A Smartphone-based System for Population-scale Anonymized Public Health Data Collection and Intervention

Andrew Clarke and Robert Steele
Discipline of Health Informatics, University of Sydney, Australia
{andrew.clarke, robert.steele}@sydney.edu.au

Abstract

The wide availability and sophisticated functionalities of current mobile devices or smartphones can provide a new form of data collection capability relevant to public health. However, current data that is collected is typically siloed on individual devices and/or specific proprietary systems, only intended for individual use, limiting possible utilization for public health purposes. Additionally, the current aggregate data collection approaches do not incorporate key public health components such as support for interventions and demographic data. To address these limitations, in this paper we introduce and evaluate a system to provide aggregate population health data via utilizing smartphone capabilities, whilst fully maintaining the anonymity and privacy of each individual. In this paper we provide a detailed architecture, a method for local processing of aggregate population health data utilizing adaptive privacy thresholds to create a multi-party flexible approach to participatory data submission and evaluate its privacy properties at large scale.

1. Introduction

The recent rapid growth in both the capabilities and uptake of mobile devices with sensors or smartphones capable of acting as sensor platforms has the potential to advance public health data collection and intervention. Whilst increasingly mobile devices and sensors are used as a tool for individual health data capture and feedback, this has not extended into a general re-usable approach. Prior work has relied on a trusted data collector or aggregation process. Interestingly, the case for public health usage doesn't require the same level of precise data that would often be required in other participatory sensing applications. For example, the exact location and time of a measured sensor value is less important than the aggregate value over a period of time or the trend or change for a community as a whole.

In this paper we introduce a novel smartphone-based system for anonymized population health data capture and intervention. Interventions [1] are a key component of future Health Participatory Sensing Networks (HPSNs), and in our approach we describe a novel methodology whereby a targeted public health intervention can be distributed, performed and evaluated without the need for the identifying details of an individual to ever leave their mobile device. This leads to the logical extension of an alternate system approach that eschews the need for a fully trusted server, which might prove impractical on population-scale applications, instead adopting an architecture that utilizes an anonymous communications layer (onion routing or mix network) in combination with de-identification of data submitted, to provide anonymous submission/interaction. Beyond de-identification this approach needs to resolve the risk of re-identification based on quasi-identifiers, in the form of information known about individuals outside the HPSN that could potentially be used to match with and re-identify the submitted data. The conventional approach to address this type of risk is to use a trusted server or aggregation point to combine and obfuscate/alter data to the point where k-anonymity [2] is assured for a data set, such that any individual is indiscernible from k other records based on quasi-identifiers.

Our approach that meets the above requirements would instead perform de-identification without a trusted aggregator or server. As such, we propose that a suitable level of anonymity can be provided by firstly locally processing collected data on the user's mobile device into an aggregated, generalized form that can still meet the purposes of public health data collection. This can be achieved by utilizing quasi-identifier scores (QISs) as a measurement of approximate risk of possible use in re-identification and support a threshold approach to privacy limits. This would allow the level of privacy disclosure an individual agrees to, to be easily managed without requiring a case-by-case approval. This in combination with our approach to specifying the data required and weighting of precision and inclusion factors allows the local device to dynamically alter the resolution and breadth of data submitted to preserve privacy and anonymity, whilst still submitting the data needed for public health data usage.

2. Related work

The use of participatory sensing is of increasing interest in a number of application areas including air quality and pollution sensing [3] through the use of external air quality sensors, urban area noise level data [4], urban traffic analysis through the use of vehicle mounted sensors [5] and vehicle fuel efficiency [6], amongst many other applications.

The rich capabilities of participatory sensing have garnered interest in its usage for a range of such applications. This has in turn spurred a number of different approaches to resolving or decreasing the implicit security and privacy concerns when involving individuals in sensing/data collection. The more conventional approach would use a trusted server, then k-anonymity [2] or a variant, to anonymize the data before it is accessible for research/analysis. Of course this approach suffers from the need for a fully trusted server as well as issues of a single point of failure in terms of privacy breaches. Alternatively, other approaches have improved on this by removing some sensitive information before submission (removal of identifiers and communications anonymity) with a central point of trust [7] to provide an anonymous approach. While this is quite effective when the sensing is collecting data on something not specific to the individual, this alone is not well-suited to a model where quasi-identifiers are a key submission component (such as in the case of collection of public health data) as de-identification protection is still implemented at a central trusted point.

To resolve the issue of requiring a fully trusted server, alternative approaches include decentralized participatory sensing networks [8] using user interaction/awareness as part of the approach or keeping the data managed by the participant [9, 10] and stringent user-definable access control mechanisms to manage sharing. While these approaches may be extensible to some aspects of HPSNs, they typically have not incorporated the need and importance of health interventions in HPSNs, a capability that does not have a direct parallel in most participatory sensing systems. Additionally, the capabilities that are beneficial in other areas may make these approaches overly complex for individuals, limiting their feasibility for a large scale implementation.

