Multiparty Computation with Statistical Input Confidentiality via Randomized Response

Josep Domingo-Ferrer, Rafael Mulero-Vellido, and Jordi Soria-Comas

Universitat Rovira i Virgili
Dept. of Computer Science and Mathematics
UNESCO Chair in Data Privacy
CYBERCAT-Center for Cybersecurity Research of Catalonia
Av. Països Catalans 26, 43007 Tarragona, Catalonia
{josep.domingo,rafael.mulero,jordi.soria}@urv.cat

Abstract. We explore a setting in which a number of subjects want to compute on their pooled data while keeping the statistical confidentiality of their input. Statistical confidentiality is different from the cryptographic confidentiality guaranteed by cryptographic multiparty secure computation: whereas in the latter nothing is disclosed about the input, in statistical input confidentiality a noise-added version of the input is disclosed, which allows more flexible computations. We propose a protocol based on local anonymization via randomized response, whereby the empirical distribution of the data of the subjects is approximated. From that distribution, most statistical calculations can be approximated as well. Regarding the accuracy of the approximation, *ceteris paribus* it improves with the number of subjects. Large dimensionality (that is, a large number of attributes) decreases accuracy and we propose a strategy to mitigate the dimensionality problem. We show how to characterize the privacy guarantee for each subject in terms of differential privacy. Experimental work is reported on the attained accuracy as a function of the number of respondents, number of attributes and randomized response parameters.

Keywords: Multiparty anonymous computation; randomized response; local anonymization; big data; privacy.

1 Introduction

There are several situations in which a number of distrusting parties wish to collaborate at evaluating functions that take as inputs private data from each party, in such a way that the *privacy* of those inputs is preserved. Two different notions of input privacy are conceivable:

- *Cryptographic input confidentiality.* The input of each party should not be disclosed to the other parties.
- *Statistical input confidentiality.* A noise-added version of the input of each party is disclosed.

Multiparty computation with cryptographic input confidentiality can be easily motivated with the following example. Each of a set of companies has collected experimental data at some cost and possibly under privacy pledge to its respondents/customers/patients. Thus, no company wishes to share its data set with any other company (as data have costed money and are industrial property). At the same time, feeling that better conclusions could be drawn from their pooled data sets than from a single data set, companies would like to engage in joint computation on their pooled data. If no third party trusted by all companies is available (that can receive all data sets in confidence, perform the computations on the pooled data and return the results to all companies), this scenario is handled with secure multiparty computation [1, 3, 14].

The problem of multiparty computation in the above cryptographic sense is that a different protocol is needed for each type of required computation. This hampers exploratory analysis, which is more and more important in our big data world. Multiparty computation with statistical input confidentiality can be far more flexible at the cost of providing somewhat weaker input confidentiality. Virtually any statistical computation can be performed without requiring specific protocols. Furthermore, the set of collaborating parties can be much larger than in cryptographic multiparty computation: there can be as many parties as respondents in a data set, with each party holding just her own record.

Contribution and plan of this paper

In this paper, we propose an approach for multiparty computation with statistical input confidentiality based on randomized response [2, 7, 13]. Specifically, local anonymization via randomized response is used by the collaborating subjects to approximate the empirical distribution of their pooled data. From that distribution, most statistical calculations can be approximated as well. Furthermore, we show how to characterize the privacy guarantee for each subject in terms of differential privacy.

Section 2 gives background on randomized response. In Section 3, we describe the proposed approach for multiparty computation with statistical input confidentiality based on randomized response. In Section 4 we propose solutions to mitigate the curse of dimensionality, that is, the decreasing accuracy of the approximated empirical distribution as the number of attributes increases. Section 5 establishes the privacy guarantees for subjects. Experimental work is reported in Section 6. Finally, conclusions and future research directions are gathered in Section 7.

2 Background on randomized response

Randomized response [7, 13] is a mechanism that respondents to a survey can use to protect their privacy when asked about the value of sensitive attribute (*e.g.* did you take drugs last month?). The interesting point is that the data collector can still estimate from the randomized responses the proportion of each of the possible *true* answers of the respondents. Randomized response is closely related to post-randomization (PRAM). They differ on who performs the randomization [10]: whereas in randomized response it is the respondent before delivering her response, in PRAM it is the data controller after collecting all responses (hence the name post-randomization).

