



THE UNIVERSITY OF
SYDNEY

COPYRIGHT AND USE OF THIS THESIS

This thesis must be used in accordance with the provisions of the Copyright Act 1968.

Reproduction of material protected by copyright may be an infringement of copyright and copyright owners may be entitled to take legal action against persons who infringe their copyright.

Section 51 (2) of the Copyright Act permits an authorized officer of a university library or archives to provide a copy (by communication or otherwise) of an unpublished thesis kept in the library or archives, to a person who satisfies the authorized officer that he or she requires the reproduction for the purposes of research or study.

The Copyright Act grants the creator of a work a number of moral rights, specifically the right of attribution, the right against false attribution and the right of integrity.

You may infringe the author's moral rights if you:

- fail to acknowledge the author of this thesis if you quote sections from the work
- attribute this thesis to another author
- subject this thesis to derogatory treatment which may prejudice the author's reputation

For further information contact the University's Copyright Service.

sydney.edu.au/copyright

Health Participatory Sensing Networks for Mobile Device Public Health Data Collection and Intervention

Andrew P. Clarke

Faculty of Health Science

University of Sydney

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

October 2015

To my parents, Larry and Frances Clarke, who have provided so much support.

To my undergraduate lecturer and supervisor, Eric Pardede for convincing me that
candidature, publications and all of this was within my reach.

Finally, to my partner, Jane Woo, whose positivity and patience got me over the final
milestone.

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

A handwritten signature in black ink, consisting of a large, stylized loop followed by a horizontal line that tapers off to the right.

Andrew P. Clarke

October 2015

ACKNOWLEDGEMENTS

First and foremost, I offer my sincerest gratitude to Prof. Robert Steele, whose support throughout my thesis has been unwavering. The path this thesis took, and the high frequency of publications would surely not have been possible without his support, guidance and regular lengthy discussions. Ultimately, it has been a pleasure to work with Robert and my progress through my candidature owes much to his hard work.

My fellow PhD candidates, of which I shared a desk area with (Johanne and Dan) have been great and they have always willingly offered advice or input when asked, and the camaraderie, banter and insight into other's candidature was of benefit to my progress and general good humor.

My mother, who without her willingness to proofread my publications and chapters, the grammar and English would have been of dubious quality. I expect even at the end of this process, despite reading everything multiple times, she still has little idea what my thesis is about. Despite this, she has consistently provided her proofreading services, and caught many grammatical errors I would otherwise have missed.

I also offer my gratitude to Assoc. Prof. Roger Fulton for his supervisory contribution, even though it was outside his area of expertise. Additionally, I would like to thank Dr Peter Kench for his contributions to my supervision.

The Faculty of Health Sciences that has provided the support and equipment I have needed to produce and complete my thesis.

Additionally, I would like to acknowledge the contributions made by the Commonwealth

Government of Australia, which has provided me with an Australian Postgraduate Award.

Finally, I thank my parents, Larry and Frances for supporting me throughout all my studies at University, I owe you both so much.

ABSTRACT

The pervasive availability and increasingly sophisticated functionalities of smartphones and their connected external sensors or wearable devices can provide new data collection capabilities relevant to public health. Current research and commercial efforts have concentrated on sensor-based collection of health data for personal fitness and personal healthcare feedback purposes. However, to date there has not been a detailed investigation of how such smartphones and sensors can be utilized for public health data collection purposes.

Unlike most sensing applications, in the case of public health, capturing comprehensive and detailed data is not a necessity, as aggregate data alone is in many cases sufficient for public health purposes. As such, public health data has the characteristic of being capturable whilst still not infringing privacy, as the full detailed data of individuals that may allow re-identification is not needed, but rather only aggregate, de-identified and non-unique data for an individual. For example, rather than details of the physical activity such as specific route, just total caloric burn over a week or month could be submitted, which is much less unique and thereby not identifying the individual.

These types of public health data collection provide the challenge of the need to be flexible enough to answer a range of public health queries, while ensuring the level of detail returned preserves privacy. Additionally, the distribution of public health data collection request and other information to the participants without identifying the individual is a core requirement, with a additional need for any approach to be extremely scalable to population levels.

An additional requirement for health participatory sensing networks is the ability to perform public health interventions. In line with the requirements above, this needs to be completed in a non-identifying and privacy preserving manner.

This thesis proposes a solution to these challenges, whereby a form of query assurance is used to provide private and secure distribution of data collection requests and public health interventions to the participants. While an additional, privacy preserving threshold approach to local processing of data prior to submission is used to provide re-identification protection for the participant.

In brief, this thesis summarizes the related research, introduces, prototypes and evaluates a new type of public health information system to provide aggregate population health data capture and public health informational or behavioral intervention capabilities via utilizing smartphone and sensor capabilities, whilst maintaining the anonymity and privacy of each individual.

In particular the key aspects of privacy, anonymity and intervention capabilities of these emerging systems are considered and a detailed evaluation of anonymity preservation characteristics is carried out.

The evaluation finds that with manageable overheads, minimal reduction in the detail of collected data and strict communication privacy; privacy and anonymity can be preserved. This is significant for the field of participatory health sensing as a major concern of participants is most often real or perceived privacy risks of contribution. Furthermore, for such a system to meet its potential a large user base is required, so the reduction of participant concerns is paramount to the system's success.

Table of contents

Table of contents	xi
List of figures	xv
Nomenclature	xviii
Publications	
<i>A list of publications that resulted from the work conducted in this dissertation</i>	1
1 Introduction	3
1.1 Synopsis of the Thesis	5
2 Related Work	9
2.1 Introduction	9
2.2 Participatory Sensing Systems for Public Health	10
2.2.1 Sensor Capabilities	10
2.2.2 Motivation and Incentivization	14
2.3 Privacy and Security in Public Health Participatory Sensing Systems	19
2.3.1 Trusted Servers	20
2.3.2 Untrusted Servers	21
2.3.3 Query Assurance	22
2.4 Discussion	26
2.5 Conclusions	27

References	37
3 Health Participatory Sensing Networks	39
3.1 Introduction	41
3.2 Classification of Health Participatory Sensing Models	43
3.2.1 Incidental Participatory Sensing	44
3.2.2 Passive Participatory Sensing	45
3.2.3 Passive Participatory Sensing with Subjective Human-sensing and Feedback	46
3.2.4 Active Participatory Sensing	47
3.2.5 Active Participatory Sensing with Subjective Human-sensing and Feedback	47
3.3 Health Participatory Sensing Network Conceptual Reference Architecture	48
3.3.1 Reference Architecture	48
3.3.2 Distribution of Data Collection Policy Rules and Micro Surveys Whilst Protecting Recipient Anonymity and Privacy	51
3.3.3 Anonymous Data Collection and Submission	54
3.3.4 Public Health Interventions	57
3.3.5 HPSN and Participant Device Interaction	58
3.4 User Scenario	60
3.5 Conclusion	61
References	67
4 Query Assurance for Distribution of HPSN Data Collection Rules and Interventions	69
4.1 Introduction	72
4.2 Problem Definition	75

4.3	Proposed Novel Approach for Secure Query Assurance	76
4.3.1	Public Health Data Query Assurance Method	77
4.3.2	Encryption of Health Information Records and Trapdoor Encryption of Index and Search Terms	85
4.3.3	Scalability	86
4.3.4	Summary	87
4.4	Experiment and Evaluation	87
4.4.1	Experiment Setup	88
4.5	Results	89
4.5.1	Computation Time	90
4.5.2	Query Data Overhead	92
4.5.3	Results Discussion	95
4.6	Conclusions and Future Work	96
	References	99
5	Privacy Threshold Approach to HPSN Data Aggregation and Collection	101
5.1	Introduction	104
5.2	Sensor Capabilities and Public Health Measures	106
5.2.1	Sensor Capabilities	106
5.2.2	Public Health Risk Factors	109
5.3	Public Health Information System Architecture	110
5.3.1	Smartphone With or Without External Sensors	112
5.3.2	Intervention Capabilities	112
5.3.3	Extension via Manual Input	113
5.4	System and Prototype Components	114
5.4.1	Health Participatory Sensing Server	114
5.4.2	Network and Anonymizing Layers	115

5.4.3	User Mobile Device	116
5.4.4	Public Health Groups	117
5.5	Privacy Threshold Approach to Public Health Data Aggregation	119
5.5.1	Data Submission Components	119
5.5.2	Data Submission Policies	121
5.6	Privacy	124
5.6.1	Location	124
5.6.2	Temporal	126
5.6.3	Demographics	127
5.6.4	Measures	127
5.6.5	Public Health Interventions and Feedback	128
5.6.6	Overall Threshold	129
5.7	Evaluation	129
5.7.1	Results	130
5.8	Discussion and Future Work	135
5.9	Conclusion	136
	References	141
6	Targeted and Anonymized HPSN Public Health Interventions	143
6.1	Introduction	145
6.2	Related Work	147
6.3	Participatory Sensing for Public Health	148
6.4	Public Health Intervention Platform	149
6.4.1	Distribution	150
6.4.2	Application	151
6.4.3	Reporting Collection	151
6.4.4	Analysis	152

6.5	Implementation and Results	152
6.6	Conclusion	156
	References	159
7	Discussion and future work	161
7.1	Query Assurance	162
7.2	Public health data collection and intervention	164
8	Conclusions	167
	References	169
	Appendix A Implementation Details and Pseudocode	171
A.1	Query Assurance Algorithms	171
A.2	Local Processing Algorithms	174
A.3	Local Processing Sample File	178
A.4	Local Processing Results	179

List of figures

3.1	Health Participatory Sensing Network Reference Architecture	49
3.2	Policy Distribution Verification Tree Structure	53
3.3	Verification Tree Accessing	53
3.4	Network and Participant Device Communication	59
4.1	Distributed Public Health Information System Architecture or HPSN	73
4.2	Verification Tree Accessing Pre Maintenance	78

4.3	Verification Tree Accessing Post Maintenance	82
4.4	Verification Tree with Metadata Node	83
4.5	Experiment Architecture	88
4.6	SQL Query Average CPU Time	91
4.7	XML Query Average CPU Time	92
4.8	SQL Data and Verification Overhead	93
4.9	SQL Verification Overhead Detail	93
4.10	XML Data and Verification Overhead	94
5.1	Public Health Information System Architecture	111
5.2	Algorithm for Data Collection Policy Processing	122
5.3	Data Collection Rule Processing Algorithm	123
5.4	Demographics Removal and Impact on k -anonymity Value	131
5.5	Local Processing Impact of k -anonymity Value	132
5.6	Comparison of Adjusted and Complete Demographic Combinations	133
5.7	Local Processing Impact of k -anonymity Value with Location Types	134
6.1	Public Health Information System Architecture (see section 5.4)	149
6.2	Public Health Intervention Distribution Verification Overhead	154
6.3	Public Health Intervention Distinct Demographic Combinations to Low k Value Combinations	155

NOMENCLATURE

Roman Symbols

ACID Atomicity, Consistency, Isolation, Durability

AES Advanced Encryption Standard

BMI Body Mass Index

ECG Electrocardiogram

EHR Electronic Health Record

EMG Electromyography

HCS Human-Centric Sensing

HPSN Health Participatory Sensing Network

HPSS Health Participatory Sensing Server

PDV Personal Data Vaults

PHR Personal Health Record

PMDS Partially Materialized Digest Scheme

QIS Quasi-identifier Score

SHA Secure Hash Algorithm

XML Extensible Markup Language

PUBLICATIONS

A list of publications that resulted from the work conducted in this dissertation

- [1] Andrew Clarke and Robert Steele. “How personal fitness data can be re-used by smart cities”. In: *Intelligent Sensors, Sensor Networks and Information Processing (ISS-NIP), 2011 Seventh International Conference on*. Dec. 2011, pp. 395–400.
- [2] Andrew Clarke and Robert Steele. “Secure and Reliable Distributed Health Records: Achieving Query Assurance across Repositories of Encrypted Health Data”. In: *System Science (HICSS), 2012 45th Hawaii International Conference on*. Jan. 2012, pp. 3021–3029.
- [3] Andrew Clarke, Eric Pardede, and Robert Steele. “External and Distributed Databases: Efficient and Secure XML Query Assurance”. In: *International Journal of Computational Intelligence Systems* 5.3 (2012), pp. 421–433.
- [4] Andrew Clarke and Robert 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*. 2012, pp. 2158–2161.
- [5] Robert Steele and Andrew Clarke. “A Real-time, Composite Healthy Building Measurement Architecture Drawing Upon Occupant Smartphone-collected Data”. In: *10th International Healthy Buildings Conference*. July 2012.
- [6] Robert Steele and Andrew Clarke. “The Internet of Things and Next-generation Public Health Information Systems”. In: *Communications and Network* (2013).

- [7] Andrew Clarke and Robert 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*. Jan. 2014, pp. 2908–2917.
- [8] Andrew Clarke and Robert Steele. “Secure query assurance approach for distributed health records”. In: *Health Systems* 3.1 (Feb. 2014), pp. 60–73. ISSN: 2047-6965.
- [9] Andrew Clarke and Robert Steele. “Local Processing to Achieve Anonymity in a Participatory Health e-Research System”. In: *Procedia - Social and Behavioral Sciences* 147 (2014). 3rd International Conference on Integrated Information (IC-ININFO), pp. 284–292. ISSN: 1877-0428.
- [10] Andrew Clarke and Robert Steele. “Health Participatory Sensing Networks”. In: *Mobile Information Systems* 10 (3 2014), pp. 229–242.
- [11] Andrew Clarke and Robert Steele. “Targeted and anonymized smartphone-based public health interventions in a participatory sensing system”. In: *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. Aug. 2014, pp. 3678–3682.
- [12] Andrew Clarke and Robert Steele. “Smartphone-based Public Health Information Systems: Anonymity, Privacy and Intervention”. In: *Journal of the Association for Information Science and Technology* (2015). ISSN: 2330-1643.

1 INTRODUCTION

The recent rapid growth in both the capabilities and uptake of mobile devices suitable to act as health sensor platforms has the potential to advance public health data collection and intervention in significant ways. However, increasingly research and development is concentrating on how mobile devices and sensors can be used as a tool for individual health data capture and feedback. Indeed, most of the major device manufacturers now ship devices with at least basic health monitoring capability as a stock application [1, 2, 3]. However, this advancement has not been extended into an investigation or implementation of how these devices can be used for public health data capture purposes.

The potential for useful public health data collection is quite extensive. Even with typical on-board or currently available external sensors, there are real possibilities of augmenting or replacing traditional public health data collection methodologies. Additionally, the collection of data from mobile devices is likely to be the only practical way to assess the detailed benefit or effectiveness of mobile device health/fitness applications to improve individual and public health.

The challenges for collection of participatory sensing public health data are varied, but the foremost would be privacy and security, especially as such a system will potentially be collecting sensitive health-related data. Interestingly, the case for public health usage does not require the same level of precise data that would often be required in participatory sensing applications [4] in other domains. 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 an individual or community as a whole. These privacy and security requirements typically are represented by two groups:

- Privacy of the individual
- Security of the device, communication and data collection server.

An additional challenge is the likelihood of such a system being a multiple data-owner system, where there are many public health groups/organizations interacting with a single health participatory sensing network (HPSN). Though this increases complexity, it is the only practical way to implement a health participatory sensing system as other approaches are likely to fracture the participants into separate platforms, creating smaller, and perhaps, not as representative participant groups.

As well as the data collection capability, the second key capability would be its usage as a platform for dissemination of public health interventions. This would typically take the form of informational/behavioral adjustment communication [5]. Potentially, such a HPSN would allow for more targeted and personalized public health interventions and the ability to track the effectiveness through the data collection capabilities to allow for analysis and improvement of intervention approaches.

The thesis as a whole aims to propose a solution to the challenges of privacy and security within a multi-owner HPSN, capable of distributing public health interventions. This will require a definition and categorization of the types of HPSNs and the overall necessary capabilities and functionality of such systems. Additionally, implementation details addressing the challenges defined above that were not suitably covered in the prior literature will be investigated, designed and evaluated.

1.1 Synopsis of the Thesis

Chapter 2 will provide an overall literature review of health-related mobile sensors, exploring the new and developing potential of HPSN and similar networks. Additionally, this section will cover incentivization of participatory sensing. In the health specific context there is limited relevant participatory sensing research, so more broad considerations of incentivization of other types of participatory sensing are reviewed and their relevance to the context discussed. Finally, the privacy and security section details approaches proposed in prior work to both provide data consistency at low levels of overhead (query assurance), and anonymity to participants that submit either observational or personal data.

Chapter 3 will provide an overview of the entire HPSN system, discussing from a high level the necessary components, interaction models, security and privacy methods that comprise this approach. This chapter serves as a broad solution chapter, bringing together all the related components of the system within an overall reference architecture. The following chapters will often refer back to this foundation work and go into further technical implementation and evaluation details.

Chapter 4 will be the first implementation/evaluation chapter focusing on the concepts, methodology, implementation and evaluation of the secure distribution approach proposed to be utilized by the HPSN. This will involve building a prototype system with example data and evaluating the efficiency of this approach. Within the overall HPSN approach this chapter provides a detailed solution description of an authenticated query assurance approach to the distribution of large amounts of disparate and multi-owner data. This supports the overall HPSN approach by providing an efficient approach to distributing data through untrusted third-parties as would be required when avoiding identification of HPSN participants. The chapter sets out to describe and evaluate a system for public health data collection rule distribution that can ensure the correctness, completeness and freshness of the data distributed

to HPSN participants without incurring overheads that would significantly increase the cost of participation.

