Visual Utility Evaluation of Differentially Private Scatterplots

Liudas Panavas *
Northeastern University

Tarik Crnovrsanin
Northeastern University

Jane Lydia Adams
Northeastern University

Ali Sargavad
University of Massachusetts, Amherst

Melanie Tory
Northeastern University

Cody Dunne
Northeastern University

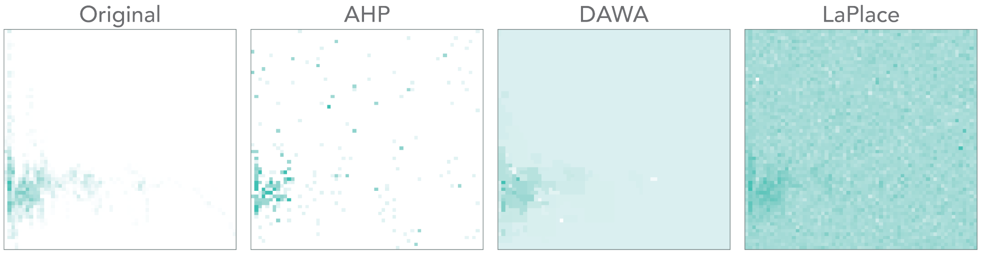

Figure 1: A binned scatterplot (leftmost plot) uses a color scale to show how many data points occur in each cell. A malicious viewer could unambiguously locate a data point by knowing only the value for one of its plotted attributes—thus, determining the likely value of the point's unknown attribute. We can reduce the amount of private information exposed by using differential privacy. Here, we show the results of three of the five algorithms tested (AHP, DAWA, Laplace) for creating differentially private scatterplots. Each has the same theoretical *privacy guarantee*, $\varepsilon = 0.2$, but each algorithm adds noise differently. The choice of algorithm clearly affects the *visual utility* of the resulting scatterplot.

Abstract

Differentially private scatterplots enable the plotting of two attributes while guaranteeing a specified level of privacy. What a user sees from the scatterplot can be affected by which privacy algorithm is used and how it adds noise to the data. However, there is no existing work that quantifies this effect. We present the results of a pilot data study that compares the visual utility of algorithms that create differentially private scatterplots. We compare five popular algorithms across a range of parameters. The results indicate that DAWA and Geometric Truncated are the best algorithms for visual utility. Future research could focus on optimizing the different parameters to maximize utility of the visual representations. A free copy of this paper along with all supplemental materials is available at osf.io (anonymous link).

Index Terms: Human-centered computing—Visualization—Empirical Studies in Visualization

1 Introduction

Scientists aim to find correlations between lifestyles and health outcomes, companies want to understand user behavior patterns, and policy makers are interested in exploring the composition of a country through its census. In all these scenarios, the data collected is personal, so individuals planning to use it must either work with obfuscated data or go through a data request process. Privacy-preserving visualizations can help the data users draw insights or direct their data requests while minimizing risks [5].

When releasing private data, there is a trade-off between *privacy* and *utility*. The greater the deviation between the released data and the original data, the less accurate it is but the more it protects the privacy of the individuals in the database. In the last decade, a common and widely-adopted strategy to privatize data has come in the form of differential privacy [4]. Differentially private algorithms add a calculated random amount of noise to aggregate statistics. In doing so, we give the individuals represented in the data plausible deniability that they are represented in the released statistics [4].

Our work investigates the intersection of differential privacy and visualization—specifically, privacy-preserving scatterplots. To make the data contained in a scatterplot private, practitioners usually first create aggregate counts by sectioning the plot into bins of a Cartesian grid. We will refer to these scatterplots of the original data as *binned scatterplots* (Fig. 1 Original). The plots with noise added we term *private binned scatterplots* (Fig. 1 AHP, DAWA, Laplace [3]).

While the *statistical privacy guarantees* are mathematically proven and quantifiable, there is less consensus on how to quantify *visual utility* and how to select the right parameters to maximize the released data's utility [5]. As seen in Fig. 2, while two private binned scatterplots may have comparable statistical metric for utility, their shape or patterns—i.e. the visual utility [5]—can vary. The utility of private scatterplots depends on a range of parameters (privacy level, scatterplot shape, task, bin size, algorithm choice), and navigating the correct selection of these parameters is a difficult task for data curators [5]. Therefore, to help data curators with the visual dissemination of their data while protecting participant privacy, we contribute:

1. A pilot evaluation of the visual utility of private binned scatterplots created by five common differential privacy algorithms. Our evaluation covers a range of tasks, bin sizes, privacy levels, and scatterplot shapes.
2. An exploration into how the results from our visual data study can be used to improve the utility of private binned scatterplots.
3. Jupyter notebooks with accompanying interactive visualizations for researchers to use to explore the parameters for themselves.

2 Supplemental Material

A copy of this paper along with all supplemental materials is available at osf.io (anonymous link). This includes dataset-generating code, study plot generation code, study website, data analysis code, and code to generate the figures used in this poster.

*Corresponding author. E-mails: [panavas.l | t.crnovrsanin | adams.jan | m.tory | c.dunne]@northeastern.edu, asarv@cs.umass.edu

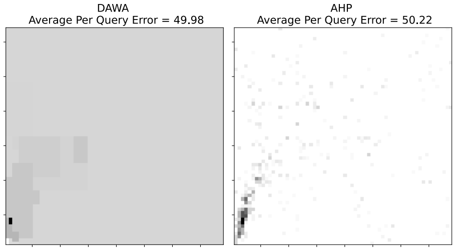

Figure 2: DAWA & AHP have comparable Average Per Query Error [3], but produce two disparate visual results.