Let us denote by X the attribute containing the answer to the sensitive question. If X can take r possible values, then the randomized response Y reported by the respondent instead of X follows a $r \times r$ matrix of probabilities

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & \vdots & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix} \tag{1}$$

where $p_{uv} = \Pr(Y = v|X = u)$, for $u, v \in \{1, \ldots, r\}$ denotes the probability that the randomized response is v when the respondent's true attribute value is u.

Let $\pi_1, \ldots, \pi_r$ be the proportions of respondents whose true values fall in each of the r categories of X and let $\lambda_v = \sum_{u=1}^{r} p_{uv}\pi_u$ for $v = 1, \ldots, r$, be the probability of the reported value Y being v. If we define $\lambda = (\lambda_1, \ldots, \lambda_r)^T$ and $\pi = (\pi_1, \ldots, \pi_r)^T$, it holds that $\lambda = \mathbf{P}^T\pi$. Furthermore, if $\hat{\lambda}$ is the vector of sample proportions corresponding to λ and $\mathbf{P}$ is nonsingular, in Chapter 3.3 of [2] it is proven that an unbiased estimator π can be computed as

$$\hat{\pi} = (\mathbf{P}^T)^{-1}\hat{\lambda} \tag{2}$$

and they also provide an unbiased estimator of the dispersion matrix. In particular, the larger the off-diagonal probability mass in $\mathbf{P}$, the more dispersion (and the more respondent protection).

3 Randomized response to achieve multiparty computation with statistical input confidentiality

Assume n subjects $i = 1, \ldots n$ each holding one record $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ containing the values for m attributes. These subjects want to engage in secure multiparty computation with statistical input confidentiality on the data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ that would be formed by their respective records. Statistical input confidentiality means that no subject wants to disclose her true record to the other subjects, even if she is ready to disclose a randomized version of it.

A possible way is for subjects to approximate the empirical distribution of $\mathbf{X}$ via randomized response. Once they have a reasonably good approximation of that distribution, most statistical calculations on $\mathbf{X}$ can be approximately computed based on the data set's approximate empirical distribution.

A first naive solution is for each subject i to separately deal with each attribute value x_{ij} for $j = 1, \ldots, m$ via randomized response. If the j-th attribute A_j can take r_j different values, then an $r_j \times r_j$ probability matrix $\mathbf{P}_j$ (see Expression (1)) can be used for each subject to report a randomized value y_{ij} for A_j instead of her true value x_{ij}.

As mentioned in Section 2, this would allow all subjects to approximate the *marginal* empirical distribution $\pi^j = (\pi^j_1, \ldots, \pi^j_{r_j})$ of each attribute A_j as

$$\hat{\pi}^j = ((\mathbf{P}^j)^T)^{-1}\hat{\lambda}^j$$

where $\lambda^j = (\mathbf{P}^j)^T \pi^j$.

The problem is that approximating the marginal empirical distributions of attributes does not yield an approximation of the joint empirical distribution of $\mathbf{X}$.

To approximate the joint distribution of $\mathbf{X}$ via randomized response, subjects must report their randomized response for the value of $A_1 \times A_2 \times \ldots \times A_m$ and proceed as above.

Once the empirical distribution of $\mathbf{X}$ has been approximated, the approximation can be made public and any subjects can perform statistical computations on it. At the same time, all subjects have preserved the confidentiality of their inputs.

However, this only works well if the number of subjects n is much larger than the number of possible values of the above Cartesian product, that is

$$n \gg |A_1| \times |A_2| \times \ldots \times |A_m|. \tag{3}$$

Otherwise, many elements of the Cartesian product are likely to have zero frequency in the empirical distribution of the reported values. Such a reported sparse distribution is unlikely to constitute a good approximation of the true empirical distribution of $\mathbf{X}$.

4 Mitigating the curse of dimensionality

If Constraint (3) does not hold, there are two alternatives:

1. Attempt to partition the set of attributes into clusters $C_1, \ldots, C_l$, for some l, such that

   $$\bigcup_{i=1}^{l} C_i = \{A_1, \ldots, A_m\}$$

   $C_i \cap C_j = \emptyset$ for $i \neq j$, and the attributes within each cluster are highly dependent/correlated and the attributes belonging to different clusters are weakly dependent or even independent. In this way, only the joint distribution of attributes within each cluster needs to be approximated. Therefore, Constraint (3) is relaxed to

   $$n \gg \max_i \prod_{A_j \in C_i} |A_j|, \tag{4}$$

   which is easier to satisfy. Furthermore, if within a cluster two or more attributes are very highly correlated, only one of them needs to be taken into account in the joint distribution approximation, with the rest being re-computed based on the approximated representative attribute. This may reduce the size of attribute clusters even more, and hence the bound on the right-hand side of Inequality (4).

