# Multi-Dimensional Randomized Response

Josep Domingo-Ferrer, *Fellow, IEEE,* and Jordi Soria-Comas

**Abstract**—In our data world, a host of not necessarily trusted controllers gather data on individual subjects. To preserve her privacy and, more generally, her informational self-determination, the individual has to be empowered by giving her agency on her own data. Maximum agency is afforded by local anonymization, that allows each individual to anonymize her own data before handing them to the data controller. Randomized response (RR) is a local anonymization approach able to yield multi-dimensional full sets of anonymized microdata that are valid for exploratory analysis and machine learning. This is so because an unbiased estimate of the distribution of the true data of individuals can be obtained from their pooled randomized data. Furthermore, RR offers rigorous privacy guarantees. The main weakness of RR is the curse of dimensionality when applied to several attributes: as the number of attributes grows, the accuracy of the estimated true data distribution quickly degrades. We propose several complementary approaches to mitigate the dimensionality problem. First, we present two basic protocols, separate RR on each attribute and joint RR for all attributes, and discuss their limitations. Then we introduce an algorithm to form clusters of attributes so that attributes in different clusters can be viewed as independent and joint RR can be performed *within* each cluster. After that, we introduce an adjustment algorithm for the randomized data set that repairs some of the accuracy loss due to assuming independence between attributes when using RR separately on each attribute or due to assuming independence between clusters in cluster-wise RR. We also present empirical work to illustrate the proposed methods.

**Index Terms**—Privacy preserving data publishing, randomized response, curse of dimensionality, local anonymization, multivariate data, differential privacy

Local anonymization is a paradigm in which each individual anonymizes her data before handing them to the data controller, thereby giving maximum agency to the individual. Several masking methods coming from statistical disclosure control (SDC [8]) can be applied locally, including generalization/recoding and noise addition. On the other hand, there are methods specifically designed for local anonymization that, in addition to helping subjects hide their responses, allow the data controller to get an accurate estimation of the distribution of responses for groups of subjects (for example, randomized response [11], [7]). Also, a number of methods have been proposed to obtain differentially private (DP) preselected statistics via local anonymization, like RAPPOR [6] and local DP [1].

While most of the local anonymization approaches were designed to obtain statistics on the set of individuals who contribute their locally anonymized input, randomized response (RR) has the attractive feature of being able to output multi-dimensional full sets of anonymized microdata (individual records) that are valid for exploratory analysis. Indeed, an unbiased estimate of the distribution of the original microdata (corresponding to the true attribute values of individuals) can be obtained from the empirical distribution of the released randomized microdata. Most exploratory analyses and statistical calculations can be performed based on this estimated distribution, including what [4] call multi-party computation with statistical input confidentiality. It is even possible to re-create a synthetic estimate of the original data set by repeating each combination of attribute values as many times as dictated by its frequency in the estimated joint distribution.

Furthermore, the privacy guarantees afforded by RR are easily expressible in terms of rigorous privacy models such as differential privacy [10] or information-theoretic secrecy [3].

Unfortunately, the picture is not as rosy as suggested by the two previous paragraphs. Like so many methods, randomized response suffers from the curse of dimensionality:

- Applying RR simultaneously to a set of attributes amounts to applying it to the Cartesian product of those attributes, which has a number of possible categories that grows exponentially with the number of attributes. Unless the number of individuals providing input is much greater than the number of categories, the accuracy and hence the utility of the estimated distribution of the original data will be poor.
- RR can certainly be applied separately to each single attribute. Nevertheless, by doing so the ability to estimate the joint distribution of the original data based on the randomized data is lost; only the marginal distributions of attributes can be estimated. This also entails a loss of accuracy and hence a loss of utility of the estimated distributions.

Dimensionality issues are common to all data anonymization techniques. However, effectively dealing with them is tougher in the local anonymization paradigm because the global picture of the data is

*J. Domingo-Ferrer is with Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics, UNESCO Chair in Data Privacy, CYBERCAT-Center for Cybersecurity Research of Catalonia, Av. Països Catalans 26, 43007 Tarragona, Catalonia. E-mail josep.domingo@urv.cat*

*J. Soria-Comas is with the Catalan Data Protection Authority, Rosselló 214, Esc. A, 1r 1a, 08008 Barcelona, Catalonia. E-mail jordi.soria@gencat.cat*

