DeID2: Comprehensive Metrics for Differentially Private Synthetic Data

Team N-CRiPT

1 Executive Summary

In this executive summary, we briefly introduce four additional metrics for the temporal data challenge, evaluating the Jaccard distance, heavy hitters, and horizontal and vertical correlations, respectively. We motivate these metrics with real applications. We believe that these additional metrics can complement the JSD metric currently used in the challenge, to provide more comprehensive evaluations. In what follows, we detail our additional metrics.

1.1 Jaccard Distance

Motivation. In the official example "PieChartMetric", there is a *misleading presence penalty (MPP)*: each time a category of incident shows up as significant in the privatized data (denoted as dp) but is actually zero in the ground truth (denoted as gt), a penalty of 0.2 is added. However, if there are many misleading presence and we repeatedly add penalties of 0.2 each, the final loss (i.e., 0.2 times the number of incidents) will dominate all other penalties combined. Therefore, we need a new metric that not only penalizes the misleading presence but also has the same scale as JSD.

Formalization. In our solution, we apply Jaccard Distance (JD) instead of MPP. Let $X = (X^{(1)}, \cdots, X^{(|\mathcal{I}|)})$ be a row in gt, where $\mathcal{I}$ denotes the set of incidents and $X^{(i)}$ denotes number of occurrences of incident i. We define a set $\mathcal{X} = \{X^{(i)} | X^{(i)} > 0, i \in \mathcal{I}\}$, and let $\hat{\mathcal{X}}$ be the corresponding set in the privatized data. Then, the JD between the i-th rows in gt and dp is defined as

$$JD(gt_i, dp_i) = 1 - \frac{\mathcal{X} \cap \hat{\mathcal{X}}}{\mathcal{X} \cup \hat{\mathcal{X}}}$$

Comparison with MPP. Suppose that the number of misleading presences is fixed. In that case, MPP gives a fixed penalty, whereas JD is more adaptive in the following sense. First, a larger size of $\mathcal{X}$ leads to a smaller JD. In addition, when there are many types of incidents, a small amount of error would have little effect on JD. However, if there is only one type of incident in gt, then one misleading presence in dp will cause 50% error.

1.2 Heavy Hitters

Motivation. Which incidents are frequently reported? For example, if *injured person* is frequently reported, then the authority or education department may raise the awareness of first aid guide among the public. Likewise, if *armed person* or *robbery:armed* is frequently reported, then the public should be informed about the situation.

Formalization. Formally, given the i-th month (or throughout a year) and neighborhood A (or the whole area), we define a k-heavy hitter incident as an incident that occurs more than k times during that time in that neighborhood. A synthetic dataset H' (output) is said to be accurate for H (ground-truth) in terms of k-heavy hitters with an error at most Δ if the following two conditions hold:

- If incident *X* is a $(1+\Delta)k$-heavy hitter in H', then *X* is a k-heavy hitters in H, and
- If incident *X* is not a $(1-\Delta)k$-heavy hitter in H', then *X* is not a k-heavy hitters in H.

Once the conditions above are violated, we add a constant penalty h to the total loss denoted as HH. To be consistent with JSD, we set the maximum penalty of one row in H' to 1. The overall HH penalty for H' is the sum of penalties across all neighborhoods in M months.

1.3 Horizontal Correlation

Motivation. Throughout a year, how does the number of *armed person* in neighborhood A correlate with the number of *armed persons* in neighborhood B? If there is a strong correlation, then it is likely that some incidents in the two neighborhoods are related. Such information could be used to decide police patrol routes, etc. Similarly, it could be useful to preserve the correlations among other incidents.

Formalization. We define a new metric, referred to as *horizontal correlation (HC)*, inspired by the well-known similarity measure *Pearson Correlation Coefficient (PCC)*. PCC can reflect the vertical correlation of two time series. The higher the PCC of two series is, the closer they are. Formally, given X^A that denotes the occurrences of type X in neighborhood A in a series of months and Y^B, the correlation between them is defined as

$$r(X^A, Y^B) = \frac{cov(X^A, Y^B)}{\sigma_{X^A}\sigma_{Y^B}}$$

$r(X^A, Y^B)$ ranges in $[-1, 1]$. In privatized data $\hat{H}$, there is a corresponding correlation $r(\hat{X}^A, \hat{Y}^B)$. The HC penalty of $\hat{H}$ is defined as the sum of absolute differences between $r(X^A, Y^B)$ and $r(\hat{X}^A, \hat{Y}^B)$ whenever $|r(X^A, Y^B)| > 0.9$.

1.4 Vertical Correlation

Motivation. In each neighborhood, what is the distribution or trend of incidents over time? Existing research [1, 4, 5] has shown that the patterns of incidents such as crime fluctuate seasonally, and different types of crime vary in their strength and seasonalities. If the emergency dispatch office learns the distribution and trend of incidents in each neighborhood, they may arrange the police force in accordance with seasonality of crime, which may improve the efficiency and effectiveness of their work. Additionally, the government may put in more security and publicity effort during the peak months or seasons of incidents, for public safety.