3. Overall system architecture

The overall system architecture (Figure 1) involves one or many central Health Participatory Sensing Servers (HPSSs) that communicate with mobile devices through a mix network or onion routing network to provide communications anonymity, and mobile devices that incorporate local processing and privacy thresholds to maintain data anonymity/privacy/de-identification.

The same HPSN and HPSS could be utilized by multiple health organizations i.e. organizations involved in public health-related activities.

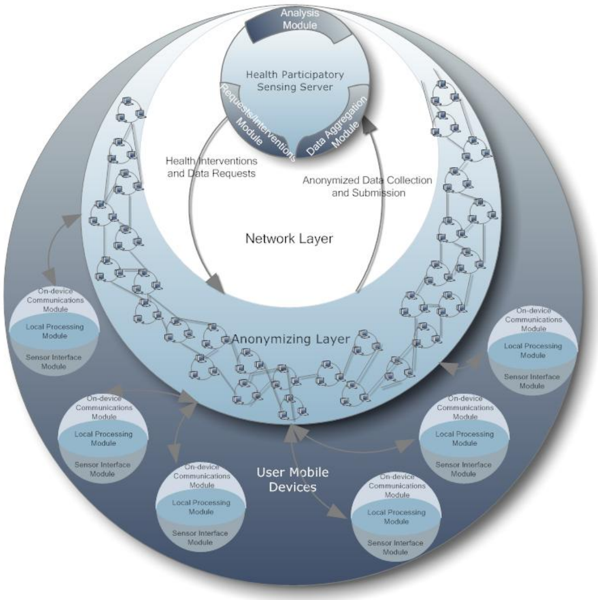
Analysis Module
Health Participatory Sensing Server
Requests/Interventions Module
Data Aggregation Module
Health Interventions and Data Requests
Anonymized Data Collection and Submission
Network Layer
Anonymizing Layer
On-device Communications Module
Local Processing Module
Sensor Interface Module
User Mobile Devices

Figure. 1. Health participatory sensing system architecture

There are two primary data transmissions from and to the HPSS respectively: (1) data requests and public health interventions are distributed from the HPSS, and (2) anonymized data collection submissions are sent to the HPSS. The core functionality components of the HPSS are (1) Data Aggregation, (2) Analysis and (3) Intervention/Data Requests.

The fundamental architecture can support different levels of both data collection and potentially public health intervention, depending largely on the capabilities of the end user mobile devices as well as the level of participation in the public health data collection task of the individual users of these devices. We introduce these configurations in the following subsections.

3.1 Smartphone-only configuration

This occurs when an individual utilizes just a smartphone without additional external sensors and the user is not required to take additional actions to participate in the public health data capture. This configuration has the advantage that it has the greatest level of existing deployment and ease of adoption – that is, smartphones without additional external sensors are the most common smartphone usage case. Various types of data can prove to be important public health or epidemiological data sources. An example would be physical activity tracking [11] which has become increasing popular in recent years, and for which we have discussed its potential secondary usage in our previous work [12].

3.2 Smartphone plus additional sensors

In this configuration an individual participating in the public health data capture is required to carry out the extra effort of using additional external sensors connected to their smartphone, but is not required to change their day-to-day behaviour.

3.3 Intervention capabilities

This configuration additionally provides inputs to the individual to alter the actions they would have taken while participating in the HPSN, in addition to the sensing capabilities arising from smartphones, with or without additional external sensors. Such participatory sensing in the health context has a somewhat different goal to that of 'active' participatory sensing in many other contexts. Whilst an 'active' participatory sensing model for a typical sensing task might focus on achieving more complete data collection in terms of spatial/temporal range, health and epidemiological-related active sensing would be more concerned with affecting a health-related action and hence have a component equating to a public health intervention. As such, the instigation to carry out 'active' sensing activities could essentially constitute a public health intervention input. Additionally for public health purposes, this allows for immediate and continuous feedback of the effectiveness of campaigns upon targeted groups.

3.4 Extension via manual input

This configuration combines the potential sensing capabilities of smartphones and external sensors with additional human-sensing capabilities, allowing for large amounts of objective sensing data to be complemented with subjective human-generated data and feedback. Further, this configuration can be implemented with or without intervention capabilities, with the combination allowing the additional capability of providing feedback in regards to interventions.

This is implemented through the addition of context-sensitive micro-surveys that are requested of users and attached to relevant collected sensor data. This allows for both data that is difficult to record through sensors alone and data that may have been missed to be added to the overall collection

4. Architectural components

The architecture includes four major components: the HPSS, network layer, anonymizing layer and user mobile device.

4.1 Health participatory sensing server

The HPSS provides the central component of the public health sensing system. In this section we will describe its key modules, which are: (1) the data requests/interventions module; (2) the data aggregation module; and (3) the analysis module.

Firstly, the data requests/ interventions module addresses the sending of data requests to end-user mobile devices, but through the intermediary of the anonymizing layer.

Secondly, the data aggregation module receives incoming sensing data, but once again via the intermediary anonymizing layer.

As our approach incorporates submissions of variable resolution (that is submissions for the same task can provide more or less detail), the aggregation module primarily works to integrate this data and provide any data cleansing as necessary.