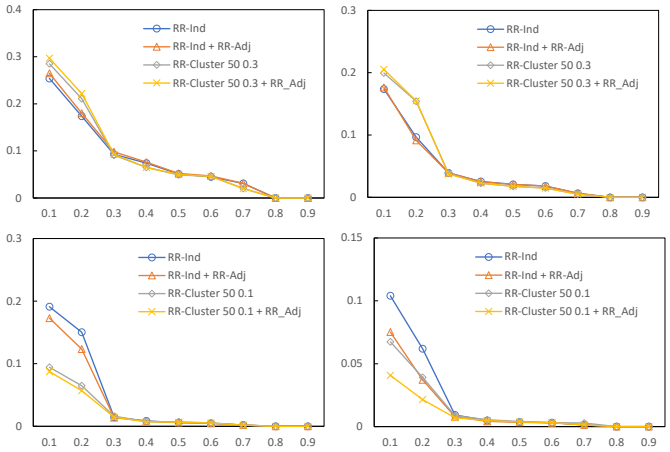



Fig. 1. Relative error in count queries on the Adult data set for different randomizations and coverages

missing. Research in this topic is ample. In particular, in differential privacy and local differential privacy, many strategies have been proposed to cope with a high number of attributes, such as dealing with $k$-way marginals [12], taking advantage of sparsity [1] or taking advantage of the dependence between attributes [9].

In this paper, we propose several complementary approaches to mitigate the curse of dimensionality in randomized response:

- We first present two basic protocols for RR and discuss their limitations: one performs separate RR for each attribute (RR-IND) and the other joint RR for all attributes (RR-JOINT).
- We then describe an intermediate approach (RR-CLUSTER) based on identifying clusters of attributes such that attributes in different clusters can be viewed as (nearly) independent. In this way, joint RR can be performed for the attributes *within* each cluster. Since our clustering algorithm requires as input the dependences between attributes, we present several methods for assessing these dependences in an RR scenario in which the true data of each individual must stay confidential to that individual.
- After that, we introduce an adjustment algorithm (RR-ADJ) of the randomized data set that "repairs" some of the accuracy loss incurred by assuming independence between attributes when using RR-IND or assuming independence between clusters when using RR-CLUSTER. The idea is to leverage the information about the relation between attributes that remains in the randomized data set in order to obtain more accurate estimates of the joint distribution.

The graphs in Figure 1 show the relative error in count queries on the Adult data set, for various randomization levels, denoted by the probability $p$ of *not* changing the original values: $p = 0.1$ (top left, high randomization), $p = 0.3$ (top right), $p = 0.5$ (bottom left) and $p = 0.7$ (bottom right, low randomization). In each graph, the $x$-axis represents the coverage (size of the set of value combinations to be counted). It can be seen that, for low to moderate randomizations ($p \geq 0.5$), combining RR-CLUSTER and RR-ADJ results in the lowest relative error between the counts on randomized data and the counts on original data. The reason is that, for moderately randomized multi-dimensional data, RR-ADJ can still reconstruct the joint distribution of the original attributes fairly well. In contrast, for high randomization ($p \leq 0.3$), the joint distribution cannot be reconstructed and RR-IND is best, albeit with a large relative error. Further details and empirical results can be found in [5].

## ACKNOWLEDGMENT AND DISCLAIMER

Thanks go to Rafael Mulero-Vellido for help in the empirical work and to Oriol Farràs for valuable comments. Partial support to this work has been received from the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer), and the Spanish Government (project RTI2018-095094-B-C21 "CONSENT"). The first author is with the UNESCO Chair in Data Privacy, but the views in this paper are the authors' own and are not necessarily shared by UNESCO.

## REFERENCES

[1] G. Cormode, T. Kulkarni and D. Srivastava. Marginal release under local differential privacy. In *SIGMOD 2018* (2018) ACM, pp. 131–146.

[2] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15 (1977) 429–444.

[3] J. Domingo-Ferrer. Big data anonymization requirements vs privacy models. In *15th Intl. Conference on Security and Cryptography - SECRYPT 2018* (2018) pp. 471–478.

[4] J. Domingo-Ferrer, R. Mulero-Vellido and J. Soria-Comas. Multiparty computation with statistical input confidentiality via randomized response. In *Privacy in Statistical Databases - PSD 2018* (2018) Springer, pp. 175–186.

[5] J. Domingo-Ferrer and J. Soria-Comas. Multi-dimensional randomized response. *IEEE Transactions on Knowledge and Data Engineering*, to appear. Online first since Dec. 18, 2020.

[6] Ú. Erlingsson, V. Pihur and A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proc. of the 2014 ACM CCS SIGSAC Conference on Computer and Communications-CCS 2014* (2014) ACM, pp. 1054–1067.

[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) (1969) 520–539.

[8] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P. P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[9] T. Wang, J. Blocki, N. Li and S. Jha. Locally differentially private protocols for frequency estimation. In *Proceedings of the 26th USENIX Security Symposium* (2017) ACM, pp. 729–745.

[10] Y. Wang, X. Wu and D. Hu. Using randomized response for differential privacy preserving data collection. In *Proceedings of the EDBT/ICDT 2016 Joint Conference* (Bordeaux, France, 2016).

[11] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association 60*, 309 (1965), 63–69.

[12] Z. Zhang, T. Wang, N. Li, S. He and J. Chen. CALM: Consistent adaptive local marginal for marginal release under local differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS'18)* (2018) ACM, pp. 212–229.