclosure risks and clustering usability. Zhang et al. [74] proposed a privacy-preserving heat map by discretizing the raw data into grid cells and using the Laplace mechanism to add noise to the data point counts in each cell. They later generalized a DP pipeline for generating privacy-preserving visualizations [75]. The pipeline includes a DP data publishing algorithm to add calibrated noise to sensitive data and render the privacy-preserving visualizations based on the privatized data. They proved that if the privatized data satisfies ε-DP, the generated visualization also satisfies ε-DP according to the DP's post-processing property [27].

The DP data publishing methods can be further divided into interactive and non-interactive [77]. The interactive methods release query answers, such as mean and quantile [25, 71], one by one on-demand. However, data custodians usually publish multiple charts when they have multidimensional data. The interactive methods become inefficient and are prone to over-noising if a variable appears in many charts. On the other hand, non-interactive approaches produce "sanitized" datasets to support subsequent operations [58, 73]. Data is generated by approximating the marginal distributions with probabilistic graphical models, such as the Bayesian network [41] and Markov random field [47]. Deep learning models, such as GAN and auto-encoder, have also been applied, but their current performance in capturing the correlations in tabular data is not satisfactory [65].

The proposed *PriVis* model belongs to the data-space sanitization approach, which provides a theoretical privacy protection guarantee by DP. We extend a Bayesian network-based DP technique, *PrivBayes* [76], to generate privatized data in one go. This non-interactive method generates multiple privacy-preserving charts efficiently. Unlike previous works, we also model user preferences for particular patterns as pattern constraints to guide the structure of Bayesian networks, which improves the utility of private visualizations.

3.2 Interactive Interfaces for Privacy Protection

Applying data privacy protection requires a specific understanding of the underlying protection schemes. Interactive techniques are thus developed to make algorithms transparent and understandable. Most of them also provide an interface to perceive better the trade-offs between privacy and utility [9]. Pioneer works generally use conventional anonymization methods. For example, Wang et al. [70] supported users in applying syntactic anonymity to tabular data and interactively evaluating each attribute's distribution loss. They introduced GraphProtector [69] for graph data, which let the data custodian set priorities and inclusion rules for different nodes. Privacy-preserving operations modify these nodes to maintain important structural features such as node degrees, hub fingerprints, and sub-graphs.

For DP-based interactive systems, a key feature is to interactively optimize the allocation of privacy budgets, usually by displaying the difference between non-private and private results. For example, PSI(Ψ) [29] shows the absolute and relative error of queries, such as mean and quantile, for the users' references to assign privacy budgets. Overlook [66] supports interactive parameter configurations and shows count queries with privacy guarantees. ViP [51] visualizes relationships between privacy budgets, accuracy, and disclosure risks with uncertainty visualization. These user-centric approaches provide more flexible and granular options for privacy protection, freeing users from tedious data inspection work and improving productivity. However, they need users to set budgets for each query. Users are responsible for carefully maintaining the global allocation scheme by analyzing and predicting the errors in individual query results; otherwise, the privacy budget would easily be exhausted. Our *DPVisCreator* uses a "sanitized" dataset with DP, allowing users to generate as many privacy-preserving visualizations as they need. *DPVisCreator* also supports users in expressing pattern preferences and visually comparing errors directly in the charts.

4 Requirements Analysis

This section summarizes the design requirements obtained through analyzing the example scenario and interviews with privacy practitioners and visualization researchers. Similar to the trade-offs made by data custodians, we formed the requirements into two perspectives: privacy protection and visualization utility.

From the **privacy protection** perspective, when data custodians generate visualizations from raw data, existing practice seldom considers privacy risks. Given the high expressiveness of visualizations, publishing them could lead to unwanted disclosure of sensitive information, similar to publishing the raw data itself.

R1 Recommend privacy protection schemes for publishing visualizations. Data custodians are aware of the importance of protecting privacy in data visualizations, as strict regulations bond them to keep valuable data safe. Unprotected charts are vulnerable to malicious attacks that could cause severe consequences. However, data custodians generally lack the skills and expertise to derive a persuasive privacy protection scheme, especially for visualizations. In addition, manual configurations from scratch are inefficient for periodic publishing demands. Automated recommendations for privacy protection schemes can significantly improve data custodians' workflow.

R2 Explain the privacy-preserving visualization generation process. Although many existing tools have provided theoretical guarantees to the protection schemes, privacy practitioners want more evidence to support the results and enhance their trust in the models. Explaining the privacy protection process helps data custodians understand how and to what extent privacy is protected. Therefore, the visualization generation process should be transparent to users and explained intuitively.

From the **visualization utility** perspective, data consumers expect to read data patterns and gain insights from the published visualizations. If privacy protection schemes severely distort such data patterns, they are detrimental to the purpose.

R3 Explore data patterns of interest. Multidimensional tabular data contains rich information under different attributes. Data custodians need to explore the publishing dataset from different aspects to identify sensitive attributes and representative data patterns. They evaluate the intrinsic values of data patterns and seek ways to sustain the important ones in the privacy protection process. We should support visual exploration because the final output is in a visual form.

R4 Examine different configurations to balance the privacy-utility trade-off. While the automatic generation process can yield privacy-preserving visualizations, the data pattern preference elicited by data custodians might not be sustained. Data custodians should review the protection results to check whether the current privacy budget constraints and pattern preferences produce satisfactory results. Then, they should adjust their preferences and optimize configurations to obtain acceptable results.

R5 **Compare different privacy protection schemes.** For different analysis purposes, data custodians may indicate different data pattern preferences. They want to compare different privacy protection schemes for an optimal solution to their goal. We should provide quantitative metrics to help them evaluate these schemes. Moreover, multiple charts might be produced under a single privacy protection scheme, and the desired data patterns might not be easily perceived among these charts. Therefore, we should also provide support for qualitative assessments of patterns.

5 Our Approach

In this section, we describe the workflow of our approach and details of the proposed model called *PriVis*.

5.1 Approach Overview

Our approach lets the user guide the automatic privacy preservation from the utility and privacy perspectives. As shown in Fig. 2, the user first elicits their requirements on the utility by creating a series of pattern constraints. Each **pattern constraint** consists of a group of data records and a weight of significance. After browsing the data overview, the user creates pattern constraints in two steps: 1) select a data group by marking the corresponding visual elements from a chart, and 2) set the corresponding weight. Based on these pattern constraints, *PriVis* recommends private visualizations to the user, which would then be evaluated to balance the trade-off between privacy and utility. If none of the recommendations are satisfactory, the user should specify other pattern constraints or set another privacy budget.

5.2 Our model

As mentioned in Sec. 2.1, significant patterns need to be preserved in private visualizations. Although existing approaches, like *PrivBayes* [76], can provide privacy preservation for tabular data, none of them consider patterns in visualizations. To address this issue, we propose *PriVis*, which extends *PrivBayes* to support pattern constraints. In this section, we first introduce *PrivBayes* and then describe *PriVis*. For clarity of description, we have listed the notations in Table 2.

5.2.1 PrivBayes

PrivBayes is a DP algorithm for publishing high-dimensional datasets. Due to the curse of dimensionality [40], processing high-dimensional data tables often creates uncontrollable noise, resulting in the degradation of usability. To solve this problem, *PrivBayes* first abstracts the high-dimensional dataset into low-dimensional marginal distributions, abbreviated as *marginals*. Sampling noise from the marginals, *PrivBayes* generates privacy-preserved high-dimensional data.