Chapter 5 will be the second implementation/evaluation chapter focusing on the concepts, methodology, implementation and evaluation of the secure de-identification approach proposed to be utilized by the HPSN. Due to the nature of health related data, simple de-identification of records is insufficient. This is due to the risk of re-identification of records based on analysis or inference. As such, an anonymizing approach is proposed which provides additional protection against re-identification with a prototype with example data developed, implemented and evaluated in this chapter. Within the overall scope of the solution this chapter serves as a detailed solution to the requirement to keep the data collected by the system on individuals within the system at a level of detail that would make re-identification of a participant based on those details highly unlikely. Due to the nature of this type of problem a probabilistic approach is taken whereby we consider the overall levels of k -anonymity protection within the system as our method for evaluation. The scope of this problem is to ensure that the system is not able to re-identify an individual to learn sensitive details based on less sensitive details that could be obtained externally to the system. Additionally, it is required that the system does not allow for re-identification and hence learning of sensitive details based on externally sensitive details of the individual that may be already known to a privileged user of a non-public system.

Chapter 6 constitutes the final chapter of the implementation/evaluation sections. This chapter will provide additional implementation detail and evaluation of the public health intervention approach. This final chapter deals specifically with the distribution, application and reporting collection of the use of public health interventions (rather than data collection) within a health participatory sensing network. Though these topics were touched on briefly in chapters 4 and 5, they are covered in greater detail in this section with an appropriate prototype implementation and results.

Chapter 7 is the discussion and future work chapter that reviews the thesis content and provides further context of the thesis contribution.

Chapter 8 is the conclusion, providing a summary and potential direction for future work or developments.

Appendix A provides additional implementation details of the query assurance, secure de-identification and public health intervention methodologies/implementations such as pseudo-code samples used for implementation.

2 RELATED WORK

2.1 Introduction

In recent years technology and new communications models have begun to be utilized for public health. Though the possibilities involving information collection through mobile or other Internet connected devices, including both traditional survey-based approaches as well as more innovative crowd-sourcing or participatory sensing [1] approaches are just beginning to be explored. Additionally, information dissemination approaches will allow public health authorities to communicate relevant health messages to the community. However, there are some significant differences in the detail, between approaches that use technology to improve current methodologies and those that are exploring entirely new opportunities - such as targeted communication and sensor data collection.

Additionally, the incentivization for participation in public health approaches/campaigns varies with distinct techniques used with different technologies. The most common approaches include rewards (monetary or non-monetary) [2, 3], gamification [4], altruistic and personal benefit [5, 6, 7].

Lastly, the area of participatory public health often raises issues of privacy and data security for the individuals taking part. Some technologies have parallels to traditional public health data collection and communication. However, increasingly the types of data and the real-time or near real-time nature of collection require more advanced privacy protections. As such, privacy protection in participatory sensing has numerous approaches/techniques

ranging from de-identification [8] to personal data vaults [9] and distributed anonymization techniques [10].

This chapter will firstly briefly summarize participatory sensing which is the focus of this thesis. Following that, the current and ever expanding sensor capabilities for health participatory sensing networks will be discussed in detail. Next, research relevant to incentivization of participation in human-centric sensing [11] systems such as HPSNs will be discussed and presented. Lastly, the literature related to privacy and security protections for participatory sensing will be summarized.

2.2 Participatory Sensing Systems for Public Health

Increasingly, technology and especially smartphones serving as mobile sensing platforms have been creating data that is relevant to public health. Known as human-centric sensing [11] (HCS) or participatory sensing, these approaches allow for the collection of highly relevant data both through sensing of the surroundings and sensing of the individual. This trend is relevant to recent research suggesting one in four adults track their health information online [12].

A key driver of the capabilities of participatory sensing systems for public health will always be the sensing capabilities available to these networks. In the rest of this section the common types of sensors that are relevant to public health, either existing or currently in development will be summarized.

2.2.1 Sensor Capabilities

The proliferation of commercial fitness and health sensors provides new mechanisms for population health data capture, even though these are currently targeted for use in relation to an individual's health and fitness. Commercially available sensors are also already able

to capture many biomedical measures collected in public health data surveys. Such sensors include wearable patches, stretchable electronic tattoos, smartwatches, other wearables and implantable sensors along with the more widely deployed smartphones and connected sensors. In addition, such public health data capture would have a number of characteristics quite distinct from traditional survey-based public health data capture approaches. These include:

- Being real-time/ near real-time
- Larger participant numbers/ proportion of population
- More detailed data
- Captured electronically
- Direct measurement, not human response
- Anonymized collection

The area of personal health sensor and software development and commercialization [13] is currently a highly active area. This is possibly due to the relevance of these individual sensors to both the rapidly developing smartphone market and technologies, and the increasing interest to leverage such technologies for personal wellness, fitness, health and healthcare purposes [14, 15].

Fitness and Physical Activity Sensors

Commercial implementations such as Fitbit [16] and Jawbone Up [17] demonstrate the potential for and achievability of continuous physical activity sensing. Jawbone Up extends beyond physical activity monitoring to include sleep patterns and sleep quality, and a nutritional diary. Other well-known examples of such wearable sensors include RunKeeper, myFitnessPal, Pebble Watch and the Basis Watch. Such fitness and health sensors are the

most contemporarily available type of sensor that can be utilized for public health purposes, because such sensors are already achieving widespread interest and a level of mass adoption. Also of significant relevance is Google Now's Activity Summary [18] which automatically provides a monthly estimate of how far an individual has walked and cycled, and comes as part of Google's Android mobile operating system – hence is already extremely widely deployed. A more recent development has been the introduction of Google Fit [19] which extends on the automatic tracking of Google Now's Activity Summary and adds goals, hardware/application integration and daily tracking. Apple's Health and HealthKit [20] are also targeting this area for the iOS platform, as are independent manufacturers such as Samsung's S Health [21]

Vital Signs Sensors

Smartwatches such as the Mio Active are able to capture heart rate; the Amiigo wristband captures blood oxygen levels; Somaxis provides ECG and EMG sensors; and the mc10 stretchable electronic tattoo can transmit heart rate and brain activity [13]. In general the use of sweat-based sensors / temporary tattoos [22] is currently of great interest, with potential uses from managing cystic fibrosis to monitoring physical exertion, hydration and performance [23]. Another example, the Sense A/S monitoring patch is able to measure blood pressure [13]. The capturing of vital signs is often beneficial for individual health care, but it also adds new capabilities for public health data systems. The greatest shift in uptake so far has been in heart rate monitors, with the technical advancements of affordable optical heart rate monitors, meaning that chest strap monitors are no longer required. The shift has provided a large variety of wrist and arm band sensors at consumer or sport grade level as well as less traditional approaches such as heartphones [24] available commercially as iRiver Heart rate headphones [25].

Blood Constituent Sensors

Increasingly, there are wireless-enabled patch technologies emerging that may be able to capture the levels of some blood constituents. Examples include the forthcoming Sano Intelligence [26] wearable patch which is touted to allow the capture of blood glucose and potassium levels, with further blood constituent capture planned for the future. Numerous continuous blood glucose monitoring systems are also currently available, particularly of relevance to the management of diabetes. Such sensor capabilities in a cheap and accurate form have the potential to revolutionize individual health care, early detection and preventative health; and by extension, also public health. Because such capabilities may be so beneficial in terms of individual health monitoring, health maintenance and early detection, they could achieve wide adoption. If so, their possible role in public health data capture can also be proportionately significant.

Ambient sensors

Other initiatives such as Riderlog [27] and the Copenhagen Wheel [28] are moving towards capturing physical activity levels and at the same time, additional contextual and environmental data. The Copenhagen Wheel goes beyond physical activity sensing, to urban environment monitoring with air quality and noise sensors included in the implementation to provide additional data beyond just the activity of the individual.

Sleep monitors

Sleep monitors/sensors are becoming commonly available and use a variety of techniques to track sleep length and/or quality. The primary techniques are actigraphy-based systems such as Actiwatch [29], Fitbit [16] and Jawbone [17], among many others. Other techniques utilize radio-frequency biomotion sensors such as SleepMinder [30] and Sleep Design [31]. The accuracy of such sleep sensors and their applicability outside the average population is

often raised in research [32] as well as the comparative strengths of actigraphy compared to radio-frequency approaches [31]. However, overall there has been a great increase in the accessibility and quality of sleep sensors.

Summary

Recent years have produced an explosion of consumer grade sensors and sensor platforms allowing individuals to quantify ever larger portions of their lives. The capabilities and accuracy of these sensors are ever evolving and represent a myriad of possibilities for public health usage. Where currently, very little of this already extant information is available for public health usage, it is foreseeable that this could change especially with the development/implementation of built for purpose HPSNs.

2.2.2 Motivation and Incentivization

Participatory sensing due to its decentralized collection methodology and opt-in participation often needs to consider the motivation and incentivization of participants. The reality of these platforms is their reliance on good quality and representative data collection, without which the value of the analysis and conclusions of any such data collection would be affected. As such, motivation and incentivization techniques are often a key component of participatory sensing design and research [33, 2, 34]. Additionally, any incentivization approach would need to take into account potential negative impacts [35] from participation; and provide adequate compensation. There are two main goals of incentivization and motivation for all participatory sensing networks and an additional goal for health participatory sensing, these are:

1. Increase the participation rate in a sensing network [33]. This meets the need of involving a sufficient number of participants to have a meaningful coverage of the

location/area/demographics the sensing network is collecting data upon. This could involve rewarding more valuable/rare contributions at different rates.

2. Increase the quality of the data [34], whereby participants who provide better quality data or do not provide poor quality data are incentivized.
3. Specific to the health domain, incentivization approaches are already common in attempts to improve an individual's health. So incentivization of health participatory sensing networks will need to add this to the existing goals listed above.

In this section, the more common incentivization approaches used in participatory sensing with examples of research/commercial applications will be covered in greater detail.

Altruistic

An altruistic system is one that encourages user participation on the basis that participation is good for society. There is no immediate reward other than the perceived (or perhaps personally realized) downstream benefits. This type of approach has a long history in other research methodologies, and is sometimes considered to be equivalent to the practice of 'citizen science'. In the realm of participatory sensing for public health, this is most relevant to creating a small sample which may or may not be fully representative of the wider community. Altruistic approaches are often made more attractive through gamification of the data collection process, such as including statistics or achievements [4], for the individual to track their personal progress and contribution.

Social Translucence

A system whereby participant's contributions and/or their associated value to the participatory sensing system are visible to one another [36]. In practice this could be implemented through a reputation-based approach or through a competitive gamification [4] approach

amongst other techniques. An example of competitive gamification is the inclusion of leaderboards or rankings/ranked participants, as in NoiseMap [4]. Another approach utilized in [37], is the use of participant-to-participant feedback-based rankings, where the individual users of the system endorse each other based on the quality or usefulness of their contribution.

Personal Benefit

This is an incentivization approach whereby the primary motivation for participation in the participatory sensing system is the direct (non-financial or reputational) benefits that a participant can gain. In a health participatory sensing context this would involve things such as improved health and more relevant health communication. In essence this refers to the ability to use the individual's information in concert with the overall collected data for a population to provide a better service to the health consumer than could be achieved through stand-alone collection. Overall, personal benefit has been an extremely successful incentivization approach in recent years, with high quality software/applications regularly being provided for free to the user, to allow data collection or advertising by the software creator. Some examples specifically in the participatory sensing field include Waze [5], a community-based traffic and navigation application, that includes sensor and contributor data (maps, petrol prices, accidents, road hazards and traffic jams) to improve the quality and usefulness of the application to the users. Another example of personal benefit is WeatherLah [38], a crowd sourced weather application which uses participatory sensing data from individuals to provide a more detailed rain map. Other possible examples include wait time applications and public transport applications [6, 7].

The health domain brings a different component to the potential benefit, whereby, participation has benefits to the individual's health, such as Astmapolis [39], a participatory data collection system for asthma inhalers that uses the aggregate data from multiple users

to help an individual manage their asthma. As such, this would be the primary benefit type approach of a health participatory sensing application. However, some additional informational/public contribution benefits have been considered in previous applications such as Riderlog [27], which as well as collecting cycling information, uses the collected information to campaign for better infrastructure. Another alternative benefit for health participatory sensing applications is the commonly found route sharing or heatmap [40] capabilities that can be provided, distributing information to participants of the system as to more convenient/better routes/areas to exercise/commute based on popularity.

Market Approaches to Incentivization

This refers to a compensation-based system that pays participants for their contribution. Such a system may pay users on a pay-per-measurement or pay-per-time or task basis [3]. There has been substantial research around this area such as [34] which utilizes a market equilibrium monetary incentivization approach to reward quality of contributed service. The length of potential participation is also of concern, with some prior work focusing on long term participation [41]. An alternate approach that focuses on minimizing the total cost of compensating participants through a reverse auction design, has been presented in previous work [42]. The advantages of this approach are that it motivates users to participate through incentives, but also provides motivation for truthful cost reporting and hence a more efficient incentivization approach, than if a monetary (or value reward) approach is to be utilized. The previous approaches are characteristic of the two different market approaches:

1. Incentivize based on the value of the contribution to the participatory sensing system
2. Incentivize based on the cost of participation to the individual.

In practice, it is likely that both approaches will have different use cases and relevance in various health participatory sensing systems.

Most of the previous work on market approaches to incentivization has been limited to providing a mathematical proof or example data set prototype of its effectiveness. However, some smaller/pilot studies including participants have been completed [42]. This study used an incentive approach based on the number and quality of contribution (based on the market approach proposed in [34]) alongside an endorsement social-cloud, whereby participants provide peer-to-peer feedback. The initial results were positive in relation to the effectiveness of an incentive model. A larger trial involved the use of pedometers and a reward system, Steptacular [2]. Steptacular was quite a large trial involving over 5000 employees at Accenture-USA. The project utilized a monetary incentive and a combination of gamification and social incentives. Their comparative analysis to the 10K Challenge previously deployed at Accenture-USA indicated that a combination of incentive approaches is more effective than monetary alone. However, a combination of monetary and gamification, whereby the monetary amount was randomized/gamified was found to be problematic due to creating confusion amongst users as to how much money they could actually win.

Mandated

A mandated system is one that requires users to participate. This includes systems in use by participants as part of their professional duties (e.g. civil service employees and first responders), private workforce unrelated to their professional duties, or indeed directly from a government body. Though these approaches would result in a simpler system than one which needs to be designed to encourage participation, it is foreseeable that this type of initiative could be potentially controversial. Typically, the literature has not gone into detail as to a mandated approach, and search of the literature did not find a relevant study/prototype of this type of approach. However, it is conceivable that it will have similar issues to other mandatory public data collections, such as censuses, that require participation. One key concern would be of course data quality, and any future consideration of mandated participation

will have to consider how to detect and handle purposefully inaccurate data contributed by unwilling participants.

2.3 Privacy and Security in Public Health Participatory Sensing Systems

Public health participatory sensing systems through their involvement of individuals in the collection and submission of data inherently pose significant privacy concerns. Firstly, there are concerns over privacy of the individual within the system. Where the individual is de-identified in a system it will not be primary concern. However in an identified system, the questions of who has access and how securely stored the individual's details are will be important. Secondly, there is the security and privacy of the collected data. This deals with access to the data, the sensitivity and how likely it is that the data could be used to re-identify an individual (if they were de-identified within the system). Lastly, is the concern over communication privacy between the participant and the system. If the user is unidentified to the system it must be possible to submit data in a non-identifiable way, as well as keep the communication secure from a third-party that may attempt to intercept. While in an approach where the user is to be identifiable by the public health information system, the former requirement for non-identifiable communication can be excluded.

The rest of this section will introduce and discuss the various privacy and security methods for participatory sensing networks presented in the prior literature. Approaches specific to health participatory sensing networks are less common, so where appropriate, the relevance for health data will be considered. The approaches are typically broken into two subgroups

1. Approaches for trusted servers
2. Approaches for untrusted servers.

2.3.1 Trusted Servers

The utilization of a trusted server approach reduces the complexity of security and privacy considerations for a health participatory sensing system by assuming a level of security at the server level. Therefore, provision of communication security, server security and encryption are needed to provide protection from the threat of malicious outside attackers or other security threats (which are similar to any other server holding secure data). However, there remains the issue of privacy of data once it is being used or distributed.

A conventional approach would be to anonymize any used data to a k -anonymity [8] or another anonymizing variant [43, 44] level, so as to anonymize the data before it is accessible for research/analysis. k -anonymity provides an approach whereby, a data set is considered to be anonymized if an individual record is indistinguishable from k other records.

Another approach is differential privacy [45, 46] which adds noise to a data set to create individual uncertainty while maintaining the integrity of the statistical dataset (that is the meaningfulness of the statistical data is not lost). This is a well developed approach to anonymizing datasets for use. Essentially, differential privacy tries to ensure that the removal or addition of any particular record in a dataset does not change the outcome of any analysis by much. As such, the presence of an individual in the dataset cannot be discerned and hence exploited through multiple data requests. However, there are some limitations, such as either needing to know the use of data in advance, limitation in data types or certain types of sums/counts. Additionally, prior work has discussed that without significant advancements it is still of limited usefulness for health data [47].

Alternatively, other approaches have taken an additional step by removing some sensitive information before submission (removal of identifiers and communications anonymity) with a central point of trust [48] to provide an anonymous approach. This approach reduces the amount of sensitive information held by the trusted server, without reducing the data

quality or precision. This approach is most effective when the participatory sensing system 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, but overall it provides a good compromise.

2.3.2 Untrusted Servers

To reduce some privacy and security concerns, it is often proposed that eschewing a trusted server approach may be appropriate, and having the participant device perform much of the anonymizing and security protection before submission of data. As such, individual devices pose much smaller targets for malicious attack, or improper use of collected data.

There are a number of participatory sensing privacy approaches [49] and the area continues to be active [50, 51]. Most approaches focus on the major challenge of participatory sensing data collection i.e. providing privacy with fine location/time data submissions. Both these data types are not core requirement for public health data collection so these advancements are of limited utility for HPSNs.

