Challenges of Visualizing Differentially Private Data

Dan Zhang
UMass Amherst
College of Information and Computer Sciences
dzhang@cs.umass.edu

Michael Hay
Colgate University
Department of Computer
mhay@colgate.edu

Gerome Miklau
UMass Amherst
College of Information and Computer Sciences
miklau@cs.umass.edu

Brendan O'Connor
UMass Amherst
College of Information and Computer Sciences
brenocon@cs.umass.edu

ABSTRACT

Differential privacy has become a primary standard for protecting individual data while supporting flexible data analysis. Despite the adaptation of differential privacy to a wide variety of applications and tasks, visualizing the output of differentially private algorithms has rarely been considered. Visualization is one of the primary means by which humans understand and explore an unknown dataset and therefore supporting visualization is an important goal to advance the practical adoption of differential privacy.

In this initial work on private data visualization we explore key challenges and propose solution approaches. We use two-dimensional location data as an example domain, and consider the challenges of plotting noisy output, the impact of visual artifacts caused by noise, and the proper way to present known uncertainty about private output.

1. INTRODUCTION

Differential privacy seeks to enable the analysis of sensitive datasets while protecting the individuals who provide the data. It has become the the state of the art standard for private data analysis and there has been a flood of research into the design algorithms that meet this guarantee.

Despite the wide variety of analysis tasks that have been studied in the context of differential privacy, there has been little or no attention to the problem of producing useful, accurate visualizations which satisfy the privacy standard.

Visualization can be used as a presentation tool, where the goal is to convey information already extracted from data. Or it may also be used in an exploratory phase, prior to more rigorous statistical analysis. In either case, we maintain that visualization plays a central role in many real-world data analysis workflows and therefore should be supported by private methods.

In this short paper, we identify some main challenges in the visualization of differentially private algorithms and also discuss potential solutions. We consider obstacles to plotting noisy output, the impact of visual artifacts caused by noise, and the proper way to convey the inherent uncertainty in the data being visualized. We hope these challenges will motivate future work on algorithm design targeted to effective visualization.

2. BACKGROUND

Informally, differential privacy is a property of an algorithm that takes as input a collection of records. It guarantees that the algorithm output is statistically indistinguishable (governed by a privacy parameter ϵ) from the output that would have been published had any one individual opted out of the collection. Formally, a randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy [5] if for all databases D and D' that differ on one record, and for any subset of outputs $S \subseteq Range(\mathcal{A})$, $Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon} \times Pr(\mathcal{A}(D') \in S)$.

Although accurate visualization of private data is important for a variety of problem domains, we focus on two-dimensional (2D) location data. Analyzing 2D location data has been studied in the privacy literature [3, 15, 18, 6, 9] and it is also a rich enough application to make visualization challenges evident. In this paper, we use a 2D dataset of taxi pickup information in Beijing during a single month [1]. Each tuple records a pickup location of a taxi ride in the form of a (longitude, latitude) pair. This data may be sensitive as it has the potential to reveal an individual's location.

In practice, an individual may contribute multiple records to this dataset (one per taxi ride taken in the month) and thus an ideal application of differential privacy would extend the definition above to encompass any set of records associated with a single individual [8]. However, a practical limitation of this dataset is that multiple pickups by an individual taxi are not linkable in the data. Thus, we apply the standard definition of differential privacy and note that this still protects individuals who take multiple rides, albeit at a lower ϵ.

A number of algorithms have been proposed for publishing 2D data. A common strategy is to construct a grid (equi-width partition) over the 2D domain and then use the Laplace mechanism [4] to compute noisy counts within each grid cell (essentially, a noisy histogram). In recent years, this simple approach has been improved upon by more sophisticated algorithms [3, 15, 18, 6, 9]. Some algorithms also take as input a workload of linear queries expressed over the histogram counts, for which the workload query answers are released as output. Even when a workload is provided, most algorithms produce as a by-product a noisy histogram suitable for visualization.

3. CHALLENGES

In this section, we present some of the main challenges

(a) Scatter plot of non-private data (b) Heat map of a private output of the Laplace mechanism

Figure 1: Plotting non-private data and private output

faced in visualizing differentially private data. We justify challenges with examples and discuss potential solution approaches.

3.1 Plotting private algorithm output

The visualization techniques used in the absence of privacy concerns may not be applicable to private algorithm output. For 2D location data, the original data is often visualized using a scatter plot, where each mark represents the exact location of an individual.

