Measurement by Proxy: On the Accuracy of Online Marketplace Measurements

Alejandro Cuevas, *Carnegie Mellon University;* Fieke Miedema, *Delft University of Technology;* Kyle Soska, *University of Illinois Urbana Champaign and Hikari Labs, Inc.;* Nicolas Christin, *Carnegie Mellon University and Hikari Labs, Inc.;* Rolf van Wegberg, *Delft University of Technology*

https://www.usenix.org/conference/usenixsecurity22/presentation/cuevas

This artifact appendix is included in the Artifact Appendices to the Proceedings of the 31st USENIX Security Symposium and appends to the paper of the same name that appears in the Proceedings of the 31st USENIX Security Symposium.

August 10–12, 2022 • Boston, MA, USA

978-1-939133-31-1

Open access to the Artifact Appendices to the Proceedings of the 31st USENIX Security Symposium is sponsored by USENIX.

A Artifact Appendix

A.1 Abstract

Our work leverages three artifacts to conduct our analyses. First, back-end data seized by Dutch National Police (DNP) from the Hansa marketplace. Second, data scraped externally by the research team. And third, code used to simulate artificial marketplaces. The back-end data is used to test the completeness and uncover biases in the scraped data. The simulation is used to explore how other scraping methodologies could achieve improved coverage, based on distributions from the Hansa back-end. To allow further research in online criminal marketplaces, we are making our public scrapes and simulation code available. However, all of our analyses of the back-end data were conducted on-site at Dutch law enforcement agencies, so we never stored nor owned the data ourselves. Due to Dutch privacy laws on law enforcement data we are thus unable to release that dataset.

A.2 Artifact check-list (meta-information)

- **Data set:** Yes, the data scraped externally is provided. The back-end data is not.
- **Publicly available (explicitly provide evolving version reference)?:** Yes, the scraped data can be visualized and queried at: https://arima.cylab.cmu.edu/markets/viewmarketplace.php?name=Hansa. The simulation code is found at: https://github.com/aledcuevas/dnm-simulation/releases/tag/v0.2
- **Security, privacy, and ethical concerns:** The scraped data contains no Personal Identifiable Information. Research using the scraped data should never seek to provide any legal proof of criminal conduct.
- **Code licenses (if publicly available)?:** The simulation has a MIT License.
- **Data licenses (if publicly available)?:** The scraped data has the following license: https://arima.cylab.cmu.edu/markets/license.php
- **Archived (explicitly provide DOI or stable reference)?:** The stable reference for the scraped data will be available at: https://www.impactcybertrust.org/dataset_view?idDataset=1498. Our DOI request is pending.

A.3 Description

A.3.1 How to access

- The simulation code can be cloned or downloaded from the following Github release: https://github.com/aledcuevas/dnm-simulation/releases/tag/v0.2.
- The anonymized version of our dataset can be queried and visualized by navigating to the following URL: https://arima.cylab.cmu.edu/markets/viewmarketplace.php?name=Hansa.

Downloads of anonymized and non-anonymized versions of our dataset are done through IMPACT Cyber Trust. A free account is required to download the dataset. Researchers pursuing legitimate R&D in a valid organization at a DHS-approved location are eligible for accounts. Accounts may take a few business days to be approved. For more details refer to: https://www.impactcybertrust.org/help_faq. Additionally, requests for the non-anonymized versions of the dataset are handled on a case-to-case basis and require the signature of a Memorandum of Agreement.

- The anonymized version of our dataset can be requested from the following URL: https://www.impactcybertrust.org/dataset_view?idDataset=1498.
- The non-anonymized version of our dataset can be requested from the following URL: https://www.impactcybertrust.org/dataset_view?idDataset=1499.

A.3.2 Hardware dependencies

N/A

A.3.3 Software dependencies

No software dependencies are needed beyond those shipped with the Python 3.8 standard library, `pandas`, and `numpy`. To use the Jolly Seber abundance estimator, `RMark` is required. Instructions on abundance estimators is available in the repository.

A.3.4 Data sets

No other data sets are required beyond those described in the "How to Access" subsection.

A.3.5 Models

N/A

A.3.6 Security, privacy, and ethical concerns

The scraped data contains no Personal Identifiable Information. Research using the scraped data should never seek to provide any legal proof of criminal conduct.

A.4 Installation

No installation is required, assuming software dependencies are met. The code can be cloned or downloaded as a `.tar.gz` or `.zip`. The simulation code is executed by calling `main.py`. The datasets don't require any installation.

A.5 Evaluation and expected results

To evaluate the artifact, users should run `main.py`. We provide a set of test files that parameterize the simulation for testing purposes. The user will observe a count of days elapsed in `stdout`. The code will also create a folder structure (described in the GitHub documentation) which will contain the results of the simulation. Upon reaching the end of the simulation, the simulation will save to disk a `.json` file with a market

transcript (e.g., a record of the day that items, vendors, and reviews were created/hidden/deleted from the market). Additionally, the market will also run a simple artificial scraper which will output a `.json` file with the pages it captured as it scraped the market. Given that we are using dummy parameters, the expected result is a market transcript which contains between 1-10,000 vendor pages, 10,000-200,000 item pages, and 50,000 to 300,000 review pages.

A.6 Notes

While we provide a stable reference to our dataset, the DOI is still pending. Furthermore, given the number of days it may require to obtain an account from IMPACT Cyber Trust, we are happy to provide reviewers access to our raw anonymized dataset for evaluation through another channel, if necessary.

A.7 Version

Based on the LaTeX template for Artifact Evaluation V20220119.