For the minimum resolution of data the aggregation is straightforward as the more detailed submission are just summarized to the same level. However, for analysis of lower resolution data, where drilldown or greater detail is required, the lower resolution data can either be excluded or extrapolated based on more precise data of other submissions and an approximation approach utilized. Additionally, there are the additional data components (see Section 5.1). Where a component has not been submitted for analysis either the data can be excluded or populated based on statistical averages and an approximation model.

Thirdly, the analysis module calculates metrics of interest for public health analysis by the health organizations from the received sensing data.

4.2 Network layer

The network layer supports communication between the health participatory sensing server and the onion routing network (or mix network).

This layer also carries the data submissions from the onion routing network to the HPSS and the data submission policies/public health interventions from the HPSS to the onion routing network, to then be delivered onwards to the distributed HPSN data nodes (see Section 4.3).

4.3 Anonymizing layer

The Anonymizing Layer consists of a mix network [13] or onion network [14], which provides for anonymity of the submitter as well as secure communication. Such approaches utilize a chain of proxy servers between the participant and HPSS, which can provide anonymity for both parties, though in this case it is only required for the mobile device user. Though this creates additional implementation complexity the potential benefit to real privacy is significant, with the only remaining significant privacy threat being the content of the data submitted.

In this system these proxy servers are referred to as HPSN data nodes.

The limitation of anonymous submission is primarily that it reduces the practicality of detecting and removing invalid or purposefully erroneous data as there is no history of submissions attached to an individual participant.

4.4 User mobile device

Software incorporating the following modules is present on the end user's mobile device. Again the user's mobile device can operate according to the different levels of configuration identified in Section 3. This would depend upon such end user choices as: the external health sensors they have chosen to use, if any; their willingness to receive occasional micro-surveys if any; and their willingness, if any, to participate in and receive public health intervention information. This level of choice would be manifested at both the application level – that is an overall opt-in or out of data collection, health interventions and micro-surveys, as well as allowing controls over specific health organization interactions. This could allow the user to opt-in for example to health interventions from one health organization on a specific topic and opt-in to just data submission with a second organization. In this section we will describe the three key modules of the user mobile device: (1) On-device communication module; (2) Local processing module; and (3) Sensor interface module

Firstly, the on-device communication module interfaces with the onion routing network. However, to complement this privacy approach the on-device communications module operates entirely on a pull approach through the distributed HPSN data nodes for requesting new data submission polices and public health interventions. This is because a push-based approach could be used to selectively distribute narrow policies for short periods of time that could potentially impact on re-identification privacy.

As such, distributed policies have associated distribution timestamps (period after which the policy should no longer be distributed) and expiry timestamps (period whereby the policy should no longer be used on the local device, and needs to be replaced). The on-device communication module checks the distribution timestamp on receipt of new data submission policies/public health interventions and if it has passed, these can be discarded. A similar approach is taken with expiry timestamps, an expired policy/intervention should be discarded and no longer used on the local device and should be replaced.

The other capability of the on-device communications module is the submission of aggregate de-identified anonymized data. The preparation of this data is handled by the Local Processing module with the on-device communication module packaging the data for submission through the onion routing network.

Secondly, is the local processing module, Section 5 describes the local processing provided by this module.

Thirdly, the sensor interface module incorporates all capabilities required to support integration of on-device sensors, external sensors and environmental sensors that may contribute to a data submission.

This module makes use of existing communications standards such as the ISO/IEEE 11073 Personal Health Data standard to carry out communication with external sensors where such standards are adopted.

5. Privacy threshold approach to data aggregation

Our system, by applying granular and modular restrictions upon data collection controlled by the user, reduces real privacy risks though high levels of user control of contribution and restrictions on data potentially usable for re-identification. Additionally, the use of a local processing approach to data submission and health interventions policies allows the on-device adaptation to achieve a data submission which matches the data request as closely as possible without breaching variable user defined privacy conditions.

In this section we will define the overall data aggregation model, the core types of data submission components, a data submission policy approach that allows prioritization of measures/components for submission adaption and a privacy threshold structure to evaluate the requests against.

5.1 Data submission components

The core concept of local processing (on the user mobile device) of health data for anonymized submission requires that individual components of a data submission have an associated quasi-identifier score (QIS). Additionally, as the components are made more generalized such as for example a submission including the city of submission rather than specific postcode, the QIS would be lower to reflect the increased generality. The approach also takes into account the case where multiple quasi-identifiers are submitted together as such a group of quasi-identifiers will have a combined QIS value that is assessed against privacy thresholds. The four core data components in determining the combined QIS are Measures, Location, Temporal and Demographic and are described below.

Measures are aggregate or calculated values that refer to a specific value to be collected. A data collection can have multiple measures for comparison. Examples of possible population-wide anonymized wellness measures are discussed in our previous work [15] and include values such as physical activity patterns and intensity, caloric burn and caloric intake, nutritional data, BMI and sleep regularity and patterns - however this is not an exhaustive list and rather just representative of current sensing capabilities. Emerging wearable patches that may be able to capture some blood constituent information [16], future lab-on-a-chip technologies, smartwatches and wirelessly-enabled ‘tattoos’, all portend to significantly extend the capabilities of the proposed smartphone-based population health data capture system.