Formalization. In order to evaluate how well the differential private data retains the trend of incidents, we define a new loss function referred to as vertical correlation (VC). Let X^A denotes the occurrences of type X in neighborhood A in a series of months. The corresponding time series in differential private data (dp) is $\hat{X}^A$. The PCC between X^A and $\hat{X}^A$, $r(X^A, \hat{X}^A)$, can measure the similarity between X^A and $\hat{X}^A$. Then $\frac{1}{2}(1 - r(X^A, \hat{X}^A))$ can describe the dissimilarity between $X^A, \hat{X}^A$ and its value locates in $[0, 1]$. To penalize the dissimilarity of all time series between gt and dp, we sum this value over all incidents and neighborhoods, and define it as the VC penalty of dp.

2 Metric Definition

2.1 Technical Background

2.1.1 Jensen Shannon Distance

The main metric, Jensen-Shannon Distance, is the square root of the Jensen-Shannon divergence. Given two probability distributions P and Q, we define the Jensen-Shannon distance $JSD(P, Q)$ as:

$$JSD(P, Q) = \sqrt{\frac{D(P||M) + D(Q||M)}{2}}$$

where $M = (P + Q)/2$ and D is the Kullback-Leibler divergence..

The form of Jensen Shannon Distance is similar to KL divergence. However, KL divergence becomes infinite if some cell is zero in the ground truth data gt but is non-zero in the differentially private data dp due to positive noise. This penalizes the difference between gt and dp too harshly. Therefore, we choose Jensen Shannon Distance but combine it with a Jaccard Distance (JD) to penalize the spurious labels that do not appear in the ground truth data.

2.1.2 Jaccard Index

Jaccard index is a measure of similarity between two sets. It compares their shared and distinct members. The higher the Jaccard index is, the similar the two sets are. Given two sets X and Y, their Jaccard index is defined as

$$J(X, Y) = \frac{X \cap Y}{X \cup Y},$$

whose range is $[0, 1]$. According to this, the Jaccard Distance (JD) is defined as

$$JD(X, Y) = 1 - J(X, Y),$$

whose value also locates in $[0, 1]$. It is worth noting that the value of Jaccard index and Jaccard Distance are sensitive to the size of the two sets.

2.1.3 Pearson Correlation Coefficient

The Pearson Correlation Coefficient (PCC) is a widely used statistic that measures linear correlation between two series X and Y. Given two series (vectors) X and Y, the formula of PCC is defined as:

$$r(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

where $cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ is the covariance between X and Y and σ_X, σ_Y are the standard deviation of X and Y.

2.2 Formal Metric Definition

In this subsection, we propose four additional metrics for the temporal data challenge, including the Jaccard distance, heavy hitter, horizontal correlation among incidents and vertical correlation between time series. We motivate these metrics with real world applications. Then we give a formal definition for each metric and demonstrate how to calculate them. These additional metrics complement the JSD.

2.2.1 Jensen Shannon Distance

Motivation. Jensen Shannon Distance or Divergence is a useful metric in many area such as machine learning [3] and social sciences [2]. It completely describes the statistical difference between two distributions. So, We reserve this metric and will briefly introduce it following the given documentation "PieChartMetric".

Formalization. One row of the ground truth and differentially private data can be regraded as a vector of non-negative incident counts. We denote them as gt_i and dp_i. Firstly, we zero out the non-significant count whose frequency is less than a Frequency Threshold (FT) in each vector. Secondly, we divide the vector by its new sum to normalize it and compute $JSD(gt_i, dp_i)$. Thirdly, we add a Bias Penalty (BP). $BP(gt_i, dp_i) = 0.2$ if the sum of the counts in dp_i is more than 500 off from gt_i, otherwise $BP(gt_i, dp_i) = 0$. Thus, the complete loss function is as follows,

$$Loss(gt, dp) = \sum_{i \in R} JSD(gt_i, dp_i) + BP(gt_i, dp_i)$$

where R denotes the number of rows.

2.2.2 Jaccard Distance

Motivation. In the official example, it adds a misleading presence penalty (MPP). That is, each time a category of incident shows up as significant in the privatized data but is actually zero in the ground truth, they add 0.2 as penalty. However, if there are many misleading presence and we continuously add 0.2, the final misleading loss (0.2x[Number of incidents]) will be much larger than other metrics. So, we need a new metric that can not only penalize the misleading presence but also have the same scale as JSD. In our solution, instead of adding MPP, we apply Jaccard Distance.

Formalization. Given a row in the ground truth data $X = (X^{(1)}, \cdots, X^{(|\mathcal{I}|)})$, where $\mathcal{I}$ denotes the set of incidents and $X^{(i)}$ denotes number of occurrences of incident i. We define a set $\mathcal{X} = \{X^{(i)} | X^{(i)} > 0, i \in \mathcal{I}\}$ and $\hat{\mathcal{X}}$ is the corresponding set in differentially private data. Then the formula of Jaccard Distance is defined as,

$$JD(gt_i, dp_i) = 1 - \frac{\mathcal{X} \cap \hat{\mathcal{X}}}{\mathcal{X} \cup \hat{\mathcal{X}}} \tag{1}$$

Now we set the loss function as,

$$Loss(gt, dp) = \sum_{i \in R} \omega_1 JSD(gt_i, dp_i) + \omega_2 JD(gt_i, dp_i) + BP(gt_i, dp_i)$$

where ω_1, ω_2 are weights of the metrics.

