Replicable Privacy:
Enabling Replication on Sensitive Internet Data[*]

Suso B. Baleato[†] James Honaker, Mercè Crosas

August 17, 2019

Abstract

The use of data retrieved from the Internet enables unprecedented precision for political analysis, but it also introduces specific ethical and legal challenges to share and replicate results. As personal communications tend to take place over the Internet, researchers are increasingly required to prove compliance with an also increasingly restrictive and intrincate body of privacy law such as the GDPR. In addition to legal compliance, researchers need to put safeguards in place to prevent harm caused by identity disclosures from their datasets. In addition to economic and moral damage, a successful re-identification attack could lead to imprisonment, torture or even life termination in the case of research datasets about subjects or groups on autocratic or violent contexts.

In this paper we present the first results of a privacy-preserving system designed to enable safe sharing and replication of statistical analysis computed from sensitive datasets. Our system is composed of three elements, all of them made available to the scientific community thanks to an effort lead by the Institute for Quantitative Social Science at Harvard University. First, we use differential privacy, a privacy-preserving technique that avoids re-identification while preserving the statistical properties of the sensitive dataset. Second, we use the Dataverse open source software to share the resulting statistics consistently with FAIR principles, including automatic citation, persistent identifiers and data provenance. Third, we apply a simplified Datatags implementation to enable access to any sensitive dataset required for replication.

We illustrate the utility of our approach reproducing a remote-sensing technique that retrieves the most precise statistics of Internet connectivity available, and making them available for the scientific community. We show that our privacy-preserving estimates match the accuracy of the original method up to the 5th decimal of the correlation coefficients when compared with country-year official statistics worldwide from 2004 to 2012. We conduct an aditional validation to check the accuracy below the country level estimating Internet connectivity for the 31 provinces of China for the same period, and we find a 0.802 correlation coefficient with the available official statistics. We make the resulting data available for the scientific community using the Harvard University Dataverse. We implement a differential privacy algorithm which is made available through an open source library written in R to facilitate the integration or our method in a conventional statistical analysis workflow.

[*] Draft, not for circulation. Prepared for the panel "Education and Replication in Political Methodology" at the 2019 American Political Science Association (APSA) Annual Meeting, held August 29 - September 1 in Washington, DC (US)..

[†] Corresponding author. Contact: <susobaleato@iq.harvard.edu>

1 Privacy Implications of Internet Data

The mass adoption of the Internet is popularizing the use of that infrastructure as a tool for scientific research. This trend is prominent in the study of political phenomena, as the increasing availability of behavioral data is opening the way for previously unfeasible research. Examples of this trend can be recognised in recent studies looking at political censorship (King, Pan, and Roberts, 2013), to estimate connectivity statistics (Baleato et al., 2015), to evaluate digital discrimination in ethnic groups (Weidmann et al., 2016), to estimate ideology from social media (Barberá, 2015), to monitor network shutdowns in contexts of political unrest (Dainotti et al., 2011), or to evaluate the incidence of denial of service attacks during contentious periods (Lutscher et al., 2018). Initiatives such as the *Social Science One* partnership (King and Persily, 2018) to open social media data to researchers illustrates the value of Internet data, and suggest that this trend can be expected to continue growing up.

While academic scholarship has showed the utility of Internet data, it has started to outline its limitations as well. For example, problems of representativity have been identified that affect much of Twitter scholarship (Pfeffer, Mayer, and Morstatter, 2018), amplifying previously identified issues on bias and reliability of the sample (O'Neil, 2017; Lazer et al., 2014). However, the major challenge that researchers need to solve is not a technical, but a legal and ethical one: the need to comply with privacy law, and to prevent the research to harm the integrity of the involved individuals by disclosing their identity.

While ethical concerns are traditionally associated with experimental research, the capacity of behavioral data to reveal the identity of individuals extends the privacy issue to observational studies, also to those applying anonymisation techniques to their datasets. For example, research using mobile phone and credit card databases showed that a sample of just four spatio-temporal points is sufficient to identify 90-95% of the involved individuals (Montjoye, Radaelli, et al., 2015; Montjoye, Hidalgo, et al., 2013). The growing availability of behavioral data increases the relevance of the issue, because the combination of different datasets increases the probability of singling out, as illustrated in a study combining anonymised data about preferences of Netflix users and the IMDB public database of film evaluations (Narayanan and Shmatikov, 2006). The literature on re-identification attacks is not short of other examples including health records, discharge data, genetic samples and others. The growing availability of data also increases the accuracy of the resulting profiles in terms of political ideology, religious beliefs, ethnicity, health habits, or sexual preferences. In addition to moral or economic consequences, individuals can be imprisoned or even terminated when the identity disclosure takes place in contexts of political violence or authoritarian rule.