2. If the above attribute clustering is not feasible (because all pairs of attributes are significantly dependent/correlated), an alternative solution is to coarsen the values of attributes $A_1, \ldots, A_m$, in such a way to reduce the right-hand side of Inequality (3). This obviously will reduce the accuracy of the approximation to the empirical distribution. Hence, it should be only used as a fallback solution.

We now describe in more detail attribute clustering:

1. Compute an approximation of all bivariate empirical distributions, by using randomized response on the Cartesian product $A_i \times A_j$, for all pairs (A_i, A_j) of attributes.
2. Construct a complete graph such that:
   (a) Nodes are attributes $A_1, \ldots, A_m$.
   (b) The edge between each pair of attributes A_i and A_j is labeled with a measure of independence between A_i and A_j.
3. Cluster attributes according to their distances in the graph. A possibility is to use the power iteration clustering (PIC) algorithm [8].

The specific measure of independence to be used must take into account the type of the attributes, as follows. If A_i and A_j are numerical and/or ordinal, we can take as a measure of independence

$$1/|r_{ij}|, \tag{5}$$

where r_{ij} is Pearson's correlation coefficient between A_i and A_j.

If one of A_i and A_j is nominal (without an order relationship between its possible values) and the other is nominal or ordinal, we can take as a measure of independence

$$1/V_{ij}, \tag{6}$$

where V_{ij} is Cramér's V statistic [4], that gives a value between 0 and 1, with 0 meaning complete independence between A_i and A_j and 1 meaning complete dependence. Cramér's V_{ij} is computed as

$$V_{ij} = \sqrt{\frac{\chi^2_{ij}/n}{\min(c_i - 1, c_j - 1)}},$$

where c_i is the number of categories of A_i, c_j is the number of categories of A_j, n is the total number of subjects/records and χ^2_{ij} is the chi-squared independence statistic defined as

$$\chi^2_{ij} = \sum_{a=1}^{c_i} \sum_{b=1}^{c_j} \frac{(o^{ij}_{ab} - e^{ij}_{ab})^2}{f^{j}_{ab}}, \tag{7}$$

with o_{ab}^{ij} the observed frequency of the combination $(A_i = a, A_j = b)$ and e_{ab}^{ij} the expected frequency of that combination under the independence assumption for A_i and A_j. This expected frequency is computed as

$$e_{ab}^{ij} = \frac{n_a^i n_b^j}{n},$$

where n_a^i and n_b^j are, respectively, the number of subjects who have reported $A_i = a$ and $A_j = b$.

Finally, if one of A_i, A_j is nominal and the other is numerical, the latter must be discretized, for example by rounding or by replacing values by intervals. After that, the contingency table between A_i and A_j can be constructed, and the measure of independence given by Expression (6) can be computed.

Since the denominators in Expressions (5) and (6) are bounded in $[0, 1]$, the outputs of both expressions are comparable when trying to cluster the nodes in the graph.

5 Privacy guarantees

The confidentiality guarantee given by randomized response results from the fact that each individual may misrepresent her data by randomly drawing from a previously fixed distribution. Thus, given the individual's randomized response, we are uncertain about what her true response would have been.

In spite of the previous intrinsic guarantee of randomized response, given the popularity of differential privacy, it may be interesting to analyze the privacy guarantees of randomized response in terms of differential privacy.

A randomized query function κ gives ϵ-differential privacy [6] if, for all data sets D_1, D_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(\kappa)$, it holds

$$\Pr(\kappa(D_1) \in S) \leq \exp(\epsilon) \times \Pr(\kappa(D_2) \in S. \tag{8}$$

In plain words, the presence or absence of any single record is not noticeable (up to $\exp(\epsilon)$) when seeing the outcome of the query. Hence, this outcome can be disclosed without impairing the privacy of any of the potential respondents whose records might be in the data set. A usual mechanism to satisfy Inequality (8) is to add noise to the true outcome of the query, in order to obtain an outcome of κ that is a noise-added version of the true outcome. The smaller ϵ, the more noise is needed to make queries on D_1 and D_2 indistinguishable up to $\exp(\epsilon)$.

