Targeted and Anonymized Smartphone-based Public Health Interventions in a Participatory Sensing System

Andrew Clarke[1] and Robert Steele[2]

Abstract— **Public health interventions comprising information dissemination to affect behavioral adjustment have long been a significant component of public health campaigns. However, there has been limited development of public health intervention systems to make use of advances in mobile computing and telecommunications technologies. Such developments pose significant challenges to privacy and security where potentially sensitive data may be collected. In our previous work we identified and demonstrated the feasibility of using mobile devices as anonymous public health data collection devices as part of a Health Participatory Sensing Network (HPSN). An advanced capability of these networks extended in this paper would be the ability to distribute, apply, report on and analyze the usage and effectiveness of targeted public health interventions in an anonymous way. In this paper we describe such a platform, its place in the HPSN and demonstrate its feasibility through an implementation.**

I. INTRODUCTION

The use of information and behavioral adjustment type public health interventions has large potential to evolve into a more targeted, measurable form of public health intervention through the use of new communications and mobile computing platforms. Advantages include the collection of real time or near real time data on the effectiveness of public health interventions, effective long term measurement of benefits and more precise targeting. Additionally, health participatory sensing systems such as HPSNs [1] allow for potential population-wide data capture, the ability to more rapidly change an intervention/collection approach and reduction of some of the biases associated with survey based methodologies.

HPSNs differ from other health communication systems such as interconnected EHR and PHRs [2] which deal with identified individuals and their personal health data and possible communication such as appointment reminders or medication adherence, by focusing on collecting data that is non-identifying, and is used for overall population measurements and behavioral or informational public health communication rather than individual specific medical communication.

However, such advances pose their own significant privacy and security challenges that need to be addressed. There are two key challenges to this type of public health intervention platform. Firstly, as the specific intervention is by necessity decided on and applied at the local device level a large number of broader interventions need to be delivered to each device efficiently. Secondly, is the need to report with as much detail as possible, as to which intervention was performed and its effectiveness without breaching privacy, or inadvertently allowing for individual re-identification at a later stage.

We propose as a solution to these problems which is an extension and combination of our prior work in relation to HPSNs [1], [3] and query assurance [4]. The query assurance architecture is adapted to reduce the quantity of health interventions that need to be delivered to participants and hence the resultant computation load on the devices. The HPSN approach is used as the data collection and distribution framework for public health interventions, as the interventions are distributed, applied, and the outcomes collected and analyzed within the existing capabilities of the HPSN framework.

II. RELATED WORK

The rich capabilities of participatory sensing have garnered interest in its usage for a range of quite disparate areas such as air quality and pollution sensing [5], to urban area noise level data collection [6] and public health data collection [1] amongst many others. 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 be to use a trusted server, then k-anonymity [7] or a variant, to anonymize the data before it is accessible for research/analysis. The main downside of this type of approach is the need for a fully trusted server, which creates 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 [8] 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 information on the participant is a key submission component (such as in the case of collection of public health intervention data) as de-identification protection is still implemented at a central trusted point. There has been some prior research to resolve the issue of requiring a fully trusted server, such as, decentralized participatory sensing networks [9] using user interaction/awareness as part of the approach or keeping the data managed by the participant [10], [11] and stringent user-definable access control mechanisms to manage sharing. The limitation of these approaches when considering HPSNs is

[1]A. Clarke is with the Faculty of Health Science, University of Sydney, Sydney, NSW 2006, Australia andrew.clarke@sydney.edu.au

[2]R. Steele is with the School of Engineering and Technology, CQUniversity, Sydney, NSW 2000, Australia robert.steele@cqu.edu.au

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

that typically they have not incorporated support for public health interventions (or an equivalent), a capability that does not have a direct parallel in most participatory sensing systems and remains an important component of HPSNs.

III. Participatory Sensing for Public Health

The growth in the potential for participatory sensing has been greatly increased through the high levels of smartphone adoption in many countries [12] and proliferation of commercial wearable devices and health sensors, leading to the pervasive availability of powerful sensing platforms that are highly human-centric, making them ideal as the center-points for health participatory sensing models.

