Now, Your Information is *Beyond* Enigmatic

Ruobin Gong is Assistant Professor of Statistics at Rutgers University. Her research interests lie at the theoretical foundations of Bayesian and generalized Bayesian methodologies, statistical modeling, inference, and computation with differentially private data, and ethical implications of aspects of modern data science. Ruobin received her PhD in statistics from Harvard University in 2018. She is currently an associate editor of the *Harvard Data Science Review.*

We've all been there: your airline company emails you and asks your opinion about their service. Still bummed out about last summer when they almost lost your luggage, you take up the invitation hoping to rant about it, only to find out three pages into the survey they want to know how old you are and what you do for a living.

Deep down inside, you know that telling the truth may eventually help the airline company understand your needs as a budding academic. In the summer months, we fly from conference to conference living out of a suitcase, and parting ways with it for even a single day will throw our talks, meetings, and travel plans into havoc. But still, that doesn't feel reason enough to warrant an all-out confession. Even if telling the truth may potentially do you (and your fellow academics) a big favor, it's not worth the risk to expose your most personal information. After all, who can guarantee what I put into the survey will be used solely for the betterment of my or others' experience, and never against my *rights* – to anonymity, confidentiality, and privacy? So, you put down some made-up age and occupation in the answer box, and quickly move on.

When it comes to surveys, the feeling of distrust that a respondent harbors against the surveyor corrodes the quality of the data. To refuse participation biases the survey at the sampling stage, and erroneous answers further harms data informativeness. In marketing surveys, faults like this at most render a company clueless about the true opinions of its customers. Yet grievous consequences await when distrust imbues surveys of more substantial significance. In just about a month, the 2020 U.S. decennial Census, the most comprehensive enumeration of the living population of America, will hit the ground running. As mandated by the U.S. Constitution, data obtained through the Census supply crucial factual evidence to economic and policy decisions. The Census serves as the basis for the apportionment of House seats, as well as the allocation of federal funding and resources (Sullivan, 2020). It is a massive and serious statistical undertaking.

The Census differs from all other surveys in one important aspect. By design, it should cover 100% of its target population, that is every single person living in the United States at the time it takes place. Therefore, when it comes to Census data releases, privacy protection carries insurmountable significance due to the sheer number of respondents involved. The 2020 Census made a revolutionary step forward by endorsing a new, and formal, standard for privacy protection, called differential privacy (Dwork et al., 2006; Abowd, 2018). Differential privacy draws a sharp distinction from the heuristic approach to privacy protection that traditional methods typically follow, such as full and partial suppression of data tables and swapping of individuals. It supplies a mathematical definition on what is meant by the *privacy* of data releases, which doubles as a metric to quantify the amount of privacy the data release gives away at most.

How does differential privacy work? Suppose that you're filling out the Census questionnaire, and let's denote your *true* answer as x_t. You are reminded of the risk of a hypothetical privacy breach, and contemplate whether to put down instead a *fabricated* answer, say x_f. Your answer, together with billions of others', constitute the enormous Census database $\mathcal{D}$, which takes the value $\mathcal{D}(x_t)$ if you supplied the true answer, or $\mathcal{D}(x_f)$ if the fabricated one. (Let's say the others' answers, whether true or fabricated, are identical in $\mathcal{D}(x_t)$ and $\mathcal{D}(x_f)$.) Finally, the Census Bureau releases the database summary generated by a probabilistic algorithm, based on the observed (and confidential) database: $S = S(\mathcal{D})$.

Suppose an ill-minded hacker is eyeing the Census data, hoping to learn about your information. Looking at the released summary S, the hacker needs to discern between two possibilities: that the data you contributed was true (H) or fabricated ($\bar{H}$). If the algorithm that generated S is sufficiently private, the information contained in S (expressed in terms of probabilities) that can sufficiently discern H from $\bar{H}$ is limited. Precisely speaking, S is ϵ-differentially private if the log ratio of its probabilities evaluated under either hypothesis (i.e. their respective likelihoods) is bounded within the ϵ-neighborhood around zero:

$$\log \frac{P(S \mid H)}{P(S \mid \bar{H})} \in [-\epsilon, \epsilon], \qquad [1]$$

and that such is true for every respondent (including you) who contributes to the Census database $\mathcal{D}$. The ϵ here, called the privacy loss budget, controls the extent to which we are willing to tolerate

discernibility among hypotheses, or leak of information. The smaller the ϵ, the more stringent the bound becomes, and the less informative S is relative to the pair of hypotheses H and $\bar{H}$.

