

Replicable Privacy: Enabling Replicable Research on Sensitive Internet Data*

Suso B. Baleato,[†] James Honaker, Mercè Crosas, [**add yourself**][‡]

July 17, 2018

Abstract

The mass adoption of the Information and Communication Technologies (ICT) is increasing the demand for quantitative social science analysis to understand the effects, and the causes, of digitalisation. For example, while some see the ICT as a ‘liberation technology’ capable to empower citizens against autocrats, others have shown how the digital technologies can be used to consolidate the authoritarian rule as well (King et al, 2017).

In order to discern this questions, the quantitative approach requires an accurate description of the Internet, because this infrastructure is increasingly turning into the backbone of digitalisation. The limitations of the official statistics on Internet have hampered the capacity of social science to develop empirical analysis, because the provided precision is limited to country-year observations, and because the methodology used to estimate the statistics is subject to different bias sources. Those limitations have been recently solved using a remote-sensing method, recently featured in *SCIENCE* (Benitez-Baleato et al., 2015; Weidmann et al., 2016) which is capable to observe variations on Internet adoption inside countries, and with temporal precision below the monthly frequency. Instead of relying on the reports provided by telecommunication regulators, the remote-sensing approach relies on direct observation of the global Internet data flow. This allows measuring digitalisation also in areas where official statistics are not available and data cannot be retrieved in the field, such as authoritarian regimes or territories experiencing long-term political violence.

While this new method can enable highly disaggregated analysis, the achieved precision can break the privacy and data protection law. The regulation of most developed countries could introduce

*Prepared for the 35th Annual PolMeth Meeting at Brigham Young University, July 2018.

[†]Corresponding author. Contact: <susobaleato@iq.harvard.edu>. Postdoctoral Fellow at Harvard University Institute of Quantitative Social Science (IQSS). Oxford Martin Associate for the Global Cybercapacity Center, University of Oxford. Civil Society Liaison for the OECD Committee on the Digital Economy, and G7/G20 Digitalisation Task Force.

[‡]1) Request access to the repository, 2) follow the indications included in the source code of this file.

potential objections at the Institutional Review Boards level that would avoid the use of the data, or even the release for replication purposes. This paper introduces the method and illustrates the utility of this data to a political methodology audience, presents the privacy and data protection challenges, and shares the ongoing efforts to make this data available for the political science community building on the resources shared by the Harvard University Institute for Quantitative Social Science. The proposed approach includes an application of *differential privacy*, privacy-preserving method used by the US Census, and companies such as Apple or Google, to allow access and analysis based on sensitive data; and *Datatags*, a formalization of privacy levels to be implemented in the *Dataverse* repository project.

1 References for panel Discussion

- Internet remote-sensing: Baleato et al., 2015. Transparent Estimation of Internet Penetration from Network Observations. In: Mirkovic J., Liu Y. (eds) Passive and Active Measurement. PAM 2015. Lecture Notes in Computer Science, vol 8995. Springer, Cham https://link.springer.com/chapter/10.1007/978-3-319-15509-8_17; Weidmann, N. and Baleato, S. et al (2016). Digital discrimination: Political bias in Internet service provision across ethnic groups. *Science* 09 Sep 2016: Vol. 353, Issue 6304, pp. 1151-1155.
- Replication: King, G., 1995. Replication, replication. *PS: Political Science & Politics*, 28(3), pp.444-452.; King, G., 2011. Ensuring the Data-Rich Future of the Social Sciences. *Science* 11 Feb 2011: Vol. 331, Issue 6018, pp. 719-721 DOI: <https://doi.org/10.1126/science.1197872>
- Dataverse: <https://dataverse.org/>, King, Gary 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, Vol 36, Issue 2, pp. 173 - 199. <https://doi.org/10.1177/0049124107306660>
- Datatags: <http://datatags.org/>, Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The Datatags System. *Technology Science*. 2015101601. October 16, 2015. <https://techscience.org/a/2015101601>
- Differential Privacy: <https://privacytools.seas.harvard.edu/differential-privacy>; Nissim, K. et al. (2018). Differential Privacy: A Primer for a Non-technical Audience. *Vanderbilt Journal of Entertainment and Technology Law* (forthcoming) https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf; Privacy tool: Gaboardi et al., 2016. PSI (): a Private data Sharing Interface (working paper) <http://hona.kr/papers/files/psi-privacy-tool.pdf>

2 Acknowledgments

Thanks to Gary King for his comments at the Political Methodology Seminar at the Institute for Quantitative Social Science (IQSS); Salil Vadhan, Cynthia Dwork, Stephen Chong, James Honaker and Mercè Crosas for their input at the Harvard/ MIT Privacy Tools Technical Meetings; to Mercè Crosas, Danny Brooke and the Harvard University IQSS Dataverse team, as well as Wendy Guan and Jeff Blossom from Harvard University Center for Geographic Analysis (CGA) for their comments at the IQSS Tech Talk; to the attendants of my presentation at the International Studies Association (ISA) 2018 Annual Meeting in San Francisco, and the Midwest Political Science Association (MPSA) 2018 Conference in Chicago.