Reviewing Sensitive Data with Differential Privacy

JEDI Workshop, November 20, 2020. Organized by Data-PASSS Alliance.

As Open as Possible, As Closed as Necessary: Empowering Transparency in Publications Based on Sensitive Research Data

Mercè Crosas, Ph.D., Harvard University

scholar.harvard.edu/mercecrosas @mercecrosas

Differential Privacy tools to explore sensitive data

- A **differentially private** algorithm introduces a minimum amount of noise to released statistics to mathematically guarantee the privacy of any individual in the dataset
- **OpenDP** (https://opendp.org) is a **community effort** to build a trustworthy and open-source suite of **differential privacy tools** to explore sensitive data
- We are currently working on the first release of **OpenDP and Dataverse integration**

What will this mean:

- Opens up sensitive data to the research community
- Sensitive datasets findable in a Dataverse repository will be explorable through **differentially private statistics**, without ever accessing the original dataset
- Statistics included: mean, histogram, quantile, median, variance, OLS regression, logistic regression, probit regression, difference of means, unbiased privacy

Openly findable data, secure computation and storage, differentially private releases

Public Repository

Sensitive datasets discoverable via repository (only metadata open)

Dataverse®

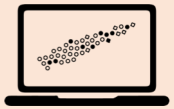
Blue | Green | Yellow

Notary Service

Digitally signed attestations confirming DUA compliance

Data owner/depositor

Approved Secure Compute Environment to run differential privacy statistics



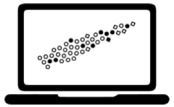
OpenDP

Trusted Remote Storage Agents (TRSA) or data enclaves

impact

Orange | Red

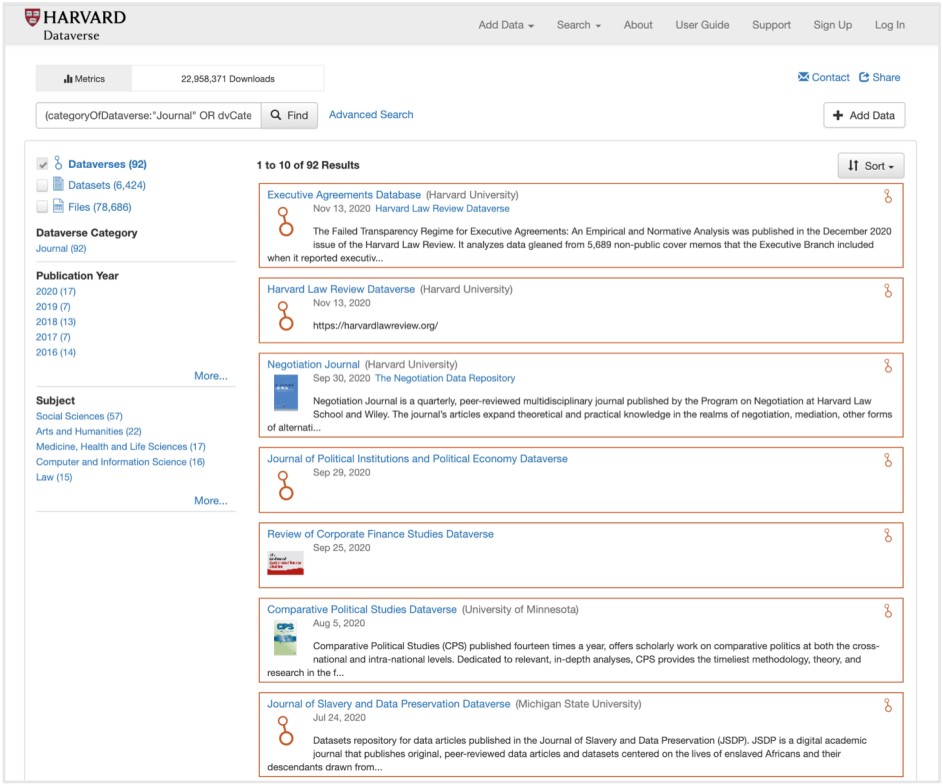
Data Reviewer



Differential Privacy release

Impact: Infrastructure for Privacy-Assured Computation
https://cyberimpact.us/

Journals' Dataverse collections will be able to support discovery, citation, and review of sensitive data



https://dataverse.harvard.edu

- **92 journal Dataverse collections** in Harvard Dataverse repository
- Replication datasets currently contain non-sensitive data
- With **OpenDP-Dataverse integration**, sensitive datasets will be archived in a secure storage
- **Data reviewers and users** will have **access to differential privacy statistics** of the sensitive datasets