To calculate marginals, *PrivBayes* employs the Bayesian network, a probabilistic graphical model that summarizes tabular data via a directed acyclic graph (DAG). In a Bayesian network N, each node represents an attribute, and each directed edge describes the conditional probability between two attributes. A linear ordering of vertices $(X_1, X_2, \ldots, X_d)$ can be obtained by topological sorting. Formally, N approximates the probability distribution of high-dimensional data by a series of distributions of attribute-parent (AP) pairs (X_i, Π_i), having

$$\Pr_N(A) = \Pr(X_1, X_2, \ldots, X_d) = \prod_{i=1}^{d} \Pr(X_i \mid \Pi_i), \tag{2}$$

where X_i represents the i-th attribute in A and Π_i is the set of parents of node X_i in network N.

The approximation degree of N to the original data distribution $Pr[A]$ can be measured by the *KL divergence*, which is defined as

$$D_{KL}(\Pr(A) \| \Pr_N(A)) = -\sum_{i=1}^{d} I(X_i, \Pi_i) + \sum_{i=1}^{d} H(X_i) - H(A), \tag{3}$$

where $H(X)$ represents the entropy of the attribute X over its domain, and $I(\cdot,\cdot)$ represents the mutual information between the two variables. When the dataset D is given, $\sum_{i=1}^{d} H(X_i) - H(A)$ is also fixed. The construction of the Bayesian network is transformed into an optimization problem that maximizes the mutual information between all AP pairs.

To protect privacy, *PrivBayes* introduces the exponential mechanism to select AP pairs by using mutual information as a score function and generates marginals with Laplace noise. The released dataset can then be synthesized by sampling from the marginals. According to the post-processing property [27] of DP, *PrivBayes* has a privacy guarantee.

Table 2. Notation definitions used in this paper.

Notation	Description
D	A multidimensional table including sensitive information
n	The number of records in D
A	The set of attributes in D
d	The number of attributes in A
N	A Bayesian network over A
X	The nodes in Baysian network
Π	The set of parents of nodes in Bayesian network
$Pr(A)$	The distribution of records in D
$Pr_N(A)$	An approximation of $Pr(A)$ defined by N
$dom(X)$	The domain of attribute X
$\mathbb{1}_{P_k}$	Indicator function for the k-th pattern
$w(k)$	The importance weight for the k-th pattern
$MW(r)$	The mixture weight value of a record r

5.2.2 PriVis

PriVis generates privacy-preserving visualizations in four steps.

Step 1: Understand pattern constraints. Users may select multiple patterns from different charts, represented as $P = \{P_1, P_2, \cdots, P_s\}$. Each pattern P_i corresponds to a subset of data records in the whole data D. For example, as mentioned in Sec. 2.1, the data custodian needs data records with both age and illness attributes from the hospital data to generate a bar chart. To emphasize the illness risk in people over 60, the data custodian can specify the pattern by selecting bars corresponding to patients over 60 and setting a weight. User preferences for different patterns can be recorded as different subsets of data records with corresponding weight $W = \{w_1, w_2, \cdots, w_s\}$. For a data record r, its final weight needs to consider its occurrence under each pattern. To this end, we give each record an initial weight of 1 and then add additional weights of w_k for each record r in pattern P_k. The mixture weight $MW(r)$ of data record r is defined as:

$$MW(r) = \sum_{k=1}^{s} w_k \mathbb{1}_{P_k}(r) + 1, \tag{4}$$

where $\mathbb{1}_{P_k}$ is indicator function which maps record in pattern P_k to 1, and all others to 0. Therefore, we translate the different pattern constraints into a mixture weight assigned to each record.

Step 2: Construct a Bayesian network with pattern constraints. To preserve user-specified patterns, *PriVis* emphasizes the correlations between the corresponding records when constructing the Bayesian network. Specifically, when selecting AP pairs, *PriVis* not only employs the exponential mechanism but also replaces the mutual information in the PrivBayes network with the weighted mutual information:

$$I_w(X, \Pi, MW) = \sum_{x \in \mathrm{dom}(X)} \sum_{\pi \in \mathrm{dom}(\Pi)} \frac{\sum_{r \in x \cap \pi} MW(r)}{|x \cap \pi|} \Pr(X = x, \Pi = \pi) \log \frac{\Pr(X = x, \Pi = \pi)}{\Pr(X = x)\Pr(\Pi = \pi)}. \tag{5}$$

With mixture weights, the constructed Bayesian network can better preserve dependence among user-selected data. Therefore, the patterns consisting of these records can be better maintained in visualizations.

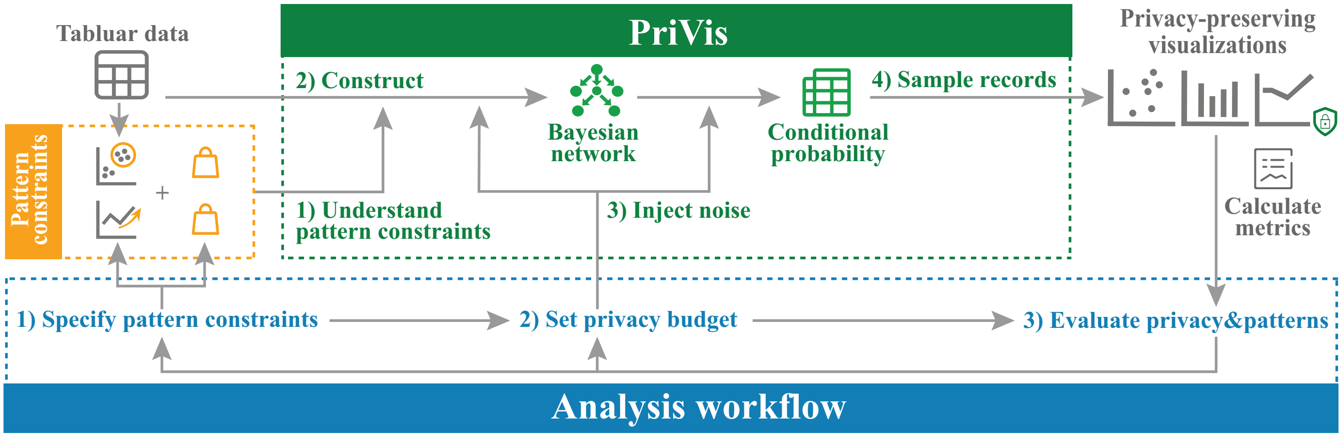


Fig. 2. The overview of our approach. Users can select patterns of interest from visualizations of tabular data and specify importance weights. The *PriVis* model considers these pattern constraints and privacy budgets to generate privacy-preserving schemes. The user can evaluate the results based on the visualization charts and utility metrics and return for iterative adjustments.

Step 3: Inject noise to the Bayesian network. The constructed Bayesian network provides conditional probability distributions $Pr(X_i|\Pi_i)(i \in [1,d])$ to approximate the distribution of the high-dimensional dataset. To inject noise, *PriVis* first calculates noisy distribution $Pr^*(X_i,\Pi_i)$ for any $i \in [k+1,d]$ by adding Laplace noise to joint probability distribution $Pr(X_i,\Pi_i)$. Then, the conditional distribution $Pr^*(X_i|\Pi_i)$ can be derived. The $Pr^*(X_i|\Pi_i)(i \in [1,k])$ can thus be directly obtained from $Pr^*(X_{k+1}|\Pi_{k+1})$.