But directly plotting any individual's information is incompatible with differential privacy. Instead, as mentioned in Section 2, a common strategy is to impose a grid, count the number of records within each grid cell, and add noise. To visualize the resulting noisy counts, one can use a heat map rather than a scatter plot.

These two approaches are compared in Fig. 1. A scatter plot of the Beijing Taxi data introduced in Section 2 is shown in Fig. 1(a) while in Fig. 1(b) a heat map is used to visualize the private output of the Laplace mechanism. Each grid cell is assigned a color scaled to the logarithm of its noisy count.

This immediately causes visual differences in the presentation of the true data and the private data and causes a loss of fidelity even in the absence of noise from the privacy mechanism. The true data consists of longitude and latitude measured to a precision of 0.00001° in a 0.46280° by 0.30534° area, plotted here over a region of 4166 px by 4166 px, resulting in maximum resolution of $0.00011^\circ \times 0.00007^\circ$ per pixel, although it is unlikely that the full resolution is perceptible in the Fig. 1(a). For the private data, we are immediately faced with a choice about the grid size of the 2D. It is 256 x 256 in Fig. 1(b). A finer grid could be selected for greater resolution, but the counts in each cell will be smaller and may be overwhelmed by noise from the privacy mechanism.

3.2 Visual artifacts

The output of a differentially private algorithm may include visual artifacts which obscure true features or lead to false conclusions. For example, the noise introduced by the algorithm may result in negative counts for grid cells (which are clearly impossible) and can have a significant impact on a visualization if not corrected. For example, the blue cells in Fig. 1(b) represent negative counts.

Negative counts can be easily corrected by rounding, but such adjustments sometimes have their own consequences. Simply rounding all negative values to zero boosts the overall sum across the grid cells leading to a biased output which may have its own visual impacts. More sophisticated ways to handle negative values have been proposed [10] and some mechanisms, like the Multiplicative Weights Exponential Mechanism [6], output non-negative counts directly. Issues such as non-negativity can sometimes be ignored when the private output is used to compute query answers, but are likely to become much more important in the context of visualization.

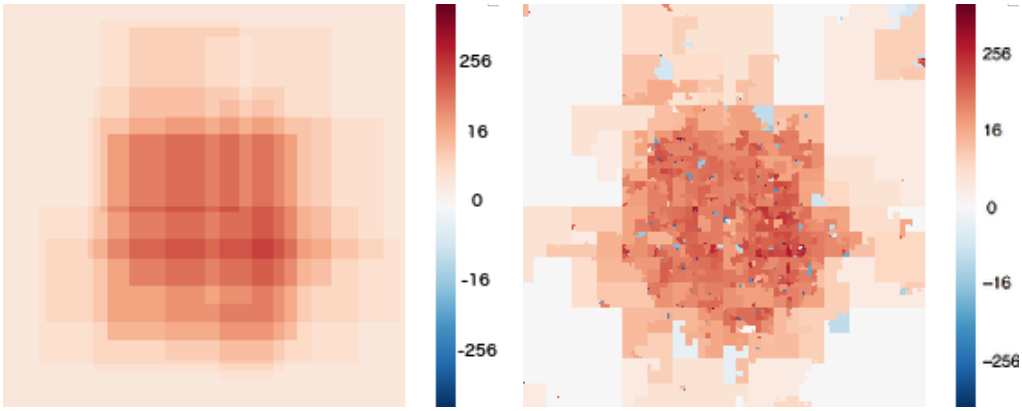
In addition to negative counts, there are other algorithm-specific artifacts that obscure the interpretation of the visualization. For example, the visualization in Fig. 2(b), produced by the DAWA algorithm, includes large blocky uniform regions, especially on the periphery of the figure where the density is lower. The algorithm intentionally estimates these regions uniformly and avoids estimating sub-regions or cells internal to the region. This feature of the algorithm is quite effective in reducing numerical measures of error that are commonly used in the research literature, but may mislead the viewer when the results are presented visually. This is especially true for the non-expert viewer unfamiliar with the algorithm's mechanics who may mistake algorithmic artifacts for structure in the data. This may call for re-thinking some of the advanced algorithmic techniques which are currently used to reduce error when measured by standard error metrics.