In [11,12], a connection between randomized response and differential privacy is established: randomized response is ϵ-differentially private if

$$e^{\epsilon} \geq \max_{v=1,\ldots,r} \frac{\max_{u=1,\ldots,r} p_{uv}}{\min_{u=1,\ldots,r} p_{uv}}. \tag{9}$$

The rationale is that the values in each column v $(v \in \{1, \ldots, r\})$ of matrix $\mathbf{P}$ correspond to the probabilities of the reported value being $Y = v$, given that the true value is $X = u$ for $u \in \{1, \dots, r\}$. Differential privacy requires that the maximum ratio between the probabilities in a column be bounded by e^{ϵ}, so that the influence of the true value X on the reported value Y is limited. Thus, the reported value can be released with limited disclosure of the true value.

6 Empirical results

Randomized response is usually performed independently for each attribute. As a result, marginal distributions are well preserved but multivariate distributions are not (see Section 3). If the aim is to preserve the joint distribution of all attributes $A_1, \dots, A_m$, we should perform a single randomized response over $A_1 \times \dots \times A_m$. However, this is usually unfeasible due to the curse of dimensionality explained at the end of Section 3.

In this section we empirically evaluate the technique proposed to solve the previous difficulties, which is based on clustering attributes in several groups (so that attributes in different groups have low correlation) and running randomized response independently for each of the groups.

6.1 Dataset

Experiments are based on the Adult dataset. This is a data set with over 32,500 records and a combination of numerical and categorical attributes. For the experiments, we only take categorical attributes into account. These attributes are: Work-class (with 9 categories), Education (with 16 categories), Marital-status (with 7 categories), Occupation (with 15 categories), Relationship (with 6 categories), Race (with 5 categories), Sex (with 2 categories), Native-country (with 42 categories) and Income (with 2 categories).

6.2 Methodology

In the test dataset there are 76,204,800 possible combinations of attribute values. This makes the clustering approach to randomized response (see Section 4) indispensable to get useful results.

We clustered attributes using the PIC algorithm with $k = 3$. The graph was constructed following the procedure described in Section 4. That is, for each pair of attributes:

- we obtained randomized responses on the Cartesian product of the two attributes, and
- the distance between the nodes corresponding to these attributes was computed as $1/V$, where V is Cramer's V statistic computed over the randomized data.

To avoid having to manually build the randomized response matrix for a Cartesian product (which would be burdensome because there are potentially many categories), we automatically build the matrix as follows:

- The probability of the cells in the main diagonal is set to a fixed value $p \in [0, 1]$.
- The probability of the off-diagonal cells is set to be inversely proportional to the number of attribute changes that the cell accounts for:

  $$p_{uv} = (1-p)\frac{d_{uv}}{\sum_k d_{uk}},$$

  where d_{uv} is the inverse of the number of attributes whose values differ between u and v.

It would seem that, to compute $\sum_k d_{uk}$, we need to loop through each of the possible combinations of attribute values and do the sum. This is not necessary, because we can compute that sum by just considering the number of attributes whose categories change between u and k:

$$\begin{aligned} \sum_k d_{uk} = &\sum_{a_1 \in \{1,\ldots,r\}} (|A_{a_1}| - 1) \\ +\frac{1}{2} &\sum_{1 \le a_1 < a_2 \le m} (|A_{a_1}| - 1)(|A_{a_2}| - 1) \\ &\ldots \\ +\frac{1}{m} &\sum_{1 \le a_1 < \ldots < a_m \le m} (|A_{a_1}| - 1)\ldots(|A_{a_m}| - 1), \end{aligned} \qquad (10)$$

where the first sum on the right-hand side of Expression (10) corresponds to changes of a single attribute (for each category a_1, we count the number $|A_{a_1}| - 1$ of alternative categories of the attribute A_{a_1} to which a_1 belongs); the second sum corresponds to the changes of two attributes and uses the same notation; and so on. We can rewrite Expression (10) in a more compact way as:

$$\sum_k d_{uk} = \sum_{w=1}^{r} \frac{1}{r} \sum_{1 \le a_1 < \ldots < a_w \le m} (|A_{a_1}| - 1)\ldots(|A_{a_w}| - 1). \qquad (11)$$

Notice that Equation (11) does not depend on u. This means that, the value of $\sum_k d_{uk}$ is constant across all u; thus, we only need to do this computation once.