In our previous work [1] we identified a number of different classifications for participation in a HPSN. The classification most relevant to public health interventions is 'active participatory sensing'. Active participatory sensing differs from other types of participatory sensing by providing inputs to the individual to alter the actions they would have taken whilst participating in the HPSN. Active participatory sensing in the health context has a somewhat different goal to that of many other active participatory sensing contexts [13]. While an active participatory model for typical sensing might focus on affecting individuals to collect a more complete data set in terms of spatial/temporal range, health and epidemiological-related active participatory 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 behavior change would be to firstly attempt to improve the sensing data captured in terms of risk and preventative factors. Additionally for public health goals, this allows for immediate and continuous feedback on the effectiveness of interventions on receiving groups. It is assumed that active participatory sensing would have similar levels of technical sensor capabilities to other classifications [1], with the focus shifted to the potential two-way communication that can be built on sensing data and an inherent feedback loop. This has the potential to be both a powerful data collection tool as well as a novel public health intervention platform. Its potential scope includes the ability, in a timely and accurate manner, to quantify precisely the effectiveness of public health interventions.

IV. Public Health Intervention Platform

As a necessity, an anonymous public health intervention platform will need to be incorporated into a larger system which provides for public health data collection. This is because without such a larger capability the effectiveness of the utilized public health interventions could not be collected and analyzed in a timely manner. Even without this larger system the intervention system can still provide a lesser but still significant improvement over traditional public health information/behavioral interventions. As such, we consider that public health interventions can be conducted as a component of a HPSN as described in section III and our previous work [1].

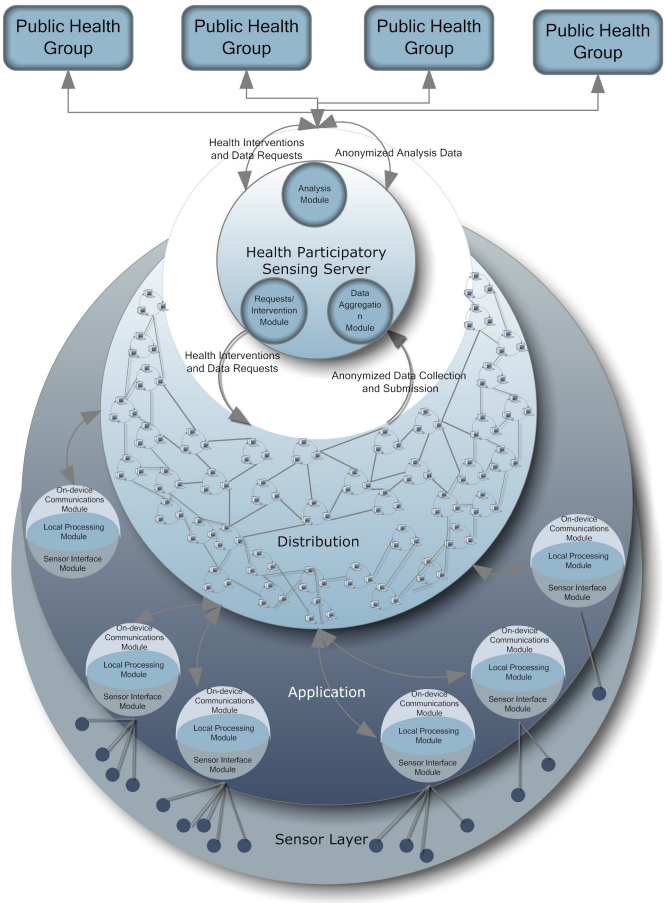

Fig. 1. Public Health Invention Platform

The platform components and their inter-relationships are illustrated in Fig 1, and serve to support the capabilities of anonymous distribution, local application of public health interventions, data collection for reporting, and analysis of results. These are described in further detail in the following subsections.

A. Distribution

The distribution of public health interventions in the HPSN comprises of two main components, the distribution network and the distribution approach.

The distribution network consists of a mix network [14] or onion network [15], which provides for anonymity of the submitter as well as secure communication. Such approaches utilize a chain of proxy servers between the participant and HPSN, 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 threats being: insecure storage of data on the local device which we consider outside the HPSN network; and re-identification via the content of the data submitted discussed below.