2 A Privacy-Preserving and Replicable Approach

As the case of the European Union's General Data Protection Regulation exemplifies, legal requirements to work with identifiable data tend to be more strict. Regulatory safeguards affect the capacity of researchers to retrieve, store and process Internet data; and also to share it with other researchers to allow the replication of their research (King, 2011; King, 2003; King, 1995). In order to conduct research on sensitive data, researchers need a privacy preserving approach capable to preserve the statistical power of privacy-less analysis, to provide quantifiable evidence about the privacy safeguards, and whose computational complexity allows for integration in a conventional statistical analysis workflow. We believe that *differential privacy* can satisfy the privacy requirements of researchers working with Internet data, using the implementation proposed by the Harvard/ MIT Privacy Tools research group[1] (Nissim et al., 2017; Gaboardi et al., 2016; Dwork, McSherry, et al., 2011; Dwork, Roth, et al., 2014).

Conventional privacy-preserving mechanisms set the focus on the dataset limiting its access -for example, locating the data in a bunker with physically access only-, or modifying the data introducing fake records or removing identifying fields such as name or social security number. While those methods can work in some circumstances they are limited by the administrative costs, and by their vulnerability to data breaches and re-identification attacks. Instead, differential privacy keeps the dataset untouched and applies the preserving mechanism at the statistical computation time, on the basis of a cryptography paradigm that avoids inferring the identity of an individual from the analysis of the released statistics.

Standard quantitative analysis produce a single set of statistics -eg mean or regression coefficients- that is a raw representation of the dataset or the computed model. With differential privacy, the researcher sets a privacy protection parameter that defines the distance between the values of the raw analysis, and those released after the privacy-preserving computation. This parameter, referred in the literature by *epsilon*, can be conceptualized as the probability between 0 and 1 of inferring the raw data from the released one. More precisely, per definition in Dwork, Roth, et al., 2014 et ϵ, δ > 0 be numeric parameters, D be the set of databases, let R be the set of possible outputs. A program M : D $\rightarrow$ D(R) satisfies ϵ, δ differential privacy if

Let ϵ be a numeric parameter, D be the set of databases, R be the set of possible ootputs. A program $M : D \rightarrow D(R)$ satisfies ϵ differential privacy if:

$$Pr(M(d)) \in S \leqslant e^{\epsilon} Pr(M(d')) \in S \tag{1}$$

for all pairs of adjacent databases $d, d' \in D$ such that $d \Phi d'$ (*adjacent*, i.e.: identical except for one record), and for every subset of outputs $S \subseteq R$.

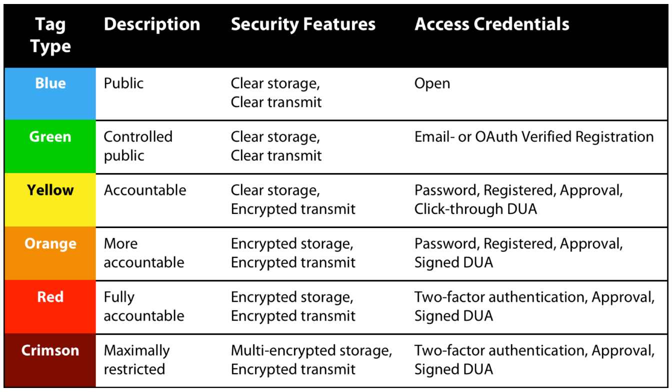
In addition to the *epsilon* parameter, the research can tailor the algorithm to match the analytical needs, and apply different mechanisms, allowing for quantitative assessment of the privacy safeguards to facilitate the ethical and legal determination of the research.

