Oracles: Authentic Data Collection Tool for Researchers

By Yan Ji (Cornell Tech)

Oracles: Authentic Data Collection Tool for Researchers

By Yan Ji (Cornell Tech)

I. DATA PROTECTION FOR RESEARCH PURPOSES

Data is the fuel in various aspects of technology development and economy. Today, the government, institutions and companies have devoted significant efforts in building particular databases for different purposes. However, access to some databases is limited and only granted to researchers in collaboration with these institutions for research purposes.

As an alternative, researchers may conduct surveys to collect research data on a large scale, assuming that participants are willing to honestly answer the survey questions. Existing online platforms such as Amazon Mechanical Turk (MTurk) [1] and Craigslist [2] are popular among social scientists for recruiting volunteers to participate in surveys. A series of works have shown the reliability of various types of data collected on MTurk since its launch in 2005. [3][4] However, there is recent evidence of bot accounts on MTurk raising concerns about the quality of collected data. [5][6][7]

Ⅱ. ORACLES: A POTENTIAL SOLUTION

Oracles (e.g., Town Crier [8], DECO [9]), originally proposed to provide trustworthy data feed to blockchains, allow an entity to collect authentic data from trusted sources under users' consent. They have a potential in mitigating the data quality issue on platforms like MTurk while providing a few additional nice properties. In particular, oracles offer the following advantages when serving as a data collection tool for researchers:

1. Data sharing based on users' will.

Access to data collected by the government, institutions and companies is limited and not managed by users themselves. With oracles, users of legacy database systems will have a chance to participate in research projects they care about and voluntarily share their data recorded in those databases for social good.

2. **Sybil prevention and trustworthy eligibility filtering.**

Oracles can help prevent an adversary, potentially using bots for automation, from participating in the same study as many times as they want, and mitigate the bot panic mentioned previously. More specifically, researchers may ask participants to register with CANDID [10], a privacy-preserving identity system using oracles as a building block, to prove their identities (de-identified) and participate only once in each research project. In addition, oracles can help filter out ineligible volunteers in a trustworthy and privacy-preserving way by flexible eligibility conditions such as age, occupation, etc.

3. **Broader access to authentic institution-owned data**

Institution-owned data is not available to individual researchers without an institutional collaboration or agreement. Oracles, unlike OAuth used by Google, Facebook and other large companies, are compatible with legacy systems and allow researchers to collect participants' data under their consent from institutional databases, so researchers can reach a broader range of high-quality datasets without modification to existing database systems. For example, Mozilla Rally [11], a platform allowing researchers to deploy studies and collect users' browsing data, asks participants to optionally provide their demographic information. There's no authenticity guaranteed regarding the data, however. Oracles can help capture an accurate demographic portrait by enabling researchers to collect authentic demographic information from trusted sources, e.g., government-issued identities including SSN and driver's license, thus better understanding or mitigating bias in sampling.

4. **Privacy of sensitive/identifiable data**

Oracles only collect data necessary for research purposes and keep other sensitive/identifiable data away from researchers. Additionally, zero-knowledge statements can be made by oracles so that researchers cannot learn information they shouldn't. For instance, if we want to filter out minorities in a project, we can implement an oracle that makes a verifiable statement on whether a volunteer is over 18 without leaking any further information such as the exact age.