The distribution approach utilizes our previous query assurance approach [4] to provide granular completeness,

correctness and freshness assurance of the public health interventions that are distributed to the HPSN clients. This approach uses an implementation of one or many sorted and digitally signed merkle hash tree/s utilizing expiring timestamps, retrieved alongside the requested data to verify the content of the retrieved data. This allows for a hash of each possible granule of retrieved data to be efficiently distributed with a single digital signature and expiring time stamp for the overall request, reducing verification overhead of both computation time and data. This is effective even where only subsets of the overall data are retrieved through a third party or untrusted distributed network. This allows for high levels of certainty of the validity of data, while allowing for flexibility in request size even though the data is distributed through untrusted nodes, while keeping verification data overhead size and processing time to acceptable levels.

B. Application

The public health interventions are performed on the local device. The decision as to the intervention to perform also must be made locally as more specific information about the individual is not transmitted to the HPSN server. As such, the specific intervention is chosen locally to most closely match the individual's demographic and health profile details, even if those details cannot be fully disclosed to the server.

C. Reporting collection

To provide an anonymous public health intervention system that also collects outcomes and the effectiveness of those executed interventions, a level of data collection is a necessity. However, if the necessary limitations on data collection are not considered, this could result, even in cases where de-identification of data is performed locally, in unwanted re-identification of data at a later stage using data external to the HPSN [3]. This potential scenario is a significant concern of HPSNs and by extension public health interventions systems on such networks. We consider that the most effective way to mitigate this risk that doesn't require a trusted server or aggregation in some form is the use of local processing of data reporting at a suitably conservative privacy setting to minimize risks.

As such our system, by applying granular and modular restrictions upon data reporting [3], reduces real privacy risks through a threshold approach to privacy and submissions. The local processing approach, considers the potential for re-identification before submission and reduces or modifies the number or detail of the demographics submitted. 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 reporting request as closely as possible without breaching variable user defined privacy conditions [3].

For public health interventions this is resolved by submitting aggregate data that is not time or location sensitive, with restrictions on the specificity of the intervention reported to be performed. That is, for example if an intervention was targeted at the entire population a certain level of demographic detail could be returned as well as the intervention type and the effectiveness of the intervention as a measure, such as any measurable change in behavior or health indicators. Alternatively, if the intervention was tightly focused on a small subset of the community, the specific intervention type may need to be reported as a broader type that is inclusive of the specific type and limited additional demographic details as prioritized by the intervention request.

D. Analysis

The analysis of public health participatory sensing data relies on collection of sufficient data for public health uses [16], which differs from what would be required in most other participatory sensing systems. As such, generally aggregate nonspecific demographic level data is needed as well as the measured values and the types of interventions performed.

V. Implementation and Results

Our implementation provides an approach that addresses the key challenges of efficient public health intervention distribution and the reporting of the application and effectiveness of public health interventions.

Public health interventions are likely to include a combination of text, images, video and audio components. Additionally, even when considering only targeting broad demographics this can result in potentially tens of thousands of different combinations for targeted interventions, when extending this to specific data about the individual stored at the mobile device level and multiple public health groups/organizations involved in the system. While much of this overhead could be reduced through conditional approaches which make a single intervention relevant to multiple targets the core problem remains. As such, we created an example data set that includes different data types and compares the data overhead of retrieving the entire data set, to retrieving a subset and a verification tree for data quality assurance and our previous approach that utilizes a more efficient verification tree [4].

The data setup involved 2000 components typical of an audio/visual intervention size and 10000 components of a text and intervention details size. These components were verified by a single verification tree [4] (see Section IV-A). Even with this limited size dataset it is apparent that it wouldn't be feasible to distribute the entirety of the interventions to any particular user, as this would represent hundreds of megabytes. Our proposed approach instead involves a user requesting a subset (approximately 8-10 megabytes). The request is broad enough that it does not expose any personally identifiable details, then uses the verification tree to authenticate the subset. This removes the need for direct communication with the source, or for the source to hash and digitally sign every possible requested combination.