6.3 Risk evaluation

Table 1 shows the levels ϵ_1, ϵ_2 and ϵ_3 of differential privacy attained in each of the three clusters for $p = 0.9$, $p = 0.8$, and $p = 0.7$, respectively. The overall level of differential privacy is the sum $\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3$, that is, the sum of levels across the three clusters. Note that the composition of clusters changed when p changed.

The overall ϵ decreases when p decreases, which means that privacy increases as p decreases. This was to be expected: the less probability mass in the diagonal of the randomized response matrix, the more privacy. It must be pointed out here that the randomized response matrix was not designed with differential privacy in mind; that is, we did not seek to minimize Equation (9). Seeking such minimization would yield yet smaller ϵ, but would probably impinge on utility.

Table 1. Cluster evaluation results

$p = 0.9$			$p = 0.8$			$p = 0.7$		
Cluster	Attributes	Epsilon	Cluster	Attributes	Epsilon	Cluster	Attributes	Epsilon
C1	2	7.309	C1	4	11.844	C1	3	7.337
C2	2	7.735	C2	2	5.171	C2	3	8.569
C3	5	11.975	C3	3	6.234	C3	3	5.695

6.4 Utility evaluation

We measured the utility of the generated randomized dataset by measuring the difference in the number of records for combinations of attribute values between the original dataset (X) and the randomized dataset (Y). In particular, we computed the difference for all combinations of values of two attributes. If $a_i \in A_i$ and $a_j \in A_j$ are attribute values, we computed the relative error as:

$$e_{ij} = \frac{X_{a_i a_j} - Y_{a_i a_j}}{X_{a_i a_j}} \times 100,$$

where $X_{a_i a_j}$ and $Y_{a_i a_j}$ are the number of records with attribute values a_i and a_j in the original dataset and in the randomized dataset.

Figures 1, 2 and 3 show the values of $e_{a_i a_j}$ when randomized response was run with parameter p equal to 0.9, 0.8 and 0.7, respectively. Both, in the x-axis and in the y-axis, we represent all possible attribute values; that is, in each axis we represent the set $\{a : a \in A_i, 1 \leq i \leq N\}$. In the intersection between column a_i and row a_j, we represent e_{ij}. From the histograms, we observe that for higher values of p the difference between the original and the randomized dataset is smaller. The light gray squares on the bottom-left top-right diagonal of the histogram represent the cases in which both attribute values are categories of the same attribute (which are impossible combinations, as we are interested in combinations of values of two attributes).

7 Conclusions and future research

We have proposed a methodology to perform computations on a dataset that offers statistical input confidentiality. Each respondent can keep her input (true answer) confidential by giving to the data collector a reported answer via randomized reponse. Doing so still allows the data collector to approximate the empirical distribution of the pooled true answers of the set of respondents. After