The above quantity looks like an incredibly simple, if not overly simple, metric to quantify the so-called "information" in S regarding H versus $\bar{H}$. But do take it seriously. The failure to maintain this log probability ratio at a small magnitude by its encrypted messages was the Achilles' heel of the Naval Enigma and the Tunny machines, a deadly giveaway that led to their heroic breaking by the genius scientists at Bletchley Park during World War II (McGrayne, 2011; Zabell, 2012, 2015). In I. J. Good's account of Alan Turing's statistical contribution during the war (Good, 1979), he defined the "weight of evidence" concerning a hypothesis H as against $\bar{H}$ provided by evidence S, written as $W(H/\bar{H}:S)$, a quantity that works out to be precisely the log probability ratio in [1]. For cipher machines such as the Enigma and the Tunny, S stands for the encrypted messages, and H, $\bar{H}$ are hypotheses concerning the different configurations of the cipher wheels. If a configuration hypothesis receives from S a disproportionately large weight of evidence relative to other hypotheses, there is reason to think that it may be the correct configuration. Turing called one unit of the log probability ratio in [1] a *natural ban*, which is equal to 4.34 units of *deciban* (ten times the base 10 logarithm of the probability ratio), the "smallest change in weight of evidence that is directly perceptible to human intuition" (Good, 1979, p394). Carrying over this calculation to the privacy context, a privacy loss budget ϵ set at or less than $1/4.34 \approx 0.23$ makes the hypotheses regarding the truthfulness of your input data probabilistically indiscernible, based on the differentially private release S. It would be fair to say, then, that the privacy algorithm behind the released summary S encrypts your personal information securely, a job much better done than the Enigma machine. In other words, your information is now *"beyond Enigmatic"!*

Differential privacy brings clarity to the meaning of privacy through a formal and verifiable definition, setting a rigorous standard for implementation, investigation and future improvement. It merits other benefits from a technical point of view. Data releases compliant with differential privacy are resistant to post-processing, and behave nicely under the compounding of multiple sources; see Dwork et al. (2014) for details and Wood et al. (2018) for an approachable introduction. Differential privacy further permits the transparent dissemination of the privacy algorithm without compromising the privacy guarantee, drawing an analogy with public-key encryption. For statisticians, this means that the data curator is free to publicize the inner specification of the privacy mechanism (as in the case of US Census Bureau, 2020), paving ways for statistical methods to account for its effect, and to maintain inferential validity based on private releases (Gong, 2019).

Just as no probabilistic promise is ever definitive, no privacy is absolute if we simultaneously demand to learn useful information. With differential privacy, however, the tradeoff between privacy and information is put in concrete terms. In the 2020 Census, we in America will collectively pay an *epsilon* price in privacy, in exchange for a large body of useful knowledge about this country we live in, and the people who live in it with us. The Census operationalizes democracy and equality through enabling fair and data-driven allocation of resources. A transparent and effective framework for privacy protection is yet another reason to overcome mistrust, and to actively and honestly participate. Time will tell whether the price we pay is money well spent.

References

Abowd, J.M. (2018). The US Census Bureau Adopts Differential Privacy. In *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867–2867. ACM.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference*, 265–284. Springer.

Dwork, C. and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. In *Foundations and Trends in Theoretical Computer Science,* **9**(3–4): 211–407.

Gong, R. (2019). Exact inference with approximate computation for differentially private data via perturbations. arXiv:1909.12237

Good, I. J. (1979). Studies in the History of Probability and Statistics XXXVII: A. M. Turing's Statistical Work in World War II. *Biometrika*, **66**(2): 393–396.

McGrayne, S. B. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy.* Yale University Press.

Sullivan, T. A. (2020). Coming to our Census: How social statistics underpin our democracy (and republic). *Harvard Data Science Review*, 2(1). https://hdsr.mitpress.mit.edu/pub/1g1cbvkv.

US Census Bureau (2020). 2010 Demonstration Data Products. https://www.census. gov/programs-surveys/decennial-census/2020-census/planning-management/ 2020-census-data-products/2010-demonstration-data-products.html. Accessed on Feb 02, 2020.

Wood, A., Nissim, K., et al. (2018). Differential Privacy: A Primer for a Non-Technical Audience. *Vanderbilt Journal of Entertainment & Technology Law* **21**(1): 209.

Zabell, S. (2012). Commentary on Alan M. Turing: The Applications of Probability to Cryptography. *Cryptologia*, **36**(3):191–214.

Zabell, S. (2015). Statistics at Bletchley Park. In Reeds, J. A., Diffie, W., and Field, J., eds, *Breaking Teleprinter Ciphers at Bletchley Park: An edition of I.J. Good, D. Michie and G. Timms: General report on Tunny with emphasis on statistical methods (1945)*, pages lxxv–ci. John Wiley & Sons.