5. **Cross-institution data aggregation**

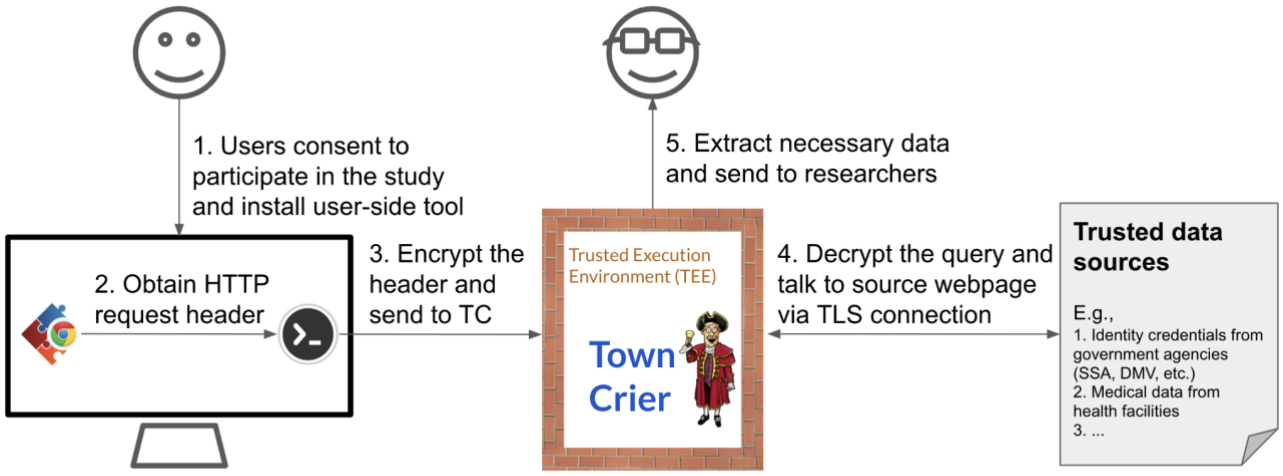
The value of institution-owned databases could be boosted when aggregated. For instance, assume Institution A owns the COVID testing results and Institution B owns vaccination records. If entries

from the two datasets can be linked to the same individuals, we will be able to study vaccine efficacy by analyzing the correlation. To do this, the data from two institutions must not be de-identified but this introduces privacy issues. Multiparty computation is a potential solution but expensive. Oracles, in contrast, allows privacy-preserving data aggregation in an efficient way. The aggregation happens on participants' side. In particular, each participant only proves to researchers that there is a person with testing result "positive/negative" authenticated by Institution A and vaccination record "yes/no" authenticated by Institution B, and no other identifiable information will be leaked.

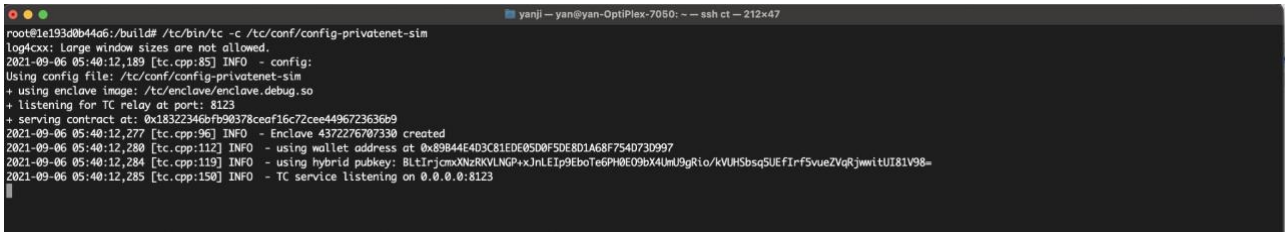
Ⅲ. OUR PROTOTYPE

We built a prototype [14] based on Town Crier (TC) due to its simplicity in implementation and efficiency. TC utilizes a trusted execution environment (TEE), e.g., Intel SGX, to communicate with the trusted source via a HTTPS connection on participants' behalf and extract the data we need. Because of the integrity and privacy properties of TEE, TC guarantees data authenticity and privacy.