Evaluation. Notably, when the number of misleading presences is fixed, the larger the size of $\mathcal{X}$ is, the smaller the Jaccard Distance is. This is intuitively understandable, when there are many types of incidents, a small amount of error has little effect. However, if there are only 1 type of incident in gt, 1 misleading presence in dp will cause 50% error. Moreover, unlike MPP, $JD(gt_i, dp_i)$ always ranges in $[0, 1]$. Even if there are many misleading presences, it will not dominate the total loss $Loss(gt, dp)$. Therefore, the Jaccard Distance is more adaptive and more suitable than MPP.

2.2.3 Heavy Hitter

Motivation. Which incidents are frequently reported? For example, if *injured person* is frequently reported, then the authority or education department may raise the awareness of first aid guide among the public. Likewise, if *armed person* or *robbery:armed* is frequently reported, then the public may need to be warned in advance, or even acquire some knowledge on how to deal with armed persons (e.g., self-defense, being precautious,etc.).

Formalization. Formally, given the i-th month (or throughout a year) and neighborhood A (or the whole area), we define a k-heavy hitter incident as an incident that occurs more than k times during that month in that neighborhood. A synthetic data-set H' (output) is said to be accurate for H (ground-truth) in terms of k-heavy hitters with an error at most Δ if the following two conditions hold:

- If incident *X* is a $(1+\Delta)k$-heavy hitter in H', then *X* is a k-heavy hitters in H, and
- If incident *X* is not a $(1-\Delta)k$-heavy hitter in H', then *X* is not a k-heavy hitters in H.

In particular, once the conditions above are violated, we add a constant penalty h to the total loss denoted as HH. With k and Δ fixed, to be consistent with JSD, we set the maximum penalty of one row in H' to 1. The overall penalty for H' is the sum of penalties across all neighborhoods in M months. The formula of HH can be defined as,

$$HH(gt, dp) = \sum_{i \in R} HH(gt_i, dp_i) = \sum_{i \in R} \min(t_i * h, 1) \tag{2}$$

where t_i counts the violation times of the above conditions over all incidents in one row of data.

Calculation. Let us consider the following two histograms H' and $\tilde{H}$ (privatized data), and the original histogram H (ground-truth). They all correspond to the number of occurrences of incidents in neighborhood A on January. The total number of incidents reported in each histogram are all 1000. In H, incident type *E1* occurs with frequency 40% while *E2* to *E5* each occurs with frequency 15%, and the rest incidents never occurs. In H', *E1* occurs with frequency 1 while in $\tilde{H}$, *E1* occurs with frequency 0 and *E2* to *E5* occurs with frequency 25%, as Figure 1 shows. According to the definition, $JSD(H'\|H) = 0.46$ while $JSD(\tilde{H}\|H) = 0.16$. Hence, H' is more favorable than $\tilde{H}$ in terms of its Jensen-Shannon divergence w.r.t. H. However, the 250-heavy hitters in H' and $\tilde{H}$ are totally different. With $k = 250$, $\Delta = 0.2$ and $h = 0.2$, the penalty for H' is 0 while the penalty for $\tilde{H}$ is $\min(5 * 0.2, 1) = 1$. This concrete example shows that JSD can not be used to evaluate the heavy hitters. In addition, we believe that heavy hitters are important information to preserver, it shall be considered as a compensation metric to the JSD.

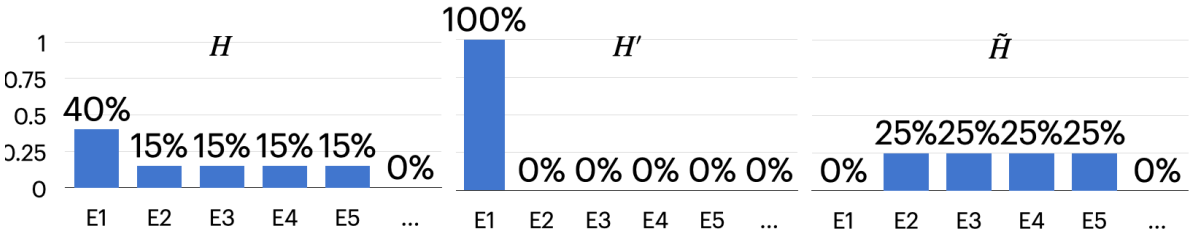


Figure 1: The ground-truth histogram H and two output histograms H' and $\tilde{H}$ for evaluation.

2.2.4 Horizontal Correlation

Motivation. Throughout a year, how does the number of *armed person* at neighborhood A correlates with the number of *armed* at neighborhood B? If there is a strong correlation, then it is likely that two reports in different neighborhoods are concerning with the same armed person (or a group of criminals). Given this information, the police may stand more chance to locate the armed person (or a group of criminals). Furthermore, if the *armed person* reported at A is correlated with abduction at B, the police may also set up barricades on the routes from B to A to stop the vehicles of the criminals or warn the resident in B in advance. Similarly, it is useful to preserve the correlations among *armed person*, *missing person*, *narcotics*, and *lost child*, etc..