Step 4: Sample records to generate private datasets. According to the topological order of the Bayesian network, $X_i(i \in [1,d])$ can be sampled in increasing order of i based on the noisy conditional distribution. For example, when all parent attributes of $X_j(j \in [2,d])$ are sampled, X_j can be sampled by conditional probability $Pr^*(X_j|\Pi_j)$. The sampled data can then be used for publishing visualizations.

Privacy Analysis. In our approach, the data custodian is located in a trust zone to elicit pattern preferences before the model performs privacy learning. The model is guided towards the right direction without additional privacy budgets. In the *PriVis* model, access to the original data is required for steps 2 and 3. In step 2, the model uses the exponential mechanism to privately select AP pairs, which converts selection from multiple candidate pairs into probabilistic sampling, which consumes an ε_1 privacy budget. In step 3, Laplace noise is added to the conditional probability distributions, which consumes an ε_2 privacy budget. According to the composition property [27] of DP, the resulting visualizations satisfy the ε-DP, where $\varepsilon = \varepsilon_1 + \varepsilon_2$.

6 System Design

In this section, we present a system overview and then introduce the details of the visual design and interaction of *DPVisCreator* [1]

6.1 System Overview

To meet the requirements in Sec. 4, we designed *DPVisCreator* with four main views, as shown in Fig. 3: the *Data View*, the *Pattern View*, the *Model View*, and the *Solution View*. We describe an analysis flow to illustrate how the four views help data custodians generate privacy-preserving visualizations with important data patterns. A data custodian first loads a dataset and looks for sensitive attributes in the *Data View* (Fig. 3-A), which shows the general data characteristics. He selects attributes of interest and specifies the desired chart settings in the *Pattern View* (Fig. 3-B). He can also highlight the data patterns of interest with the data selection tools (**R3**). Then, these preferences are processed by *PriVis* to propose a privacy protection scheme (**R1**).

The recommended scheme is visualized in the *Model View* (Fig. 3-C), with its structure and impact explained (**R2**). He can examine the relationships between different pattern constraints and adjust the importance weight for different results (**R4**). He turns to the *Solution View* (Fig. 3-D) to evaluate the privacy protection scheme in terms of privacy and utility (**R4**). All generated schemes are recorded in this view for comparison (**R5**). He can revisit the *Model View* to generate more candidates and strike a desirable balance between privacy and utility (**R4**). Iterative optimizations can find a privacy-preserving scheme that maintains important patterns with an acceptable privacy budget.

6.2 Data View

The *Data View* (Fig. 3-A) provides an overview of the loaded dataset by showing its attributes and the corresponding data distribution. It supports data range filtering and allows users to choose attributes of interest for subsequent analysis (**R3**).

The view allows the user to upload tabular data with the number of records and attributes displayed below. Each attribute can expand to show its data distribution and be selected accordingly. Selected attributes are colored in blue, and unselected ones in gray. For categorical attributes, their data records are aggregated and displayed in bar charts. Every unique data value can be selected by checking the selection box underneath. For numerical attributes, their probability density distributions are first calculated using kernel density estimation and then presented using a curve plot. A slider indicates the data value range and is used to filter the range of interest. After finishing the data selection process, users can click on the confirm button to proceed.

6.3 Pattern View

The *Pattern View* (Fig. 3-B) consists of two parts: (1) The *chart settings panel* lets users configure visualization charts; (2) The *pattern selection view* supports the interactive selection of data patterns (**R3**).

The *chart settings panel* (Fig. 3-B1) lets users select the chart type, color, and attributes for the x- and y-axis of the charts. Three common charts (i.e., scatter, line, and bar charts) are supported. Users can quickly specify chart types by mapping selected attributes to appropriate visual encodings via drop-down boxes. In order to reduce the visual clutter for large amounts of data, the view supports step adjustment for the x-axis and aggregation calculation for the y-axis.

The *pattern selection view* (Fig. 3-B2) contains an information card, a visualization chart, and a pattern list. Users can select the data patterns of interest from the visualization chart, which is rendered according to the above configuration. Due to the different shapes of the visualized data patterns, we provide a set of interactive selection methods. For cluster patterns in scatter plots, use the box and lasso to select regions of interest. For correlation patterns in line charts, horizontal selection indicates intervals that match a specific trend. For order patterns in bar charts, the bar being clicked is chosen. Gray backgrounds are overlaid on all selected areas. The information card shows the statistics of

[1] https://github.com/wanghuanliang/DPVisCreator

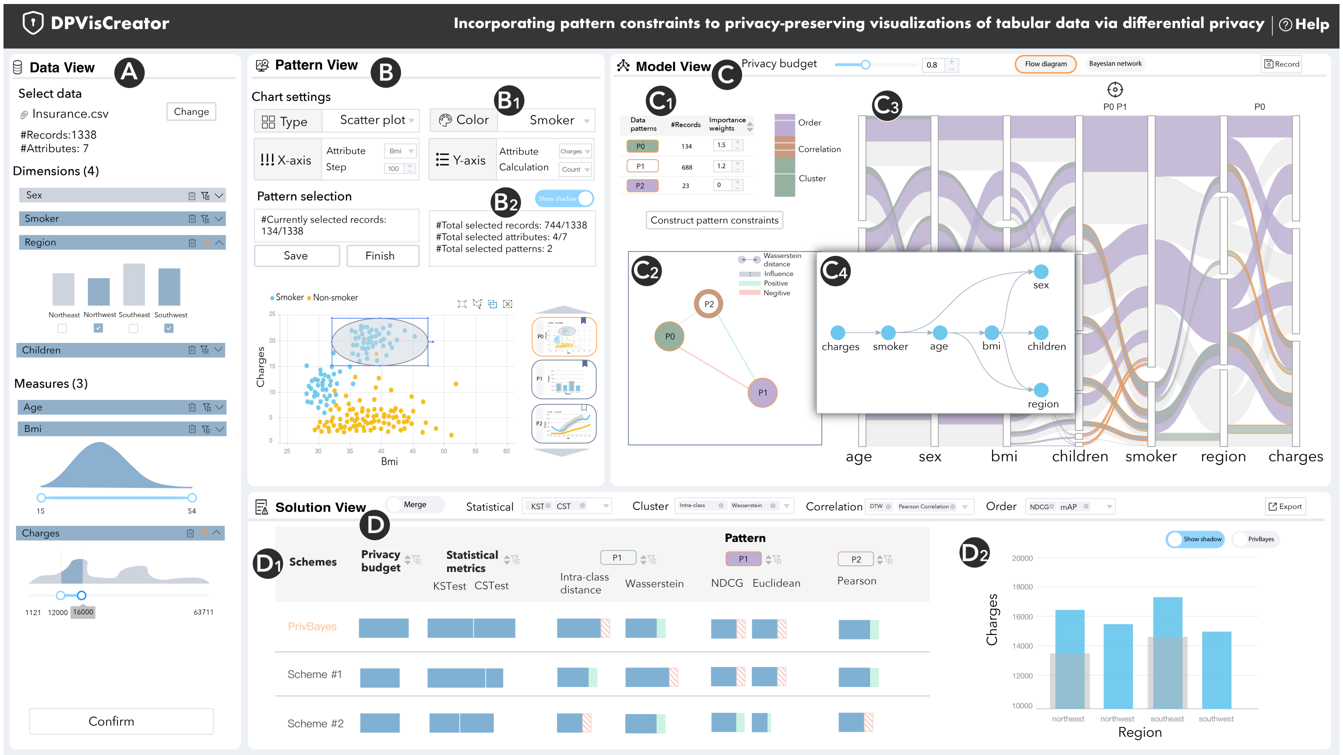


Fig. 3. *DPVisCreator* facilitates publishing privacy-preserving visualizations. (A) The *Data View* displays the distribution of attributes. (B) The *Pattern View* supports the pattern specification. (C) The *Model View* provides the relationship between data constraints and the underlying Bayesian network structure. (D) The *Solution View* assists comparison and assessment of privacy-preserving schemes.

current and all patterns, and the pattern list records all selected patterns to track analytic provenance. Users can add the current pattern to the list and revisit previously saved patterns by clicking it.