An additional component of our approach is optimization of the verification tree based on historic usage [4]. As such we perform our implementation pre and post optimization

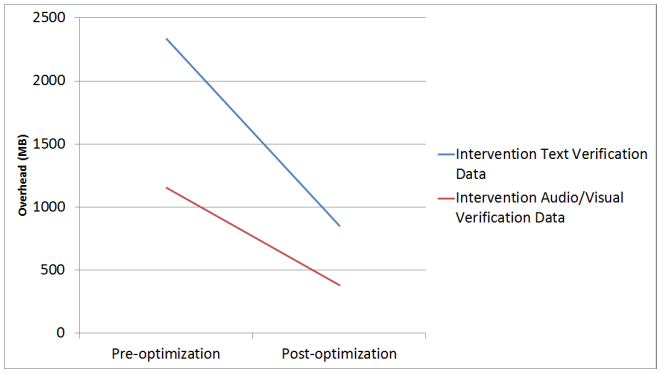

Fig. 2. Public Health Intervention Distribution Verification Overhead

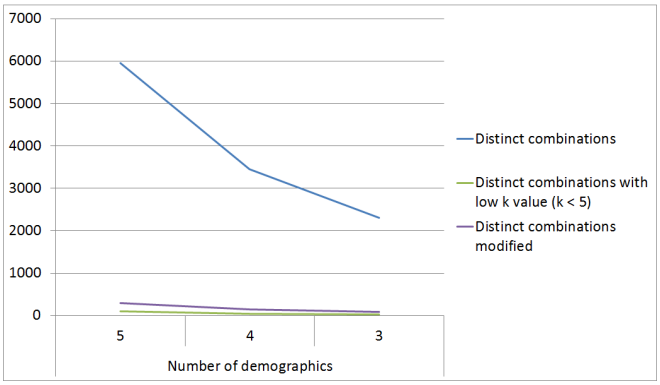

Fig. 3. Public Health Intervention Distinct Demographic Combinations to low k value combinations

incorporating 5000 requests for each. The results of the verification overhead are displayed in figure 2.

The reporting of the application of public health interventions can raise some issues relating to the potential for re-identification of the individual through their submission. As such to address this issue we utilize our previous approach for public health data collection [3], extended and modified for public health interventions, which uses a threshold and priority approach to decide what information is reported for analysis before locally processing the result and submitting through an anonymous communications network [3] (see Section IV-C for details). The implementation involves applying specific example public health interventions at the client levels, utilizing the privacy threshold approach to process the data for submission. This is followed by analysis of the submissions for their potential re-identification risk as a k-anonymity value and compared to the number of example interventions that were returned with less specific detail (for example with fewer demographic details).

To demonstrate the operation of this approach we constructed a prototype that creates a set of clients each with randomized demographics, interventions, location and time records. These clients then process a set of 100000 reporting submission requests which are submitted to the prototype server and evaluated for privacy considerations. The prototype evaluation used population distributions from the Greater Sydney Metropolitian [17] area to generate the individual clients demographics including age, gender, ethnicity, income and education. The prototype client and server are both developed in Java (1.6), the client uses SQLITE for its data storage and the server uses Microsoft SQL Server Enterprise Edition for its data storage.

The results of the evaluation are displayed in Figure 3, whereby the number of distinct demographic combinations that were collected and hence of possible use for re-identification are contrasted against the number of distinct combinations with a low k anonymity value if local processing modification did not occur. We contrast this to the number of modifications our approach made at a local processing level to decrease low k value occurrences to nil. In our implementation results it is apparent that even with a quite high number of distinct combinations it is only a small percentage that needs to be modified/changed to improve privacy, typically in the range of 1-2%, though as this is achieved through a local processing approach a safety buffer is necessary. In the case of our implementation between 3-5% of distinct combinations were modified to remove low k value combinations. This demonstrates two components of our approach, firstly that it is possible to retrieve public health interventions based on demographic grouping to reduce the overall data retrieval requirement without a significant risk of re-identification, and secondly that with minor local processing modification or partially reducing demographics based on local processing as implemented in our approach quite detailed public health intervention feedback can be provided without a privacy risk.