Some alternative approaches include decentralized participatory sensing networks [52] using user interaction/awareness as part of the approach and a "web of trust" where the interactions of individual participants define their extended trust network and hence breadth of dataset and sensitivity of data.

Alternatively, it is proposed that keeping the data managed by the participant [9, 53] and stringent user-definable access control mechanisms to manage sharing could be used to allow the use of human-centric sensing data.

A more recent approach [10] using distributed differential privacy appears promising and the area of distributed privacy continues to be an active research area [54, 55]. However, the same limitations that affect the trusted server differential privacy approaches exist for the

distributed approach, including the current lack of suitability for health data [47].

2.3.3 Query Assurance

In addition to the issues of individual privacy and security is that of certainty of the accuracy of the information distributed from the server to an individual. That is, that the information distributed meets the requirements of correctness, completeness and freshness. Techniques that can provide this certainty over a database/datasource are known as query assurance approaches, and can be typically described according to two categories - probabilistic and authenticated. The probabilistic approach uses known, fake or duplicate data to provide strong evidence that the data is queried correctly. The authenticated approach usually employs a more traditional technique to provide assurance, using a combination of hashing/digital signing and timestamps to ensure correct query replies. Most of the previous literature has focused on relational databases. Whether the data is encrypted or not varies between schemes. In some cases, encryption is required to provide query assurance.

Authenticated Verification

Authenticated verification, where verification is a methodology to provide proof of integrity or authenticity of data, typically uses various verification object models that have been proposed to provide query assurance - for example some form of hashing/signing and timestamps in combination with database sorting/indexing.

As part of a query response, a verification object is generated and returned. This object contains hashes that verify the data is correct, and a digital signature (signed by the data owner) of the hashes that ensures the authenticity of the verification object. To address freshness a timestamp is also signed [56].

Nested B+ Merkle Tree [57] is a form of embedded merkle hash tree that is specifically tailored to XML data. The root tree is a path tree which preserves the path information,

though path order is not preserved and equivalent paths are collapsed into an element in the tree. Leaf nodes in the path tree contain a value tree and a parent tree. A value tree contains the index to search element by value. The parent tree stores data relating to parent elements.

Through this conceptually clear method of deciding how to nest trees, the overall verification object size should be reduced. However, as with other solutions, if a path has an extremely high number of nodes, large verification objects will be created to verify data.

One approach to remedy the issue of large verification object size is Embedded Merkle Hash Trees [56], which provides an improvement to create more control of fan-out and depth of merkle hash trees. It allows a more efficient tree structure through a higher control of the size of the verification objects that will be needed.

A further approach suggested to improve the efficiency of authenticated query assurance known as Partially Materialized Digest Scheme (PMDS) [58] like other authenticated query assurance approaches, it is based on merkle hash trees. However, PMDS takes a novel approach of considering that as creating a hash value is faster than performing file reads from a hard drive, it is preferable to not instantiate the entire hash tree at any one time. The digital signature on the root of the tree is created based on the temporary hash values, and some of the lower branches of the tree are kept intact, but the rest of the hash tree is then discarded rather than stored. When a query is made, the hash tree values are recreated to verify the result, packaged with the digital signature and returned to the client. The limitation of this approach is that to efficiently recreate the hash tree, the stored database and the PMDS must be on the same server.

A more distinct form of authenticated verification is proposed in [59] where the data is encrypted. Completeness and correctness are preserved by adding the node ID to the data before encryption, and no queries apart from block retrieval are supported. Freshness is addressed by adding a timestamp to the root node and disseminating the timestamp to users directly.

Though the final approach provides data confidentiality in addition to query assurance, it is only capable of responding to very limited queries. This would likely affect the applicability of this approach in health information system architectures.

Probabilistic Verification

Probabilistic query assurance is based on a fairly simple premise [60, 61]: if the data-owners and clients have knowledge of a reasonable portion of the remote database, they can test whether queries are being executed correctly, by batching a set of queries that execute over both the known and unknown portions of the database. The query assurance provided is not absolute, but it is considered that a reasonably high chance of detection of incorrect results can be sufficient.

However, probabilistic systems do have distinct advantages, namely that in most cases, no modification of the database server needs to be made, and that increases in overhead are both easy to predict and quite reasonable, as the overhead is just a linear relation to database size. A further advantage is that probabilistic query assurance can provide variable levels of certainty based on how much extra data is allowed as overhead. More overhead allows for a higher certainty of the returned queries being correct.

A non-trivial issue of the implementation of probabilistic query assurance relates to the creation of non-pattern based data over the long-term that will not be detected. Dual Encryption [62], avoids this problem by encrypting the entire database in blocks with a symmetrical key, then encrypting redundant data with a second key. Therefore, as the service provider has no knowledge of the contents of any of the database, and as the key used for each portion cannot be determined by the server, the redundant data cannot be detected. Thus the client can, by querying from parts encrypted with both keys, provide a high level of query assurance.

However, though a probabilistic approach seems well suited to query assurance over

encrypted data, there are drawbacks that make it less optimal for health data. Firstly, the assurance for the majority of records is based on an assumption that if the checked records are correct, all returned records will be correct. This allows incorrect data to be returned but not detected. While this could be acceptable for some applications, it is likely that health information systems will require a higher level of assurance. Secondly, the approach has to go to significant effort to mitigate query correspondence and distribution attacks. The mitigations include for example query batching and delaying assurance checks. In the likely health information architecture where distributed and numerous data repositories exist, this could create a higher level of complexity that needs to be further investigated. Lastly, due to the nature of this scenario, whereby there are multiple users of the system and they need to have knowledge of the known portions of the database, keeping that knowledge from the server providers and data sources is problematic, if they are also the owners of portions of the health data.

Summary

Though there have been quite varied and distinct approaches in previous work, the more relevant approach within the electronic public health domain is that of encrypted storage with authenticated query assurance. This motivation was explored in our previous work [63] that indicates the need to strengthen confidentiality beyond the use of purely legislative mechanisms, in relation to insider activities. However, there is no previous solution that provides searchable encryption and robust authenticated query assurance. This chapter introduces a model that incorporates both query assurance through an authenticated method and searchable encryption. Further, we extend the freshness coverage to be more effective in cases of data repositories with multiple data-owners, and elaborate its applicability to distributed health information systems, including those where some data components may be stored with consumers.

2.4 Discussion

As covered in the earlier sections, the capabilities and potential for health participatory sensing networks are quite unique and different from previous public health data collection methodologies - even those where Internet technologies have also been utilized. However, the nature of the data collection method creates significant privacy and security concerns, beyond those which would be encountered in a regular participatory sensing network.

Review of the relevant literature for privacy and security approaches for participatory sensing networks did not uncover an approach suitable for the type of data collected and the interaction model for HPSNs described in this thesis. Typically, the limitation is that most participatory sensing privacy approaches are focused on location privacy which is of low usefulness for HPSNs which do not require fine-grained location access. Additionally, most approaches to participatory sensing privacy don't consider that the actual non-location sensor data may be the most sensitive part of the collection and that de-identification of the data and protection from re-identification is a key concern. This limitation continues when considering data anonymization approaches - most of which require a trusted server to act as a data aggregation point, which would be undesirable in a large scale health participatory sensing system due to privacy risks and a single point of security failure.

This privacy concern appears to be a significant research gap - which will be addressed within the implementation chapters that follow. Additionally, while the use of participatory sensing or human-centric sensing for health or public health has often been proposed, details of how this could occur and the capabilities, functionality and architecture to support such an approach has not been sufficiently covered in prior research and will be introduced and detailed in the following chapters.

2.5 Conclusions

This chapter summarized the capabilities of health participatory sensing networks in the context of existing public health data collection research, which were detailed as being advanced and transformative over traditional data collection methods. Additionally, the current state of the literature for the key challenges to health participatory sensing networks: incentivization and privacy and security were summarized and gaps in the current research were identified.

REFERENCES

- [1] Jeff Burke et al. “Participatory sensing”. In: *In: Workshop on World-Sensor-Web (WSW’06): Mobile Device Centric Sensor Networks and Applications*. 2006, pp. 117–134.
- [2] Naini Gomes et al. “Steptacular: An incentive mechanism for promoting wellness”. In: *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*. Jan. 2012, pp. 1–6.
- [3] Buster O. Holzbauer, Boleslaw K. Szymanski, and Eyuphan Bulut. “Incentivizing Participatory Sensing via Auction Mechanisms”. In: *Opportunistic Mobile Social Networks*. Ed. by Jie Wu and Yunsheng Wang. CRC Press, 2014, pp. 339–375. ISBN: 978-1-4665-9494-4.
- [4] Immanuel Schweizer et al. “Noisemap: Multi-tier Incentive Mechanisms for Participative Urban Sensing”. In: *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones*. PhoneSense ’12. Toronto, Ontario, Canada: ACM, 2012, 9:1–9:5. ISBN: 978-1-4503-1778-8.
- [5] Waze: *Free GPS navigation with turn by turn*. 2014. URL: <https://www.waze.com/>.
- [6] Buuuk. *Mana Rapid Transit*. May 2012. URL: <https://itunes.apple.com/sg/app/mana-rapid-transit/id501961346?mt=8>.

- [7] J.K.-S. Lau, Chen-Khong Tham, and Tie Luo. “Participatory Cyber Physical System in Public Transport Application”. In: *Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on*. Dec. 2011, pp. 355–360.
- [8] Panos Kalnis and Gabriel Ghinita. “Spatial k-Anonymity”. In: *Encyclopedia of Database Systems*. Ed. by Ling Liu and M. Tamer Özsu. Springer US, 2009, p. 2714. ISBN: 978-0-387-35544-3, 978-0-387-39940-9.
- [9] Min Mun et al. “Personal data vaults: a locus of control for personal data streams”. In: *Proceedings of the 6th International Conference. Co-NEXT '10*. Philadelphia, Pennsylvania: ACM, 2010, 17:1–17:12. ISBN: 978-1-4503-0448-1.
- [10] Vibhor Rastogi and Suman Nath. “Differentially Private Aggregation of Distributed Time-series with Transformation and Encryption”. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. SIGMOD '10. Indianapolis, Indiana, USA: ACM, 2010, pp. 735–746. ISBN: 978-1-4503-0032-2.
- [11] Mani Srivastava, Tarek Abdelzaher, and Boleslaw Szymanski. “Human-centric sensing”. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 370.1958 (2011), pp. 176–197. ISSN: 1364-503X.
- [12] Susannah Fox and Sydney Jones. “The social life of health information”. In: *Washington, DC: Pew Internet & American Life Project* (2009), pp. 2009–12.
- [13] Melanie Swan. “Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0”. In: *Journal of Sensor and Actuator Networks* 1.3 (2012), pp. 217–253. ISSN: 2224-2708.
- [14] Robert Steele. “Social media, mobile devices and sensors: Categorizing new techniques for health communication”. In: *Sensing Technology (ICST), 2011 Fifth International Conference on*. Nov. 2011, pp. 187–192.

- [15] Robert Steele et al. “Elderly persons’ perception and acceptance of using wireless sensor networks to assist healthcare”. In: *International Journal of Medical Informatics* 78.12 (2009). Mining of Clinical and Biomedical Text and Data Special Issue, pp. 788–801. ISSN: 1386-5056.
- [16] Fitbit Inc. *Why Fitbit*. URL: <https://www.fitbit.com/au/whyfitbit>.
- [17] Jawbone. *UP by Jawbone with MotionX Technology Empowers You to Live a Healthier Life*. 2011. URL: <http://content.jawbone.com/static/www/pdf/press-releases/up-press-release-110311.pdf>.
- [18] MobiHealthNews. *Google adds activity tracking to Android app*. 2012. URL: <http://mobihealthnews.com/19551/google-adds-activity-tracking-to-android-app/>.
- [19] Dan Graziano. *The complete guide to Google Fit*. 2014. URL: <http://www.cnet.com/how-to/the-complete-guide-to-google-fit/>.
- [20] Apple Inc. *Health*. URL: <https://www.apple.com/ios/whats-new/health/>.
- [21] Samsung Electronics Co. Ltd. *S Health*. URL: <http://content.samsung.com/us/contents/aboutn/sHealthIntro.do>.
- [22] Jason Heikenfeld. *Sweat Sensors Will Change How Wearables Track Your Health*. 2014. URL: <http://spectrum.ieee.org/biomedical/diagnostics/sweat-sensors-will-change-how-wearables-track-your-health/>.
- [23] Meghana Keshavan. *Electrozyme’s metabolism-monitoring temp tattoos backed by Mark Cuban*. 2014. URL: <http://medcitynews.com/2014/09/electrozymes-metabolism-monitoring-temp-tattoos-backed-mark-cuban/>.
- [24] Ming-Zher Poh et al. “Heartphones: Sensor Earphones and Mobile Application for Non-obtrusive Health Monitoring”. In: *Wearable Computers, 2009. ISWC '09. International Symposium on*. Sept. 2009, pp. 153–154.

- [25] Ben Coxworth. *Bluetooth fitness headset plays tunes and tracks your stats*. 2013. URL: <http://www.gizmag.com/iriver-on-audio-biometrics-headset/29440/>.
- [26] Sano Intelligence. 2012. URL: <http://rockhealth.com/accelerator/portfolio-companies/sano-intelligence/>.
- [27] Bicycle Network. *Riderlog*. 2012. URL: <http://www.bv.com.au/general/ride-to-work/91481/>.
- [28] Christine Outram, Carlo Ratti, and Assaf Biderman. “The Copenhagen Wheel: An innovative electric bicycle system that harnesses the power of real-time information and crowd sourcing”. In: *EVER Monaco International Exhibition & Conference on Ecologic Vehicles & Renewable Energies*. 2010.
- [29] Ronald J Gironda et al. “Preliminary evaluation of reliability and criterion validity of Actiwatch-Score”. In: *Journal of rehabilitation research and development* 44.2 (2007), p. 223.
- [30] Alberto Zaffaroni et al. “SleepMinder: an innovative contact-free device for the estimation of the apnoea-hypopnoea index”. In: *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE. 2009, pp. 7091–9094.
- [31] Emer O’Hare et al. “A comparison of radio-frequency biomotion sensors and actigraphy versus polysomnography for the assessment of sleep in normal subjects”. In: *Sleep and Breathing* 19.1 (2015), pp. 91–98. ISSN: 1520-9512.
- [32] Hawley E. Montgomery-Downs, Salvatore P. Insana, and Jonathan A. Bond. “Movement toward a novel activity monitoring device”. In: *Sleep and Breathing* 16.3 (2012), pp. 913–917. ISSN: 1520-9512.

- [33] Siavash Aflaki et al. “Evaluation of incentives for body area network-based health-care systems”. In: *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. USA: IEEE, Apr. 2013, pp. 515–520.
- [34] Chen-Khong Tham and Tie Luo. “Quality of Contributed Service and Market Equilibrium for Participatory Sensing”. In: *Distributed Computing in Sensor Systems (DCOSS), 2013 IEEE International Conference on*. May 2013, pp. 133–140.
- [35] M.R. Wigan and R. Clarke. “Big Data’s Big Unintended Consequences”. In: *Computer* 46.6 (June 2013), pp. 46–53. ISSN: 0018-9162. DOI: 10.1109/MC.2013.195.
- [36] Ioannis Krontiris and Nicolas Maisonneuve. “Participatory Sensing: The Tension Between Social Translucence and Privacy”. In: *Trustworthy Internet*. Ed. by Luca Salgarelli, Giuseppe Bianchi, and Nicola Blefari-Melazzi. Springer Milan, 2011, pp. 159–170. ISBN: 978-88-470-1818-1.
- [37] Fang-Jing Wu and Tie Luo. “A generic participatory sensing framework for multi-modal datasets”. In: *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2014 IEEE Ninth International Conference on*. Apr. 2014, pp. 1–6.
- [38] Buuuk. *WeatherLah: Singapore Weather App with PSI Trend Widget*. Nov. 2014. URL: <https://itunes.apple.com/us/app/weatherlah/id411646329?mt=8>.
- [39] Robert Wood Johnson Foundation. *Asthmapolis*. 2010. URL: <http://www.rwjf.org/en/about-rwjf/newsroom/newsroom-content/2010/06/asthmapolis.html>.
- [40] Strava. *Strava Labs Global Heatmap*. Nov. 2014. URL: <http://labs.strava.com/heatmap/>.
- [41] Lin Gao, Fen Hou, and Jianwei Huang. *Providing Long-Term Participation Incentive in Participatory Sensing*. Tech. rep. Cornell University, 2015.
- [42] Iordanis Koutsopoulos. “Optimal incentive-driven design of participatory sensing systems”. In: *INFOCOM, 2013 Proceedings IEEE*. Apr. 2013, pp. 1402–1410.