6.4 Model View

The *Model View* contextualizes the privacy-preserving process to foster users' understanding and supports users in making trade-offs between data pattern maintenance and privacy protection. The *Model View* (Fig. 3-C) contains four parts: (1) The *weight configuration panel* lets users adjust importance weights to each pattern **(R4)**; (2) The *pattern constraint relationship* represents the overall impact between different pattern constraints **(R4)**; (3) The *pattern constraint flow* shows the detailed pattern constraint distribution on each attribute and highlight them for comparisons **(R4)**; (4) The *Bayesian network view* visualizes the structure in the underlying *PriVis* model **(R2)**. The latter two views can be switched with a button.

The *weight configuration panel* (Fig. 3-C1) uses a stacked bar chart to provide an overview of the importance weight distribution. Colors encode the pattern constraint's type, having green for the cluster, brown for correlation, and purple for order. This color encoding is applied throughout the system. The corresponding pattern ID involving data amount and importance weights are displayed on the data pattern list and hovers on the corresponding bar. Users can adjust the corresponding pattern constraint weights in the data pattern list.

The *pattern constraint relationship* (Fig. 3-C2) encodes pattern constraints as circles. The position of a circle is determined by multidimensional scaling (MDS) [18], where the Wasserstein distance [68] is used to measure the similarity between patterns. The projection reflects the differences in distributions. Edges between circles encode the impact of different pattern constraints by affecting the network structure. Their colors encode the type of influence of changing weights. Green edges imply positive correlations, where increasing one pattern's weight will promote the other. In contrast, red represents negative correlations that the other pattern will be weakened. The magnitude of influence is encoded by the thickness of the edge, which is calculated by the sum of the difference between the intersection and the symmetric difference of AP pairs. Clicking a pattern constraint will highlight it and its adjacent edges, together with the corresponding entities in other views.

The *pattern constraint flow* (Fig. 3-C3) improves the basic Sankey diagram with sub-flows representing constraints in each attribute. The x-axis lists all attributes, and the y-axis indicates the discrete interval of each attribute. Continuous attributes are discretized by k-means with the elbow method [46], while categorical attributes use their discrete domain as the interval. The length along the y-axis represents the amount of data belonging to this interval. The specific interval information can be viewed by hovering. Different flows represent different data distributions, and the width of a flow is proportional to its data volume. By default, all data flows are grayed out to provide a clear background, while the selected data pattern will be highlighted correspondingly for a more intuitive comparison.

The data flow between attributes follows three visual patterns: focus, convergence, and divergence. Focus means that data flowing from a single interval mostly goes to a specific interval of another attribute; convergence means that data flowing from multiple intervals goes to the same interval; divergence means data flowing from a single interval goes to multiple intervals. This difference in data distribution reflects, to some extent, the impact of different data patterns. Suppose that two data patterns have the same attribute pairs of the focus relationship. In this case, they have a high probability of choosing the same AP pairs when constructing the Bayesian network; that is, the network can be a good approximation of the data distribution of both of them. If two data patterns have a significant difference in the attribute pairs of the focus relationship, increasing the weight of one pattern will likely reduce the Bayesian network's approximation effect on the other pattern.

The *Bayesian network view* (Fig. 3-C4) shows the structure of the *PriVis* model. Since the network is a DAG, we use a hierarchical layout [4] to show the dependencies of attributes. Specifically, the depth of each node is obtained by topological sorting. We draw each layer in turn along the x-axis, and nodes in the same layer are evenly distributed along the y-axis. Nodes represent attributes, and edges reflect conditional dependencies. The probability density distributions before and after noise addition are displayed by superposition when hovering over a node, making it easy to compare and understand the exact degree of privacy protection.

6.5 Solution View

The *Solution View* (Fig. 3-D) measures privacy-preserving schemes from different perspectives **(R4)** and shows comparisons for selecting an appropriate scheme **(R5)**. It contains two parts: (1) The *schemes ranking list* displays all privacy-preserving schemes. Inspired by lineup [31], the schemes' metrics are organized into a multidimensional table. (2) The *pattern comparison view* uses the superposition technique of visual comparisons [30] to highlight the scheme's effect.

The *schemes ranking list* (Fig. 3-D1) shows evaluation metrics for each privacy protection scheme in detail, including privacy budgets, statistical indicators, and pattern retention metrics. KSTest [7] and CSTest [32] analyze the differences in statistical properties of data. As to utility, suitable metrics are adopted for different patterns, such as Wasserstein distance for the cluster, dynamic time wrapping (DTW) for correlation, and Euclidean distance for order. These metrics reflect the degree of pattern retention. Users can choose different metrics to measure the scheme's utility according to their needs. The length of the horizontal bar encodes the metrics' value. We explicitly encode the differences in the metrics before and after privacy preservation for comparing pattern-related metrics. The increased proportion is represented by green bars, while the decreased proportion is represented by red striped bars. Users can rank schemes based on metrics with sorting and filtering functions.

The *pattern comparison view* (Fig. 3-D2) shows the visualization chart after privacy protection. The gray border represents the previous selection area, and the gray visual elements represent the original data. This visual comparison helps users qualitatively perceive the difference in visualization utility before and after. Users can also switch off the pattern constraints, i.e., to compare the chart with the baseline.

7 Evaluation

We implemented a quantitative experiment to assess *PriVis* and described two case studies to verify the effectiveness of *DPVisCreator*. We finally report subjective feedback gathered from domain experts.

7.1 Quantitative Evaluation

Experiment settings. The purpose of this experiment is to test the ability of *PriVis* to maintain preferred data patterns in visualization while preserving privacy. We chose *PrivBayes* [76] as the baseline method, which has been shown to perform well on tasks such as statistical queries and classification [65]. We compare *PriVis* and *PrivBayes* on three types of data patterns separately under various privacy budgets and also evaluate the effect of *PriVis* on multiple patterns under different weight configurations. The same condition has 25 replicate runs to mitigate the effects of randomness introduced by noise. The detailed experimental information can be found in the `https://osf.io/ugw29/`.

Datasets We tested these two methods on five datasets with different distribution characteristics, numbers of data records, and attribute types. *Shopping* [2], *Adult* [3] and *Bank* [4] datasets are from the UCI machine learning repository, and *Insurance* [5] and *Loan* [6] from Kaggle.

Evaluation metrics. We chose the corresponding metrics for different data patterns. For cluster, all data before and after privacy protection are used to calculate the metrics since the selection area may be inaccurate and the cluster position could move out of the selected region due to noise. For correlation and order, data in the user-selected range are used. The specific metrics are as follows:

- **Cluster.** The Wasserstein distance [68] is used to measure the distribution distance of data points in scatter plots.
- **Correlation.** The difference of Pearson correlation coefficients [6] is used to measure the change in trend, and DTW [1] is applied to measure the distance between the selected lines.
- **Order.** The NDCG [11] is used to reflect the correctness of ranking, while the amount of variation between selected bars is calculated by the Euclidean distance.