VI. Conclusion

This paper describes the public health intervention capabilities of a smartphone-based participatory sensing system for population-scale public health data capture and intervention. In particular, we describe the new and powerful capability that 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. Additionally we have considered the efficiency, privacy and anonymity of the intervention capabilities. The smartphone-based public health information systems include an approach based on local processing to aggregate data for public health use that utilizes privacy thresholds and an adaptable approach to public health interventions and reporting. To this end we provided a detailed evaluation of the privacy preserving characteristics of such intervention systems, and an analysis of the overheads and efficiency of the public health intervention distribution model.

References

[1] A. Clarke and R. Steele, "Health participatory sensing networks," *Mobile Information Systems*, vol. 10, 2014.

[2] R. Steele, K. Min, and A. Lo, "Personal health record architectures: Technology infrastructure implications and dependencies," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, pp. 1079–1091, 2012.

[3] A. Clarke and R. Steele, "A smartphone-based system for population-scale anonymized public health data collection and intervention," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pp. 2908–2917, Jan 2014.

[4] A. Clarke and R. Steele, "Secure query assurance approach for distributed health records," *Health Systems*, vol. 3, pp. 60–73, Feb 2014.

[5] B. Predic, Z. Yan, J. Eberle, D. Stojanovic, and K. Aberer, "Exposuresense: Integrating daily activities with air quality using mobile participatory sensing," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pp. 303–305, March 2013.

[6] M. Wisniewski, G. Demartini, A. Malatras, and P. Cudr-Mauroux, "Noizcrowd: A crowd-based data gathering and management system for noise level data," in *Mobile Web Information Systems* (F. Daniel, G. Papadopoulos, and P. Thiran, eds.), vol. 8093 of *Lecture Notes in Computer Science*, pp. 172–186, Springer Berlin Heidelberg, 2013.

[7] P. Kalnis and G. Ghinita, "Spatial k-anonymity," in *Encyclopedia of Database Systems* (L. Liu and M. T. Özsu, eds.), p. 2714, Springer US, 2009.

[8] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, "Anonysense: privacy-aware people-centric sensing," in *Proceedings of the 6th international conference on Mobile systems, applications, and services*, MobiSys '08, (New York, NY, USA), pp. 211–224, ACM, 2008.

[9] D. Christin, "Impenetrable obscurity vs. informed decisions: privacy solutions for participatory sensing," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, pp. 847–848, 2010.

[10] M. Mun, S. Hao, N. Mishra, K. Shilton, J. Burke, D. Estrin, M. Hansen, and R. Govindan, "Personal data vaults: a locus of control for personal data streams," in *Proceedings of the 6th International COnference*, Co-NEXT '10, (New York, NY, USA), pp. 17:1–17:12, ACM, 2010.

[11] H. Choi, S. Chakraborty, Z. Charbiwala, and M. Srivastava, "Sensorsafe: A framework for privacy-preserving management of personal sensory information," in *Secure Data Management* (W. Jonker and M. Petkovic, eds.), vol. 6933 of *Lecture Notes in Computer Science*, pp. 85–100, Springer Berlin Heidelberg, 2011.

[12] Gartner, "Gartner says worldwide smartphone sales soared in fourth quarter of 2011 with 47 percent growth," 2011.

[13] J. Rula and F. E. Bustamante, "Crowd (soft) control: Moving beyond the opportunistic," in *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, HotMobile '12, (New York, NY, USA), pp. 3:1–3:6, ACM, 2012.

[14] K. Sampigethaya and R. Poovendran, "A survey on mix networks and their secure applications," *Proceedings of the IEEE*, vol. 94, pp. 2142 –2181, Dec. 2006.

[15] S. Mauw, J. Verschuren, and E. de Vink, "A formalization of anonymity and onion routing," in *Computer Security ESORICS 2004* (P. Samarati, P. Ryan, D. Gollmann, and R. Molva, eds.), vol. 3193 of *Lecture Notes in Computer Science*, pp. 109–124, Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-30108-0_7.

[16] A. Clarke and R. Steele, "Summarized data to achieve population-wide anonymized wellness measures," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 2158–2161, 2012.

[17] Australian Bureau of Statistics, "Census community profiles greater sydney," 2011.