Location is a pivotal component - the place a measure occurred can be of material relevance to public health. Examples include places physical activity occurred, active transport data (where physical activity is combined with commuting/ transportation) etc. A fine location resolution would have a high QIS score, whilst a more general location would have a lower QIS score.

The Temporal component indicates the period of time in which a measure occurred. Often the overall temporal range would be set by the data request, however to keep the QIS value low, keeping the temporal value of the returned result less precise is preferred.

The demographic component includes all the other data about the participant that may be additionally submitted for data analysis for example gender, age, ethnicity etc.

5.2 Data submission policies

Data submission policies will have:

1. Core data requirements – typically a Measure value. If this is not submittable without breaching a privacy threshold the submission is not possible.
2. Supplementary data requirements – additional data components that can be submitted alongside the core data requirements. To allow for the calculation of the highest level of data that can be submitted without breaching the threshold, the additional data components will be weighted by importance and whether a less specific data submission is acceptable for a data component as a secondary weighting.

Our algorithm (see Section 5.3) will calculate the inclusion of data components versus the resolution (the detail) of data to create the most suitable data submission (based on weightings) that can be achieved. This will allow beyond the inclusion decision, the level of detail that is submitted e.g. for time data, reducing the resolution down to a larger time period rather than an exact time could avoid a location/time threshold limit as well as lowering the overall submission QIS to meet the overall threshold allowing for more detailed data for other data components.

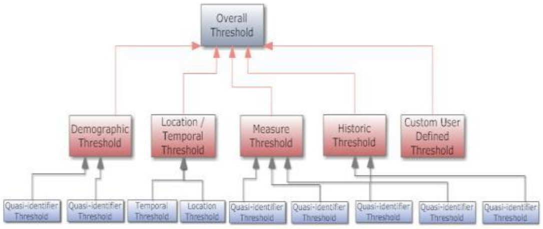
Overall Threshold
Demographic Threshold
Location / Temporal Threshold
Measure Threshold
Historic Threshold
Custom User Defined Threshold
Quasi-identifier Threshold
Quasi-identifier Threshold
Temporal Threshold
Location Threshold
Quasi-identifier Threshold
Quasi-identifier Threshold
Quasi-identifier Threshold
Quasi-identifier Threshold
Quasi-identifier Threshold

Figure 2. Privacy threshold structure

Additionally, as shown in Figure 2 a tree-based approach to the privacy threshold structure is utilized, where all lower level thresholds as well as the overall threshold cannot be exceeded by a data submission QIS. Apart from the threshold related to the data components we identified in the previous subsection, there are the additional thresholds of ’Historic’ and ’Custom User Defined’. Historic relates to a limitation as to how often and how many times a mobile device will submit similar data (typically based on the same measure for a specific temporal range) to a given data requester or to all requesters generally. Finally, the user defined threshold allows for the limitation of certain contexts or combinations of data components that they would like to restrict in addition to the standard thresholds.

5.3 Algorithm for data collection policy processing

The data request is proceed by the local processing module by adapting the data submission request to the anonymous submission settings on the local device. Firstly (Figure 3), it is confirmed that the required minimum data can be submitted (data components with an inclusion weighting greater than the required inclusion threshold), at the minimal level of precision, without exceeding any privacy constraints.

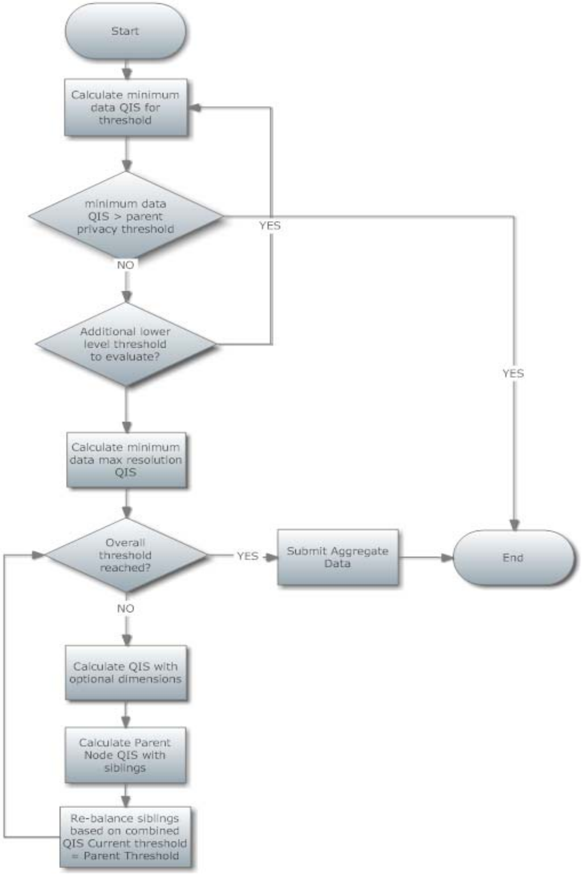
Start
Calculate minimum data QIS for threshold
minimum data QIS > parent privacy threshold
YES
NO
Additional lower level threshold to evaluate?
YES
Calculate minimum data max resolution QIS
Overall threshold reached?
YES
Submit Aggregate Data
End
NO
Calculate QIS with optional dimensions
Calculate Parent Node QIS with siblings
Re-balance siblings based on combined QIS Current threshold = Parent Threshold

Figure 3. Data collection rule processing algorithm

Secondly, the level of precision of the required minimum data is increased based on the resolution rating up to the level that the maximum precision or privacy threshold is met.