Results. We conducted ANOVA and posthoc tests with Bonferroni correction to assess the effectiveness of the different methods for pattern maintenance. As shown in Fig. 4, for three types of patterns, different methods perform similarly when the degree of privacy protection is strict ($\varepsilon < 0.5$), mainly due to a large amount of noise that severely obscures the original data. When the protection level is moderate ($\varepsilon \in [0.5, 5]$), *PriVis* outperforms *PrivBayes*, and the pattern retention effect can be improved by increasing the weights, which shows the improvement of the visualization utility by optimizing the bayesian network structure. When the privacy need is relaxed ($\varepsilon > 5$), noise interference is less, and the difference in the performance of methods gradually narrows. The maintenance effect of cluster is more significant, while correlation and order are mainly reflected in the reduction for distance deviation. As shown in Table 3, we compare the performance of *PriVis* and *PrivBayes* in generating three charts with different patterns on the adult dataset. The results show that the maintenance of the different data patterns can be improved by adjusting the corresponding weight configuration of *PriVis*.

Table 3. The mean of pattern metrics for different schemes. Metrics are reported for each pattern in the order mentioned above. ($\varepsilon = 2$)

Schemes	Cluster	Correlation	Order
PrivBayes	264.05	1.61 281.20	0.86 531.40
PriVis W=(1, 1, 4)	277.92	1.49 279.20	**0.96** **510.80**
PriVis W=(1, 9, 4)	287.08	**1.13** **252.80**	0.88 514.20
PriVis W=(9, 4, 1)	**256.85**	1.27 287.40	0.88 517.00

Limitations. Although these results demonstrate the advantages of *PriVis* in pattern retention, there exist some factors that can affect the validity, such as the data selection and weights settings. Since *PriVis* improves the pattern utility by influencing the network structure, it is difficult to change the direction of structure learning when the amount of selected data is small and the weights are low. Also, when the distribution of the selected data and the whole data are similar, the current network structure may already be optimal and cannot be further improved. The results of *PriVis* and *PrivBayes* are comparable in the above conditions. In particular, since DP introduces random noise, *PriVis* may be less effective when epsilon is small.

7.2 Case Studies

We introduce two case studies to show the privacy-preserving visualizations produced by *DPVisCreator*.

7.2.1 Health Insurance Dataset

As shown in Fig. 3, the dataset contains attributes for 1,338 U.S. insured persons, including age, sex, BMI, number of children, smoker, region, and charges. The data custodian wanted to analyze the impact of different attributes on premium charges, so he uploaded the table into the *Data View*. The attribute distribution revealed that the costs mainly concentrate around 15k and 40k. To investigate the reasons for this difference, he analyzed the relationship between charges and other attributes in charts of the *Pattern View*. He quickly generated a scatter plot of BMI and charges by selecting the x- and y-axis attributes via

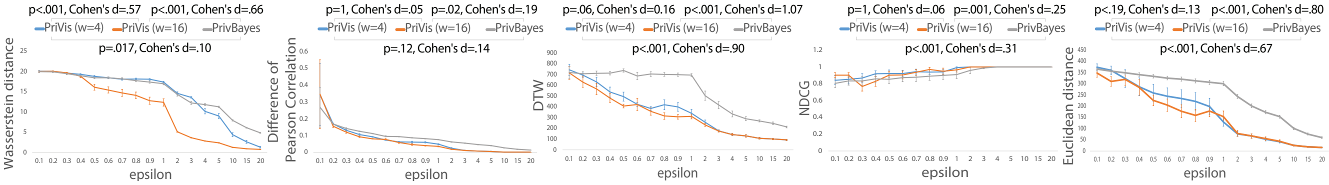
[2] `https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset`

[3] `https://archive.ics.uci.edu/ml/datasets/adult`

[4] `https://archive.ics.uci.edu/ml/datasets/bank+marketing`

[5] `https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset`

[6] `https://www.kaggle.com/datasets/burak3ergun/loan-data-set`



(a) Cluster (Wasserstein) (b) Correlation (Pearson) (c) Correlation (DTW) (d) Order (NDCG) (e) Order (Euclidean)

Fig. 4. The comparison of *PriVis* (weight=4/16) and PrivBayes [76] in maintaining three data patterns under various epsilon conditions on the adult dataset. Except for NDCG, the smaller the value, the better. The p-values and effect sizes are labeled next to the legend. Error bars represent 95% confidence intervals.

the drop-down boxes. He found that points with BMI between 30 and 45 and charges between 30,000 and 50,000 formed a cluster. After color encoding the attribute of smoker, he found all the people in the cluster were smokers, reflecting those obese people who smoked had high charges. It was an important finding, so he saved it as data pattern P0. He then intended to see if there was a geographical difference in charges by generating a bar chart of regions and charges. He found that the southeast and northeast regions ranked first and second in average charges and were significantly higher than other regions. This data pattern P1 also reflected the differences in medical conditions across regions and was therefore saved.

If original charts were published directly, an attacker could directly read sensitive information from charts or infer the privacy of the insured through background knowledge. Therefore, he decided to treat them with privacy protection. In the *Model View*, he configures the same importance weight of 0.5 for both patterns to show his preference. Then, he clicks the record button to have the system generate a privacy protection scheme#1 accordingly. By looking at the *Solution View*, he found that the order of the selected bar was well maintained, but the cluster in the scatter plot was not concentrated. So he returned to the *Model View* to make adjustments. The *pattern constraints relationship* shows a negative influence edge between P0 and P1, and the *pattern constraints flow* indicates that their data flows co-focus on a small percentage of attributes, which prevents the network structure from being able to approximate the distribution of the two data patterns simultaneously. To ensure the effectiveness of the cluster pattern, he adjusts the weights of cluster and order to 1 and 0.1, respectively, thus obtaining a new scheme#2. As shown in Fig. 1, the effect of the cluster had improved, and the order had not been affected much, which was a satisfactory result. To further confirm, he looked at the network structure of the *PriVis* model for scheme#1 and scheme#2 and found that its structure had changed, which in turn led to the change. The quantitative metrics also reflected the utility of these visualization charts, which he was satisfied with and finally exported.

7.2.2 Loan Dataset

As shown in Fig. 5, the dataset contains detailed information about customers applying for loans, including gender, marital status, education, etc. The data custodian wanted to analyze which factors could help identify customer segmentation and produce reports to aid loan qualification approval. In order to know the number of loan origination for different loan amounts, a line chart was generated. He found that the loan amount around 100 have the highest origination number, but the number declined rapidly with the loan amount increased, as shown in Fig. 5-(d). He saved the pattern as P0. He further analyzed the income of co-applicants whose loan amount ranged from 100-200. By quickly generating a scatter plot and color encoding education, he found that co-applicants who have not graduated formed a cluster. He recorded this pattern of the stratification of the loan population as P1. Similarly, he reviewed a bar chart showing the number of dependents and loan amount and saved another data pattern P2 which indicates that the fewer relatives, the lower the loan amount.