Formalization. Hence, it is important to keep the correlation between different incidents. We define a new metric called HC, inspired by the well-known similarity measure Pearson Correlation Coefficient (PCC). PCC can reflect the vertical correlation of two time series. The higher the PCC of two series is, the closer they are. Formally, given two incidents X and Y and two neighborhoods A and B, the correlation between X reported at A and Y reported at B is calculated as:

$$r(X^A, Y^B) = \frac{cov(X^A, Y^B)}{\sigma_X^A \sigma_{Y^B}} = \frac{1}{M} \sum_i^M \frac{(X_i^A - \bar{X}^A)(Y_i^B - \bar{Y}^B)}{\sigma_{X^A} \sigma_{Y^B}}$$

where X_i^A denotes the number of occurrences of type X in neighborhood A in the i-th month, $\bar{X}^A$ is the corresponding average over M months, and σ_{X^A} denotes the corresponding standard deviation. $r(X^A, Y^B)$ ranges between -1 and $+1$. In privatized data, there is a corresponding correlation $r(\hat{X}^A, \hat{Y}^B)$. Given M months data happened in a neighborhood A, the metric HC is defined as,

$$\begin{aligned} HC(gt, dp) &= \sum_{A \in \mathcal{P}} HC^A(gt, dp) \\ &= \sum_{A \in \mathcal{P}} \frac{1}{|\mathcal{I}|} \sum_{X \in \mathcal{I}} \frac{1}{|\mathcal{C}(X^A)|} \sum_{Y^B \in \mathcal{C}(X^A)} \frac{M}{2} \left\| r(X^A, Y^B) - r(\hat{X}^A, \hat{Y}^B) \right\|_1 \end{aligned} \tag{3}$$

where $\mathcal{P}$ denotes the set of all neighborhoods, $\mathcal{I}$ denotes the set of all incidents and $\mathcal{C}(X^A) = \{Y^B \in \mathcal{P} \times \mathcal{I} \mid |r(X^A, Y^B)| > 0.9\}$ denotes the set of the [neighborhood]×[incident] that are highly related to X^A. The loss of X^A is 0 if $|\mathcal{C}(X^A)| = 0$.

Evaluation. Considering a neighborhood A, an incident type X and $Y^B \in \mathcal{C}(X^A)$, the penalty of H' is defined as absolute difference of H' and H in terms of $r(X^A, Y^B)$. Similar to the heavy hitters, we need to normalize the this penalty to be consistent with JSD. We propose to firstly divide the penalty by 2 to make it in range $[0, 1]$. Next, we multiply the sum by M since $r(X^A, Y^B)$ is calculating by M months (M rows). Finally, we take the average of all penalties across $\mathcal{C}(X^A)$ and $\mathcal{I}$ in the neighborhood.

Calculation. The calculation of HC has two steps: constructing $\mathcal{C}(X^A), X^A \in \mathcal{P} \times \mathcal{I}$ (one-time processing) and calculating $r(\hat{X}^A, \hat{Y}^B)$ (for each privatized data-set). Assume there are 200 neighborhoods and 200 incident types. Although constructing $\mathcal{C}(X^A)$ for all X^A costs $2 \cdot \binom{200}{2} \cdot \binom{200}{2} \approx 8 \cdot 10^8$ operations, it is a one-time process that can be done in advance. The time complexity of calculating all $r(\hat{X}^A, \hat{Y}^B)$ is $O(\sum_A \sum_X |\mathcal{C}(X^A)| \cdot M) \approx O(M\,|\mathcal{P}|\,|\mathcal{I}|)$, where $|\mathcal{C}(X^A)|$ is usually a very small number (< 10) in practice. Therefore, the time consumption of

HC is comparable to other metrics like JSD. The reason that JSD can not express correlations between incidents is similar to the previous case. It is possible that *E1* and *E2* are positively correlated in both H (the ground-truth) and H' (output), while they are negatively correlated in $\tilde{H}$. However, it is possible that $\tilde{H}$ has a smaller JSD w.r.t H compared with H'. The construction of such a counter example is in spirit similar to the previous one. In conclusion, JSD can not express heavy hitters in a neighborhood at a time nor correlations among incidents. Considering the importance of both heavy hitters and correlations, we should use these additional metrics to JSD.

2.2.5 Vertical Correlation

Motivation. In each neighborhood, what is the distribution or trend of incidents over time? Many researches such as [1, 4, 5] have shown that the patterns of incidents like crime fluctuate seasonally and different types of crime vary in their strength and seasonalities. If the emergency dispatch office learn the distribution and trend of incidents in each neighborhood, they can arrange the working hours of the police seasonally and rationally dispatch police strength in Baltimore, which will greatly improve the efficiency and effectiveness of their work. Additionally, the government can increase security and publicity during the month or season of high incidents to protect people from danger.