Thirdly, if there is additional QIS margin to the threshold at this point, optional data components are included. The inclusion of the optional components is calculated based on the inclusion weighting and precision weighting giving an optimal inclusion structure. This approach is performed for all the lower level thresholds of the privacy threshold structure individually then adjusted to meet and balance at the parent node threshold, then adjusted to meet the root threshold and re-balanced. This process is illustrated in Figure 3.

This provides for a personalized adjustment of the submission requirement to meet the previously described privacy rules on specific data or overall data components. This facilitates an easy to manage system of user-level privacy control that does not remove the usefulness of the data for public health data collection.

6. Privacy analysis

Unlike many other types of participatory sensing networks, the proposed public health information collection system does not require high resolution data in relation to each and every component of the data submission.

In fact the goals are of a different nature than those of such participatory sensing networks that for example provide noise maps [4] or air quality data [3].

6.1 Location

While exact GPS location information is typically used as a component for on-device calculation of physical activity, essentially all of this location information can be dispensed with before submission to the public health data system. For example, while the on-device data shows that an individual cycled 50km along a particular route between town A and town B, the aggregate data to be submitted for this event can be simply the physical activity level or caloric burn of cycling 50km – no start point of end-point need be given. This is because it is overall physical activity levels or alternately sedentary behaviour levels that are of interest in relation to public health. By submitting only 50km of cycling, the re-identification level can be shown to be close to zero. This is because the measure (50km cycling), location and temporal details are unlikely to be statistically unique. For example in Australia which has a fairly low cycling participation rate compared to the international community, in an average week 3.6 million Australians (18% of the population) [17] use a bicycle for transport of recreation. Though this is more or less common based on particular demographic groups, with that level of scale it is unlikely that an individual contribution would be statistically unique. Additionally, using the known demographic distributions from previous research the submission can be adapted to minimize the re-identification risk for rarer demographics (see analysis in Section 7).

An example of adaptive local processing of location privacy is demonstrated in previous work [18]. However, as public health data submission does not require location data at all for submissions this means our approach can be shown to provide complete location privacy at the most conservative privacy setting. However, though not required for the core purpose of public health data collection, there are niche analyses that could benefit from more detailed location information which would operate on a privacy threshold approach. Using a simple calculation such as:

$$L_{QIS} = 1/d * \lambda$$

In the above formula d is the population density of the area and λ is the location resolution.

6.2 Temporal

While a participatory network seeking to capture food intake, might in theory involve capturing this information per meal and submitting this, for the purposes of public health data capture, such time-specific data is not required. For example, simply submitting the aggregate nutritional intake for a week may be more than sufficient for public health measures, and significantly more detailed than provided by current public health data approaches.

Based upon known population food intake distributions [19] such a level of detail will not result in re-identification possibilities.

Knowing more specifically the detail of when a measure started, ended or narrowing the period of time it occurred during (through granules or shorter reporting length), can be considered to affect risks of re-identification. As such, we identify the following characteristics of a temporal period to be considered in terms of calculating its QIS:

- Length – the duration of the time period. Longer periods will have a higher number of potential submissions and as such are less likely to result in re-identification.
- Granules – is it possible to break the total period (and the associated measure) into its component parts and at what resolution.
- Start time – whether the start time is standard or targeted (standard would imply typical data submission breakdowns such as start of day, start of week, morning, evening, night etc) e.g. 00:00am or 9:15am
- End time – whether the end time is standard or targeted e.g. 23:59pm or 9:33am

As such we use the following formula to calculate the Temporal QIS:

$$T_{QIS} = T_{calc}(L/(1 - G), S, E)$$

In the above formula L is the length, G is the granules, S is the start time and E is the end time privacy assessment value.

6.3 Demographics

In public health data capture systems, the types of demographic data needed such as age or age range, gender, major ethnicities, city or zip/postcode are typically non-identifying so long as they represent a large enough share of the population. The population demographics of regions and countries are already collected for public planning and research due to collection of census data or similar large scale data collections. Additionally, in some cases averages are known for specific activities that may be used in measures, such as the cycling example discussed in relation to the Location component [17]. As such, based on this existing data the probability of a combination of demographics can be calculated and compared against a privacy threshold setting. Such as in the formula below where λ is the individual demographic details.

$$D_{QIS} = 1 - Pr(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_n)$$

6.4 Measures

The identifiability arising from specific measures can be decreased to near zero simply by decreasing the location and temporal resolution as described above. Additionally, in most public health data submissions that do not require specific location or temporal details, the only potentially privileged data that would be at risk is the measure value. Therefore, if re-identification is achieved through external knowledge of an individual's measure value no actual leak of information has occurred.