Using the three pattern constraints, the data custodian proceeded to privacy protection processing before publishing the visualizations. The *pattern constraints relationship* revealed that P0 had a positive influence on both P1 and P2, so he increased the weight of P0 from 0.5 to 1. The results revealed that the adjustment led to an improvement in the metrics for all three patterns. The underlying Bayesian network (Fig. 5-(a-c)) showed that the network structure changed as the weights increased, leading to an improvement of correlation in the private line chart. This case demonstrated that *DPVisCreator* could help users adjust and optimize privacy protection schemes.

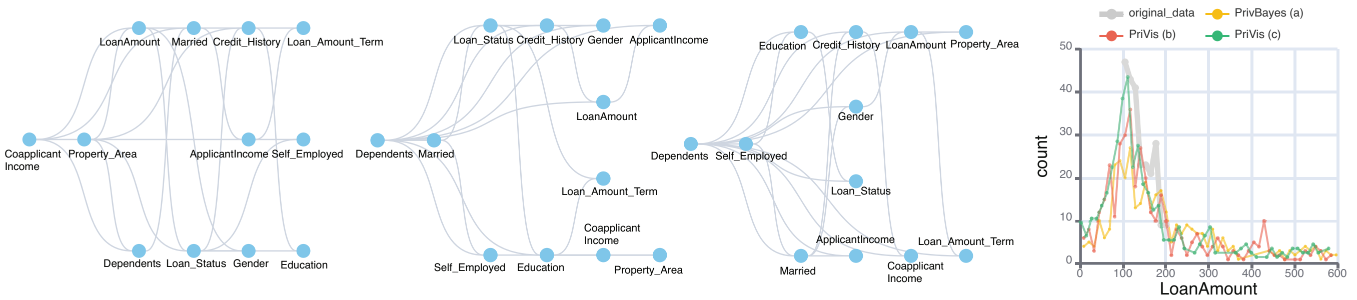
7.3 Expert Reviews

DPVisCreator was reviewed by four domain experts (E_A-E_D) who were potential target users with basic knowledge in DP and data visualization.

Procedure. For each interview, we first introduced our project background and requirements (10 minutes). Then the analysis process of the system was demonstrated through case studies (20 minutes), followed by free exploration in a think-aloud manner (15 minutes). Finally, we collected the feedback (15 minutes) summarized below.

Visualization and interaction. All experts agreed that *DPVisCreator* meets analysis requirements. We have observed that they might not immediately remember all the visual encodings and interactions when first encountering the system. However, after a demo explanation, they were able to understand. E_A appreciated that the multiple selection options (click, box, and lasso) in the *Pattern View* gave him the freedom to express a preference on specific data. At the same time, he admitted that this selection might also have errors and needs to be combined with the filtering function in the *Data View* to select the important pattern more accurately. For the *Model View*, E_A and E_C commented that *"the pattern flow visually show the 'flow' of different patterns across attributes, making it easy to see differences or consistency."* E_B adds, "Stacked bar chart provide a nice overview to help adjust the weights interactively, compensating for the current one-size-fits-all privacy protection approach". Regarding the *Solution View*, there are different comments. E_A believed that *"Quantitative metrics-based schemes ranking list is consistent with most existing data-centric privacy processes, which is more familiar for me to use."* E_D emphasized that *"overlaying data before and after protection with the user-selected area provides a visual reference for comparing pattern retention...It might be more intuitive to compare charts of different schemes with a small multiple techniques."*

System usability. All experts agreed that the overall analysis flow of the system was clear. The combination of automated model and interactive visualizations helps them publish privacy-preserving visualizations more efficiently and make a trade-off between data pattern maintenance and privacy protection. E_B mentioned that *"previous approaches to privacy protection required writing a lot of code and manually checking the results, which are time consuming and laborious. This user-centric approach to privacy was much more controllable, allowing users to compare and adjust privacy protection schemes interactively."* E_C said, *"If I need to visualize some sensitive data for the public, I would be interested in using this system for processing because it makes the complex DP model more transparent and intuitive"*. Both E_A and E_B pointed out that adjusting both importance weights and privacy budget parameters required some expertise. If the parameters can be set automatically, it will help to simplify the process. They also admit that in current practice, finding an optimal privacy protection scheme was inherently tricky, and joining the human involvement would help to find an acceptable scheme more efficiently.



(a) PrivBayes (b) PriVis (w_correlation=0.5) (c) PriVis (w_correlation=1) (d) Line chart with a correlation pattern

Fig. 5. The development of the Bayesian network in Case 2. (a-c) The Bayesian network structures of different schemes. (d) The correlation maintenance is getting better with the corresponding weight increase.

Suggestions. Experts also provided valuable suggestions based on our research. E_D suggested that data tables before and after privacy protection could be displayed to provide more detailed information. E_A and E_C, on the other hand, noted that when the system handles more than 15 attributes, it does not give timely interactive feedback. Hence the computational efficiency needs to be improved. There were also suggestions to include domain-specific evaluation metrics. *"The definition of utility is task-specific. The current evaluation metrics are the basis for many analysis tasks, but more users could benefit from support for customizable and more complex utility evaluations. For example, metrics for periodic patterns are important when publishing time-series visualization charts of shopping behavior,"* said E_C.

8 Discussion

In this section, we summarize the implications of our work and lessons learned during our interdisciplinary collaboration with data privacy experts. We also discuss the generalizations, limitations, and future work of *PriVis* and *DPVisCreator*.

Implications. To the best of our knowledge, our work is the first to extend DP to multiple visualization chart publishing. Previous work [75] has offered preliminary suggestions for preserving salient visual patterns in private visualizations. We took a step forward in proposing a Bayesian network-based DP data publishing method, taking into account user preferences for data patterns in different charts. Our approach can help data custodians make trade-offs on privacy protection and visualization utility in privacy-critical areas such as healthcare and finance. We hope our work will move privacy-preserving visualization research forward and inspire broader cooperation between data privacy experts and visualization researchers.

Lessons learned. We have gained valuable experience from our interdisciplinary research in visualization and privacy. First, choosing an effective solution requires knowledge and practical experience in both domains. We initially sought DP algorithms for basic query operations, such as calculating averages and sum. However, the visualization generation process involves many interrelated data transformations, such as binning, aggregating, and sorting, which either do not have corresponding privacy-protected versions or only have theoretical derivations awaiting further experiments to illustrate their efficacy. As such, we chose a more general DP data generation approach to obtain private visualizations.

Second, incorporating the user's prior knowledge into the privacy-preserving process can lead to more refined and customized results. The traditional automatic data-centric methods do not consider the different emphasis of users on data. In contrast, the human-centric approach allows users to analyze and compare the impact of privacy protection on charts through a visualization interface, which facilitates users to interactively guide the model until their needs are met, thus improving the utility of the results.

Generalizability. The generalizability of our approach is twofold. Firstly, *PriVis* is designed for tabular data. Since we use pattern constraints and original data to generate Bayesian networks, privacy-preserving charts based on tables can be generated as long as data corresponding to the chosen patterns is available. However, applying it to other data types such as graphs and geospatial data requires additional effort. These data have a particular concept of DP [39, 52]. For example, graphs need to consider information that may be leaked by sub-structures such as nodes and edges, which is beyond the scope of this paper. Secondly, for different utility metrics, we realize that the definition of utility changes with different tasks. As *DPVisCreator* is developed in a modular way, users can replace the evaluation metrics according to their needs to adapt to more diverse working scenarios.