The use of differential privacy can help to overcome many of the ethical challenges involved in the use of Internet data for research, but it does not remove other concerns such as security or intellectual property. With or without differential privacy, a standard workflow involves the need to to lawfully and safely retrieve and deposit the data, run the computations, and share the resulting statistics. The latter is specially relevant for scientific research, as code and data needs to be made available for reviewers to replicate results, and the shared data needs to be properly referenced to facilitate reuse and citation by other researchers. In addition to privacy, replication and reuse of Internet data needs to address the intellectual property rights implications as well. For example, contents from social media may be freely retrieved but not redistributed, and auxiliary datasets required for the analysis -such as health records or geolocation databases- may be purchased for research, but cannot be made available to third-parties. In order to facilitate the replication and reuse of privacy preserved analysis, we propose to combine

[1] `http://privacytools.seas.harvard.edu/`

the use of differential privacy with *Datatags* framework in the implementation proposed by the Harvard University *Dataverse*.

Dataverse is an Open source research data repository software[2], developed by Harvard University Institute for Quantitative Social Science, that facilitates data reuse and replication consistently with the FAIR principles for data management (Wilkinson et al., 2016; Merc?? Crosas et al., 2015; Mercè Crosas, 2013; Mercè Crosas, 2011; King, 2007). For each deposited dataset, the software generates a standardized citation linked to a persistent identifier, that is updated to automatically keep track of any changes made to the included files (Altman and Mercè Crosas, 2013; Altman and King, 2007; Altman, 2008). The software allows researchers to customise their data distribution needs allowing the use of custom license or data use agreement, and requiring authentication to download data if needed. The current support for customised data access levels is evolving to implement the more comprehensive *Datatags* system, a taxonomy that identifies the security features and the access credentials required to share the different types sensitive datasets (Sweeney, M. Crosas, and Bar-Sinai, 2015).



Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Fig 1: The Datatags System. Sweeney, Crosas and Sinai (2015).

3 Internet Connectivity Statistics

The utilization of the Internet as a research tool generally involves a data retrieval process which makes use of techniques primarily developed for engineering purposes, as exemplified by the extensive use of the API of social media platforms to retrieve their data[3]. Another prominent example is the analysis of the Internet routing traffic channeled through the Border Gateway Protocol (BGP). When a message is sent over the Internet, it travels the network being relayed by the routers that link the source and destination devices. At each step, the receiving router calculates the shortest remaining path and relays the message to the corresponding router, establishing a relaying sequence which is repeated until the package eventually arrives to destination. The BGP takes part in that process standardizing the communication between the routers that connect the segments of the Internet addressing space grouped by Internet based on Autonomous Systems (ASes) (Rekhter and Li, 2006).

The analysis of BGP data is useful because it provides a global view of the Internet connectivity, and a very precise one as it allows for highly disaggregated spatio-temporal analysis. This technique has been successfully used to detect Internet shutdowns during times of political unrest (Dainotti et al., 2011), to improve the spatio-temporal precision of Internet connectivity statistics (Baleato et al., 2015), and to prove the discrimination of politically excluded ethnic groups in the access to the Internet (Weidmann et al., 2016). As much of the observed traffic comes from the digital interaction of connected individuals, researchers working with BGP data need to

[2] `http://dataverse.org/`

[3] An *Application Programming Interface* (API) is an abstraction of a computer system that standardizes the input/ output operations to facilitate its programming. This standardizatio includes the process to request and retrieve data, a feature used by researchers to acquire data, remarkably from social media platforms such as Twitter.

ensure compliance with privacy law and provide safeguards to their ethical and institutional review boards. The privacy risks increase together with the sampling precision, which can go beyond standard demographic protection, such as the census tract in the US, and is exacerbated in areas with less population and lower industrial development, especially in contexts involving authoritarian regimes or political violence.

The estimation of Internet connectivity statistics is a particularly useful application of BGP analysis which would require a privacy preserving approach to fully harness its utility. As the Internet becomes the backbone of digitalisation, accurate descriptions of the infrastructure are required to enable sound analysis on digitally mediated phenomena. The limitations of the official statistics on Internet have hampered the capacity of social science to develop highly disaggregated 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. The analysis of BGP data can help in solving those limitations, producing statistics 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 or data cannot be retrieved, such as in the case of authoritarian regimes or territories experiencing long-term political violence.

