Data-NoMAD
A Tool for Boosting Confidence in the Integrity of Social Science Survey Data

Sanford C. Gordon, NYU Politics
Cyrus Samii, NYU Politics
Zhihao Su, NYU Center for Data Science

February 27, 2025

What is Data-NoMAD?

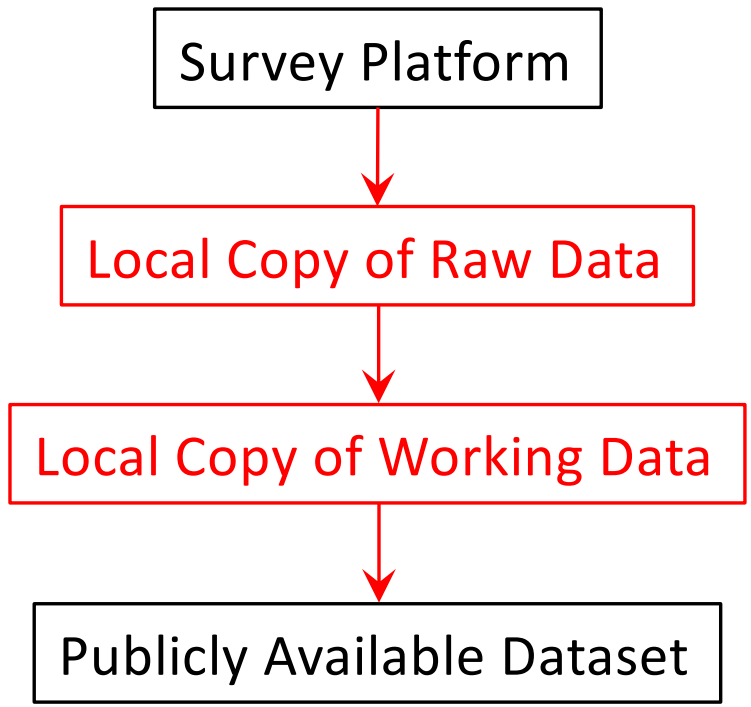
Data- Non-Manipulation Authentication Digest (Data-NoMAD): web app using SHA-256 hashing to allow researchers to certify veracity of their data and permit third parties to verify integrity of archived datasets

- *Digest* mode (for original researcher)
- *Verify* mode (for third parties)

Currently works with Qualtrics and SurveyCTO

Data-NoMAD + best practices block many avenues for fraud

Data “chain of custody” and where intervention is needed



Manipulating Data

Cheating

1. Response editing
2. Row deletion
3. Column deletion
4. Fake row addition
5. Fake column addition

Legit

1. Redactions for anonymity
2. Data transformations
3. Deletion of extraneous fields
4. Deletion of “problematic” responses
5. Data joins

Data-NoMAD: Digest Mode

name	color	age	drink
What is your first name?	What is your favorite color?	How old are you?	Do you prefer coffee or tea?
{"ImportId": "QID1_TEXT"}	{"ImportId": "QID2_TEXT"}	{"ImportId": "QID3_TEXT"}	{"ImportId": "QID4"}
Robert	green	52	coffee
Jane	Topaz	27	tea
Herman	Chartreuse	36	tea
⋮	⋮	⋮	⋮
Frida	taupe	19	coffee

Digested data stored at Data-NoMAD

survey_id	column_name	column_hash
SV_0q9y0TA1fUsvrkG	StartDate	1c01f6f7d24c54f81b2eefbc5cbb7b50 2233ca2a2d031abec3f90ea41155128c
⋮	⋮	⋮
SV_0q9y0TA1fUsvrkG	IPaddress	7894eafa3f9cae1a19048dae4a2982c7 f696ea6071304a620a6bc906515c525f
SV_0q9y0TA1fUsvrkG	name	96c6304ae92f64fb8c6eea32665b16d 430539ac3cf169fdc6ce2fc96baf9fa5b
SV_0q9y0TA1fUsvrkG	color	527da150cdfd38bc7c12709da41d1cb 08ec9267632d5298b3ccd027ba14e92fe
⋮	⋮	⋮

Data-NoMAD: Verify Mode

Used by third parties to authenticate replication dataset or original investigator to protect against accidental changes

1. Third party uploads archived data to portal with identifier
2. Data-NoMAD produces a report with names of columns that were:
    - deleted
    - added
    - altered

Data-NoMAD report, Verify Mode

Changes detected.

- Removed columns: IPaddress, hookParams, name, RecipientID, hookType, PanelID, EmbeddedData, PanelMemberID,
- Modified columns: color

The ideal

Changes detected.

- Removed columns: IPaddress, name

Ongoing and future work

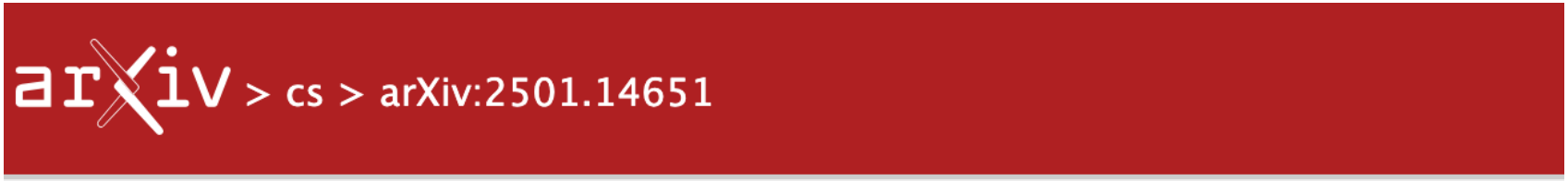
Addressing remaining vulnerabilities (e.g. row deletion within platforms)

Cell-by-cell hashing for special cases (open ended responses with sensitive content)

Elaborating best practices to complement the tool

Adapting to other types of data (e.g., observational, administrative)

For more details



Computer Science > Cryptography and Security

[Submitted on 24 Jan 2025]

Data-NoMAD: A Tool for Boosting Confidence in the Integrity of Social Science Survey Data

Sanford C. Gordon, Cyrus Samii, Zhihao Su

`https://arxiv.org/abs/2501.14651`