3.3 Specifying and achieving "visual utility"

The above discussion shows that effective data visualization is a utility goal which is potentially very different than the utility goals considered to-date in the literature on differential privacy. It is not clear how to make a notion of "visual utility" precise. For now, we stay with an informal definition based on *perceived visual similarity to the true data.*[1]

The extent to which recent algorithm advancements will prove beneficial to visualization is unknown. For our setting, there are many differentially private algorithms that can be applied to 2D data. Some algorithms produce a noisy his-

[1]Ultimately, we believe this notion should probably be task-based, in which a specific task is specified for the user to carry out with the visualization and success rates are compared across the true data and the private data [13].



(a) Output of MWEM (b) Output of DAWA

Figure 2: Outputs with equal query-based error

togram targeting a general class of queries (e.g. sums over all rectangles in the 2D domain) [4, 7, 15, 16] and some accept as input a user-specified workload of queries and tailor the output to accuracy for the workload queries [6, 9]. In either case, error is commonly measured using metrics like L_1 or L_2 error on some set of queries of interest. We refer to this as *query-based error*.

Query-based error is not a reliable measure of visual utility. We show in the following example that two noisy outputs with query-based equal error may have very different visual utility.

For plotting 2D data, we measure query-based error as the average, per-cell L_2 error of the 256×256 histogram output. This seems like the most natural metric because the user is seeing a colored representation of noisy values in each cell and we are comparing this representation to the true heat map. Fig. 2(a) and Fig. 2(b) are noisy outputs of two different algorithms named MWEM [6] and DAWA [9] with the same input data as shown in Fig. 1(a). We have used *different* epsilons for each algorithm in order to make the query-based error of the algorithms equal (here MWEM uses $\epsilon = 1$, while DAWA uses $\epsilon = 0.0065$). Clearly these two figures have very different visual properties, demonstrating that visual utility is not captured by query-based error.

3.4 Visualizing uncertainty

The examples above plot a single output of a randomized private algorithm. Although the effects of noise are visible, the uncertainty in the output is not presented to the user in a manner that can be properly interpreted. Appropriate visualization of uncertainty is a key challenge in visualizing differentially private data. This is similar to the challenge of visualizing *statistical* uncertainty, in which a practitioner is encouraged to not directly trust data (since there is uncertainty in statistical inference), or forecasts from a computational model like climate simulations [17] (since there is uncertainty in the model's accuracy).

Fig. 3 illustrates some of these challenges with the Beijing taxi data. Part (A) shows original data, plotted using the heat map approach described earlier, with cell colored mapped to the counts on a log scale. Dense road networks can be seen in the city, as well as some less-traveled roads in less dense areas, such as those that connect to the airport. The Laplace mechanism ($\epsilon = 0.1$) is used to obtain the noisy version (B) which would be given to a data analyst. Some of the high-frequency structures are preserved, but low-density regions are substantially changed in a random manner. For three selected cells, plot (C) shows the original true values (red triangles), versus noisy versions (blue dots). The fact that cell (1) has a higher frequency than (2) is preserved in the noisy data. But cells (2) and (3) have a *sign error*—their relative ordering is flipped in the data.

For 1d data, uncertainty can be summarized with error bars. Fig. 3(C) shows 95% intervals as vertical lines. These are constructed from a noisy data point $\hat{x}_i$ as $[\hat{x}_i + F^{-1}(0.025), \hat{x}_i + F^{-1}(0.975)]$, where F^{-1} is the inverse CDF of the Laplace (yielding intervals of approximately $\hat{x}_i \pm 30$ for this setting of ϵ); by construction, these intervals contain the true value 95% of the time. These error bars could be presented to a user, to be interpreted in a similar manner as confidence intervals from statistical inference; and helpfully, unlike the case of statistical inference where modeling assumptions may not hold, in this setting the confidence intervals are guaranteed to have correct coverage since the noise distribution is known.

But for 2D data, uncertainty visualization is less straightforward due to limitations on space and visual channels in a 2D setting (e.g. (A) or (B)). Researchers have explored methods to represent uncertainty on the same 2D figure with the data, such as summary plots [14], modifying the color to use hue or saturation to encode uncertainty [12], and showing uncertain data out of focus [11]. Alternatively, one can use interactivity. For example, in a linked-displays approach [2], a user could click to select one or a few cells from the (B) map, then be shown the cells' values in a second display (like (C)) with room to show error bars. These approaches deserve further consideration for visualizing private data.

