Internet Connectivity Statistics

Using the Dataverse to Enable Replication of Sensitive Data Research

HARVARD UNIVERSITY

Sensitive Data Research

Data-driven analysis enables unprecedented empirical research but it introduces legal and ethical challenges for researchers working with sensitive data. Individuals and groups can be singled out as the result of statistical re-identification attacks–even from "anonymized" records. In addition to moral or economic damage, privacy violations can lead to imprisonment, torture or even death in authoritarian or political violence contexts. As funding and academic credit increasingly requires the research data to validate results, researchers need a privacy-preserving approach to publish replication of sensitive datasets.

We show how to enable privacy-preserving replication combining three tools developed at Harvard University:

- **Differential Privacy** to prevent re-identification,
- **Datatags** to ensure security and access compliance,
- **Dataverse** to safely release replicable results.

We illustrate our approach using differential privacy (equation 1) to release privacy-preserving data to replicate analysis involving a potentially sensitive dataset (figure 2). We apply a dual access policy (green and yellow Datatags in figure 3) to release the data using the Dataverse (figure 4). As differential privacy adds noise to prevent singling out, we measure the error between the original and the privacy-preserving analysis, finding a negligible effect (figures 5).

`https://dataverse.harvard.edu/dataverse/ics`

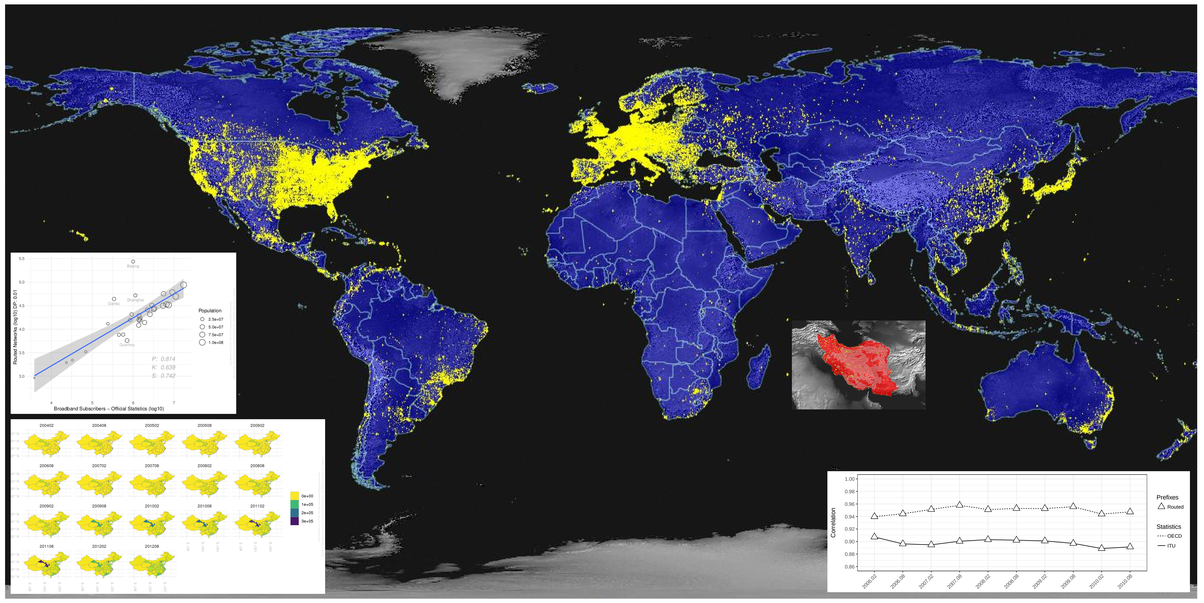

Figure 2: Internet Connectivity Statistics. Each yellow dot in the maps represents an Internet connection observed by geolocating network routing traffic. The statistics are estimated counting the observed connections for each delimited area, such as country borders (blue map), provinces or ethnic settlement boundaries (Persian areas of Iran in red map). The estimation method Baleato, Weidmann, et al., 2015 matches official country level statistics (comparison with ITU and OECD in the bottom right lines plot), and multiplies their reach and precision as they can be applied to estimate connectivity in areas where official data is not available such as authoritarian regimes, or territories with violent conflicts. Internet connectivity statistics are scarce. The bottom left plot shows previously unavailable statistics of the semestral growth of connectivity for mainland provinces of China from 2004 to 2012, and the scatter plot above shows a 0.814 Pearson coefficient when correlated with the 2010 China census data.