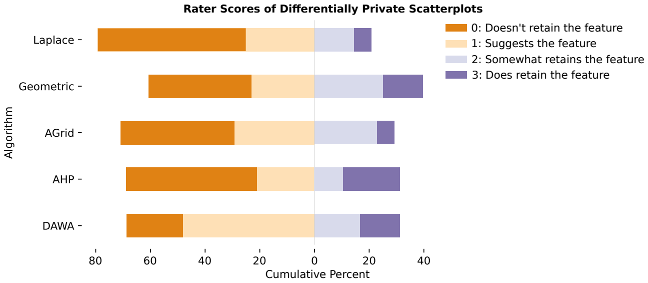

Figure 3: The distribution of ratings from the pilot results.

3 Study Design

Our pilot study evaluates four different datasets across the parameters of bin size, task, privacy level, and privacy algorithm. Three visualization experts each independently rated the how well the private binned scatterplots retain the correlation, distribution, and clustering of the original binned scatterplot. They judged 240 plots on an ordinal scale (the private plot [0 - doesn't, 1 - suggests, 2 - somewhat, 3 - does] retain the user's ability to complete the task) resulting in the ratings seen in Fig. 3.

4 Results

Our results can help data curators select the best parameters when releasing binned scatterplots of their data. Fig. 3 shows the distribution of the ratings made by the expert reviewers for the five algorithms tested. These results demonstrate that the different algorithms have different visual utility when using the same parameters listed earlier. From these results we have several recommendations. **Use DAWA if avoiding total loss in utility is the main priority.** It had the fewest 0 ratings where the visualization didn't retain the feature. **Use Geometric Truncated [2] for the highest likelihood that a private binned scatterplot retains its utility** (most 2's and 3's, which are the highest utility scores given by raters). While AHP and Agrid often perform similarly to the other algorithms, **avoid using Laplace as it consistently had the lowest utility ratings.**

5 Discussion/Directions

Our pilot study of visual utility evaluated by experts has yielded a set of design guidelines and recommendations that also present future research directions in the area of visual communication of privacy-preserving scatterplots.

Statistical Metrics: Using the expert utility evaluation of the privacy-preserving scatterplots as ground truth, we can compare common statistical utility metrics used to benchmark the algorithms [3] against our results. Knowing which statistical metric most closely corresponds to visual utility can help future researchers compare differentially private algorithms' visual utility more accurately. Also, these metrics can serve as a starting point to help optimize the parameters of bin size, task, privacy level, and privacy algorithm.

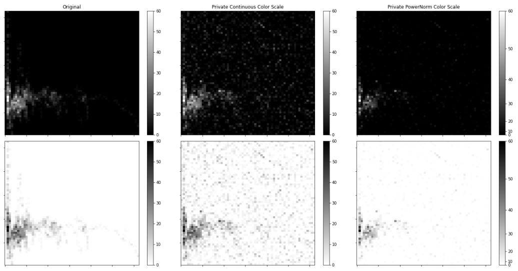

Figure 4: Adjusting the color scale can change the the amount of perceptible noise. The underlying data retains the same amount of privacy but the graph may have higher utility.

Noise Patterns: An evident and prevailing theme in Fig. 1 is the diversity of patterns with which the noise is added. For many of the graphs generated, as long as the noise does not outweigh the signal, an individual looking at the private data can likely perceive visual patterns close to underlying data. A data curator looking to release many scatterplots without visually checking each one may want to use a statistical metric to ensure that the data has not lost nearly all of its visual utility.

Parameter Optimization: The parameters evaluated have a large impact on the utility of the private scatterplots. Bin size is a difficult parameter to accurately set aside from using a guess-and-check system. Using a statistical metric, we can quickly test a variety of bin sizes to find one that maximizes the signal-to-noise ratio and therefore utility. This can save the data curator time when selecting the bin size parameter and provides the the data user with a clearer visual output.

In addition to bin optimization, data curators can adjust the color scale to present the data more clearly. Previous literature in the visualization field has stressed the importance of selecting an appropriate color scale for the data being presented [1]. In much the same way, a data curator can select different colors or bin the color scale in varying ways to help increase the signal-to-noise ratio of the final image.

6 Conclusion

In this work we explore how to more effectively visually communicate the released statistics of a dataset with private, personal information. After evaluating a variety of parameters that affect the visual utility of private binned scatterplots, we discuss a number of ways the results can be used to continue deepening our understanding of privacy-preserving plots. We hope this work inspires further analysis into the largely unexplored intersection of differential privacy and data visualization.

References

[1] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister. Evaluation of artery visualizations for heart disease diagnosis. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2479–2488, 2011. doi: 10.1109/TVCG.2011.192

[2] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012. doi: 10.1145/1536414.1536464

[3] M. Hay, A. Machanavajjhala, G. Miklau, Y. Chen, and D. Zhang. Principled evaluation of differentially private algorithms using dpbench. *Proceedings of the 2016 International Conference on Management of Data*, 2016. doi: 10.1145/2882903.2882931

[4] A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. R. O'Brien, T. Steinke, and S. Vadhan. Differential privacy: A primer for a non-technical audience. *Vand. J. Ent. & Tech. L.*, 21:209, 2018. doi: 10.2139/ssrn.3338027

[5] 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.