- [43] ChristineM. O’Keefe and Natalie Shlomo. “Applicability of Confidentiality Methods to Personal and Business Data”. English. In: *Privacy in Statistical Databases*. Ed. by Josep Domingo-Ferrer. Vol. 8744. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 350–363. ISBN: 978-3-319-11256-5. DOI: 10.1007/978-3-319-11257-2_27. URL: http://dx.doi.org/10.1007/978-3-319-11257-2_27.
- [44] Benjamin C. M. Fung et al. “Privacy-preserving Data Publishing: A Survey of Recent Developments”. In: *ACM Comput. Surv.* 42.4 (June 2010), 14:1–14:53. ISSN: 0360-0300.
- [45] Cynthia Dwork. “Differential Privacy: A Survey of Results”. In: *Theory and Applications of Models of Computation*. Ed. by Manindra Agrawal et al. Vol. 4978. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, pp. 1–19. ISBN: 978-3-540-79227-7.
- [46] Cynthia Dwork and Adam Smith. “Differential privacy for statistics: What we know and what we want to learn”. In: *In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, volume 4052 of LECTURE NOTES IN COMPUTER SCIENCE*. Springer, p. 1.
- [47] Fida Kamal Dankar and Khaled El Emam. “The Application of Differential Privacy to Health Data”. In: *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. EDBT-ICDT ’12. Berlin, Germany: ACM, 2012, pp. 158–166. ISBN: 978-1-4503-1143-4.
- [48] Cory Cornelius et al. “Anonymsense: privacy-aware people-centric sensing”. In: *Proceedings of the 6th international conference on Mobile systems, applications, and services*. MobiSys ’08. Breckenridge, CO, USA: ACM, 2008, pp. 211–224. ISBN: 978-1-60558-139-2.
- [49] Mohannad A. Alswailim, Mohammad Zulkernine, and Hossam S. Hassanein. “Classification of participatory sensing privacy schemes”. In: *Local Computer Networks*

- Workshops (LCN Workshops), 2014 IEEE 39th Conference on.* Sept. 2014, pp. 761–767.
- [50] Emiliano De Cristofaro and Claudio Soriente. “Participatory privacy: Enabling privacy in participatory sensing”. In: *Network, IEEE* 27.1 (Jan. 2013), pp. 32–36. ISSN: 0890-8044.
- [51] Delphine Christin et al. “A survey on privacy in mobile participatory sensing applications”. In: *Journal of Systems and Software* 84.11 (2011). Mobile Applications: Status and Trends, pp. 1928–1946. ISSN: 0164-1212.
- [52] Delphine 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.* 2010, pp. 847–848.
- [53] Haksoo Choi et al. “SensorSafe: A Framework for Privacy-Preserving Management of Personal Sensory Information”. In: *Secure Data Management*. Ed. by Willem Jonker and Milan Petkovic. Vol. 6933. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 85–100. ISBN: 978-3-642-23555-9.
- [54] Slawomir Goryczka, Li Xiong, and Vaidy Sunderam. “Secure Multiparty Aggregation with Differential Privacy: A Comparative Study”. In: *Proceedings of the Joint EDBT/ICDT 2013 Workshops*. EDBT ’13. Genoa, Italy: ACM, 2013, pp. 155–163. ISBN: 978-1-4503-1599-9.
- [55] T-H. Hubert Chan, Elaine Shi, and Dawn Song. “Optimal Lower Bound for Differentially Private Multi-party Aggregation”. In: *Proceedings of the 20th Annual European Conference on Algorithms*. ESA’12. Ljubljana, Slovenia: Springer-Verlag, 2012, pp. 277–288. ISBN: 978-3-642-33089-6.

- [56] Feifei Li et al. “Dynamic Authenticated Index Structures for Outsourced Databases”. In: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. SIGMOD '06. Chicago, IL, USA: ACM, 2006, pp. 121–132. ISBN: 1-59593-434-0.
- [57] Viet Hung Nguyen and Tran Khanh Dang. “A Novel Solution to Query Assurance Verification for Dynamic Outsourced XML Databases”. In: *Journal of Software* 3.4 (2008).
- [58] Kyriakos Mouratidis, Dimitris Sacharidis, and Hweehwa Pang. “Partially Materialized Digest Scheme: An Efficient Verification Method for Outsourced Databases”. In: *The VLDB Journal* 18.1 (Jan. 2009), pp. 363–381. ISSN: 1066-8888.
- [59] Tran Khanh Dang. “Ensuring Correctness, Completeness, and Freshness for Outsourced Tree-Indexed Data”. In: *Inf. Resour. Manage. J.* 21.1 (Jan. 2008), pp. 59–76. ISSN: 1040-1628.
- [60] Radu Sion. “Query Execution Assurance for Outsourced Databases”. In: *Proceedings of the 31st International Conference on Very Large Data Bases*. VLDB '05. Trondheim, Norway: VLDB Endowment, 2005, pp. 601–612. ISBN: 1-59593-154-6.
- [61] Min Xie et al. “Providing Freshness Guarantees for Outsourced Databases”. In: *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*. EDBT '08. Nantes, France: ACM, 2008, pp. 323–332. ISBN: 978-1-59593-926-5.
- [62] Haixun Wang et al. “Dual Encryption for Query Integrity Assurance”. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA: ACM, 2008, pp. 863–872. ISBN: 978-1-59593-991-3.

-
- [63] Michael Czapski and Robert Steele. “Strengthening Privacy and Confidentiality Protection for Electronic Health Records.” In: *Web Technologies, Applications, and Services*. Ed. by M. H. Hamza. IASTED/ACTA Press, Oct. 27, 2005, pp. 35–40. ISBN: 0-88986-485-3.

3 HEALTH PARTICIPATORY SENSING NETWORKS

Preamble

This chapter is based on a journal paper that was published in the Journal of Mobile Information Systems [1], it has been included as a chapter of this thesis with only minor formatting changes to align with the thesis format. This chapter forms a foundation for the following chapters, introducing the classification of health participatory sensing models, and identifying distinct levels of interaction between individuals and a health participatory sensing network. Additionally, a reference architecture describing the key capabilities and requirements of the system to meet its purpose, provide privacy and security to individual participants and finally, an example user scenario. These foundations are built on in the latter chapters which provide technical details of the implementations and capabilities.

ABSTRACT

The use of participatory sensing in relation to the capture of health-related data, is rapidly becoming a possibility, due to the widespread consumer adoption of emerging mobile computing technologies and sensing platforms. This has the potential to revolutionize data collection for population health, aspects of epidemiology, and health-related e-Science applications and as we will describe, provide new public health intervention capabilities, with the classifications and capabilities of such participatory sensing platforms only just beginning to be explored. Such a development will have important benefits for access to near real-time, large-scale, up to population-scale data collection. However, there are also numerous issues to be addressed first: provision of stringent anonymity and privacy within these methodologies, user interface issues, and the related issue of how to incentivize participants and address barriers/concerns over participation. To provide a step towards describing these aspects, in this chapter we present a first classification of health participatory sensing models, a novel contribution to the literature, and provide a conceptual reference architecture for health participatory sensing networks (HPSNs) and user scenario example.

Keywords: Participatory Sensing - Public Health - Epidemiology - Mobile Health

3.1 Introduction

The use of health participatory sensing as a data collection methodology is rapidly becoming a reality that will revolutionize the scale and types of data that can be aggregated for a

number of population health, epidemiological, statistical and data analysis purposes. Participatory sensing is the act of using mobile devices to allow public and professional users to collect, analyze and submit or share local knowledge to a larger interactive participatory sensing network [2]. The range of possibilities for participatory sensing is large [3], however previous work in participatory sensing has not considered in detail the different participatory models that are likely to occur in the health context. As these models have the potential to scale to millions or nation-wide levels, both the potential and complexities of these data systems are also substantial. This raises the need for a categorization of health participatory sensing models and a description and analysis of the basic architectures possible. Interestingly, we also suggest in this work that beyond sensing alone, the possible models of health participatory sensing may also often include elements of public health intervention or two-way interaction.

The growth in the potential for participatory sensing has been largely accelerated through the high levels of smartphone adoption in many countries [4], leading to the proliferation of powerful sensing platforms that are highly human-centric, making them ideal as the center-points for health participatory sensing models [5, 6, 7]. The potential capabilities are further extended with the addition of ubiquitous external sensing components such as activity monitoring [7] and other wearable consumer health sensors.

Contemporary commercial implementations such as Nike Fuel and Jawbone Up [7] demonstrate the achievability and potential for continuous physical activity sensing. Jawbone Up extends beyond physical activity monitoring to include sleep pattern and quality, and a nutritional diary. Other initiatives such as Cykelscore [5], Riderlog [6] and the Copenhagen Wheel [8] are moving towards participatory sensing for specific usage groups. Cykelscore acts as an active transportation data collection tool and incentive framework for cycling, while Riderlog acts as just a collection tool. Alternatively, the Copenhagen wheel goes beyond physical activity sensing, to urban environmental monitoring with air quality

and noise sensors included in the implementation to provide additional data beyond just the activity of the individual.

In the research domain, attempts to make data collection more automatic or less invasive are being explored [9] and there also efforts to automate nutrition and dietary intake information capture [10], with proposals including in relation to acoustical dietary intake [11], specialized hardware [12, 13], and food image analysis [14]. Additionally, there are attempts to make sensors more unobtrusively wearable, such as with projects like Heartphones [15] – a coupling of a heart rate monitor with headphones and a mobile device.

Overall, this has led to a great improvement in potential capability, but with most research focused on the use of such data collection by the individual, there has been less attention given to the potential for and challenges to usage to provide population wellness measures or for wider population health or population epidemiological usage. With the growth in sensor capabilities now and in the near future, an analysis of health participatory sensing capabilities, models and architectures is timely. In section two we provide a classification of health participatory sensing models, in section three we describe a conceptual reference architecture for HPSNs, in section four we provide an example user scenario and section five is the Conclusion.

3.2 Classification of Health Participatory Sensing Models

The classification of health participatory sensing models has not previously been formalized, and the types of interaction not systematically identified or described. In this section we classify health participatory sensing models according to five distinct categories and briefly discuss the potential privacy, security, interface and incentivization aspects.

3.2.1 Incidental Participatory Sensing

Incidental participatory sensing requires the lowest levels of participation by an individual. It is defined as contributing sensor information for population health measures, which the individual would have already collected for their own use or benefit [16]. An example of incidental participatory sensing is physical activity self-tracking that has become increasingly popular in recent years. Due to the nature of physical activity level as a risk or preventative factor for a number of lifestyle-related diseases, it is also an important item for population health and epidemiological data collection. In our previous work we have discussed in greater depth the possibilities for such data collection for numerous secondary uses [17].

Demonstrating the low perceived privacy risk of sharing this data at least for some individuals, is the recent practice of sharing such physical activity or fitness data, often in real-time, publicly via social media. Additionally, it is reasonable to predict that as mobile device sensors evolve, the collection and utilization of data by individuals will also expand to quantify further details of their activities, leading to increasingly rich and useful incidental participatory data collection.

Due to the intrinsic self-motivation that users already have to collect some types of sensor data, this category of participatory sensing benefits from the lowest ‘barrier to entry’, since individuals already contribute their own resources such as effort, time, CPU processing, network bandwidth and device battery usage to achieve data capture. This likely indicates that incentives to participate would be unneeded or minimal, if privacy concerns/barriers can be avoided or minimized as indicated in previous work [18]. Sensors and devices that are already acceptable in terms of user experience are already evidenced via numerous commercially available sensor products.

3.2.2 Passive Participatory Sensing

Passive participatory sensing is defined as sensor usage that requires explicit additional effort to enable data collection, which an individual would not have done unless they were explicitly participating in population health data capture, but does not attempt to and does not require any change of the day-to-day behavior of the individual. For example, this could include the individual using additional sensors that they would not have otherwise used, that collect data, for example, relating to physical activity, diet, heart rate, sleep cycles or environment, specifically to contribute to population data capture. This allows for a potentially more complete data collection in comparison to what is possible through purely incidental data collection. This brings to the fore the question of user motivation, once an individual is required to contribute any significant additional effort to collect sensor data. Previous work has discussed the idea of incentive schemes that use a market-based system [19] or the preservation of social translucence [20]. Inherently incentives will continue to pose a challenge to participatory sensing systems. In the health domain there is also the possibility of further incentives – greater self-knowledge of health, information and risk assessment, improved care, diagnostics, and possibly even an interest to contribute in a ‘citizen scientist’ capacity.

However, passive participatory sensing also increases privacy concerns. Since the data collection is intrinsically something that would not otherwise have occurred, a higher level of privacy responsibility rests with the collector. In a purely incidental approach, it is necessary that any data that leaves the mobile device either be kept strictly de-identified (with re-identification avoided) or securely and privately transmitted and stored. However, in a passive participatory model, responsibility for the secure storage on the device is also of import. This indicates that the type and characteristics of the HPSN architecture will vary depending on the category of health participatory sensing that is deployed.

3.2.3 Passive Participatory Sensing with Subjective Human-sensing and Feedback

This model combines the potential sensing advantages of passive participatory sensing with ‘human-sensing’ capabilities, allowing for the large amounts of objective sensing data to be complemented with subjective human-generated data and feedback. By human-sensing we refer to manual information inputs or responses provided by individuals.

This could easily be implemented through the addition of context-sensitive micro-surveys that are displayed to the user and attached to relevant collected sensor data. This would allow for both collection of data that is difficult to record through sensors alone, and also allow data that may have been missed via sensor collection to be added to the overall collection. Additionally, where anomalous data has been collected by sensors, human-sensing and feedback allows for validation to be performed by requesting subjective details or clarification of the data collected. As discussed below, the load on the user to provide such manual input would obviously need to be low.

Human-sensing has potential as a complement to sensor-based participatory sensing, able to bridge the gaps or some limitations of sensing technology [21]. While it would intrinsically increase the effort required to participate, and would require a further challenge to motivation, incentives and potential participation levels, it does not pose additional security concerns. This is apparent as it would not require direct one-to-one communication, rather generic context rules and micro-surveys could be broadcast to all participants with the processing of context and hence triggering of the additional data collection to occur locally. This is described in further detail in section 3.3.2.

3.2.4 Active Participatory Sensing

Active participatory sensing provides 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 [22]. 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. The instigation to carry out ‘active’ sensing activities could essentially constitute a public health intervention action. As such, the behavioral 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 of the effectiveness of campaigns on recipient groups. It is assumed that active participatory sensing would have similar levels of technical sensor capabilities to passive, 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.

3.2.5 Active Participatory Sensing with Subjective Human-sensing and Feedback

This final category incorporates the goals and framework of active participatory sensing but adds subjective human-sensing and feedback - hence incorporating the most complete level of data collection with public health intervention.

As the most all-encompassing level of HPSN system that we have considered, it may correspondingly have greater challenges in relation to motivation. However, it should be noted that security and privacy provisions would be required to be no more stringent than for that of either of its component parts. The interaction of these two components adds the capability to give human-sensing or feedback related to a specific intervention. This is a higher level of capability than that available in the other models. It also allows for more complete and useful information to be collected by enhancing what can be detected through sensors.

While this model represents the most complete functionality, in many cases it is likely that not all its capabilities would be required or that the potential motivation challenges may lead to a less comprehensive participatory sensing model being utilized for a particular goal. This is expected, and suggests the material distinctions existing between the different categories of health participatory sensing networks.

3.3 Health Participatory Sensing Network Conceptual Reference Architecture

3.3.1 Reference Architecture

The reference architecture proposed in this paper is able to support all models of health participatory sensing described in Section 3.2. However, the emphasis in discussion is on the most comprehensive models of HPSN such as described in Sections 3.2.4 and 3.2.5. In the discussion of this reference architecture, we also describe and emphasize how HPSNs can achieve strict privacy and anonymity for all individual participants, via such technologies and techniques as mix networks [23], k -anonymity, de-identification and submission of only ‘aggregate’ data.

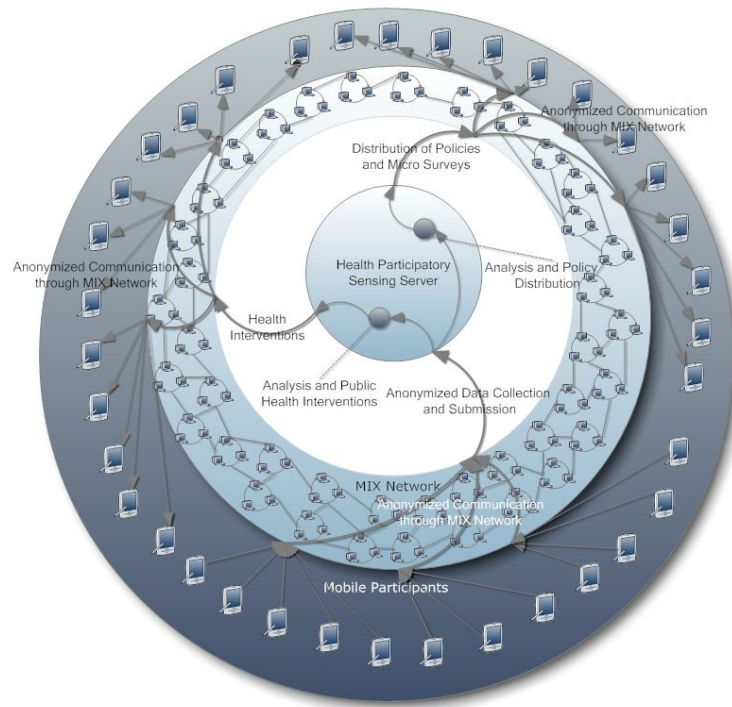


Fig. 3.1 Health Participatory Sensing Network Reference Architecture

The HPSN as a whole allows for the collection of various types of health-related data and various user interactions depending on the category of HPSN being adopted. The user will have granular control of the three types of interactions within the HPSN, as displayed in Fig. 3.1 which are: (1) data collection and submission, (2) micro surveys and (3) health interventions. At the coarsest level, each of these three functionalities can be disabled or enabled, with more specific options also being provided in each area. For example, a HPSN of the types described in Sections 3.2.1 and 3.2.2 would require no micro-survey or health intervention capabilities. An active HPSN as described in Section 3.2.4 would require data collection and health intervention capabilities, but not micro-surveys. An active HPSN with human-sensing as described in Section 3.2.5 would require all three capabilities.

Fig 3.1 shows a diagram of our conceptual reference architecture that would allow the higher levels of privacy required to support health participatory sensing, along with the distributed and multi-party capabilities of the overall platform. This architecture would also

include a number of different public health organizations utilizing and sharing a single distributed HPSN server group that communicates to participants via mix networks to preserve the anonymity of the participants. A public health organization/health organization would be defined as an organization involved in population data collection and/or public health intervention.