Formalization. Therefore, in order to evaluate how well the differential private data keeps the trend of incidents, we define a new loss function called VC. Given an incident X lasted M months in ground truth data (gt), we let $X_t^A, t = 1, \cdots, M$ denotes the proportion of type X of the overall incidents happened at time t in neighborhood A. Let time series $X^A = (X_1^A, \cdots, X_M^A)$. The corresponding time series in differential private data (dp) is $\hat{X}^A = (\hat{X}_1^A, \cdots, \hat{X}_M^A)$. Then the Pearson Correlation Coefficient $r(X^A, \hat{X}^A)$ that measures the similarity between X^A and $\hat{X}^A$ is:

$$r(X^A, \hat{X}^A) = \frac{cov(X^A, \hat{X}^A)}{\sigma_{X^A}\sigma_{\hat{X}^A}}$$

Then $\frac{1}{2}(1 - r(X^A, \hat{X}^A))$ can describe the dissimilarity between $X^A, \hat{X}^A$ and its value locates in $[0, 1]$. Therefore, to penalize the dissimilarity of all time series between gt and dp, we define the loss of the whole data as follows:

$$VC(gt, dp) = \frac{1}{|\mathcal{I}|} \sum_{A \in \mathcal{P}} \sum_{X \in \mathcal{I}} \frac{M}{2}(1 - r(X^A, \hat{X}^A)) \tag{4}$$

There are some special cases that PCC is inapplicable ($\sigma_{X^A} = 0$ or $\sigma_{\hat{X}^A} = 0$), we treat them as follows.

- If $\max_t |X_t^A - \hat{X}_t^A| <= \frac{FT}{2}$, $VC(X^A, \hat{X}^A) = 0$.
- If $\frac{FT}{2} < \max_t |X_t^A - \hat{X}_t^A| <= FT$, $VC(X^A, \hat{X}^A) = 0.5M$.
- If $\max_t |X_t^A - \hat{X}_t^A| > FT$, $VC(X^A, \hat{X}^A) = M$.

where FT is the frequency threshold, $A \in \mathcal{P}, X \in \mathcal{I}$.

Evaluation. We time the original dissimilarity with M and divide it by $\mathcal{I}$ to make it consist with JSD since the calculation of JSD includes all incident types and is computed row by row

(month by month). Here ω_5 is a weight used to adjust the importance of VC compared with other losses such as JSD.

Calculation. We give an example in Table 1. The columns in H are ground truth and the columns in $\hat{H}$ and $\widetilde{H}$ are outputs of two differential private algotithms. Each column can be regarded as a time series except for the first column denoting months from Jan. to May. For simplicity, we consider only two incidents I_1 and I_2 and each cell in the table denotes the proportion of the incident in the corresponding month (we normalize each vector by dividing its sum). According to the data in Table 1, $JSD(\hat{H}||H) = 0.75$ while $JSD(\widetilde{H}||H) = 0.37$. However, for our metric VC (setting $\omega_5 = 1$), we notice that $VC(\hat{H}, H) = 0$ while $VC(\widetilde{H}, H) = 1.88$! In addition, if the data contains longer time series, the VC may be even larger. Obviously, the data $\hat{H}$ is in compliance with the variation of the ground truth data H while the data $\widetilde{H}$ completely deviates from the trend of H. The Jensen Shannon distance, unfortunately, cannot capture this important feature. So, the metric VC is necessary.

Table 1: Data of the ground truth H and two output $\hat{H}, \widetilde{H}$ in 5 months at some neighborhoods

Month	H		$\hat{H}$		$\widetilde{H}$	
	I_1	I_2	I_1	I_2	I_1	I_2
1	0.3	0.7	0.15	0.85	0.4	0.6
2	0.3	0.7	0.15	0.85	0.2	0.8
3	0.3	0.7	0.15	0.85	0.4	0.6
4	0.4	0.6	0.25	0.75	0.4	0.6
5	0.3	0.7	0.15	0.85	0.4	0.6

2.2.6 Metric Combination

Combing all the metrics mentioned above, the final metric can be defined as follows,

$$Loss = (\omega_1 JSD + BP) + \omega_2 JD + \omega_3 HH + \omega_4 HC + \omega_5 VC \tag{5}$$

where $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5$ denote the weights of the metrics.

2.3 Explanation of Metric Parameters

In this subsection, we provide explanation for each of the parameters that affect the performance of the metrics. We cover all parameters that can affect the discriminative power of the metric.

2.3.1 Data Configuration Parameters

Our metrics are based on a histogram of record types. In the 2019 Baltimore Police Incident and 911 data-set, record type refers to incident descriptions.

2.3.2 Tuning Parameters

Heavy Hitter: Δ, k, h. When applying Heavy Hitter metric, we recommend setting the Δ around 20%. If the Δ is set too big, this metric would never take effect. If it is too small, most of rows will get a loss near 1 so that the metric does not have any discriminative power. The motivation of Heavy Hitter focus on the frequent incidents, whose occurrences are prominent.

As a result, for one row $X = (X^{(1)}, \cdots, X^{(|\mathcal{I}|)})$ in the data, we set k equal to $\lfloor \frac{1}{10} \sum_{i \in \mathcal{I}} X^{(i)} \rfloor$, i.e., 1/10 of the sum of all incidents' occurrences.

Frequency Threshold: FT. In both ground truth and differentially private data, we set the incident count to 0 when its frequency proportion is less than FT%. FT% is usually set to $\leq 5\%$. The official document "PieChartMetric" has demonstrated the impact of tuning FT so we will not repeat in this report.

