Social and Behavioral Aspects
SaTC 2019 PI meeting breakout group report
Co-leads: April Edwards and Vivek Singh
Scribe: Sanjay Goel

1. Problem/Domain Summary

Research groups working in the social and behavioral aspects of Secure and Trustworthy computing rely heavily on data from a variety of sources. These data can be scraped from publicly available websites (ex. Twitter), collected (with informed consent) as part of a research study, or passed from one group to the next, with or without annotation/labeling.

Some examples of repositories with links to a variety of corpora include:

- *The Data is Plural archive: https://data.world/jsvine/data-is-plural-archive. This archive lists over 800 datasets for direct download, with some short description of what each corpus contains. Examples include: data on Oklahoma female prisoners, taxi rides in DC from 2015-2017, and 30,000 letters to "Dear Abby"*
- *ChatCoder.com. Data annotated and labeled for Internet Sexual Predation and Cyberbullying. https://www.chatcoder.com/drupal/DataDownload. Requests for access to these data continues to increase.*
- *A query service for a new corpus of text collected from a cell phone of 70 youths, ages 10-14. Over 800,000 records are available, and the query system allows regular expression and keyword queries. The data is presented in aggregated form only but can be summarized by race, gender, time of day or day of week.*
- *There are many other dataset catalogs available with just a bit of searching.*

As researchers we have a responsibility to protect our research subjects from harm, and to performing good science, including providing ways for our work to be replicated.

So, the questions become:

1. *How do we collect, monitor and analyze data streams for content that might be exposing an individual or organization to harm?*
2. *How do we collect, label and manage datasets for research purposes?*
3. *How do we validate that machine learning techniques that are applied to these data sets are accurately identifying the right instances?*

2. Key Research Challenges

The challenges in this domain include (in no particular order):

- *Understanding terms of use – collection of publicly available data (via webcrawling, for example) may violate website policy*
    - *How much should we care about terms of use for profit-making companies when the company's goals may not align with research (and the company may be using the data itself to further its own interests)?*
    - *We are competing both with companies and with foreign entities who may not follow the same standards*

    - *Securing data sharing from social media companies remains an ongoing issue.*
- *Often the problem is less one of collection and more one of filtering/labeling for huge volumes of data.*
- *We need a good definition of exposure to harm. Can we categorize certain activities and identify best practices so that new social media sites or studies can rely on previously defined terms and rules? Must everything be handled case by case? How do we educate IRBs as this domain shifts rapidly?*
- *A lot of security relevant communication (misinformation, cyber harassment, etc.) are happening inside private encrypted messaging (What's App, FB Messenger, etc.)*
- *Recognizing that true anonymity is almost impossible to achieve, how do we balance the needs of the research with the needs of the participant or study subject? Can we find ways to organize and store data so a participant can opt-out after the fact? Does allowing opt-out after the fact introduce bias into the data (are those with more social capital are likely to exercise this option)?*
- *What do we mean by informed consent? How and when do we obtain it? What happens when a participant who previously gave consent drops out? Who can reasonably give consent/assent (ex. children, developmentally disabled adults, other at-risk groups)? Should we ask for additional consent when data is to be re-used?*
- *Data can persist for a long time. How can we tell when a corpus is robust enough to be considered a "gold standard?" A dataset that meets that threshold at one point in time may no longer meet that definition as technology changes. Machine learning and AI will continue to improve, how can we tell when data is no longer "good."*
- *The perception of harm depends a great deal on both context and culture. How can we develop a more universal definition of harm in the context of data collection?*
- *We need to collaborate longer with participants – to improve our understanding of the data and follow-up when new concerns/issues arise.*
- *We need a greater emphasis on data collection in our proposals to NSF. There are good bodies of literature on structuring questions, the order of questions, validating questions, etc.*
- *We need to think about the experimental design before we collect the data, and plan for data collection that meets the needs of the study we are proposing.*
- *How do we detect and handle bias in the data? Can the bias change over time?*
- *Does simulated/synthetic data have a role in this domain? Have we studied the benefits and pitfalls?*
- *Where does explainable AI fit in?*

3. Potential Approaches

- *Peer review of datasets*
- *Shifting the norms around data sharing*
- *Development of repositories of data that are well-defined, peer reviewed, and meet certain standards for quality and reproducibility. Put resources (NSF) into independent verification.*

    - *AoIR (Association of Internet Researchers) has guidelines for data collection*
    - *Associating data with revocable certificates*
- *Disincentivize use of data for secondary purpose (i.e. data that has not be vetted for this particular use).*
- *Some conferences are already beginning tracks for describing new data sets and reproducibility with independent verification of results (ex. ICWM, ACM Multimedia).*
- *Keeping up with new trends as users move to the next big thing – ex. TikTok and Discord.*
- *Developing our own tools for data collection, replacing the corporate apps. Participants use these apps knowing that data is being collected, and we have tracking mechanisms for later opt-outs.*
- *Use the work of Casey Fiesler (*https://pervade.umd.edu/*) and others who have studied informed consent and understand the social science behind it.*
- *Need to partner across disciplines. Of those in the room (~30), only 6-8 with PhDs in social sciences.*
- *How can NSF support collecting large scale datasets in a privacy sensitive way?*
- *Can we encourage infrastructural proposals? To build benchmark datasets with limitations that are defined and documented.*
- *Research on the ethical considerations and potential risks related to collection and dissemination of large data.*

4. Long-Term (> 10 years) Significance
    - *We are confident this problem will remain relevant for a long time to come!*

5. Other Important Aspects of This Topic *(specify)*
    -