In order to overcome the legal and ethical challenges of BGP-based analysis, we improve the original estimation method (Baleato et al., 2015) applying a differential privacy algorithm implemented in R (Gaboardi et al., 2016). In order to evaluate the accuracy of the privacy-preserving statistics, we run the validation test proposed in the original paper comparing the computed estimates with official statistics at country-level between 2006 and 2010, and measure the error between the privacy estimates and the original ones. For robustness, we conduct a sensitivity analysis running the test for 15 different *epsilon* values, repeating the analysis 21 times to remove issues of randomness quality. As Figure 2 shows, our privacy-preserving statistics match the accuracy of the conventional analysis up to the 5th decimal of the Pearson correlation coefficients under the most restrictive configurations.

Once the privacy-preserving safeguards are in place, it is safer to publish the resulting analysis. We show how to share the data for reuse and replication deploying the Harvard University *Internet Connectivity Statistics* (ICS) Dataverse [4] consistently with the FAIR principles for Open Science. In order to use the latest stable version of the Dataverse code, we restrict our Datatags implementation to a simplified version with two permission levels only to allow for public access to the analysis, and restricted access to the source sensitive data for replication purposes. Next steps include refactoring of the R code; global sensitivity analysis automation; additional country level validations. We aim to extend the coverage of the dataset through Open Science collaboration with researchers interested in the connectivity of additional administrative, economic and ethno-cultural areas.

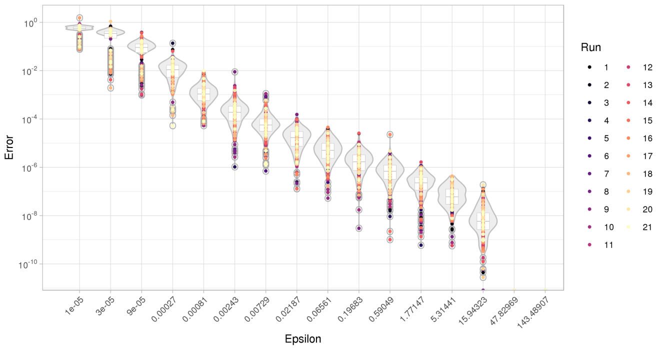


Fig. 2: Error between raw and privacy-preserving statistics for different epsilons, in 21 simulations.

[4] https://dataverse.harvard.edu/dataverse/ics

4 Conclusion

We have proposed a privacy-preserving approach based on differential privacy, the Dataverse software and the Datatags system to address the ethical and legal challenges of using the Internet as a research tool. We apply our approach to a BGP-based Internet connectivity estimation method, and we compare the accuracy of our privacy-preserving statistics with the original ones, showing a match up to the 5th decimal under the most strict conditions. We share the data deploying the Harvard University Internet Connectivity Statistics Dataverse, and we call for collaborations to validate and improve our data, and to use it for Open Science research.

5 Acknowledgements

The corresponding author wants to thank to Gary King for his comments at the Political Methodology Seminar at the Institute for Quantitative Social Science (IQSS); to Salil Vadhan, Cynthia Dwork, Stephen Chong and my coauthors James Honaker and Mercè Crosas for their input at the Harvard/ MIT Privacy Tools Technical Meetings; to all of the Dataverse team, remarkably to Danny Brooke and Gustavo Duran; to the Harvard University Center for Geographic Analysis (CGA) for their comments at the IQSS Tech Talk; to the attendants and of my presentation at the International Studies Association (ISA) 2018 Annual Meeting in San Francisco, and to the Midwest Political Science Associacion (MPSA) 2018 Conference in Chicago; and to Tina Freyburg and Lisa Garbe, University of St. Gallen for their insights about the importance of mapping the Internet infrastructure in autocratic regimes; and to my coauthors of the PAM and Science paper, specially to Nils Weidmann, ETH Zürich engineers Xenofontas Dimitropoulos and Eduard Glatz, and CAIDA engineers Alberto Dainotti and Alistair King.

References

Altman, Micah (2008). “A fingerprint method for scientific data verification”. In: *Advances in Computer and Information Sciences and Engineering*. Springer, pp. 311–316.

Altman, Micah and Mercè Crosas (2013). “The Evolution of Data Citation: From Principles to Implementation.” In: *IAssist quarterly* 37.

Altman, Micah and Gary King (2007). “A proposed standard for the scholarly citation of quantitative data”. In: *D-lib Magazine* 13.3/4.