However, in cases of multiple measures in a single submission, it would be possible that one measure could provide re-identification and exposure of an additional measure. As such it is required to impose a threshold on the measure component of the submission, which can require obfuscation or exclusion of measures from the submission.

$$M_{QIS} = \lambda_A A_1 + \lambda_B B_2 + \lambda_C C_3 + ... + \lambda_D D_n$$

In the above formula A, B, C and D are individual measures and λ_x is the resolution for the measure x.

6.5 Public health interventions and feedback

Although other participatory sensing applications do not have a public health intervention component, parallels can be drawn between some interventions and participatory sensing that involves tasking. The use of targeted or personalized tasks/interventions would usually involve the HPSS knowing enough detail about the individual to provide this capability. However, to provide a higher level of privacy, targeting/personalization can be performed on the local device based on the much more specific detail available there. Additionally, the use of an onion routing network restricts the risk of the HPSS being aware of which individual mobile devices have received particular interventions.

After interventions are performed on a mobile device, feedback regarding the effectiveness and suitability of the intervention would be required for public health usage. For example for a specific public health campaign it may be necessary to know which interventions were initiated and what effect they had on an individual over a 3 month period. As with other data submissions the type of intervention and the metrics of success can be considered the 'measure' and the other details, the additional data components. The same approach can be taken in regard to privacy thresholds to ensure that whilst a very specific intervention can be issued, it is not reported as the specific intervention type, if to do so would violate a privacy threshold.

6.6 Overall threshold

The overall threshold is calculated by combining the L_{QIS}, T_{QIS}, D_{QIS} and M_{QIS}.

Stage 1: $\theta_{L/T} > \omega_L L_{QIS} + \omega_T T_{QIS}$
Stage 2: $\theta_{LTDM} > \omega_{LT}\theta_{L/T} + \omega_D D_{QIS} + \omega_M M_{QIS}$

In the above formulae θ_x refers to the threshold for x and ω_y refers to the weighting on individual thresholds/QIS components of a higher level threshold.

7. Privacy evaluation

To demonstrate the operation of this approach we evaluate an example data submission for the Greater Sydney Metropolitan area based on real data distributions. This fits the purpose of public health data collection well as typically such initiatives are targeted to a large area. Additionally, our approach aims to provide high levels of privacy for participants at a significant scale of submissions. As such, we consider the use of known population data and the analysis of the likely *k* values of data submissions at varying levels of submissions will provide a straightforward approach to compare the effective privacy in terms of the risk of re-identification.

This area of Greater Sydney has a population of 4,391,674 as of last census. Using the Australian Bureau of Statistics census population statistics [20] we generated a random data set based on the relative size of the demographics, specifically looking at gender, age range and ancestry based on parents' country of origin. Additionally, to create plausible activity measures we then generated activity averages and cycling participation based on real data [17].

Assessing our local processing approach we generated the data set out to a specific number of participants submission numbers: 50,000, 100,000, 200,000, 400,000 and 800,000. We then tallied the number of unique demographic types that had a number of submissions *k*, under the threshold of 20, 10 and 5. Having a small *k* value for a specific demographic is undesirable, as it can allow for potential re-identification or inference based attacks to be used against the data set.

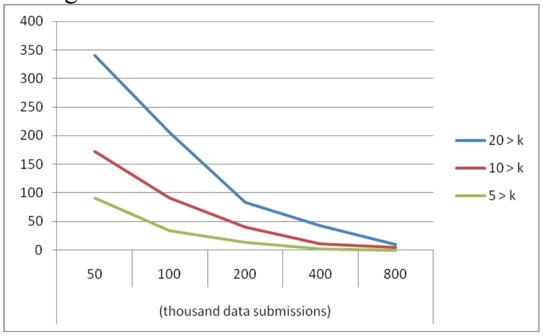

Figure. 4. Demographic *k* value without local processing

As can be seen in Figure 4, there were high numbers of unique demographic combinations at 50,000 submissions, that had low *k* values with nearly 350 different groups having a *k* value lower than 20 and nearly 100 having a *k* value lower than 5. In practice this would be problematic in ensuring anonymity and privacy of data submissions. As, for example, if additional knowledge that an individual participates in the population data submission is available, it may be enough to perform re-identification of some individuals. As the data submissions are increased to 800,000 these risks diminish but there is still a reasonable potential chance of re-identification even at significant data collection levels of 400,000.

To improve this result we implemented our demographic formula and set a reasonably conservative threshold value for D_{QIS}. The other QIS scores M_{QIS}, T_{QIS}, and L_{QIS} were not significant values of the θ_{LTDM} and were not adapted. As ancestry was the optional value in this submission it was adjusted. If a D_{QIS} value for an individual was over the threshold based on known population demographics ancestry details were excluded from the submission.