Bias Penalty Amount: BP. When we compute the JSD between two rows, we only compare their vector after normalization. However, the sum of a vector could vary largely while its elements keep their proportion invariant. So, we need this penalty to evaluate the sum of each row and add it to JSD. In practise, we set BP in $[0, 0.1]$ to avoid a too harsh penalty.

Metric weight: $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5$. For the five metrics mentioned above (JSD, JD, HH, HC, VC), JSD has better statistical properties and it can capture every detail of the data while the other four additional metric can be regarded as a compensation of JSD. So, we increase the importance of JSD and set the their weights $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) = (0.3, 0.15, 0.15, 0.15, 0.15)$. We will explain why we apply this setting later. The weights will not affect the performance of each metric. It only affects the final score. Moreover, how the weight are set depends on our analysis purpose — what features we need in the data-set. So, we will not explore the effect of $\boldsymbol{\omega}$.

2.4 Snapshot and Deep Dive Modes

2.4.1 Snapshot Mode

JSD, BP, JD, and HH are calculated by row. JSD, JD and HH range in $[0, 1]$ while BP range in $[0, 0.1]$. Although the Vertical Correlation is calculated by column, if we look at it in another light, it assigns identical loss ranging in $[0, 1]$ for every row in the column. Therefore, the final loss of each row has a minimum score of 0 and a maximum score of $(0.3 + 0.15 + 0.15 + 0.15 + 0.15) * 1 + 0.1 = 1$. For the whole data, it has a minimum score of 0 and a maximum score equal to [total number of map segments] $\times$ [total number of time segments].

2.4.2 Deep Dive Mode

Since all of our metrics scores map or time segment separately, it provides clear utility for deep dive investigation.

For the metrics calculated by row, the **Temporal Scores Chart** allows you to select any neighborhood and see the change in scores of all metrics in that neighborhood over each of the time segments. The HC and VC metric scores are identical during 12 months in each neighborhood since their calculation takes 12 months data together. Table 2 gives an sketch of our metric scores in Remington and Reservoir Hill over months in 2019. The last column, "Total", score is calculated by equation 5 with $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) = (0.3, 0.15, 0.15, 0.15, 0.15)$.

Table 2: A sketch of Temporal Scores Chart at neighborhood 213 (Remington) and neighborhood 214 (Reservoir Hill) in 2019. Ground truth: the 2019 Baltimore Police Incident and 911 data-set. Privatized data: additional Laplace noise with $\epsilon = 4$.

Neighborhood	Month	JSD	BP	JD	HH	HC	VC	Total
213	1	0.372	0.100	0.667	0.080	0.121	0.277	0.383
213	2	0.268	0	0.667	0.080	0.121	0.277	0.252
213	3	0.190	0	0.500	0.080	0.121	0.277	0.203
...	...	...	...	...	...	...	...	...
214	1	0.335	0	0.400	0.080	0.171	0.290	0.242
214	2	0.243	0	0.250	0.100	0.171	0.290	0.195
214	3	0.018	0	0	0.160	0.171	0.290	0.099

3 Metric Defense

3.1 Exploration of Parameter Tuning

3.1.1 Data

We use a publicly available data provided by Sprint 1 of the NIST Differential Privacy Temporal Map Challenge which records 911 calls that occurred in Baltimore in 2019. The privatized data used for our analyses is generated by the baseline differential privacy algorithm from that challenge's competitor's pack. There are privatized data in poor quality ($\epsilon = 0.5$) and moderate quality ($\epsilon = 4$). To explore the impact of the parameters, we focus on the distribution of scores of the metrics.

3.1.2 Metric Parameters

Unless otherwise stated, the parameters of our metrics have been set to the defaults given in the Metric Definition, where $\boldsymbol{\omega} = (0.3, 0.15, 0.15, 0.15, 0.15)$, $\Delta = 0.2$, $k = \lfloor \frac{1}{10} \sum_{i \in \mathcal{I}} X^{(i)} \rfloor$, $h = 0.02$, FT=0.05, BP=0.1.

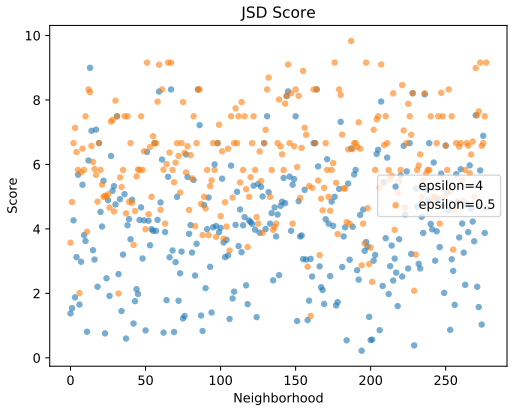
3.1.3 Scores Composition

The total score has six components as described in the Metric Definition section above: JSD, BP, JD, HH, HC, VC. Here we look at the impact each component has on the total score, depend on the quality of data. We sum the scores of each metric in 12 months in each neighborhood and Figure 2 shows the results. Here we do not focus on the HC metric since it costs a lot of preprocessing time. We can find out that our metrics have a strong discriminative power in Figure 2. It is obvious in all our metrics that the scores from the poor quality privatized data-set ($\epsilon = 0.5$) are much larger than those from the moderate quality data-set($\epsilon = 4$).