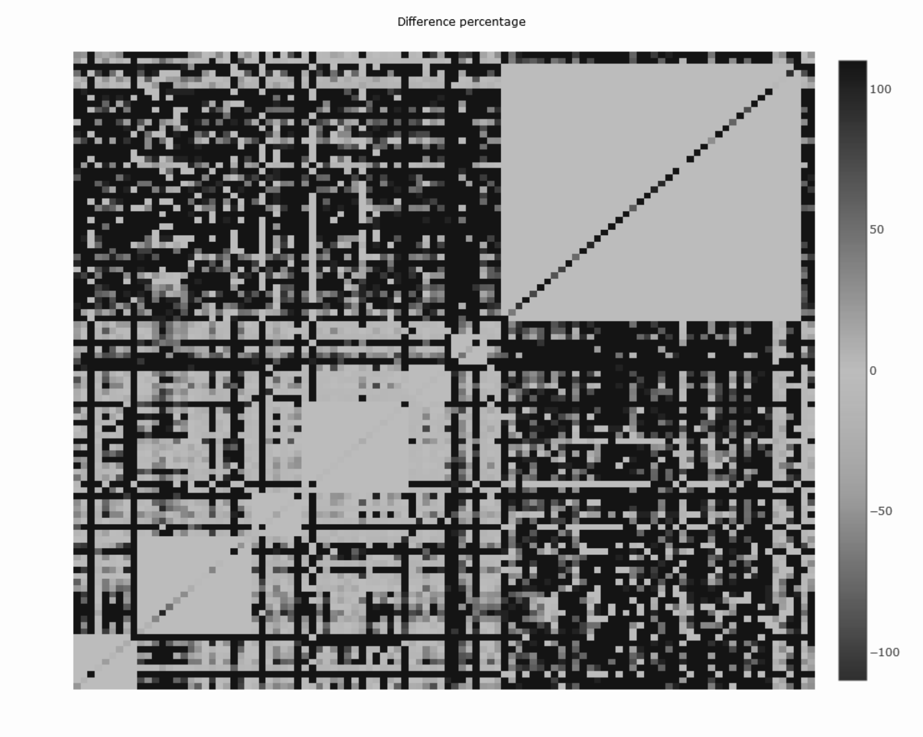

Fig. 1. Relative errors in randomized response for $p = 0.9$

that, statistical computations can be performed on the approximated distribution.

Randomized response is only feasible when the number of possible categories is small compared to the number of records. For this reason, this technique is usually applied on an attribute-by-attribute basis. However, separately dealing with each attribute does not allow approximating the joint empirical distribution of the data. In this work, we have proposed a way to overcome this issue. We cluster attributes so that attributes in different clusters are independent (or nearly so) from each other. In this way, we can perform randomized response independently for each cluster without severely impairing the approximation of the joint empirical distribution.

We have experimentally validated the proposed methodology on a standard data set. In the experimental section, we have also described how to automatically construct a randomized response matrix that is based on the number of categories that are altered in the randomization process.

As future research, we plan to develop a new clustering procedure that requires less information (the current procedure needs users to run randomized response for each pair of attributes to measure the dependency between them).

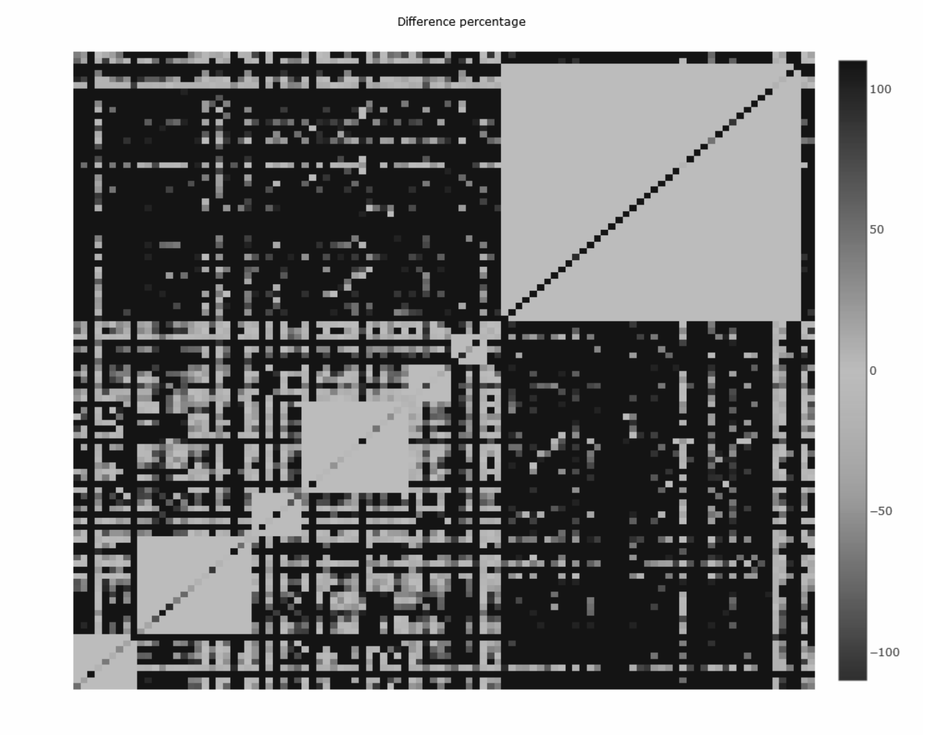

Fig. 2. Relative errors in randomized response for $p = 0.8$

We will also investigate a quantification of the privacy guarantees that does not depend on differential privacy.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: European Commission (H2020-700540 “CANVAS”), Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer) and Spanish Government (projects TIN2014-57364-C2-1-R “SmartGlacis” and TIN2015-70054-REDC). The views in this paper are the authors’ own and do not necessarily reflect the views of UNESCO or any of the funders.