Data collection is the process of automated collection and submission of anonymized data. For the user, there is the ability to customize what data submission policy providers (health organizations) they are subscribed to, as well as what types of data may be submitted. An example would be of opting-in to submission of physical activity data to one health organization, but physical activity and sleep patterns to a second organization. Additionally, this can be adjusted on an individual policy level-basis rather than on specific categories. In a similar way, micro-surveys and health interventions can be opted into, with specific categories of interventions and types of micro-surveys being customizable. The distribution of data policies, micro-surveys and health interventions takes this granular nature into account, only distributing the selected categories/policies to an individual's device.

Data policies, health interventions and micro-surveys are distributed to an individual's device based on the customizations discussed above. This is achieved by the user mobile device periodically polling for updated policy sets based on the specific customization. However, beyond broad user customization for health interventions and micro-surveys, once on the device, only the most personalized/suited of those distributed is enacted on the mobile device, based on local processing so as to further maintain privacy (see Section 3.3.4). The best fit surveys/interventions are displayed/completed and responses then stored for data submission/analysis.

The functionality introduces a feedback loop - not for the individual, but for a generalized group. Once data is submitted to the server, including the automated data collection, micro-surveys and specific details of health interventions deployed, this can then be used

to adapt the later policies, survey and health interventions distributed. This can have both automated and manually supervised aspects to it.

3.3.2 Distribution of Data Collection Policy Rules and Micro Surveys Whilst Protecting Recipient Anonymity and Privacy

The HPSN data collection methodology has a number of advantages due to being able to collect near real-time population-wide data, and adjust the data collected when needed. As such, this requires a robust process for distribution of updated data collection policies, micro-surveys and health interventions, with the key requirements that updates be distributed in a timely and accurate manner. An addition to this, that is specific to anonymous data collection, is that of the critical need to preserve individual user privacy. The conventional approach would be to distribute personalized encrypted data updates through a mix network. While this would meet the requirement for anonymous distribution, as no detail of the individual is known, it also provides the data policy owner with specific details of the policies distributed to individual participants. To allow for greater efficiencies and provide a higher level of privacy, a generic distribution approach using the mix nodes themselves as distributed servers would provide an elegant solution.

To support the collection of anonymized data, distribution should be achieved through transmission of general policies packaged and distributed broadly. Additionally, the use of a multi-purpose HPSN with multiple public health organizations involved also raises the issue of quality assurance of the messages throughout the network that may have come from various different senders and sent through a distributed multi-user network. As such, the completeness (all the requested data is retrieved), correctness (the data returned is accurate and has not been modified) and freshness (the data is the most up to date version available) need to be assured.

A granular approach utilizing verification objects constructed by hashing and digital

signing of distributed content can assure the completeness and correctness of data. Additionally, the inclusion of timestamps and expiry rates can assure the freshness of distributed data without direct communication to the associated public health organization. Our previous work found that user CPU time and data overheads of this type of approach can be quite minimal [24]; without significant additional overheads for the data owners/distributor.

This approach would additionally allow for dissemination and retrieval of data through the anonymous communications network, with users retrieving policy updates and interventions relevant to them only, without breaching their anonymity.

To do this efficiently while improving privacy, we have considered the use of an adapted query assurance process suitable for distributed data sources [24]. This would allow for the nodes in the mix network to cache general data and just distribute the requested subset to requests - without the policy owner being aware or involved in this efficient distribution.

Going into further detail, this approach uses signed verification objects distributed with policy data that allow the policy content to be verified in a granular way - that is, even if only a small portion of the policy data is delivered to an individual user, verification can be assured at reasonable levels of overhead. This is achieved by the use of sorted and signed merkle hash verification trees. Each policy block will be hashed and stored as a leaf node in the merkle hash tree with the branches to the root comprised of a hash of the branches/leaves below. Finally, the root is signed by the policy owner with a digital timestamp attached and included in the signed value. Due to the signed nature of the verification tree, the policy data distributed can be authenticated as correct. As the tree is sorted, completeness can be assured by attaching border values with any policy distribution. Finally, due to the expiring timestamp, the distribution of old/expired policies can be detected and discarded.

In Fig. 3.2 the organization of a policy distribution verification tree is displayed. It allows for many layers of branches depending on the size of the data set. Further, based on the data requested, a subset of this tree would be packaged as a verification object and

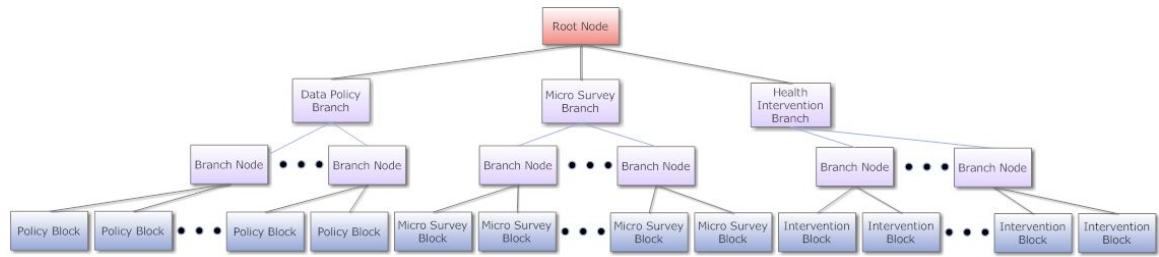


Fig. 3.2 Policy Distribution Verification Tree Structure

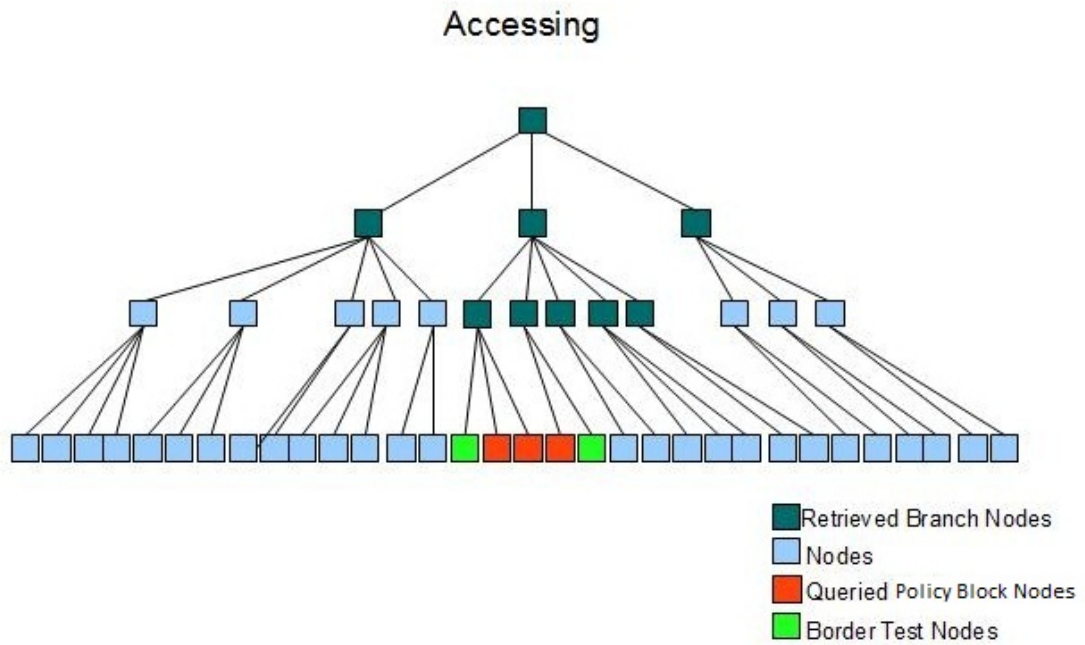


Fig. 3.3 Verification Tree Accessing

distributed with the policy data to allow for verification. The verification tree contains three specific types of nodes - root, branch and leaf.

The verification tree structure is shown in Fig. 3.3 and contains four primary elements as follows:

- $Root_node(S(T_s, H((Child_node_1), \dots, (Child_node_{n-1}), (Child_node_n))), T_s)$: Given S means the contents are digitally-signed, T_s is a time stamp with expiry duration, H means the contents are hashed and $Child_node$ refers to the H value stored in $Branch_node$ or $Leaf_node$ directly below the root.

- $Branch_node(H((Child_node_1), \dots, (Child_node_{n-1}), (Child_node_n)), Index_value)$: Given H means the contents are hashed, $Child_node$ refers to the H value stored in $Branch_node$ or $Leaf_node$ directly below the current branch and $Index_value$ is the sort value of the branch.
- $Branch_S_node(S(T_s, H((Child_node_1), \dots, (Child_node_{n-1}), (Child_node_n))), T_s, Index_value)$: Similar to the $Root_node$, the $Branch_S_node$ contains a digital signature and an expiring time stamp. Additionally, it also contains the $Index_value$ as the sort value of the branch. The intention is that by providing signed branch nodes throughout the verification tree, in proximity to often retrieved leaves or clusters of leaves, the overall efficiency of the verification scheme can be improved overtime based on usage.
- $Leaf_node(H(M, Record_Path), Record_Path, Read_Count, Write_Count)$: Given H is the hash of its contents and M is the stored record. $Record_Path$ is the direct path to the record, $Read_Count$ and $Write_Count$ are metrics to track the activity on individual leaves of the verification object.

3.3.3 Anonymous Data Collection and Submission

The utilization of a methodology for anonymous submission of collected data is necessary in all implementations that do not incorporate a trusted server, and this capability alone should remove much of the potential and perceived privacy risks of a health participatory sensor network. Two such possible extant examples are mix networks [23] (Fig 3.1) and onion routing networks [25]. Either would be capable of allowing for anonymous submission, with onion routing having some additional advanced capabilities. With the addition of an anonymous submission network, the remaining primary privacy concern relates to the data submitted.

Secure de-identified approaches have gained specific coverage in the health context [26] due to the often highly sensitive nature of the data details. Conversely, such data has great

importance as a source for research and epidemiology. Typically, de-identification works by removing or making less specific identifiers and quasi-identifiers. This leads to a k -anonymity type approach where as long as k number of individuals are indistinguishable from each other and do not have sensitive details in common, data privacy is considered to be preserved. This has been extended into the sensing context with spatial k -anonymity explored in [27]. However, assured k -anonymity approaches in most cases require a trusted server or aggregator to perform this analysis and decision making based on data received. This is hard to achieve in an anonymous submission distributed network that attempts to reduce the need for highly secure trusted components. The alternative is to make de-identification decisions locally without external knowledge of the potential k value of collected values, though this is not as assuredly secure as a trusted model. The advantage of reducing the sensitivity at the initial level of the mobile device before submission is promising.

A novel infrastructure approach is required to address the inherent property of detailed sensor data that even if de-identified could still act as a privacy risk through later re-identification of the individual with other known data. This risk can be reduced by submitting only less detailed and aggregate data [28]. In the public health domain, this raises the question of ‘what is sufficient data for public health uses?’. It seems likely that more detailed information than is currently collected by traditional methodologies could be submitted without significant privacy risks, due to the very broad nature of population-wide health measures as potential preventative and risk factors.

This is made possible through storing detailed sensor data locally, then performing anonymizing/aggregation on the data before submission. This includes removing any overly rare and identifying demographic information, making any temporal or location details more general and generating calculated aggregate measures based on the detailed data. This allows potentially long-term analysis to be performed on the device and submitted without

the server needing to know the sensitive detailed history.

The types of data that could be collected through this approach are quite varied, especially with the potential capabilities of extending the sensing component through micro-surveys and feedback, and through future advancements of health sensor technologies. Some examples that are often of interest in previous large-scale health data collection are:

- **Physical Activity Patterns and Intensity** - Due to its significance as a preventative factor in a number of lifestyle diseases, it is of high importance when considering a population-wide health participatory sensing model. The physical activity can usually be split between the following three categories.
 - **Work Related Activity:** The amount and intensity of physical activity completed during work.
 - **Recreational Activity:** Activity that is not associated with work or transportation.
 - **Transportation Activity:** Active transportation (walking, cycling or similar) as a form of physical activity that coincides with travel. Active transportation is a focus of public policy in many regions [29].
- **Nutritional, Caloric Burn and Caloric Intake** - This type of data could provide more detailed information on overall energy expenditure and nutritional intake of individuals, segments of the population and the population overall [30].
- **Body Mass Index (BMI) and change over time** - This would allow for a current snapshot of BMI, with the potential for trend analysis, based on the individual/community change over time.
- **Sleep Patterns and Regularity** - Sleep patterns are both an indicator and a preventative/risk factor for a number of conditions.

As indicated, all these types of data can be submitted using the fully privacy-preserving and anonymization mechanisms described. With that result that only population aggregate measures are to be stored at the public health organizations for population health purposes.

3.3.4 Public Health Interventions

A major area of potential usefulness of HPSNs is the ability to distribute targeted or personalized public health interventions to individuals (see Sections 3.2.4 and 3.2.5). It may appear that to do so implicitly requires potentially identifying details to be known by the server, which may affect willingness and motivation to participate in such a network.

However, the sensitivity of health related information and the capability of the proposed reference architecture suggests a novel approach to non-identifying targeted public health interventions.

Additionally, it seems likely that there will be a number of different public health organizations that would be interested in participating in these types of networks, whereby individuals are able to subscribe or opt-in to partake in passive or active participation with each individual organization.

The novel approach introduced, involves broadcasting larger generalized public health intervention packages to the entirety of the participants or subsets. Then based on local processing, the correct information or intervention information is displayed or actioned on individual devices. This would allow for communication with individuals that could be meaningful and personalized without risk of re-identification of the individual. This approach could also be used for the dissemination of micro-surveys to individuals for data collection for the applicable models of HPSN.

3.3.5 HPSN and Participant Device Interaction

The participant device will have the following capabilities to interact with the network (Fig. 3.4):

- **De-identification** – This component manages the de-identification of data prior to submission by making specifics of the data less sensitive and individually identifying if necessary. This is actioned by a consideration of the organization subscription policies which will identify what data is to be submitted and at what particular level of detail.
- **Micro-surveys** – An opt-in additional component (see Sections 3.2.3 and 3.2.5) that would allow the public health organizations to enact additional data collection by context-aware micro-surveys to complement or enhance the sensor data collection.
- **Local Aggregate Processing** – Calculates aggregate measures [28] according to rules delivered through the participatory network. This also takes into account local rules and is conscious of potential re-identification risks when allowing data submission.
- **Intervention Delivery** – An opt-in additional component that retrieves general interventions targeted at particular demographic groups (see Sections 3.2.4 and 3.2.5). These interventions are then locally processed according to the individual's data to present the correct targeted intervention to the end user.
- **Subscription policies** – This component will allow an individual to control which public health organizations they interact with, and the opt-in and opt-out of capabilities and collection offered. For example, an individual may choose to collect data and receive interventions from an organization X, while only collecting data and opting out of interventions from organization Y. This could be provided to a detailed granular

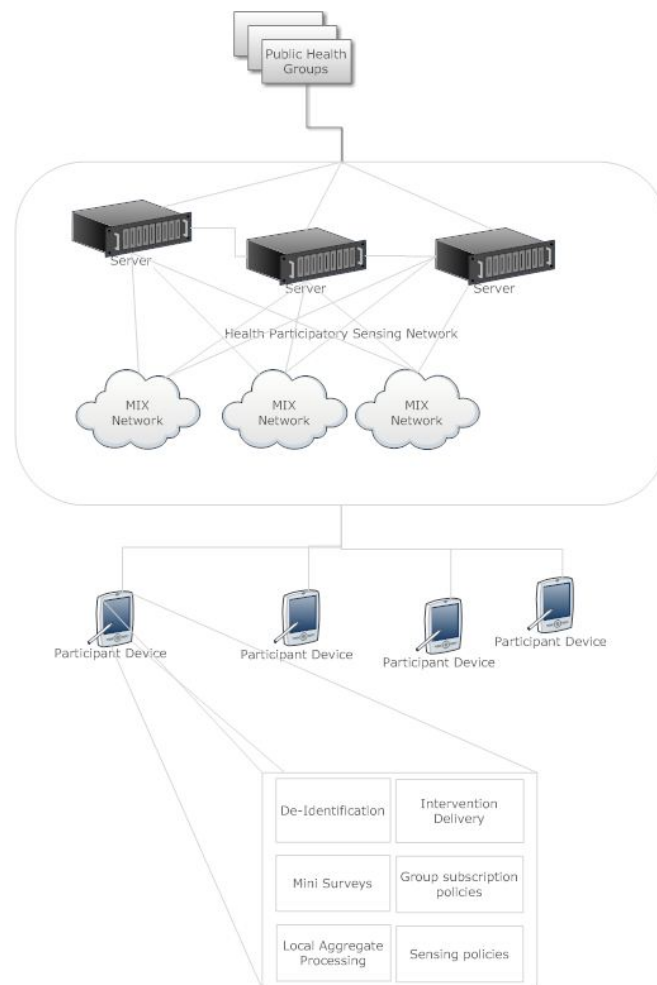


Fig. 3.4 Network and Participant Device Communication

level where individuals can collect for only a specific measure or receive communication on a very narrow topic.

- Sensing policies – Local policies defined by the user that set limitations as to what and how they will participate in the HPSN. This would include which sensors are used, battery management policy, time and detail limitations, amongst others.

3.4 User Scenario

For our example purposes, let us consider user David who is a participant in a HPSN. David is subscribed to the Department of Health, and to the Active Transport Initiative Inc. data collection policies. Additionally, he has also subscribed to micro-surveys from Active Transport Inc. and health interventions from the Department of Health.

The subscribed to policies, micro-surveys and health interventions are updated periodically. This is controlled by either the client checking for updates, with the maximum valid period of a set of policies set by an expiring timestamps; one referring to the distribution process (maximum time before the data should no longer be distributed) and a second referring to the expected validity period (maximum period before the policy needs to be updated).