As demonstrated in Figure 5 this resulted in a dramatic decrease in the number of unique demographic groups that had low *k* values, with less than a tenth of the unadjusted submission approach. This differentiation increased as the number of data submissions increased with the adjusted submission approach reaching a safe level much sooner at ~400,000 and comprising as low as 2.5% of the relevant demographic groups below the threshold at the 200,000 submission level.

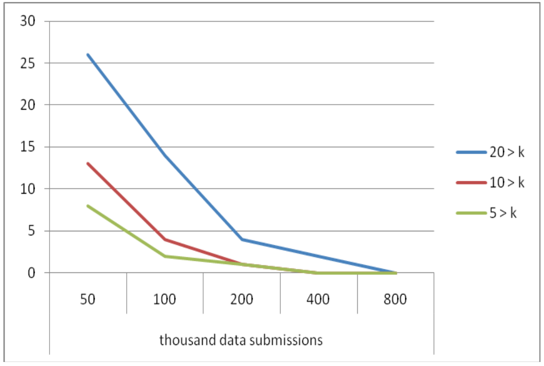

Figure. 5. Demographic *k* value with local processing

The threshold at the local device level could of course be adjusted either higher or lower based on the expected submission numbers. However, it performed quite respectably at the initial level with a significantly lower level of risk at the 50,000 submission level and close to no statistical risk at the 400,000 level which represents 9.1% of the area total population.

The limitation of this local processing approach as compared to a trusted server approach that performs *k*-anonymity, is that the number of other submissions cannot be known with certainty by the local device. As such, the privacy threshold is set at a conservative value to preserve privacy. However this means that when there are high levels of submissions more records are obfuscated/adjusted than was required. This relationship is displayed in Figure 6 where for 50,000 records the number of additional records adjusted and the miss rating was extremely low. This diverges as the number of data submissions increases, since the adjustment level remains fairly constant at around 7.8% of data submissions for the example data set.

Overall, this wouldn't pose a serious problem, as the priority of the demographic detail is controlled by the data requestor and trade-offs are to be expected for increased detail in other sections.

In summary, for the example data set the local processing aggregate data approach performed favorably compared to the defined public health requirements and privacy limitations.

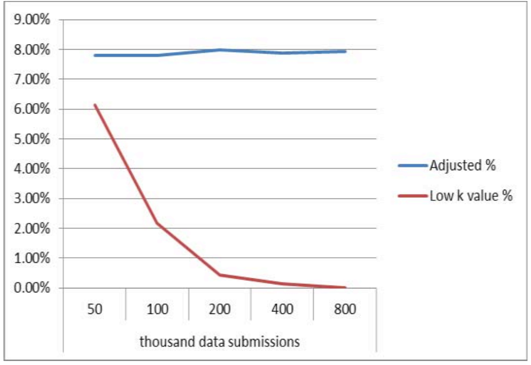

Figure. 6. Adjusted submission compared to low *k* value submissions

8. Discussion and future work

The on-going development of participatory sensing technologies and the greater understanding of participant values and requirements of systems gathered from early adopters will continue to influence and extend the types of participatory sensing possible and its potential in the health context. Of significance to health participatory sensing is the development of new and advanced sensors that continue to extend the range of what can be sensed and detected [21]. Additionally, the growth in smart-device ownership and personal electronic health tracking will continue to drive the potential of health participatory sensing.

Our approach focused on alleviating privacy issues that would be inherent in developing public health data collection capabilities from participatory sensing and personalized intervention platforms. As such, the system would be quite resilient to extension via new sensors or sensor systems as they would present just an additional data measure, where the key privacy restrictions are demographic, temporal and spatial-based. However, the extension of sensor capabilities potentially may reach the point where sensor systems are diagnostic in nature which would result in the measure itself being of a sensitive nature, in a similar manner to portions of a private electronic health record. These considerations can also be resolved within the bounds of the existing approach.

However, privacy and public perceptions of such participatory sensing approaches need to be further explored. As such, future work could include studies of perceived privacy of participatory sensing applications specific to the health domain. A useful extension of this approach would be to consider incentivization, adoption and health organization acceptance of such approaches.

9. Conclusion

This paper presents a smartphone-based system for population-scale anonymous capture of public health data and interventions, with the new and powerful capability that data requests and public health interventions can be distributed, performed and evaluated without the need for identifying details of an individual participant to ever leave their mobile device. This includes an approach based on local processing of aggregate public health data that utilizes privacy thresholds and an adaptable approach to data submission that supports the data collection model for HPSNs, utilized for the purpose of public health data collection and interventions. To this end we included an approach to submission rules/health intervention rules that allows a compromise between individual privacy and public health application requirements and an algorithmic approach to computing QIS to compare to threshold privacy values. Additionally, we provided an evaluation of the privacy preserving characteristics of the system at the level of large user numbers.