Limitations and future work. Two limitations are observed in this study. First, an important concern is scalability. After the preliminary test, the computation time of *PriVis* is around 2 minutes with 1,000 data records and 15 dimensions, which means that it may be difficult for users to get timely feedback. Extending privacy protection to big data is still a challenging problem [37]. In the future, we plan to use parallelization [45] or progressive visualization [63] to accelerate the learning of Bayesian networks and visualization rendering.

Second, our approach may be too complicated to understand. Although we have some prerequisites for data custodians using *DPVisCreator*, it is nontrivial for them to understand DP, the Bayesian model, and visualization design. At the same time, the current system lacks some guidance in parameter settings, which relies on the data custodians' domain knowledge. We have also collected feedback on the desire to add an automatic recommendation for privacy protection schemes. However, since obtaining an optimal privacy protection solution is an NP-hard problem [3], we currently use a human-in-the-loop method to give users the initiative to generate solutions iteratively. In the future, we plan to add guides and annotations to the system, collect the needs of more general users and study the relationship between descriptive privacy-utility requirements and parameter settings.

9 Conclusion

This paper presents a novel visual analytics approach that facilitates the data custodian to publish multiple privacy-preserving visualizations and make tradeoffs between data pattern maintenance and privacy protection. We propose *PriVis*, an extended Bayesian network-based DP model, and introduce pattern constraints to model user preferences for different patterns, which are then incorporated into the privacy-preserving process. We develop a visual analytics system, *DPVisCreator*, with multiple interactive and coordinated views, which assists the data custodian in specifying pattern constraints, understanding the privacy model, and evaluating privacy protection schemes. The effectiveness of our approach is demonstrated through quantitative evaluations, case studies, and expert reviews.

Acknowledgments

We would like to thank all the reviewers for their constructive comments. This work was supported by the National Natural Science Foundation of China (62132017 and 62106218) and Zhejiang Lab (2022NF0AC01).

References

[1] *Dynamic Time Warping*, pp. 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. doi: 10.1007/978-3-540-74048-3_4.

[2] N. Andrienko, G. Andrienko, G. Fuchs, and P. Jankowski. Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*, 15(2):117–153, 2016. doi: 10.1177/1473871615581216.

[3] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure limitation of sensitive rules. In *Workshop on Knowledge and Data Engineering Exchange*, pp. 45–52, 1999. doi: 10.1109/KDEX.1999.836532.

[4] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[5] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, et al. Quality metrics for information visualization. In *Computer Graphics Forum*, vol. 37, pp. 625–662, 2018. doi: 10.1111/cgf.13446.

[6] J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. 2009. doi: 10.1007/978-3-642-00296-0_5.

[7] V. W. Berger and Y. Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014. doi: 10.1002/9781118445112.stat06558.

[8] E. Bertini and G. Santucci. Visual quality metrics. In *Proceedings of the AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–5, 2006. doi: 10.1145/1168149.1168159.

[9] K. Bhattacharjee, M. Chen, and A. Dasgupta. Privacy-preserving data visualization: reflections on the state of the art and research opportunities. In *Computer Graphics Forum*, vol. 39, pp. 675–692, 2020. doi: doi.org/10.1111/cgf.14032.

[10] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, 2015. doi: 10.1109/TVCG.2015.2467732.

[11] R. Busa-Fekete, G. Szarvas, T. Elteto, and B. Kégl. An apple-to-apple comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain. In *European Conference on Artificial Intelligence: Preference Learning: Problems and Applications in AI Workshop*, vol. 242, 2012.

[12] J.-K. Chou, C. Bryan, J. Li, and K.-L. Ma. An empirical study on perceptually masking privacy in graph visualizations. In *Symposium on Visualization for Cyber Security*, pp. 1–8, 2018. doi: 10.1109/VIZSEC.2018.8709181.

[13] J.-K. Chou, C. Bryan, and K.-L. Ma. Privacy preserving visualization for social network data with ontology information. In *IEEE Pacific Visualization Symposium*, pp. 11–20, 2017. doi: 10.1109/PACIFICVIS.2017.8031573.

[14] J.-K. Chou, Y. Wang, and K.-L. Ma. Privacy preserving event sequence data visualization using a sankey diagram-like representation. In *SIGGRAPH ASIA Symposium on Visualization*, pp. 1–8, 2016. doi: 10.1145/3002151.3002153.

[15] J.-K. Chou, Y. Wang, and K.-L. Ma. Privacy preserving visualization: A study on event sequence data. In *Computer Graphics Forum*, vol. 38, pp. 340–355, 2019. doi: 10.1111/cgf.13535.

[16] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984. doi: 10.2307/2288400.

[17] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. In *IEEE International Conference on Data Engineering Workshops*, pp. 88–93, 2013. doi: 10.1109/ICDEW.2013.6547433.

[18] M. A. Cox and T. F. Cox. Multidimensional scaling. In *Handbook of data visualization*, pp. 315–347. 2008. doi: 10.1016/B978-012099975-0.50005-1.

[19] D. Craig, S. Ketterer, and M. Yousuf. To post or not to post: Online discussion of gun permit mapping and the development of ethical standards in data journalism. *Journalism & Mass Communication Quarterly*, 94(1):168–188, 2017. doi: 10.1177/1077699016684796.

[20] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. In *Computer Graphics Forum*, vol. 31, pp. 1015–1024, 2012. doi: 10.1111/j.1467-8659.2012.03094.x.

[21] A. Dasgupta, M. Chen, and R. Kosara. Measuring privacy and utility in privacy-preserving visualization. In *Computer Graphics Forum*, vol. 32, pp. 35–47, 2013. doi: 10.1111/cgf.12142.

[22] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1017–1026, 2010. doi: 10.1109/TVCG.2010.184.

[23] A. Dasgupta and R. Kosara. Adaptive privacy-preserving visualization using parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2241–2248, 2011. doi: 10.1109/TVCG.2011.163.

[24] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 202–210, 2003. doi: 10.1145/773153.773173.

[25] C. Dwork. Differential privacy in new settings. In *Proceedings of the annual ACM-SIAM symposium on Discrete Algorithms*, pp. 174–183, 2010. doi: 10.1137/1.9781611973075.16.

[26] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284, 2006. doi: 10.1007/11681878_14.

[27] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042.

[28] G. Ellis and A. Dix. The plot, the clutter, the sampling and its lens: occlusion measures for automatic clutter reduction. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 266–269, 2006. doi: 10.1145/1133265.1133318.

[29] M. Gaboardi, J. Honaker, G. King, J. Murtagh, K. Nissim, J. Ullman, and S. Vadhan. PSI (Ψ): a private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016. doi: 10.48550/arXiv.1609.04340.

[30] M. Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2017. doi: 10.1109/TVCG.2017.2744199.

[31] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, 2013. doi: 10.1109/TVCG.2013.173.

[32] P. E. Greenwood and M. S. Nikulin. *A guide to chi-squared testing*, vol. 280. John Wiley & Sons, 1996.

[33] L. Harrison, K. Reinecke, and R. Chang. Infographic aesthetics: Designing for the first impression. In *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems*, pp. 1187–1190, 2015. doi: 10.1145/2702123.2702545.

[34] K. Healy and J. Moody. Data visualization in sociology. *Annual review of sociology*, 40:105–128, 2014. doi: 10.1146/annurev-soc-071312-145551.

[35] J. Henriksen-Bulmer and S. Jeary. Re-identification attacks—a systematic literature review. *International Journal of Information Management*, 36(6):1184–1192, 2016. doi: 10.1016/j.ijinfomgt.2016.08.002.