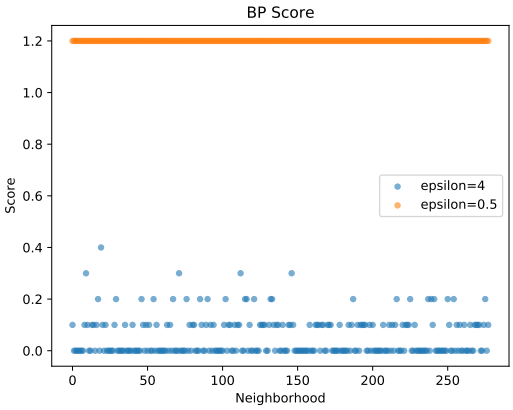
3.1.4 Effect of Δ in HH

We now briefly explore the effect of increasing and decreasing the Δ from the default value $\Delta = 0.2$.

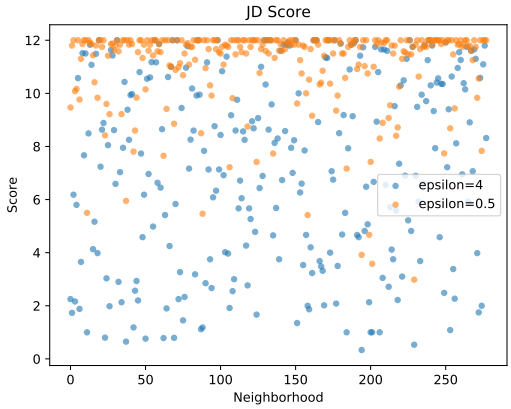
When $\Delta = 0.1$ or $\Delta = 0.2$, the scores are nearly uniform in the range $[0, 1]$. However, when it increases to $\Delta = 0.4$, the restrictions of HH are loosen so that the frequency of high score decreases a lot. We can observe that there is a slope between low and high scores. In this



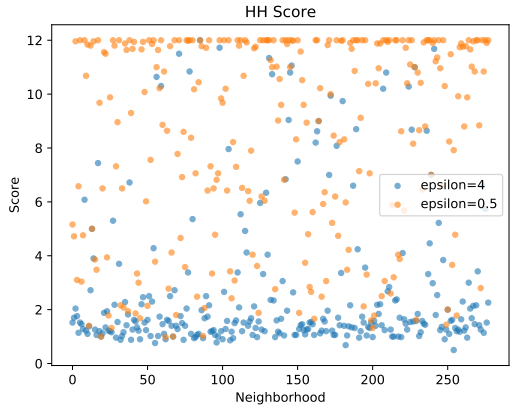
(a) JSD Score



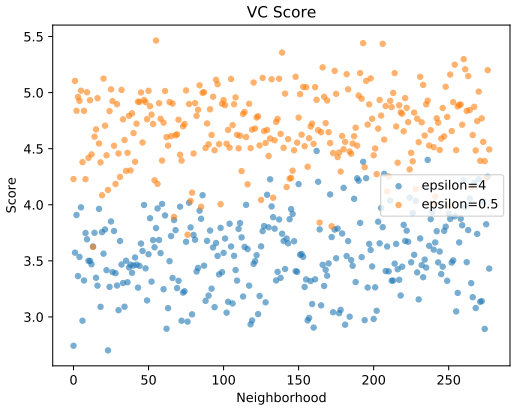
(b) BP Score



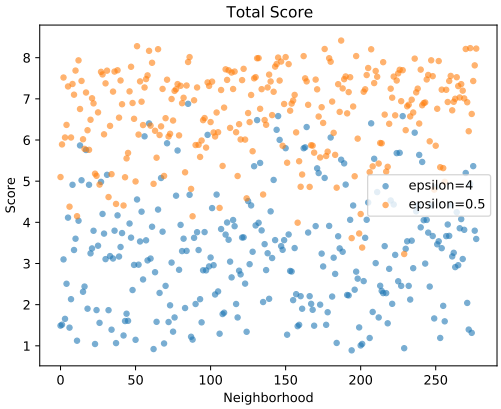
(c) JD Score



(d) HH Score



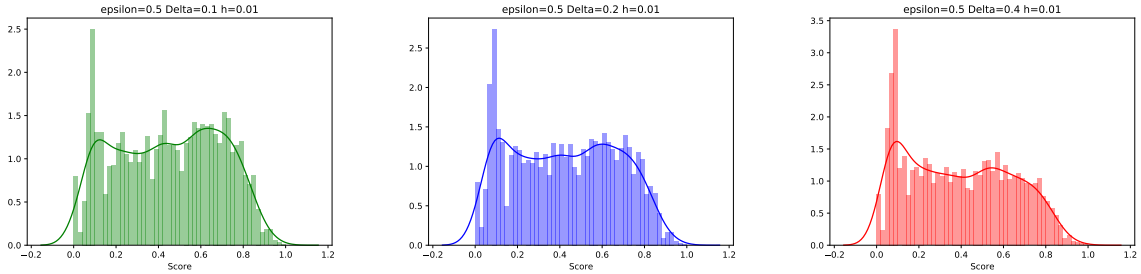
(e) VC Score



(f) Total Score

Figure 2: Sum of the scores in 12 months (y-axis) of each metric over all neighborhoods (x-axis). The orange points represent the poor quality privatized data-set ($\epsilon = 0.5$) and the blue points represent the moderate quality data-set($\epsilon = 4$)

condition, the discriminative power of HH metric is not enough. Therefore, we recommend to set the $\Delta \leq 0.2$ to make the metric effective.



