Kai Malcolm
BMIF 7380
Project Proposal
Due 3/17/22

Motivation

As wearable devices become increasingly common, and as biometric sensors are embedded into more devices that join the IOT-verse, the amount of data and the uniqueness of that data will only continue to grow. While these devices can bolster preventative and responsive measures to a number of diseases and emergencies, they also leave a very high-resolution, fine-grain trail of data extremely specific to each individual. Wearable devices are not only well-positioned to categorize an individual's baseline resting heart rate, sleep, and daily activity, but also can be used to identify minor changes in the user's data that may indicate changes in health (e.g. diseases, COVID, etc.). From today's location-tracking wearables (e.g. smartphones), real-time location data can be pivotal in emergency situations, but it is already being used to track users unknowingly, as is seen in some free store-provided WiFi networks which track movement as individuals browse merchandise. This has obvious implications for revealing certain sensitive attributes, such as if an individual spends time in front of the displays for family planning or certain types of medicines/treatments, which may reveal medical conditions. It has been shown that simply removing personally identifiable information (PII) like SSN is not enough to ensure privacy, and furthermore, Biel et. al.[1] have shown that electrocardiography (ECG) data can be used to achieve 100% reidentification. While certain security measures have been taken (namely, encryption and access control), they do little to protect against either an adversary with prior information, or against someone within the organization. Ideally, a method would exist such that each individual's dataset could be sanitized and released for use.

Pilot / Proof of Concept

The main aims of this project are to assess the privacy risk for sharing activity data using current standard privacy practices (e.g. anonymization simply by removing PII), and then proposing an effective differential privacy solution. Explicitly, the goal is to develop and validate a non-interactive differential privacy solution, which implies that the data holder/curator (myself) sanitizes the data and releases the sanitized dataset. The main difference between interactive and non-interactive is that in interactive, the data curator hosts the data and answers queries, as long as said queries do not expend the privacy budget, whereas for non-interactive, the data curator releases a sanitized dataset for which privacy loss cannot accumulate, if the sanitized dataset is indeed differentially private. There is a wealth of information on privacy risk analysis for various datasets, and, as for releasing a santizied non-interactive differentially private dataset, Dr. Bonomi is currently working on a paper in regards to sharing ECG data in a privacy preserving manner, specifically employing non-interactive differential privacy.

1 Biel, L., Pettersson, O., Philipson, L., & Wide, P. (2003, June 3). ECG Analysis: A New Approach in Human Identification. IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=930458

The first step would be to collect a complete dataset, which can be found at https://www.kaggle.com/arashnic/fitbit, which features 30 user's Fitbit data, with entries everyday for a full month, specifically for daily/hourly/minute-level output for physical activity, heart rate, and sleep monitoring, as well as calories, steps, and weight. The focus of this project will be on daily physiological data (e.g. resting heart rate, minutes asleep, number of steps). The timeline is shown below.

Timeline

For this project, I will be having weekly meetings with Dr. Luca Bonomi in order to obtain guidance and feedback. I suspect in the first 2-3 weeks I may be able to accomplish my goals quite a bit faster, and the last 3 weeks may take a bit longer than expected. The sanitization of datasets towards the end is not fully fleshed out as the exact steps/direction will likely change based on Dr. Bonomi's guidance, as I do not have as confident a handle on non-interactive differential privacy and this will require further research.

Week of March 21st

- Determine most informative variables, on the basis of entropy.
- Determine how much external information (in this case, splitting the data into training and testing, and comparing the testing to the training to identify matches) is required.

Week of March 28th

- Quantify the relationship between uniqueness and information available, mainly how this varies with number of data points available, as well as types of data (step count, heart rate, etc.) and sequences of data (e.g. data from X many days in a row, or in Y time period are required to get a confident match).

Week of April 4th

- Finalize presentation for the 6th.
- Finalize conclusions about data privacy risks from the data in its current form, as well as explicitly define relationships that may expose users to higher privacy risks (e.g. routinely posting calories burned once a week may lead to easy identification within a broader dataset in which that specific data does not appear, etc. for other variables / combinations / frequencies).

April 6th Project Status Report Due

Week of April 11th

- Explore sanitization methods for datasets.

Week of April 18th

- Apply various sanitization methods to truncated dataset.

Week of April 25th

- Evaluate privacy risks of each sanitized dataset.

Week of May 2nd

- Write final project report.
- Quantify and report results of privacy solutions applied.

May 4th Final Project