Another approach to the faithful representation of uncertainty is to match the imprecision inherent to visual perception to the imprecision introduced by the privacy mechanism. The proposed principle is that *statistically indistinguishable counts should be visually indistinguishable*. The visual limitations in a heatmap stem from the fact that the human eye cannot distinguish colors that are close to one another in the colormap. For algorithms like the Laplace mechanism, in which independent noise is added to each

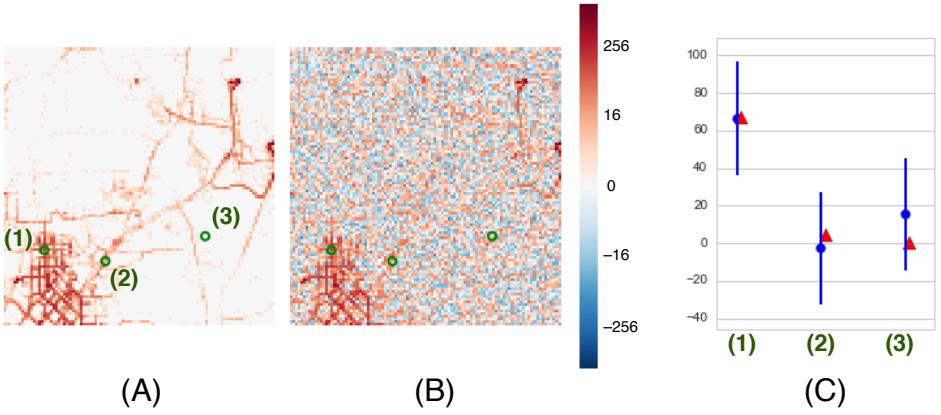


Figure 3: Illustration of uncertainty due to the Laplace mechanism, on taxi frequency data from northeast Beijing (Section 3.4). (A) Original data. (B) Noisy output, which preserves some structures but introduces spurious phenomena. (C) For three selected cells, original data values (red triangles), noisy versions (blue dots), and 95% confidence intervals (vertical lines). Cell (2) has a negative valued output, and the comparison between cells (2) and (3) has a sign error.

count, it is straightforward to impose a threshold on the probability that two different noisy counts reflect a true difference in the underlying data. Then, to obey the above principle, we seek a color mapping in which differences that do not meet the distinguishability standard are mapped to imperceptible color differences.

Overall, for visualizing uncertainty, we hope to benefit from the fact that the error in estimates is coming from a well understood process (the privacy mechanism). Yet for some state-of-the-art algorithms, reliable error bounds are hard to establish because these algorithms adapt the noise distribution to the data. While it is possible to release noisy measure of error, this adds an additional level of uncertainty that must be reconciled.

3.5 Visualization for exploration

The previous challenges are faced when producing a single static plot. A range of additional challenges will be faced in supporting interactive data exploration while satisfying differential privacy. Data exploration is an iterative process in which a sequence of visualizations must be produced privately from the data. Multiple views of the data will tend to consume the privacy budget and require increased noise. In addition, users may begin exploring data with only a vague idea of what interests them, making ineffective the algorithmic techniques which specialize the output to a known workload.

4. REFERENCES

[1] Taxi trajectory open dataset. http://sensor.ee.tsinghua.edu.cn/, 2009.

[2] A. Buja, D. Cook, and D. F. Swayne. Interactive high-dimensional data visualization. *Journal of computational and graphical statistics*, 5(1):78–99, 1996.

[3] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu. Differentially private spatial decompositions. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 20–31. IEEE, 2012.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006.

[5] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[6] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2012.

[7] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1-2):1021–1032, 2010.

[8] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.

[9] C. Li, M. Hay, G. Miklau, and Y. Wang. A data-and workload-aware algorithm for range queries under differential privacy. *Proceedings of the VLDB Endowment*, 7(5):341–352, 2014.

[10] C. Li, G. Miklau, M. Hay, A. McGregor, and V. Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *VLDB Journal*, 2015.

[11] A. M. MacEachren. Visualizing uncertain information. *Cartographic Perspectives*, (13):10–19, 1992.

[12] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[13] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[14] K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. Visualizing summary statistics and uncertainty. In *Computer Graphics Forum*, volume 29, pages 823–832. Wiley Online Library, 2010.

[15] W. Qardaji, W. Yang, and N. Li. Differentially private grids for geospatial data. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 757–768. IEEE, 2013.

[16] W. Qardaji, W. Yang, and N. Li. Understanding hierarchical methods for differentially private histograms. *Proceedings of the VLDB Endowment*, 6(14):1954–1965, 2013.

[17] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1421–1430, 2010.

[18] J. Zhang, X. Xiao, and X. Xie. Privtree: A differentially private algorithm for hierarchical decompositions. *arXiv preprint arXiv:1601.03229*, 2016.