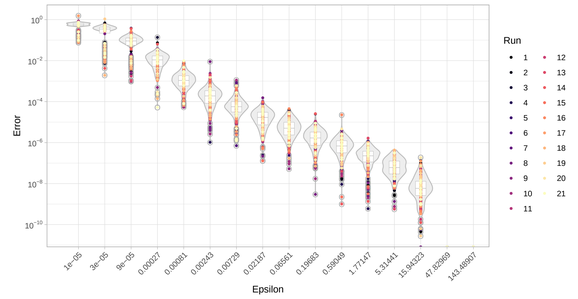

Figure 5: Error (vertical axis) introduced in the country level correlations by 16 different privacy protection levels (horizontal axis), tested 21 times. Lower epsilon values signify increased privacy protection, and values below 0.5 are considered to give significant privacy guarantees. The error is generally below the 5th decimal even for greater privacy protection levels.

Differential Privacy

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

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

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 $S \subseteq R$.

Equation 1: Differential privacy works by adding noise to the statistical computation so the released data cannot identify any individual record from the original dataset. Different from other approaches, the sensitive data is not modified, and researchers can tailor the trade-off between noise and accuracy to meet their specific research needs.

`https://github.com/privacytoolsproject/`

Datatags

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

Figure 3: Datatags is a standardized taxonomy for sensitive data sharing requirements. Each *tag* (in rows) defines storage, transmit, and access standards according to the sensitivity of the data, from the openly accessible (blue) to the more restrictive one (crimson).

`https://datatags.org/`

Dataverse

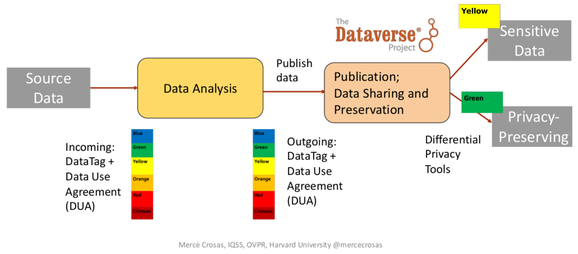

Figure 4: Dataverse is a leading research data sharing repository platform. Engineered as an Open Source project, it facilitates academic credit and citation through tools such as downloadable formal scholarly bibliographic references, persistent identifiers and automatic tracking of dataset changes. Dataverse currently integrates sensitive data sharing utilities such as data access control and will further expand to support integration with Datatags and privacy-preserving statistical analysis.

`https://dataverse.org/`

References

Baleato, Suso, James Honaker, and Merce Crosas (2019). "Replicable Privacy: Enabling Replication of Sensitive Data Research". In: *Working Paper*.

Baleato, Suso, Nils B. Weidmann, Petros Gigis, Xenofontas Dimitropoulos, Eduard Glatz, and Brian Trammell (2015). "Transparent Estimation of Internet Penetration from Network Observations". In: *Lecture Notes in Computer Science, vol 8995*. Springer, pp. 220–231.

Crosas, Mercè, Gary King, James Honaker, and Latanya Sweeney (2015). "Automating Open Science for Big Data". In: *The ANNALS of the American Academy of Political and Social Science* 659.1, pp. 260–273.

Gaboardi, Marco, James Honaker, Gary King, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan (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.

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

Questions? Proposals!

Suso Baleato
SusoBaleato@iq.harvard.edu

This work is the result of a collaborative effort. Big thanks to James Honaker, Mercè Crosas and Gary King (Institute for Quantitative Social Science at Harvard University) as their input and contributions provide the building blocks for the replicable privacy model. Special thanks to Salil Vadhan and all at the Privacy Tools Project; to Danny Brooke, Gustavo Durand and all of the Dataverse Project for their continuous support and for such a great engineering work; and to the Center for Geographic Analysis at the IQSS. Thanks also to Nils Weidmann (Konstanz University, Germany) and Xenofontas Dimitropoulos (ETH Zurich, Switzerland) for their seminal work on Internet connectivity, and to all of those who have contributed with comments including the Polmeth, MPSA, APSA and ISA spaces and the Max Planck Institute for Demographic Research (Germany). Suso Baleato CC BY-NC-SA (2019).

1737 Cambridge Street
CGIS Knafel Building, Room 350
Cambridge, MA 02138 US

The Institute *for* Quantitative Social Science at Harvard University