Throughout his daily schedule, David's mobile device automatically collects physical activity data and relevant data to each organization is submitted intermittently throughout the day utilizing the privacy-preserving mechanisms of the HPSN. Again, this can be totally anonymous and does not allow re-identification – it can just provide a valuable input when combined with the mass of other individual's data, for population health measures.

More specifically, the Department of Health is interested in overall physical activity in a day with age bracket and coarse location information also submitted. Alternatively, Active Transport Initiative Inc. is only concerned with physical activity related to transportation (e.g. walking/cycling commuting) with the additional data of age bracket and start coarse location and end coarse location submitted. Occasionally during the day/ very infrequently, David is prompted to complete a micro-survey related to active transport. This is a 30 second survey, for example, asking for a ranking of the five most significant factors as to whether on a specific day he would cycle/walk to work. This micro-survey was presented to David based on the locally stored data relating to his travel habits – although it would have been sent to a much larger group. As an average two day per week cycle commuter, he is

a prime target for increase in active transport modal share. David's preferences also restrict how often micro-surveys from individual organizations and the overall HPSN system can request micro-surveys.

Based on his personal preferences and previous trend data, David is also prompted by a health intervention from the Department of Health, suggesting the health benefits of cycling. This suggestion is computed on his local device based on the typical time of day that David exercises, current weather patterns including UV rating and potential vitamin D intake levels. David chooses not to take up the suggestion of this intervention, triggering a wait-time (set by David's preferences) as to how soon a new intervention can be received.

David is also interested to track his own health-related data and finds this a benefit that also assists his motivation to participate in the HPSN. The data he displays to himself for this purpose is kept securely on his own device, and is never transmitted to the HPSN – for this reason, this data for self-use, can be more detailed and can be viewed by David without aggregation processing [28] having first occurred if desired.

3.5 Conclusion

The utilization of health participatory sensing networks for population health and epidemiological data collection is a development that can greatly increase the capability and scale of health data collection.

The challenges to this approach however are not insignificant, with privacy and incentives to participate, being core concerns affecting the level of participation and therefore potential data collection.

In this chapter, we have introduced the first classification of health participatory sensing networks, and discussed these classifications with particular reference given to the required level of effort for individuals, privacy concerns as well as potential realizable benefits and incentives.

In this work, we have also defined a conceptual reference architecture that supports the capabilities required for the more highly detailed health participatory sensing categories, that includes anonymous submission and secure de-identification - this would have significant advantages that may positively affect participation, as compared to a more traditional trusted server methodology. We have also described a user scenario in detail.

REFERENCES

- [1] Andrew Clarke and Robert Steele. “Health Participatory Sensing Networks”. In: *Mobile Information Systems* 10 (3 2014), pp. 229–242.
- [2] Jeff Burke et al. “Participatory sensing”. In: *In: Workshop on World-Sensor-Web (WSW’06): Mobile Device Centric Sensor Networks and Applications*. 2006, pp. 117–134.
- [3] Roberta Kwok. *Personal technology: Phoning in data*. 2009. URL: <http://www.nature.com/news/2009/090422/%20full/458959a.html>.
- [4] Gartner. *Gartner Says Worldwide Smartphone Sales Soared in Fourth Quarter of 2011 With 47 Percent Growth*. 2011. URL: <http://www.gartner.com/it/page.jsp?id=1924314>.
- [5] European Cyclists’ Federation. *Track that Bike: How Can Electronic Chips Boost Cycling?* 2012. URL: <http://www.ecf.com/news/track-that-bike-how-can-electronic-chips-boost-cycling/>.
- [6] Bicycle Network. *Riderlog*. 2012. URL: <http://www.bv.com.au/general/ride-to-work/91481/>.
- [7] Jawbone. *UP by Jawbone with MotionX Technology Empowers You to Live a Healthier Life*. 2011. URL: <http://content.jawbone.com/static/www/pdf/press-releases/up-press-release-110311.pdf>.

- [8] Christine Outram, Carlo Ratti, and Assaf Biderman. “The Copenhagen Wheel: An innovative electric bicycle system that harnesses the power of real-time information and crowd sourcing”. In: *EVER Monaco International Exhibition & Conference on Ecologic Vehicles & Renewable Energies*. 2010.
- [9] Koji Yatani and Truong N. Khai. “BodyScope: A Wearable Acoustic Sensor for Activity Recognition”. In: *Proceedings of the 14th international conference on Ubiquitous computing*. UbiComp ’12. New York, NY, USA: ACM, 2012.
- [10] Robert Steele. “An Overview of the State of the Art of Automated Capture of Dietary Intake Information”. In: *Critical Reviews in Food Science and Nutrition* (2013).
- [11] S. Passler and W.-J. Fischer. “Acoustical method for objective food intake monitoring using a wearable sensor system”. In: *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. May 2011, pp. 266–269.
- [12] Jonathan Lester et al. “Automatic classification of daily fluid intake”. In: *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on*. Mar. 2010, pp. 1–8.
- [13] Keng-hao Chang et al. “The diet-aware dining table: observing dietary behaviors over a tabletop surface”. In: *Proceedings of the 4th international conference on Pervasive Computing*. PERVASIVE’06. Dublin, Ireland: Springer-Verlag, 2006, pp. 366–382. ISBN: 3-540-33894-2, 978-3-540-33894-9.
- [14] Gregorio Villalobos et al. “A personal assistive system for nutrient intake monitoring”. In: *Proceedings of the 2011 international ACM workshop on Ubiquitous meta user interfaces*. Ubi-MUI ’11. Scottsdale, Arizona, USA: ACM, 2011, pp. 17–22. ISBN: 978-1-4503-0993-6.

- [15] Ming-Zher Poh et al. “Heartphones: Sensor Earphones and Mobile Application for Non-obtrusive Health Monitoring”. In: *Wearable Computers, 2009. ISWC '09. International Symposium on*. Sept. 2009, pp. 153–154.
- [16] Predrag Klasnja et al. “Using Mobile & Personal Sensing Technologies to Support Health Behavior Change in Everyday Life: Lessons Learned”. In: *AMIA Annual Symposium Proceedings*. 2009, pp. 55–59.
- [17] Andrew Clarke and Robert Steele. “How personal fitness data can be re-used by smart cities”. In: *Intelligent Sensors, Sensor Networks and Information Processing (ISS-NIP), 2011 Seventh International Conference on*. Dec. 2011, pp. 395–400.
- [18] Predrag V. Klasnja et al. “Exploring Privacy Concerns about Personal Sensing”. In: *Pervasive*. Ed. by Hideyuki Tokuda et al. Vol. 5538. Lecture Notes in Computer Science. Springer, 2009, pp. 176–183. ISBN: 978-3-642-01515-1.
- [19] Juong-Sik Lee and Baik Hoh. “Sell your experiences: a market mechanism based incentive for participatory sensing”. In: *PerCom*. IEEE Computer Society, 2010, pp. 60–68.
- [20] Ioannis Krontiris and Nicolas Maisonneuve. “Participatory Sensing: The Tension Between Social Translucence and Privacy”. In: *Trustworthy Internet*. Ed. by Luca Salgarelli, Giuseppe Bianchi, and Nicola Blefari-Melazzi. Springer Milan, 2011, pp. 159–170. ISBN: 978-88-470-1818-1.
- [21] Robert Steele and Andrew Clarke. “A Real-time, Composite Healthy Building Measurement Architecture Drawing Upon Occupant Smartphone-collected Data”. In: *10th International Healthy Buildings Conference*. July 2012.
- [22] John Rula and Fabián E. Bustamante. “Crowd (Soft) Control: Moving Beyond the Opportunistic”. In: *Proceedings of the Twelfth Workshop on Mobile Computing Sys-*

- tems & Applications*. HotMobile '12. San Diego, California: ACM, 2012, 3:1–3:6. ISBN: 978-1-4503-1207-3.
- [23] Krishna Sampigethaya and Radha Poovendran. “A Survey on Mix Networks and Their Secure Applications”. In: *Proceedings of the IEEE* 94.12 (Dec. 2006), pp. 2142–2181. ISSN: 0018-9219.
- [24] Andrew Clarke and Robert Steele. “Secure and Reliable Distributed Health Records: Achieving Query Assurance across Repositories of Encrypted Health Data”. In: *System Science (HICSS), 2012 45th Hawaii International Conference on*. Jan. 2012, pp. 3021–3029.
- [25] Sjouke Mauw, J. Verschuren, and Eric de Vink. “A Formalization of Anonymity and Onion Routing”. In: *Computer Security - ESORICS 2004*. Ed. by Pierangela Samarati et al. Vol. 3193. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2004, pp. 109–124. ISBN: 978-3-540-22987-2.
- [26] K. El Emam. “Risk-Based De-Identification of Health Data”. In: *Security Privacy, IEEE* 8.3 (May 2010), pp. 64–67. ISSN: 1540-7993.
- [27] Panos Kalnis and Gabriel Ghinita. “Spatial k-Anonymity”. In: *Encyclopedia of Database Systems*. Ed. by Ling Liu and M. Tamer Özsu. Springer US, 2009, p. 2714. ISBN: 978-0-387-35544-3, 978-0-387-39940-9.
- [28] Andrew Clarke and Robert 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*. 2012, pp. 2158–2161.
- [29] Francesca Racioppi et al. *A physically active life through everyday transport*. 2002. URL: http://www.euro.who.int/%5C_%5C_data/assets/pdf%5C_file/0011/87572/E75662.pdf.

-
- [30] Robert Steele. “An Overview of the Role of Informatics-based Systems in Furthering an Integrated Paddock to Plate Food Supply System.” In: *Progress in Industrial Ecology*. Vol. 10. 2. 2013.

**Faculty of Health Sciences
Author Contribution Statement**

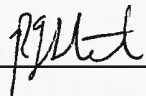
Candidate Name: Andrew Clarke

Degree Title: Doctor of Philosophy

Paper Title: Health Participatory Sensing Networks

As the corresponding author of the above paper, I confirm that the above candidate has made contributions to the following:

- Conception and design of the research
- Analysis and interpretation of the findings
- Writing the paper and critical appraisal of content

Signed  Name Prof Robert Steele Date 10th March 15

4 QUERY ASSURANCE FOR DISTRIBUTION OF HPSN DATA COLLECTION RULES AND INTERVENTIONS

Preamble

This chapter extends on a journal paper that was published in the Journal of Health Systems [1], it has been included as a chapter of this thesis with a significant rewrite to align more closely with the overall theme of the thesis, as part of the re-write the related work section has been moved to Chapter 2 and the future work moved to Chapter 7. In addition, formatting changes were made to align with the thesis format. This chapter is the first implementation/evaluation chapter, expanding on the solution first introduced in subsection 3.3.2 in the previous chapter. The present chapter is concerned with the technical details of distribution of data collection policy rules, micro-surveys and public health interventions within the health participatory sensing network. That is, this chapter covers the content distribution component of the overall architecture, while later implementation chapters will cover the anonymity and privacy protections, specifically the utilization of onion routing and content aware de-identification to reduce privacy risks (Chapter 5), and the implementation of public health interventions (Chapter 6).

The distribution of health participatory sensing content is a key challenge, within a health participatory sensing network, as it is highly likely that any successful system would have multiple public health organizations/groups utilizing the data collection and public health intervention capabilities - requiring a multiple data owner system, where accuracy

can be assured. Additionally, a distributed system is favorable as it provides scale and potentially privacy benefits, as potential breaches of privacy would require collusion between multiple distributed nodes. These issues are discussed further in chapter 5, including a prototype implementation and results.

ABSTRACT

Health information system architectures inherently include distributed systems and data repositories across multiple organizations, health providers and with potentially some data stored with the health consumer. This is part of the shift to more fully integrated electronic health systems. Due to the varied stakeholders of these systems, it will become more important to provide a high level of query quality assurance for the parties utilizing these distributed and shared data repositories. A specific example of a distributed health data model is that of a HPSN (Health Participatory Sensing Networks), where the data collection rules and public health interventions are created and distributed by multiple stakeholders in the form of public health groups/organizations and the detailed data is stored with individual participants. A core consideration of the HPSN approach is providing data confidentiality, including protecting against insider security threats. As such, it will often be desirable that communication between the health groups and the individuals be stored/transmitted in an encrypted format. In this chapter, we present and describe the implementation and evaluation of a query assurance model that implements the three requirements of query assurance across sources of searchable encrypted data. Further, we consider the issue of freshness and data persistence in a multiple data-owner environment, including a discussion of the characteristics of consumer interfacing health information systems.

Keywords: Consumer Health - Query Assurance - System Architecture - Health Sensors

4.1 Introduction

Public health information systems intrinsically entail complex distributed systems and data repositories, due to the increasingly integrated nature of national and global health systems [2]. This shift will consolidate public health information across multiple organizations, health providers and in the case of participatory sensing data stored with the health consumer, to provide more efficient and effective public health. In the model for health participatory sensing systems shown in Figure 4.1, we consider that there will be a number of health participatory sensing servers owned and operated by individual groups/organizations distributing information and collection information from individual HPSN participants either directly or through Web services.

Under this multiplication of disparate data sources, contributors and data-owners, where complete trust between all parties cannot be assumed, the challenge is to have an efficient and effective way to verify and combine all data that is distributed to the HPSN participant. This suggests the need for a unifying approach to query assurance, where query assurance is defined as the data source accurately responding to queries by meeting the requirements of correctness, completeness and freshness.

In this HPSN scenario, it quickly becomes clear that there will be significant new concerns not previously considered in traditional database models. Parallels can be drawn between this model and similar new database models [3, 4], which consider among their security concerns, data confidentiality and query assurance or integrity. Within the domain of health, information accuracy and confidentiality are of primary concern, due to the sensitive nature of the data stored. As such, a scheme that combines data confidentiality and query assurance, without substantially diminishing the usability of the information, is an important initial step. Further, we consider that it should be applicable over heterogeneous types of health data and highly scalable.

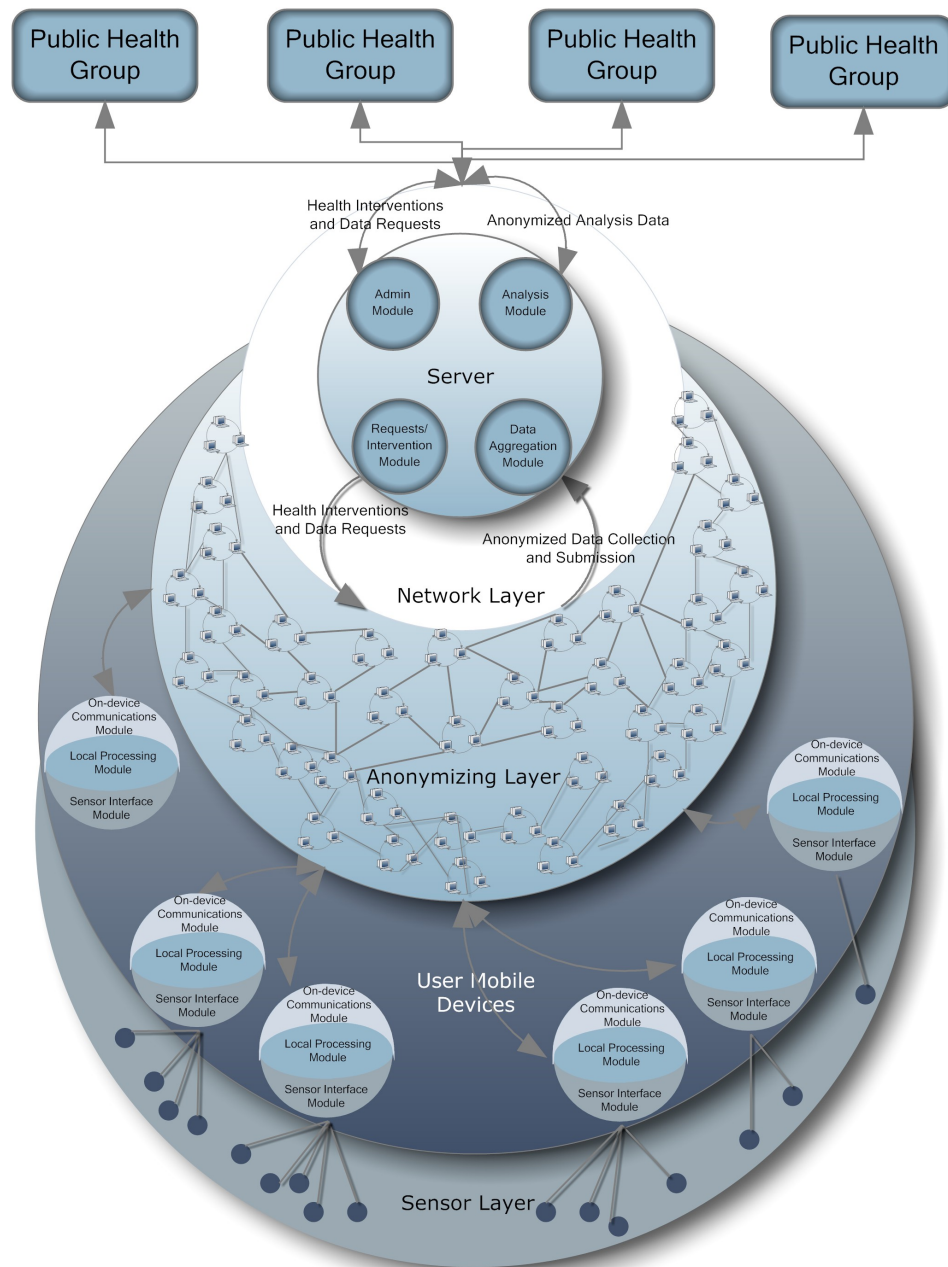


Fig. 4.1 Distributed Public Health Information System Architecture or HPSN

Query assurance covers the requirement that the returned query result should provide assurance of [5, 3, 6]:

- Correctness - The data returned is accurate, created by the data-owner and not modified in any way.
- Completeness - The data returned is the entirety of the matching result set.
- Freshness - The data returned is the current up to date version of a record.