(a) Delta=0.1 (b) Delta=0.2 (c) Delta=0.4

Figure 3: Effect of Δ on HH score on the poor quality privatized data-set ($\epsilon = 0.5$). Let $k = \lfloor \frac{1}{10} \sum_{i \in \mathcal{I}} X^{(i)} \rfloor, h = 0.01$.

3.2 Description of Discriminative Power

In this subsection, we briefly outline the capabilities and limitations of our final metric with respect to its discriminative power, i.e., how well it can distinguish between a privatized data and the ground truth.

3.2.1 Capabilities

- Our metric identifies disparities between the distribution of high frequency record types by JSD and Heavy Hitter.
- Our metric adaptively penalizes spurious record types that appear to be high frequency in the privatized data caused by rare large noise values by JD.
- Our metric specifically penalizes positive or negative bias in total record counts by BP.
- Our metric captures the correlation between any two record types (or two numerical features) in any two areas so that provides a broad pattern to evaluate the privatized data by HC.
- Our metric can measure the trends of record types (or two numerical features) across time by VC.
- The weights of the metrics can be flexibly tuned for distinct analysis purposes.
- Our metric can respond to small changes in the value of epsilon, and allows us to meaningfully understand the impact of those changes on different aspects or dimensions of data.

3.2.2 Limitations

- Our metric does not capture the relative ranking of high frequency record types. It is not motivated to know which record type causes the loss.

- When calculating scores of the JSD metric, less frequent record types are discarded during the frequency thresholding and numerical features need to be partitioned into bins. (The HC and VC metrics do not have these limitations.)

3.3 Description of Coverage

In this subsection, we briefly outline the capabilities and limitations of our final metric with respect to its coverage, i.e., how well it represents a breadth of possible use cases.

3.3.1 Capabilities

- Our final metric has five components that provide the reader with various perspectives to evaluate the privatized data. Those components are commonly used in machine learning tasks, social sciences, etc.
- Our metric can cover applications that study large patterns/trends across time or geography by HC and VC.
- Our metric not only counts the overall distribution difference (by JSD), but also focus on the record types with relatively high frequency (by HH) and close correlation (by HC).
- Our metric maintains accuracy across repeated numerical operations over privatized data by HC. The HC metric finds the most close correlations and evaluates how those correlations change in the privatized data.

3.3.2 Limitations

- Our metric does not cover more complex analytics (ex: regression, classification) which may have different sensitivities to added noise.
- Our metric attaches less importance to the infrequent record types. In fact, it is difficult to guarantee both privacy and accuracy of them since even a small amount of additional noise will largely affect the their frequencies.

3.4 Scalability/Feasibility

Suppose the data contains $|\mathcal{P}|$ neighbourhoods (map segments), $|\mathcal{I}|$ incidents and M months (time segments). In the 2019 Baltimore Police Incident and 911 data-set, $|\mathcal{P}| = 277, |\mathcal{I}| = 174, M = 12$. Same as JSD, the JD, HH and VC metrics are constant time operation in terms of the number of incidents, map segments and time segments. Computing them needs a time complexity of $O(M\,|\mathcal{P}|\,|\mathcal{I}|)$. Moreover, in this data-set, computing HC also costs $O(M\,|\mathcal{P}|\,|\mathcal{I}|)$ time. In our experiments, they can be computed within seconds on a typical laptop.

3.5 Generalizability, Alternate Use Cases

In this report, we have demonstrated the definition of our metrics including motivation, formulation, evaluation and how our four additional metrics can apply to event records. We also explore the impact of tuning parameters on metric scores. Finally, we summarize the discriminative power and coverage of the metrics. Here are a few additional examples how the metrics could be applied.

- Our five metrics can apply to various kinds of data with binning numerical features such as demographic data and financial data since "record type" can be easily defined in these data-sets.
- With default settings, our metrics do not capture trends across geography. However, it could be applied to capture more information by setting record types to be "Events that happened in region A (region A covers many neighbourhoods $A_1, A_2, \cdots$)". Then we can apply the metrics in the same way as discussed above.

References

[1] M. A. Andresen and N. Malleson. Crime seasonality and its variations across space. *Applied Geography*, 43:25–35, 2013.

[2] S. Dedeo, R. Hawkins, S. Klingenstein, and T. Hitchcock. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *CoRR*, abs/1302.0907, 2013.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *arXiv e-prints*, page arXiv:1406.2661, June 2014.

[4] C. R. Herrmann. The dynamics of robbery and violence hot spots. *Crime Science*, 4(1):1–14, 2015.

[5] S. J. Linning, M. A. Andresen, and P. J. Brantingham. Crime seasonality: Examining the temporal fluctuations of property crime in cities with varying climates. *International journal of offender therapy and comparative criminology*, 61(16):1866–1891, 2017.