References

1. M. Ben-Or, S. Goldwasser and A. Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In: *STOC*, 1988.
2. A. Chaudhuri and R. Mukerjee. *Randomized Response: Theory and Techniques.* Marcel Dekker, 1988.

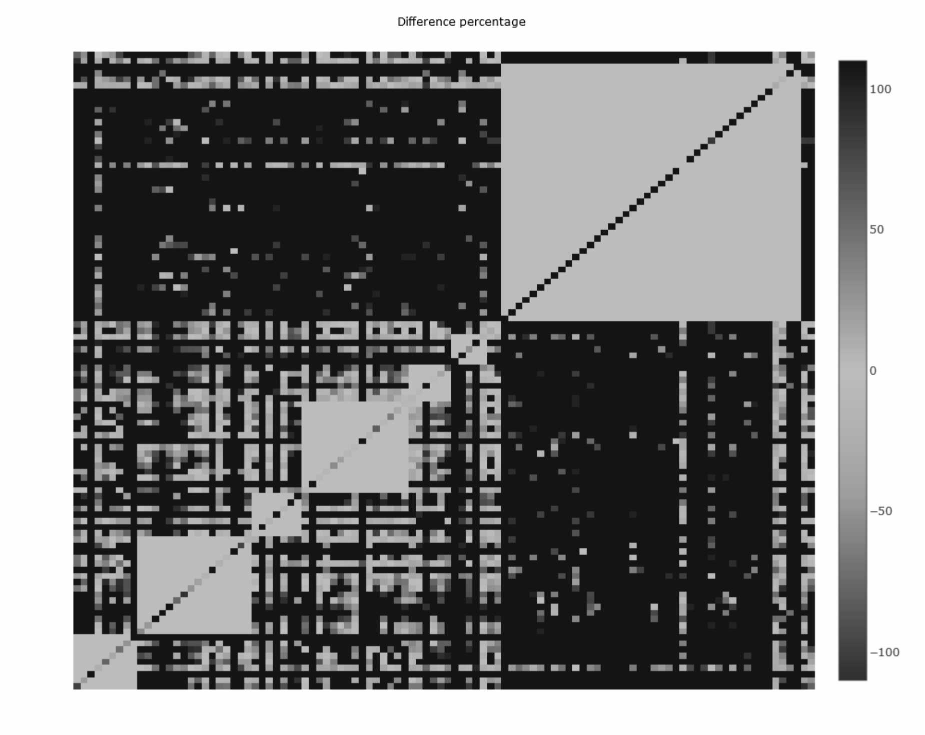

Fig. 3. Relative errors in randomized response for $p = 0.7$

3. D. Chaum, C. Crépeau and I. Damgaard. Multiparty unconditionally secure protocols. In *STOC*, 1988.
4. H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.
5. J. Domingo-Ferrer and J. Soria-Comas. Connecting randomized response, post-randomization, differential privacy and t-closeness via deniability and permutation. Arxiv preprint, March 6, 2018. `https://arxiv.org/pdf/1803.02139v1.pdf`.
6. C. Dwork. Differential privacy. In: *ICALP 2006*, LNCS 4052, pp. 1-12, 2006.
7. B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons and D. G. Horvitz. The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64(326):520-539, 1969.
8. F. Lin and W. W. Cohen. Power iteration clustering. In *Proc. of the 27th International Conference on Machine Learning-ICML 2010*, 2010.
9. C. E. Shannon. Communication theory of secrecy systems. *Bell Labs Technical Journal*, 28(4):656-715, 1949.
10. A. Van den Hout. Analyzing Misclassified Data: Randomized Response and Post Randomization. Ph.D. Thesis, University of Utrecht, 2004.
11. Y. Wang, X. Wu, and D. Hu. Using randomized response for differential privacy preserving data collection. In *Technical Report DPL-2014-003*. University of Arkansas, 2014.

12. Y. Wang, X. Wu, and D. Hu. Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT 2016 Joint Conference*, Bordeaux, France, 2016.
13. S. L. Warner. Randomised response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63-69, 1965.
14. A. Yao. Protocols for secure computations. In: *FOCS*, 1982.