For encrypted distributed data that does not provide query support (just retrieval of a single indexed record), completeness and correctness are inherently provided. Since the data will only be retrieved by an index or block number, it follows that if a record is returned that corresponds to the ID requested, completeness is assured [7]. Further, the use of secure encryption over all data ensures the record was created by the data-owner and that there was no third-party modification.

However, in the domain of health participatory sensing networks, more than a minimum rudimentary level of querying (such as that provided by [7, 8] that provided encrypted block retrieval) of outsourced data will be required, such as search term, attribute or range queries. Additionally, queries will bring together information from multiple sources, making completeness a concern. It is also an issue, that without assurance of the freshness of the data component of the record, the returned results may reflect out-of-date or stale data. Further, we consider that cooperative groups within the health sector may prefer to combine health information, thus creating the scenario where a health participatory sensing server will have multiple data-owners, thereby potentially complicating these issues further.

The following sections of the chapter discuss these issues in greater detail. The problem definition section identifies the key components required for health participatory sensing server data storage in current and emerging health information systems. Related work discusses the most relevant prior approaches to query assurance and identifies any issues that

may affect the suitability of prior approaches for application with HPSN data. In the Proposed Novel Approach to Secure Query Assurance section, we propose a new approach that is tailored to meet the requirements in the problem definition and extends from our previous work to provide an analytical evaluation of scalability. Finally, we present the details of our test implementation and evaluation results discussion in the Experiment and Evaluation, Results and Discussion sections.

4.2 Problem Definition

The distribution of data collection rules and public health interventions within HPSNs necessitates the use of scalable and distributed systems, due to the interaction of multiple groups and the potential number of participants. This is an emerging reality of the multi-site, multi-organization, health participatory sensing systems. However, provision of a secure implementation with assurance of the quality of the information remains an issue. This is especially true due to the expectation that information will be moving between different organizations/groups that are owners of data within the larger system. We consider a central issue of this problem to be that of encrypted data storage and retrieval.

In this model, we consider that the information system is vulnerable to both external attacks as well as internal attackers at the distributed data repositories who have privileged access to parts of the system. External attackers are already quite well addressed with access control and communication security. In our assessment of the problem, we consider that insider attackers have full access to the system, but do not have access to individual data owner's private encryption keys (this holds for both symmetrical and public key cryptography used in the approach).

Based on our previous work and consideration of the most closely related database models namely multi data-owner, outsourced or cloud data repository approaches we conclude that the following three key components will need to be present in the database implemen-

tation:

- Data confidentiality - The data stored must only be retrievable by authorized users. Further, the data should not be viewable by the data repository operators at any point.
- Query Assurance - The data returned from a query is the complete, unaltered and freshest version of the stored records at any particular point in time.
- Efficient and secure storage - The overheads created by providing the other requirements should not significantly impact on the efficiency of the data storage.

The proposed novel solution detailed later in this chapter will address these three points through the use of searchable encryption, authenticated query assurance and efficiency adjustments that will be demonstrated through the implementation and evaluation.

4.3 Proposed Novel Approach for Secure Query Assurance

We consider that to provide query assurance to public health information data repositories, an authenticated query assurance approach is preferred. This is due to the level of integrity being more strongly guaranteed for all data as opposed to probabilistic approaches. Additionally, as it is external to the data repository it can easily be applied to data spread across heterogeneous data sources or scaled out independently to the data repositories themselves. Lastly, it doesn't require the participants of the HPSN to have additional knowledge of the data stored as in the case of a probabilistic approach. Further, previous work has applied it across various types of database; SQL [9], XML [10, 11] and tree-indexed data [7]. In our approach to this issue, we consider that our previous work [11] provides a good starting point. In addition to previous functionality, support for the following is required:

1. Trapdoor encryption of Index keywords to limit or avoid inference attacks

2. A more robust freshness approach for multiple data-owners, that not only ensures the up to date nature of the record, but that the previous update was correctly applied and persisted.
3. Encryption of public health information data collection rules/interventions

In the following subsections, we first describe the methodology by which query assurance integrity is provided. We then discuss the other important areas of the methodology: encryption, granularity, timestamps and digital signatures, maintenance and additional freshness assurance. The considerations that were made in applying these components to the implementation and any issues and limitations are detailed.

4.3.1 Public Health Data Query Assurance Method

In our approach to storing electronic public health information, we propose to extend our previous query assurance solution to address point 2 of the required functionality with additional freshness protection provided. The approach is primarily comprised of a self-balancing merkle hash tree with additional information stored in the root, branch and the leaf nodes to provide the query verification required.

This approach is differentiated from previous approaches, through the use of additional signed nodes at common branch nodes of regularly accessed records to decrease the time required and verification object size of queries on average.

The verification tree structure is shown in Figure 4.2. The verification tree is sorted based on an *Index_vvalue*. This could be a unique path to the retrievable element, such as the details of repository, table/location and unique primary key/identifier, or a search term or attributes of the record could also be used. In general, it is expected a number of verification trees would be created in many cases to support more flexible querying. The verification tree structure contains four primary elements as follows:

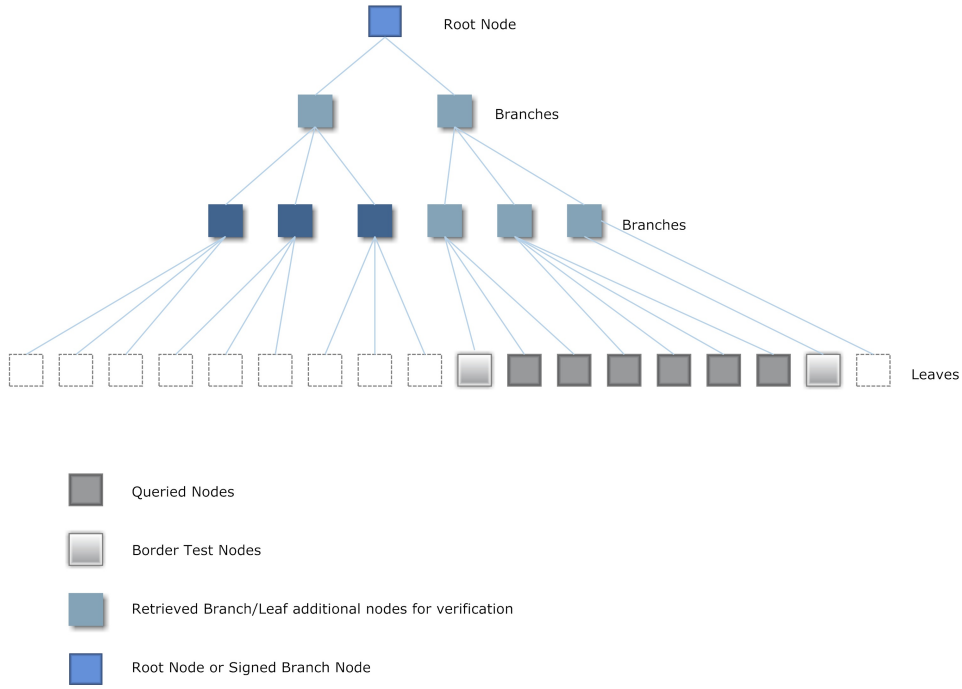


Fig. 4.2 Verification Tree Accessing Pre Maintenance

- $Root_node(S(T_s, H((Child_node_1), \dots, (Child_node_{n-1}), (Child_node_n))), T_s)$: Given S means the contents are digitally-signed, T_s is a timestamp with expiry duration, H means the contents are hashed and $Child_node$ refers to the H value stored in $Branch_node$ or $Leaf_node$ directly below the root.
- $Branch_node(H((Child_node_1), \dots, (Child_node_{n-1}), (Child_node_n)), Index_value)$: Given H means the contents are hashed, $Child_node$ refers to the H value stored in $Branch_node$ or $Leaf_node$ directly below the current branch and $Index_value$ is the sort value of the branch.
- $Branch_S_node(S(T_s, H((Child_node_1), \dots, (Child_node_{n-1}), (Child_node_n))), T_s, Index_value)$: Similar to the $Root_node$, the $Branch_S_node$ contains a digital signature and an expiring timestamp. Additionally, it also contains the $Index_value$ as the sort value of the branch. The intention is that by providing signed branch nodes throughout the verification tree, in proximity to often retrieved leaves or clusters of leaves, the overall

efficiency of the verification scheme can be improved overtime based on usage.

- *Leaf_node*($H(M, Record_Path), Record_Path, Read_Count, Write_Count$): Given H is the hash of its contents and M is the stored record. *Record_Path* is the direct path to the record, *Read_Count* and *Write_Count* are metrics to track the activity on individual leaves of the verification object.

The verification tree provides correctness, completeness and freshness in the following manner:

- **Correctness:** The hash of the stored data is kept in a leaf node, which through being propagated up the tree through multiple levels of hashing before being signed at the root node or a signed branch, ensures that the returned data can be verified. This approach is taken because digitally signing each individual element can be quite expensive in terms of CPU resources. Attempts have been made in previous works [9, 12] to improve the efficiency.
- **Completeness:** To ensure completeness when the verification object is returned with the query, the sibling leaf to either side of the result is returned as shown in Figure 4.2. Since the verification tree is sorted based on the index/keyword queried, this provides certainty that the entire result set is returned.
- **Freshness:** An expiring timestamp is stored in the root node and any signed branches. This allows the data owner to vary the length of expiry and hence the required frequency of updates/keep-alives required on the verification tree. This places a limit on the period of time that old data and its associated verification object can be used.

Finally, the verification tree can be used to sync just the records that have changed since a previous sync, by navigating the verification tree, rather than querying against temporal or logging data within the database.

In the following subsections, we go into further details into the query assurance methodology, identifying methodology decisions and their justifications.

Timestamps and Digital Signatures

Timestamps and digital signatures play a significant role in query assurance and efficiency of distributed public health information database solutions. All verification objects need to be time-stamped and digitally signed at their root to prove query correctness/completeness/freshness. This approach is quite well researched and understood [13, 7, 10]. However, previous approaches only considered a single timestamp/signature, with no consideration given to the following:

1. Propagation of the timestamp to users in set timestamp systems.
2. Calculation of an appropriate expiry rate in expiring timestamp systems.

As such, it seems beneficial that elements that are modified frequently should be covered by a timestamp with a shorter duration, while less frequently updated elements only need a basic coverage. Further, more critical areas of the database may be considered more time sensitive beyond the standard usage metrics. So, in cooperation with a hash granularity scheme, it is possible for a database to be separated into correct granularity, and then timestamps + signature added to particularly high throughput sections to enable a higher possible efficiency for the bulk of the database operations.

As a further consideration, as each individual timestamp may now only cover a smaller number of nodes - (that may be accessed often, but not modified), the performance of verification retrieval can be improved by needing smaller verification objects, without the need to refresh many timestamps at short expiry rates.

In our approach we use expiring timestamps, with expiry rates and placement of signed nodes adjusted during the maintenance portion of the implementation.

Maintenance

Of importance to this approach is the use of maintenance to adjust the verification trees in order to improve efficiency. These measures attempt to optimize the verification objects so that the most common queries are the most efficient based on the read rates.

The operations possible to increase efficiency that we have identified in our approach are:

1. Increase or decrease hash granularities to more closely match the size of the queried elements.
2. Increase or decrease the duration of timestamps based on the modification rate.
3. Placing or removing additional signed nodes in proximity to high access leaf nodes.
4. Adjust the depth and breadth of the tree by increasing or decreasing the maximum child nodes per branch.

The high access nodes have digitally signed nodes added to a common ancestor, as shown in Figure 4.3. The outcome of this is that the size of the verification object (the amount of data that needs to be returned) is significantly reduced. Please see [11] for further information on possible algorithms that can be used in carrying out this maintenance operation.

In the extended approach in this chapter that focuses on encrypted public health data, only 2 and 3 from above are incorporated. The switch to encrypted records, as well as limiting the types of queries that can be performed against the records, also removes the necessity for hash granularity modification. Similarly, the adjustment of child nodes per branch is problematic to optimize when incorporating searchable trapdoor encryption.

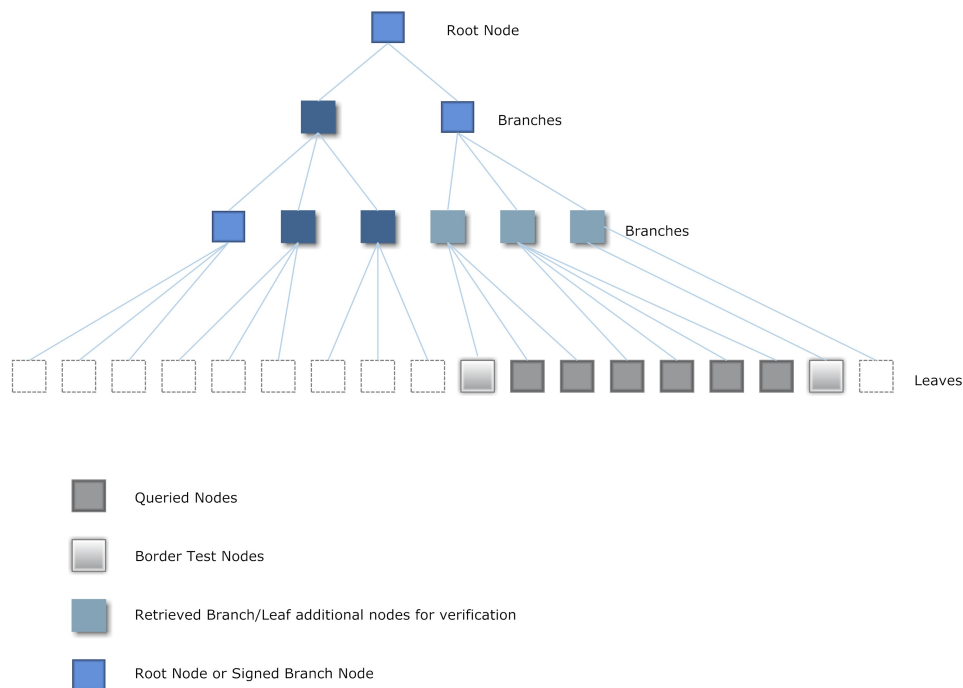


Fig. 4.3 Verification Tree Accessing Post Maintenance

Additional Freshness Assurance

In our approach, we digitally sign and include a timestamp at the root and specific branches of the merkle hash tree that are in proximity to highly accessed leaf nodes. An additional problem of shared databases with multiple data-owners, is that though the insert/update may appear to have been correctly applied and can then be subsequently queried against, there is limited protection against the database later reverting to an earlier state, as the server could be resigned by a different data-owner to give the appearance of a fresh data source. To alleviate this issue, the following approach can be used. Store the last signed time and signature for each data-owner in a separate metadata node directly below the root as shown in Figure 4.4. In cases where there are a significant number of data-owners, the metadata node could be expanded to become a branch rather than a node to allow more efficient re-signing.

Through this implementation, each individual data-owner, when next resigning the root

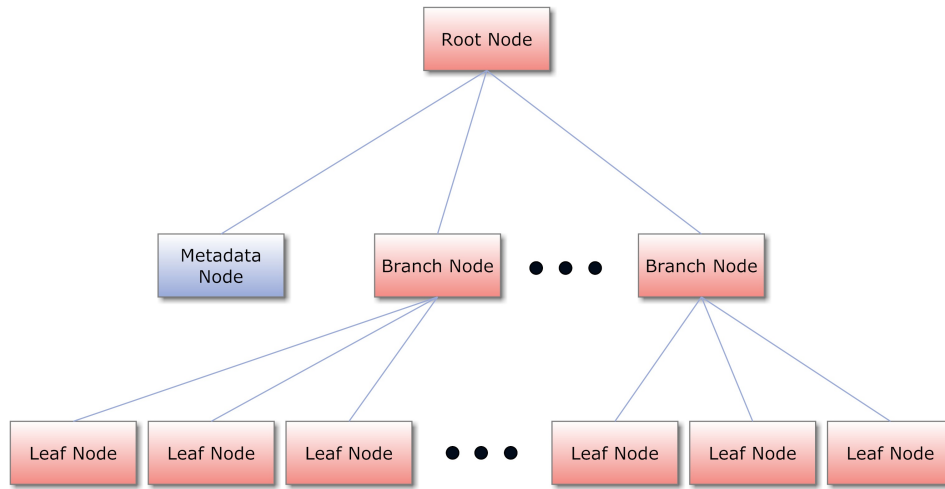


Fig. 4.4 Verification Tree with Metadata Node

node can check that the previous changes have persisted. This would be conducted by checking that the latest signature of a particular data-owner is still stored in the metadata node, as the individual data-owners metadata is signed by their unique private key. This provides assurance of persistence; however, it is still possible for a different data-owner to make changes to the stored data. Alternatively, the data-owner can intermittently check this value through the same method. In any case, in the event of data having not been persisted in the database, the event is detectable, and any data that has not been recorded is auditable and with sufficient logs/cache recoverable.

Granularity

Granularity has been a concern for authenticated query assurance [14]. Ideally data queried from a database would neatly coincide with the portions of the database that are hashed/digitally signed together. Merkle hash tree and other hash objects neatly addressed this problem by hashing uniform sections of the database and sorting them in a tree where only the root node needs a digital signature. This eliminated the need for most of the expensive public key cryptography that occurred in other works [15, 16]. This type of hashing will be referred to as uniform database hashing.

Uniform database hashing, depending on its usage, has the potential for inefficiencies. Coarse granularity and more of the database than is actually required to answer the clients query will need to be retrieved and transmitted, and fine granularity where there is the potential for the hashes/verification object to be a significant portion of the returned data. However, if a database granularity is closely matched to how the database is used (using knowledge of its application/uses) an efficient granularity level can be chosen. However, it still has the weakness of needing outside intervention to adjust granularity if database applications change over time.

