Programming Project: Safe Sharing of Population Descriptors

Many people use tracking devices to monitor their activities, such as Apple Watch, Google Watch, or Fitbit. People also use apps to monitor their diet and other lifestyle parameters. Example parameters include age, sleep patterns, physiological parameters and health indicators, exercise patterns, etc.A key concern with cloud-based services are security and privacy, which may result in poor adoption. Another concern is that it is difficult or impossible to move from one provider to a different one. We may also want to compare own parameters with those of different populations, which is generally not possible with current systems.

A blockchain-based service that maintains a distributed "ledger" with population parameters could be more appealing. The service would allow users to compare their own parameters to those of the current user populations. Each user would trade in own datapoint, such as physiological parameters (blood pressure, heartrate) or exercise data, and receive the current population descriptors. The user then would then compare his or her standing within the specified population. This relative standing can be complemented with an absolute standing derived from general guidelines or government statistics. Government-provided population statistics are usually too general, and may be incomplete or outdated. A blockchain-based service could provide real-time data and the user would be able to select different populations for the comparison. For example, the user could request a comparison the population in a given geographical region, or to a given age group.

To acquire his or her own physiological parameters, each user would need a tracking device, such as Apple Watch, Google Watch, or Fitbit. Dietary data could be entered manually by the user. An issue that needs to be solved is exporting the device-acquired data to the blockchain-based service. The users would may continue using their cloud-based services, and at the same time benefit from sharing information on the blockchain-based service. The service users do not need to keep their personal data on a private home drive. They may use unrelated cloud services for this purpose. The point is that there will be no data concentration that would make it an attractive target for security breaches.

Here are some initial points to learn more about the blockchain technology:

https://www.tutorialspoint.com/blockchain/index.htm

https://www.guru99.com/blockchain-tutorial.html

https://www.investopedia.com/terms/b/blockchain.asp

https://modus.medium.com/blockchain-and-the-promises-of-web-3-0-fe798cea1893

Issues or Subproblems to Solve

Issue #1: Exporting the participant data from different devices, such as Apple Watch, Google Watch, or Fitbit. For data that will be entered manually, such as dietary calories and sleep patterns, we need to provide a user interface. In addition, there are websites that allow people to track their dietary

information, and we need to consider how to enable the user to export this information and submit it to our blockchain-based service.

Issue #2: We need to design the data formats for submitting user's personal parameters, and ensure that the same type of data is submitted using the same data format and same measurement units. Of course, each message will be encrypted, but the original data must be formatted uniformly before the encryption is applied.

Issue #3: We need to specify the query message format that each user will use to select the type of population parameters he or she wishes to receive and the population from which these parameters are derived. At the same time, we also need to specify the format of the encrypted data maintained in the distributed "ledger," so that it is easy to extract the specific data requested by each user.

Issue #4: What data will be stored in the ledger? At one extreme, we would store every individual datapoint that was submitted; at the other extreme, we use each received datapoint to update the population statistics and immediately discard this datapoint. It is impractical and probably unnecessary to maintain all datapoints forever (as is done for financial ledgers, where each transaction must be recorded). A compromise solution may be more useful. For example, it may be useful to keep some past window of datapoints for calculating trends and detecting patterns.

Issue #5: Deciding on what type of wellness and lifestyle parameters will be tracked by our blockchain-based service. The list of currently tracked parameters should be available to all users for lookup. In addition, it should be possible to expand or shrink the list of the types of tracked parameters, depending on what proves to be important (based on findings from medical studies) or popular (based on frequent queries by the current participants). We must design a protocol for introducing new types of parameters for tracking, or removing the existing types if they prove not to be popular (i.e., when very few participants are uploading this type of information). Also, how will the current users learn about newly introduced types of population parameters, and will they be solicited to contribute their own data of this type? When the type and scope of possible queries are modified (such as new types of parameters or new population subsets), how will the users learn about these and will their current client program be able to generate new queries and accept newly introduced descriptors? Or, will they need to download and install a new version of the client program?

Issue #6: Dealing with potential conspiracies or other adversaries. What if a participant or a clique of participants conspires to systematically introduce unreasonably high or low values of parameters that in a long run may skew and distort the population parameters? Can this be detected and prevented? [Note that there is no "central server" and all computation must be performed in a distributed manner using blockchain.]

Issue #7: User data must be permanently encrypted for privacy, but they also need to be used for updating population parameters in the *encrypted form*. To calculate the population parameters from the continuously updated individual user datapoints that are encrypted and cannot be decrypted, we will need to use arithmetic operations on encrypted data. Here are some example sources on this topic:

Chin-Chen Chang & Sun-Min Tsu, *International Journal of Computer Mathematics*, Volume 56, Issue 1-2, 1995. https://www.tandfonline.com/doi/abs/10.1080/00207169508804382?journalCode=gcom20

Homomorphic encryption – Wikipedia: https://en.wikipedia.org/wiki/Homomorphic_encryption

Homomorphically Encrypted Arithmetic Operations over the Integer Ring:
https://eprint.iacr.org/2017/387.pdf

Issue #8: if each datapoint submission is anonymous and does not identify the sender, then it will not be possible to track and compare user histories or track individual trends. We will be able only to track the trends for the entire population *in aggregate*. Note, however, that the total population and individuals will change over time as new participants join and existing members leave the system. Should this fact be accounted in the population parameters calculation? Would it still be possible to track individual parameters based on their IP address of each datapoint submission? What if a user uses different devices to submit their datapoints?

Issue #9: Does the service need to encrypt the query response that contains the population parameters? —Probably not, but this issue needs to be considered and justified.

Issue #10: If the blockchain technology is used, then each participant will have the population parameters at all times because the distributed "ledger" is maintained in a *distributed* manner—at all participant computers. Why query others? When will all participants be queried to keep the data up to date?
Answer: Not all participants need to keep the distributed ledger; only "supernodes" will, and others only supply data and query the population descriptors. The reliance on "supernodes" is particularly important if the user population grows very large, at which point low-resource nodes would not be able to support the distributed ledger.

Issue #11: Why do we need to maintain population data if we can use the generally available statistical data as absolute guidelines (these data generally are available from government websites)?

Answer: To support finer-granularity queries, more diverse queries, more local-scope queries, for the ability to track new descriptors that have not been tracked officially, etc. Government-provided population statistics are usually too broad and general, and may be incomplete or outdated. They can be used when lacking more accurate or up-to-date data, or as supplement to our blockchain-based service.