References

[1] Klasnja, P., and Pratt, W., "Methodological Review: Healthcare in the Pocket: Mapping the Space of Mobile-Phone Health Interventions", J. of Biomedical Informatics, 45(1), 2012, pp. 184-198.
[2] Kalnis, P., and Ghinita, G., "Spatial K-Anonymity", in (Liu, L., and Özsu, M.T., 'eds.'): Encyclopedia of Database Systems, Springer US, 2009, pp. 2714-2714.
[3] Predic, B., Zhixian, Y., Eberle, J., Stojanovic, D., and Aberer, K., "Exposuresense: Integrating Daily Activities with Air Quality Using Mobile Participatory Sensing", Pervasive Computing and Communications Workshops, 2013 IEEE International Conference on, 2013, pp. 303-305.
[4] Wisniewski, M., Demartini, G., Malatras, A., and Cudré-Mauroux, P., "Noizcrowd: A Crowd-Based Data Gathering and Management System for Noise Level Data", in (Daniel, F., Papadopoulos, G., and Thiran, P., 'eds.'): Mobile Web and Information Systems, Springer Berlin, 2013, pp. 172-186.
[5] Ganti, R., Mohomed, I., Raghavendra, R., and Ranganathan, A., "Analysis of Data from a Taxi Cab Participatory Sensor Network", in (Puiatti, A., and Gu, T., 'eds.'): Mobile and Ubiquitous Systems: Computing, Networking, and Services, Springer, 2012, pp. 197-208.
[6] Ganti, R.K., Pham, N., Ahmadi, H., Nangia, S., and Abdelzaher, T.F., "Greengps: A Participatory Sensing Fuel-Efficient Maps Application", Proceedings of the 8th international conference on Mobile systems, applications, and services, 2010, pp. 151-164.
[7] Cornelius, C., Kapadia, A., Kotz, D., Peebles, D., Shin, M., and Triandopoulos, N., "Anonysense: Privacy-Aware People-Centric Sensing", 6th international conf. on Mobile systems, applications, and services, 2008, pp. 211-224.
[8] Christin, D., "Impenetrable Obscurity Vs. Informed Decisions: Privacy Solutions for Participatory Sensing", Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on, 2010, pp. 847-848.
[9] Mun, M., Hao, S., Mishra, N., Shilton, K., Burke, J., Estrin, D., Hansen, M., and Govindan, R., "Personal Data Vaults: A Locus of Control for Personal Data Streams", Proceedings of the 6th International Conference on emerging Networking EXperiments and Technologies, 2010, pp. 1-12.
[10] Choi, H., Chakraborty, S., Charbiwala, Z.M., and Srivastava, M.B., "Sensorsafe: A Framework for Privacy-Preserving Management of Personal Sensory Information", Proceedings of the 8th VLDB international conference on Secure data management, 2011, pp. 85-100.
[11] Klasnja, P., Consolvo, S., Mcdonald, D.W., Landay, J.A., and Pratt, W., "Using Mobile & Personal Sensing Technologies to Support Health Behavior Change in Everyday Life: Lessons Learned", AMIA Annu Symp Proc, 2009, pp. 338-342.
[12] Clarke, A., and Steele, R., "How Personal Fitness Data Can Be Re-Used by Smart Cities", Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on, 2011, pp. 395-400.
[13] Sampigethaya, K., and Poovendran, R., "A Survey on Mix Networks and Their Secure Applications", Proceedings of the IEEE, 94(12), 2006, pp. 2142-2181.
[14] Mauw, S., Verschuren, J.H.S., and Vink, E.P., "A Formalization of Anonymity and Onion Routing", in (Samarati, P., Ryan, P., Gollmann, D., and Molva, R., 'eds.'): Computer Security – Esorics 2004, Springer Berlin Heidelberg, 2004, pp. 109-124.
[15] Clarke, A., and Steele, R., "Summarized Data to Achieve Population-Wide Anonymized Wellness Measures", Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, 2012, pp. 2158-2161.
[16] Schwartz, A., "No More Needles: A Crazy New Patch Will Constantly Monitor Your Blood", Co.EXIST, 2012, http://www.fastcoexist.com/1680025/no-more-needles-a-crazy-newpatch-will-constantly-monitor-your-blood, accessed 11th July, 2013
[17] Austroads, "Australian Cycling Participation "2011, http://austroads.com.au/abc/images/pdf/AP-C91-11.pdf
[18] Agir, B., Papaioannou, A., Narendula, R., Aberer, K., and Hubaux, J.-P., "Adaptive Personalized Privacy in Participatory Sensing"2012, http://infoscience.epfl.ch/record/176115,
[19] Haubrock, J., Nöthlings, U., Volatier, J.-L., Dekkers, A., Ocké, M., Harttig, U., Illner, A.-K., Knüppel, S., Andersen, L.F., Boeing, H., and Consortium, O.B.O.T.E.F.C.V., "Estimating Usual Food Intake Distributions by Using the Multiple Source Method in the Epic-Potsdam Calibration Study", The Journal of Nutrition, 141(5), 2011, pp. 914-920.
[20] ABS, "Census Community Profiles Sydney"2011, http://www.censusdata.abs.gov.au/census_services/getproduct/census/2011/communityprofile/1GSYD, accessed 28/03/2013
[21] Swan, M. "Sensor mania! The Internet of Things, wearable computing, objective metrics, and the Quantified Self 2.0." Journal of Sensor and Actuator Networks 1.3, 2012, pp. 217-253.