Baleato, Suso B. et al. (2015). “Transparent Estimation of Internet Penetration from Network Observations”. In: *Lecture Notes in Computer Science, vol 8995.* Springer, pp. 220–231.

Barberá, Pablo (2015). “Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data”. In: *Political Analysis* 23.1, pp. 76–91.

Crosas, Mercè (2011). “The dataverse networkő: an open-source application for sharing, discovering and preserving data”. In: *D-lib Magazine* 17.1, p. 2.

— (2013). “A data sharing story”. In: *Journal of eScience Librarianship* 1.3, p. 7.

Crosas, Merc?? Et al. (2015). “Automating open science for big data”. In: *The ANNALS of the American Academy of Political and Social Science* 659.1, pp. 260–273.

Dainotti, Alberto et al. (2011). “Analysis of country-wide internet outages caused by censorship”. In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, pp. 1–18.

Dwork, Cynthia, Frank McSherry, et al. (2011). “Differential privacy–a primer for the perplexed”. In: *Joint UN-ECE/Eurostat work session on statistical data confidentiality* 11.

Dwork, Cynthia, Aaron Roth, et al. (2014). “The algorithmic foundations of differential privacy”. In: *Foundations and Trendső in Theoretical Computer Science* 9.3–4, pp. 211–407.

Gaboardi, Marco et al. (2016). “PSI (Ψ): a Private data Sharing Interface”. In: *Working Paper*.

King, Gary (1995). “Replication, replication”. In: *PS: Political Science & Politics* 28.3, pp. 444–452.

— (2003). “The future of replication”. In: *International Studies Perspectives*.

— (2007). *An introduction to the dataverse network as an infrastructure for data sharing*.

— (2011). “Ensuring the data-rich future of the social sciences”. In: *Science* 331.6018, pp. 719–721.

King, Gary, Jennifer Pan, and E. Roberts Margaret (2013). “How Censorship in China Allows Government Criticism but Silences Collective Expression”. In: *American Political Science Review* 107.2, pp. 1–18.

King, Gary and Nathaniel Persily (2018). “A New Model for Industry-Academic Partnerships”. In:

Lazer, David et al. (2014). “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343.14 March, pp. 1203–1205.

Lutscher, Philipp M. et al. (2018). “Politics with Digital Means: Denial-of-Service Attacks During Contentious Periods in Non-democratic Regimes”. In: *Proceedings of the Midwest Political Science Association Annual Convention 2018*.

Montjoye, Yves-Alexandre de, César A Hidalgo, et al. (2013). “Unique in the crowd: The privacy bounds of human mobility”. In: *Scientific reports* 3, p. 1376.

Montjoye, Yves-Alexandre de, Laura Radaelli, et al. (2015). “Unique in the shopping mall: On the reidentifiability of credit card metadata”. In: *Science* 347.6221, pp. 536–539. ISSN: 0036-8075. DOI: `10.1126/science.1256297`. eprint: `http://science.sciencemag.org/content/347/6221/536.full.pdf`. URL: `http://science.sciencemag.org/content/347/6221/536`.

Narayanan, Arvind and Vitaly Shmatikov (2006). “How to break anonymity of the netflix prize dataset”. In: *arXiv preprint cs/0610105*.

Nissim, Kobbi et al. (2017). “Differential privacy: A primer for a non-technical audience”. In: *Privacy Law Scholars Conf.*

O’Neil, Cathy (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pfeffer, Jürgen, Katja Mayer, and Fred Morstatter (2018). “Tampering with Twitter’s Sample API”. In: *EPJ Data Science* 7.1, p. 50.

Rekhter, Yakov and Tony Li (2006). *S. Hares," A Border Gateway Protocol 4 (BGP-4)*. Tech. rep. RFC updates: 6286, 6608, 6793, 7606, 7607. RFC 4271, January.

Sweeney, L., M. Crosas, and M. Bar-Sinai (2015). “Sharing Sensitive Data with Confidence: The Datatags System”. In: *Technology Science* 2015101601.

Weidmann, Nils B et al. (2016). “Digital discrimination: Political bias in Internet service provision across ethnic groups”. In: *Science* 353.6304, pp. 1151–1155.

Wilkinson, Mark D et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.