Data Management Plan

Our proposed experiments, system development, and outreach activities will produce a number of different data artifacts. As dissemination of these artifacts is essential to the goal of broad impact that is central to this proposal, we have carefully formulated distinct data management policies according to artifact types and sensitives. We anticipate that our proposal will result broadly in the following three data types:

(1) Data sets resulting from measurement studies and experiments;

(2) Open-source software; and

(3) Curriculum materials.

Data stewards. Each task supervisor will be a designated steward for all data resulting from the task, assuming responsibility for the management of the data and determining an appropriate classification (public, conditionally sharable, or private) for particular data artifacts. Should a data steward be unable to assume continuing responsibility for certain datasets, due to departure from his institution or some other event, he will transfer stewardship to another PI**[To-Do: Just one PI.]** on this project or to a senior researcher at his university and will furnish the instruction and documentation required for the new steward to assume continuing responsibility.

Data-handling policies. We will make our data artifacts available to other researchers as well as the general public to the greatest possible extent, as consistent with privacy considerations. Our approach will be to identify datasets as *public* (P), *conditionally releasable* (S), *confidential* (C), or *educational* (E). We specify the associated policies for each below. We will apply policy (P), (S), (C), or (E) to category (1) data as deemed appropriate by the associated data steward. We will apply policy (P) to category (2) data and policy (E) to category (3) data by default. Our data-handling policies are as follows:

- *Policy (P): Public data.* Public datasets will be those suitable for posting online, e.g., data derived from public sources, or the outputs of experiments (e.g., data, source code) that themselves do not involve any privacy-sensitive data. Our policy will be to retain these data for five years from the date of publication of any paper relying on the data. We will retain data for a longer period of time if possible, giving explicit priority to the goal of ensuring long-term scientific reproducibility. Public data will be made available via a project website or a public cloud. Larger data sets that cannot be disseminated by either such means will be stored locally and instructions will be published for interested researchers and others to obtain access to the data. We will adhere to a policy of releasing all source code resulting from the proposal as open-source software under suitable nonrestrictive licenses, and will make use of repositories, e.g., GitHub, that support this practice.

- *Policy (S): Conditionally releasable data.* Some data artifacts produced by our work will carry either temporary sharing limitations (e.g., individually requested moratoria on the release of personal data) or permanent ones. We will retain such data for the same duration of time as specified in policy (P). These data will not be made public, but stored locally with appropriate access-control mechanisms to restrict both external and internal access or in a cloud with protections that are suitable to the sensitivity of the data, e.g., a HIPAA-compliant cloud. Should researchers or others submit appropriate requests for data access, we will confirm that the request is appropriate (e.g., under the aegis of IRB-approved work) and will determine a practicable minimal-release strategy, specifically exploring time-limited and sanitized data-sharing approaches, as well as whether data should be released directly or through a query interface. We will release the data as expeditiously as possible, consistent with resource and policy constraints.

- *Policy (C): Confidential data.* A data steward may deem some data temporarily or permanently unsuitable for release outside his institution. University network packet traces such as we expect to collect in the course of this proposal will be deemed confidential in all cases, while derived data such as non-personally identifiable aggregate statistics or anonymized packet headers may be categorized as (S) or (P). Other data may additionally be deemed confidential by the data steward. At the time of data collection, the steward will determine whether it is appropriate to erase the data. (For example, highly sensitive data not employed in research may be summarily deleted.) Otherwise, the data will be preserved according to Policy (S), but with no access granted outside the institution of the data steward.

- *Policy (E): Educational data.* The data produced in curriculum development in the context of this project will be handled under Policy (P). These data will be made publicly accessible on the website of the data steward or in an appropriately locatable and accessible public archive.

Data storage and lifetime. The volume of data produced in this proposal will be small enough to permit handling within the existing data storage facilities of our respective universities. At a minimum, data will be stored for the duration of the project. We anticipate storing most data for a considerably more extended period of time, however, and will store for as long as is practical both data required to reproduce published experiments and data of public value. We will store all data in suitable standard formats and will confirm that university facilities include access controls and encryption as suitable for the handling of specific data artifacts.

Vulnerability disclosures. This research project does not explicitly encompass vulnerability assessments. It is very well possible, however, that we will discover security vulnerabilities or inappropriate data disclosures in the course of our work. For example, in our statistical modeling of data we may uncover inadvertent leakage of personally identifiable information; in our study of censorship circumvention we may discover vulnerabilities that expose confidential data to censors.

We will adhere broadly to community-standard responsible disclosure practices. Specifically, we will follow the following steps in disclosing a vulnerability:

1. *We will identify stakeholders.* We will identify primary stakeholders, entities developing or managing the affected systems or data, as well as secondary stakeholders, those potentially harmed by the vulnerability, e.g., users of the impacted system or subjects of the relevant data. We will work as advocates for secondary stakeholders throughout the disclosure process.

2. *We will privately disclose the vulnerability.* We will notify primary stakeholders of the vulnerability and provide tangible evidence so that they can confirm and assess its scope. We will seek to make this disclosure as expeditiously as possible.

3. *We will assist in vulnerability remediation.* We will advise primary stakeholders on technical remediation strategies, as appropriate.

4. *We will create a public disclosure plan.* In consonance with research community practice, we will by default make a public disclosure that specifically identifies the vulnerability, modifying this approach if it may bring about harm to secondary stakeholders and working with primary stakeholders to determine the appropriate level of detail to disclose about the vulnerability. Upon discovery of the vulnerability, we will set a target date for public disclosure. By default, this will be 90 days from private disclosure of the vulnerability.

5. *We will conduct a review with primary stakeholders.* We will circulate drafts of the public disclosure to primary stakeholders, soliciting their feedback and working with them to ensure that details are correct and amending the disclosure as appropriate, taking into account any harm that may affect primary and secondary stakeholders as a result of disclosure.

6. *We will make a public disclosure of the vulnerability.* We will publish the disclosure, including both technical detail and explanations accessible to secondary stakeholders, as warranted by the vulnerability.