[36] R. Hongde, W. Shuo, and L. Hui. Differential privacy data aggregation optimizing method and application to data visualization. In *IEEE Workshop on Electronics, Computer and Applications*, pp. 54–58, 2014. doi: 10.1109/IWECA.2014.6845555.

[37] P. Jain, M. Gyanchandani, and N. Khare. Differential privacy: its technological prescriptive using big data. *Journal of Big Data*, 5(1):1–24, 2018. doi: 10.1186/s40537-018-0124-9.

[38] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):993–1000, 2009. doi: 10.1109/TVCG.2009.153.

[39] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference*, pp. 457–476, 2013. doi: 10.1007/978-3-642-36594-2_26.

[40] M. Köppen. The curse of dimensionality. In *Online world conference on soft computing in industrial applications*, vol. 1, pp. 4–8, 2000. doi: 10.1097/ALN.0000000000002350.

[41] H. Li, L. Xiong, L. Zhang, and X. Jiang. Dpsynthesizer: differentially private data synthesizer for privacy preserving data sharing. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, vol. 7, p. 1677, 2014. doi: 10.14778/2733004.2733059.

[42] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE International Conference on Data Engineering*, pp. 106–115, 2007. doi: 10.1109/ICDE.2007.367856.

[43] Y. Lin, K. Wong, Y. Wang, R. Zhang, B. Dong, H. Qu, and Q. Zheng. Taxthemis: Interactive mining and exploration of suspicious tax evasion groups. *IEEE Transactions on Visualization and Computer Graphics*,

27(2):849–859, 2020.

[44] 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–es, 2007. doi: 10.1145/1217299.1217302.

[45] A. L. Madsen, F. Jensen, A. Salmerón, H. Langseth, and T. D. Nielsen. A parallel algorithm for bayesian network structure learning from large data sets. *Knowledge-Based Systems*, 117:46–55, 2017. doi: 10.1016/j.knosys.2016.07.031.

[46] D. Marutho, S. H. Handaka, E. Wijaya, et al. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *International Seminar on Application for Technology of Information and Communication*, pp. 533–538, 2018. doi: 10.1109/ISEMANTIC.2018.8549751.

[47] R. McKenna, D. Sheldon, and G. Miklau. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*, pp. 4435–4444, 2019.

[48] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Annual IEEE Symposium on Foundations of Computer Science*, pp. 94–103, 2007. doi: 10.1109/FOCS.2007.66.

[49] A. Mojsilovic and B. Rogowitz. Capturing image semantics with low-level descriptors. In *Proceedings of International Conference on Image Processing*, vol. 1, pp. 18–21, 2001. doi: 10.1109/ICIP.2001.958942.

[50] A. Mojsilovic and B. E. Rogowitz. Semantic metric for image library exploration. *IEEE Transactions on Multimedia*, 6(6):828–838, 2004. doi: Semantic metric for image library exploration.

[51] P. Nanayakkara, J. Bater, X. He, J. Hullman, and J. Rogers. Visualizing privacy-utility trade-offs in differentially private data releases. *arXiv preprint arXiv:2201.05964*, 2022. doi: 10.48550/arXiv.2201.05964.

[52] Y. Nie, L. Huang, Z. Li, S. Wang, Z. Zhao, W. Yang, and X. Lu. Geospatial streams publish with differential privacy. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 152–164, 2016. doi: 10.1007/978-3-319-59288-6_14.

[53] S. I. O'Donoghue, B. F. Baldi, S. J. Clark, A. E. Darling, J. M. Hogan, S. Kaur, L. Maier-Hein, D. J. McCarthy, W. J. Moore, E. Stenau, et al. Visualization of biomedical data. *Annual Review of Biomedical Data Science*, 1:275–304, 2018. doi: 10.1146/annurev-biodatasci-080917-013424.

[54] S. I. O'Donoghue, A.-C. Gavin, N. Gehlenborg, D. S. Goodsell, J.-K. Hériché, C. B. Nielsen, C. North, A. J. Olson, J. B. Procter, D. W. Shattuck, et al. Visualizing biological data—now and in the future. *Nature methods*, 7(3):S2–S4, 2010. doi: 10.1038/nmeth.f.301.

[55] J. Oksanen, C. Bergman, J. Sainio, and J. Westerholm. Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, 48:135–144, 2015. doi: 10.1016/j.jtrangeo.2015.09.001.

[56] A. Perer and J. Sun. Matrixflow: temporal network visual analytics to track symptom evolution during disease progression. In *AMIA annual symposium proceedings*, vol. 2012, p. 716, 2012.

[57] W. N. Price and I. G. Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019. doi: 10.1038/s41591-018-0272-7.

[58] W. Qardaji, W. Yang, and N. Li. Priview: practical differentially private release of marginal contingency tables. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1435–1446, 2014. doi: 10.1145/2588555.2588575.

[59] B. Saket, A. Endert, and Ç. Demiralp. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2505–2512, 2018. doi: 10.1109/TVCG.2018.2829750.

[60] B. Saket, A. Endert, and J. Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 133–142, 2016. doi: 10.1145/2993901.2993903.

[61] L. Shao, M. Behrisch, T. Schreck, T. von Landesberger, M. Scherer, S. Bremm, D. A. Keim, et al. Guided sketching for visual search and exploration in large scatter plot spaces. In *EuroVis*, 2014.

[62] E. Steel and G. Fowler. Facebook in privacy breach. *The Wall Street Journal*, 18(1), 2010.

[63] C. D. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662, 2014. doi: 10.1109/TVCG.2014.2346574.

[64] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002. doi: 10.1142/S0218488502001648.

[65] Y. Tao, R. McKenna, M. Hay, A. Machanavajjhala, and G. Miklau. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238*, 2021. doi: 10.48550/arXiv.2112.09238.

[66] P. Thaker, M. Budiu, P. Gopalan, U. Wieder, and M. Zaharia. Overlook: Differentially private exploratory visualization for big data. *arXiv preprint arXiv:2006.12018*, 2020. doi: 10.48550/arXiv.2006.12018.

[67] L. J. Trautman and P. C. Ormerod. Corporate directors' and officers' cybersecurity standard of care: The yahoo data breach. *Am. UL Rev.*, 66:1231, 2016.

[68] S. Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974. doi: 10.1137/1118101.

[69] 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. doi: 10.1109/TVCG.2018.2865021.

[70] 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. doi: 10.1109/TVCG.2017.2745139.

[71] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. doi: 10.1198/jasa.2009.tm08651.

[72] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Information Visualization, IEEE Symposium on*, pp. 21–21, 2005. doi: 10.1109/INFVIS.2005.1532142.

[73] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on knowledge and data engineering*, 23(8):1200–1214, 2010. doi: 10.1109/TKDE.2010.247.

[74] D. Zhang, M. Hay, G. Miklau, and B. O'Connor. Challenges of visualizing differentially private data. *Theory and Practice of Differential Privacy*, 2016:1–3, 2016.

[75] D. Zhang, A. Sarvghad, and G. Miklau. Investigating visual analysis of differentially private data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1786–1796, 2020. doi: 10.1109/TVCG.2020.3030369.

[76] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems*, 42(4):1–41, 2017. doi: 10.1145/3134428.

[77] T. Zhu, G. Li, W. Zhou, and S. Y. Philip. *Differential privacy and applications*. Springer, 2017. doi: 10.1007/978-3-319-62004-6.