The workflow of the tool is demonstrated in the following graph:



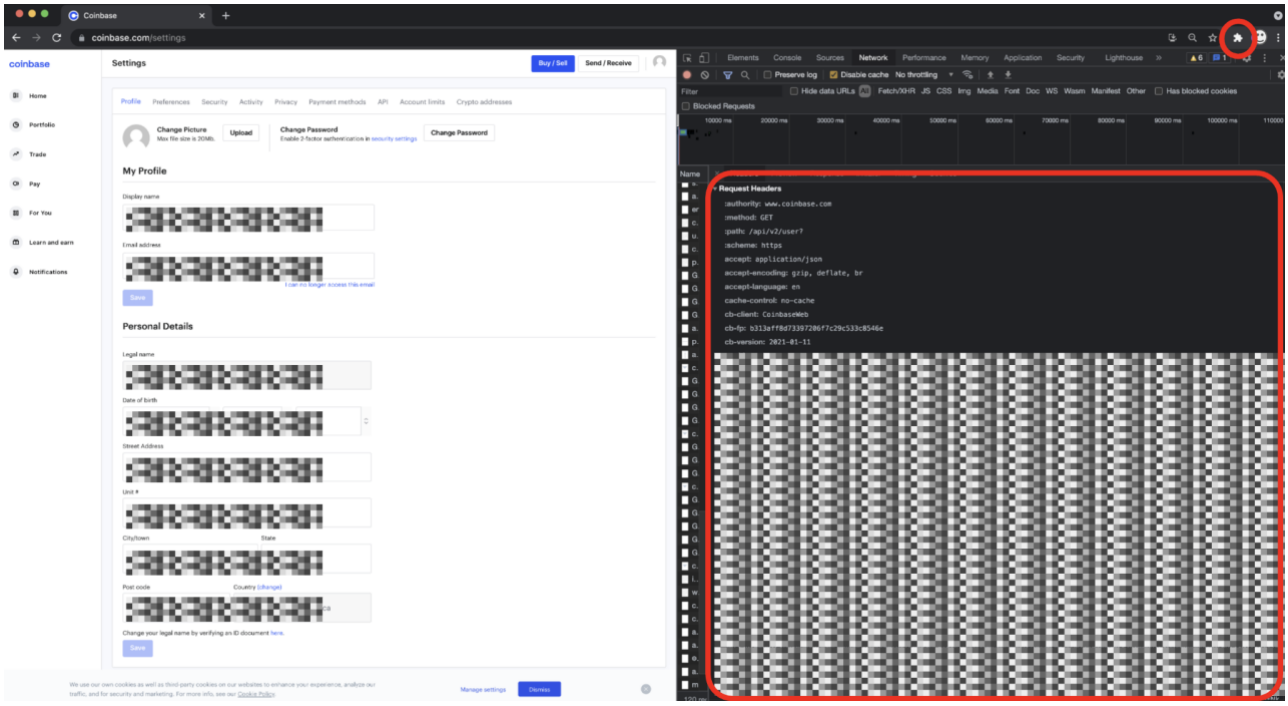
In particular, we have a TC server running inside a SGX enclave listening for messages from participants:



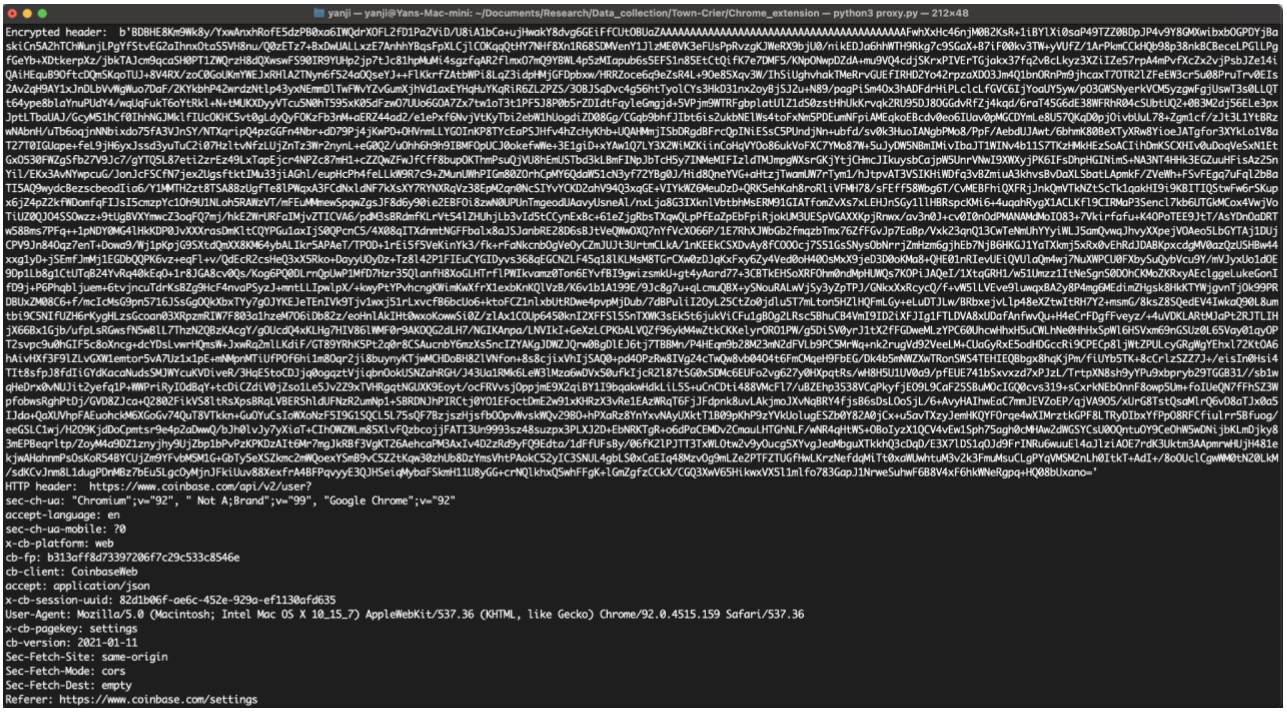
And the tool works in the following way:

1. A volunteer consents to participate in a research project and installs the participant-side tools, a Chrome extension and a python program.

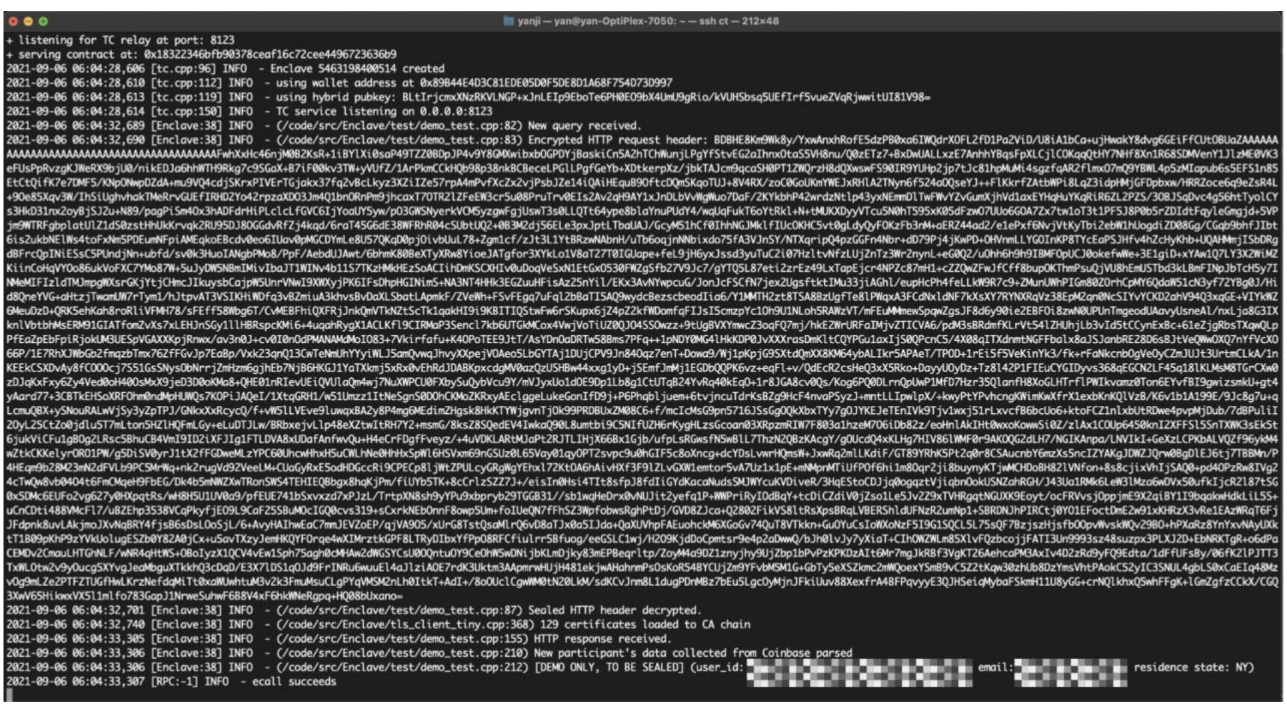
2. The Chrome extension works in the background to obtain the HTTP request header, the response to which will contain the data to be collected. If we need to know which state a participant comes from, for example, we require the participant to login to a website that contains his/her residency information, such as Coinbase which verifies a user's government-issued ID:



3. Participants also need to run locally a python program which listens for HTTP request headers passed by the extension. Upon receiving a request header of interest, the program encrypts it with the secret key of the TC server, and then sends it to TC:



4. Upon receiving an encrypted HTTP request header, the TC server decrypts it and then sends the query to the target website, Coinbase in our example, via a TLS connection:



5. Upon receiving a response from Coinbase, the TC server parses the message and only sends data that are necessary for the study under participants' consent to researchers.

Ⅳ. FUTURE WORK

As future work, we plan to further examine the effectiveness of oracles in terms of collecting high-quality data for researchers. Any comments, suggestions and discussions are welcome. We are also open to collaborating on research projects using oracles for collecting data.

REFERENCES

[1] Amazon Mechanical Turk, www.mturk.com/.

[2] "New York City Jobs, Apartments, for Sale, Services, Community, and Events." Craigslist, newyork.craigslist.org/.

[3] Behrend, Tara S., et al. "The viability of crowdsourcing for survey research." Behavior research methods 43.3 (2011): 800-813.

[4] Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?." (2016).

[5] Casler, Krista, Lydia Bickel, and Elizabeth Hackett. "Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing." Computers in human behavior 29.6 (2013): 2156-2160.

[6] Kees, Jeremy, et al. "An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk." Journal of Advertising 46.1 (2017): 141-155.

[7] Stokel-Walker, Chris. "Bots on Amazon's Mechanical Turk are ruining psychology studies." New Scientist (2018).

[8] Dreyfuss, Emily. "A Bot Panic Hits Amazon Mechanical Turk." Wired, Conde Nast, 17 Aug. 2018, www.wired.com/story/amazon-mechanical-turk-bot-panic/.

[9] Chmielewski, Michael, and Sarah C. Kucker. "An MTurk crisis? Shifts in data quality and the impact on study results." Social Psychological and Personality Science 11.4 (2020): 464-473.

[10] Maram, Deepak, et al. "CanDID: Can-do decentralized identity with legacy compatibility, Sybil-resistance, and accountability." 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021.

[11] "Mozilla Rally." Mozilla Rally, rally.mozilla.org/.

[12] Zhang, Fan, et al. "Town crier: An authenticated data feed for smart contracts." Proceedings of the 2016 aCM sIGSAC conference on computer and communications security. 2016.

[13] Zhang, Fan, et al. "Deco: Liberating web data using decentralized oracles for tls." Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020.
[14] Iseriohn. “Iseriohn/Town-Crier at summercamp21.” GitHub, github.com/iseriohn/Town-Crier/tree/summercamp21.

Yan Ji
Cornell Tech
yj348@cornell.edu