Another approach is to hash at every granularity, creating verification trees for each granularity. With this approach the granularity should always match the size of the query returned. However, it comes at the cost of much more verification data being stored, and a higher initial processing cost.

An alternative to uniform database hashing is variable database hashing. The difference is focused on how the hashed/signed portions are created. In more detail, granularity is adjusted based on database usage habits, rather than pre-set. Effective use could be in keeping the less used parts of a large database at a comparably coarse granularity. The wasted data retrieval is offset by the decreased verification data storage/sorting. On the other hand, heavily accessed portions of a database should have a granularity that most closely resembles the majority of query requests. In this case, the portion that is smaller than the granularity has to be weighed against the increased verification object size resulting from including a greater number of hashes.

In our approach, we previously utilized variable hash granularity to improve efficiency [11]. However, this approach is not applicable when utilizing encryption, and so, therefore, in this application we used uniform database hashing based on the encrypted record size.

4.3.2 Encryption of Health Information Records and Trapdoor Encryption of Index and Search Terms

Due to the sensitive nature of electronic public health data, and the need for data confidentiality, we consider that the implementation of any scheme requires, as a minimum, compatibility with data encryption (to address point 3 of the requirements). In our approach we allow encryption of the records, specifically in the implementation we utilize a AES encryption for record storage – this could be accomplished with one or many keys based on the data-owners and usage model. Key management would be applied through the application tiers/web services rather than at the data repository level. The downside of this encryption implementation is the limitation of the types of queries that can be performed.

For the electronic health information stored within these data repositories to be usable, a form of searchable encryption needs to be provided (to address point 1 of the requirements). There have been a number of different approaches suggested in previous work [17, 18, 19]. In our approach we use a simple form of searchable symmetrical encryption using a trapdoor of the search keyword or index:

$T(w, s_w)$: Given a keyword w and the secret key s_w .

When the data is stored or updated or additional keywords/indexes are created, the $T(w, s_w)$ is then stored in the index trees as the *Index_value*. When a query is attempted, the client encrypts the keyword, sends it to the server and the server processes the query by matching the encrypted keyword to the one previously stored. Therefore, the server can accurately answer the query without having knowledge of the search keyword or the contents of the record. The specific type of searchable encryption utilized in this chapter, is limited in the types of queries that can be performed, i.e. queries and hence query assurance is limited to exact match keyword/index based queries.

4.3.3 Scalability

The size of the returned verification object will have the largest effect on efficiency and scalability of querying operations. The other components of the verification module being:

1. Hashing the retrieved data $O(1)$,
2. Verifying the digital signature and timestamp $O(1)$,
3. Searching the balanced verification tree and retrieving results – similar complexity to the size of the verification object, but less computationally expensive. Specifically, the depth of the verification tree is decisive in the number of hashing and comparison operations that need to occur. As such, the hash tree size can be evaluated as a min/max value (1) based on the whether the leaves of the tree are full (f entries) or in the worst case all half full ($f/2$ entries).

$$Leaf_{min} = \lceil n/f \rceil, \quad Leaf_{max} = \lceil 2 * n/f \rceil \quad (1)$$

Based on these calculations we can evaluate the height of the tree (2) and (3).

$$Height_{min} = \lceil \log_f Leaf_{min} \rceil + 1 \quad (2)$$

$$Height_{max} = \lceil \log_{f/2} Leaf_{max} \rceil + 1 \quad (3)$$

So for a given n number of elements in the verification tree, there is an associated height based on the number of leaves per branch f . This using the max height from (3) gives a $O(\log_{f/2}(2 * n/f))$ value for verification, where that number of hashes that will have to be calculated and compared in addition to $O(1)$ digital signatures operations. Additionally, the efficiency improvements aim to decrease the height of the tree that needs to be retrieved and evaluated, however as this is a dynamic process (and heavily reliant on usage patterns), it is only evaluated in the implementation.

4.3.4 Summary

In summary, this section detailed the methodology components that are utilized to provide the required functionality as defined in the problem definition section. Data confidentiality is addressed through the utilization of record encryption and searchable encryption. Query assurance addressed through the use of an authenticated query assurance approach utilizing variable timestamp values and placements. Efficient and secure storage addressed via our evaluation of the scalability of the approach and will be further evaluated in the following experiment and evaluation section.

4.4 Experiment and Evaluation

Our aim is to show the initial compatibility of authenticated query assurance combined with searchable encryption in a single model, given its strong benefits for integrated public health data repositories. In order to investigate the feasibility of our approach, our implementation uses relevant technologies and methods, and then measures the resulting metrics relating to efficiency - in this case CPU time and data overhead. For a data format, we chose to use XML records for our text based component of our data feasibility approach. As XML files are a standardized format for the transfer of structured information between data sources, acting as an interoperability layer that will likely be required between individual public health information systems and HPSN mobile applications. Additionally, to simulate the storage and retrieval of multimedia public health interventions, we utilized the JSRT [20] digital image database. To provide breadth of database storage examples, we utilized Apache Cassandra for the XML records due to its highly scalable nature and low overheads, while the multimedia files were stored in a Microsoft SQL Server 2012 Enterprise edition database.

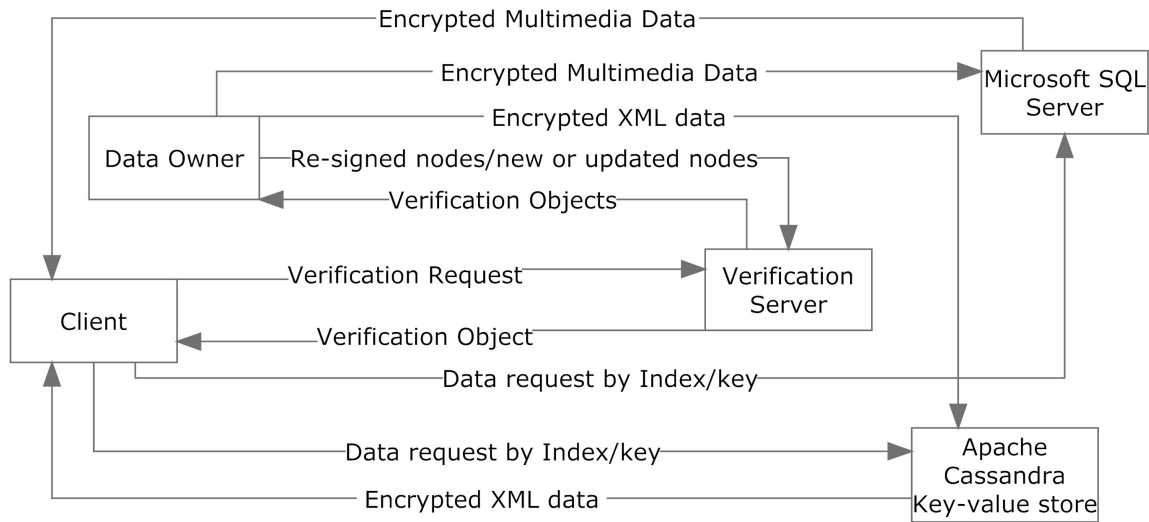


Fig. 4.5 Experiment Architecture

4.4.1 Experiment Setup

The experiment setup was conducted on a x3 AMD 710 with 3GB RAM. The client and all server implementations (verification, SQL and Cassandra) were run on the single machine. The architecture used for the implementation is modeled in Figure 4.5. A key-value store was used to store the encrypted XML records, in this case an Apache Cassandra database running as a single node. A more typical SQL database (Microsoft SQL Server) was used to store the encrypted multimedia data. The client and verification server were developed in Java with 1.6.0 runtime library. The verification tree nodes were a custom class built on top of the DefaultMutableTreeNode Java library. The tree follows a structure similar to that of a B+ tree - that is, the tree is self-balancing and all records are stored in leaf nodes. The experiment utilizes a single verification tree for both types of records; XML files and multimedia data are hashed at the document level and inserted based on trapdoor encrypted search term.

To perform the encryption of the data 128-bit AES specification was used. While a combination of RSA encryption with a 1024 key length and SHA1 were used to provide the hashing and digital signing of the verification tree.

To provide a reasonable affinity to public health data in this experiment, we used a collection of XML documents and a collection of high resolution digital images. The XML records were created with example data. In total 100000 XML records were created and encrypted then stored in a key-value database: Apache Cassandra. The approximate size of the data stored in XML format is 1.9 GB. Before storing in the key-value database, the XML records are encrypted using 128-bit AES, encoded into base64 then stored between CDATA tags in XML format to avoid issues with unparsable XML records. The images were stored in a Microsoft SQL Server 2012 database, in a blob (binary large object) format. The use of a blob format allows for file storage with additional metadata that can also take advantage of the security and reliability features of SQL databases including ACID (Atomicity, Consistency, Isolation and Durability) properties. The JSRT image database is composed of 247 images. To further test the capabilities of this approach, we duplicated the images to create 2000 total stored images by appending unique header data prior to encryption, so that once stored in the SQL database the records would appear unique. The storage process involved reading the image data from each file, appending additional header information, followed then by encrypting the binary data using 128-bit AES encryption and storing in the database table.

Due to the nature of storing the XML records and multimedia data in an encrypted format, only basic query functionality is available - in this case, index exact match and path based queries. If required, the verification tree keyword encryption could be removed to allow greater flexibility in search indexes/values.

4.5 Results

In our implementation, we measure the metrics of method execution time and data overhead. In this section we will present our results. These metrics were chosen for this initial evaluation as the goal is to demonstrate the feasibility and functioning of the novel query

assurance method and also to demonstrate that there is not a significant execution time and data overhead hit resulting from this approach, which at the same time has desirable properties for secure public health information systems. The results are organized into XML and SQL queries, though the two data repositories share the same verification tree. Additionally, for each dataset three blocks of queries were performed:

1. Initial block of 500000 XML data queries and 50000 SQL queries.
2. Repeat block of queries that perform the same set of queries as in the initial block after the maintenance process had completed. This is used to assess the effectiveness of the maintenance optimization in ideal circumstances.
3. Random block of queries that perform a random set of 50000 SQL data queries and 500000 XML data queries. This is used to demonstrate the effectiveness of optimization even where the past usage trends are not indicative of future queries.

4.5.1 Computation Time

As shown in Fig. 4.6, the average SQL query time compared to the verification time was similar. As such, the verification component of the over 50000 queries was not a major issue, and if completed concurrently could be further minimized. However, the average verification time was higher than the XML data – despite operating from the same verification tree. This is largely due to the increased hashing time related to the large file size of the multimedia data – as after SQL data retrieval, it is hashed and compared to the retrieved verification data. This also resulted in the maintenance optimization not being significant in improving CPU times – as the majority of CPU time was spent in large hash operations and SQL data retrieval.

We measured verification as being 54.4% of the total query time after maintenance optimization for the repeat data set, 54.4% for the random data set and 54.2% prior to opti-

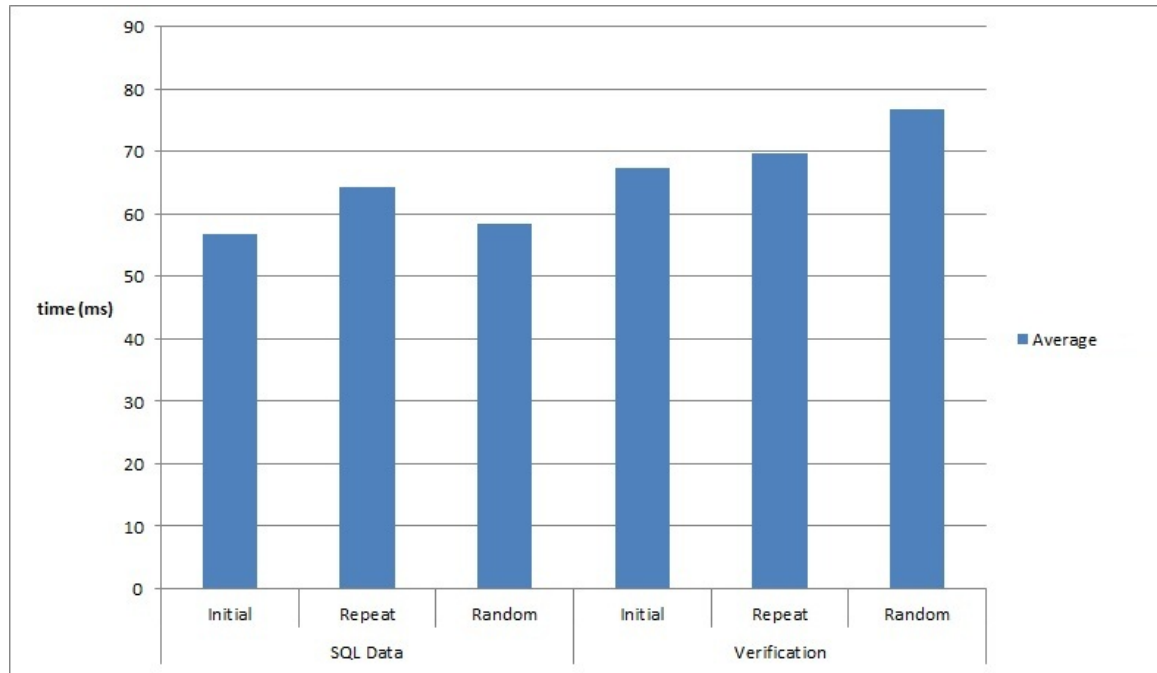


Fig. 4.6 SQL Query Average CPU Time

mization. The average query time over the 50000 queries performed during the pre and post maintenance repeat and random rounds was 124ms, 141ms and 128ms respectively.

Alternatively, the XML query sets as shown in Fig. 4.7 had relatively higher verification time compared to Cassandra data retrieval. This relates to the comparable size of data. In general the query run did not show major improvement in average CPU time pre and post maintenance.

We measured verification as being 47.7% of the total query time after maintenance optimization for the repeat data set, 52.88% for the random data set and 55.88% prior to optimization. The average query time over the 500000 queries performed during the pre and post maintenance repeat and random rounds was 1ms , .92ms and .97ms respectively.

Conversely, the verification tree initialization time was quite high. This is due to the verification tree being built incrementally, as would often occur in a database as additional data is stored. On average an insert in the verification tree took 28.99ms. This time is largely accounted for by the branch node splits and re-balancing that occurs across the verification

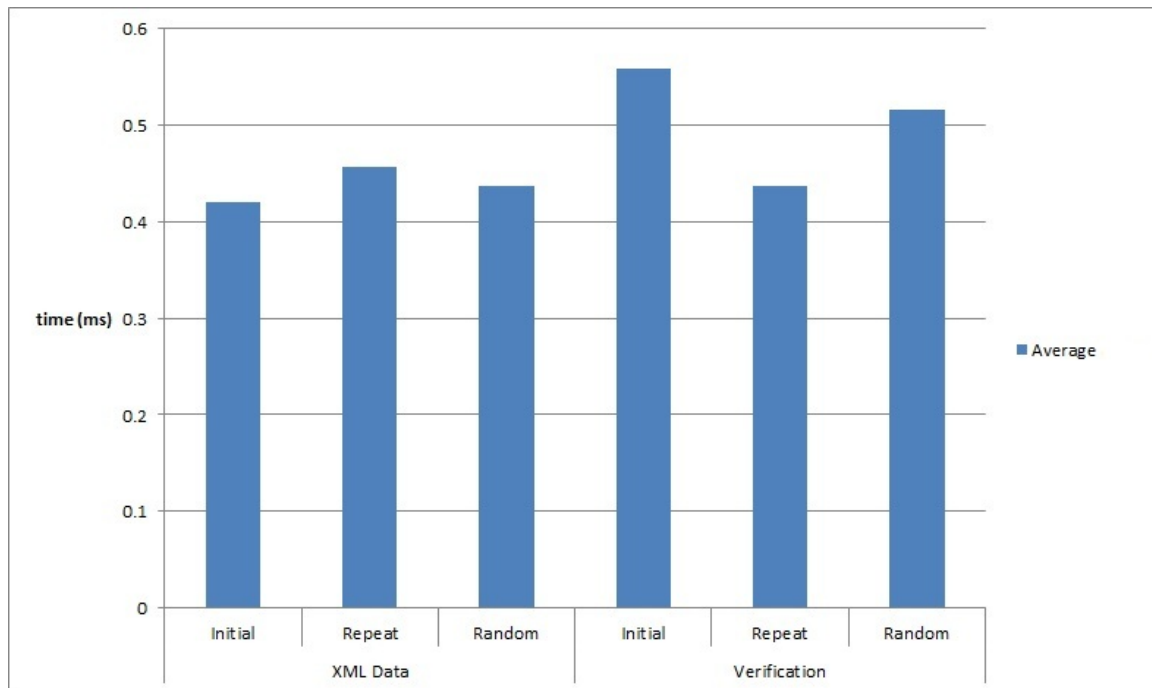


Fig. 4.7 XML Query Average CPU Time

tree. An unbalanced tree would ultimately be more efficient for inserts but less efficient for queries.

Additionally, maintenance was a significant cause of CPU overhead, with an average time of 25ms per verification leaf node. This is expected as the verification process includes large changes to the verification tree including re-hashing and re-signing of nodes. The maintenance operation could be further tweaked to provide a trade-off between data-owner CPU time and improvements in client efficiency.

4.5.2 Query Data Overhead

As shown in Fig. 4.8, the verification data required to authenticate the retrieved SQL data is minimal in relation to the SQL multimedia data. Approximately 0.018% of all data retrieved for the SQL queries was verification data pre-maintenance with that value dropping to 0.011% for post-maintenance random queries and 0.005% for post-maintenance repeat queries.

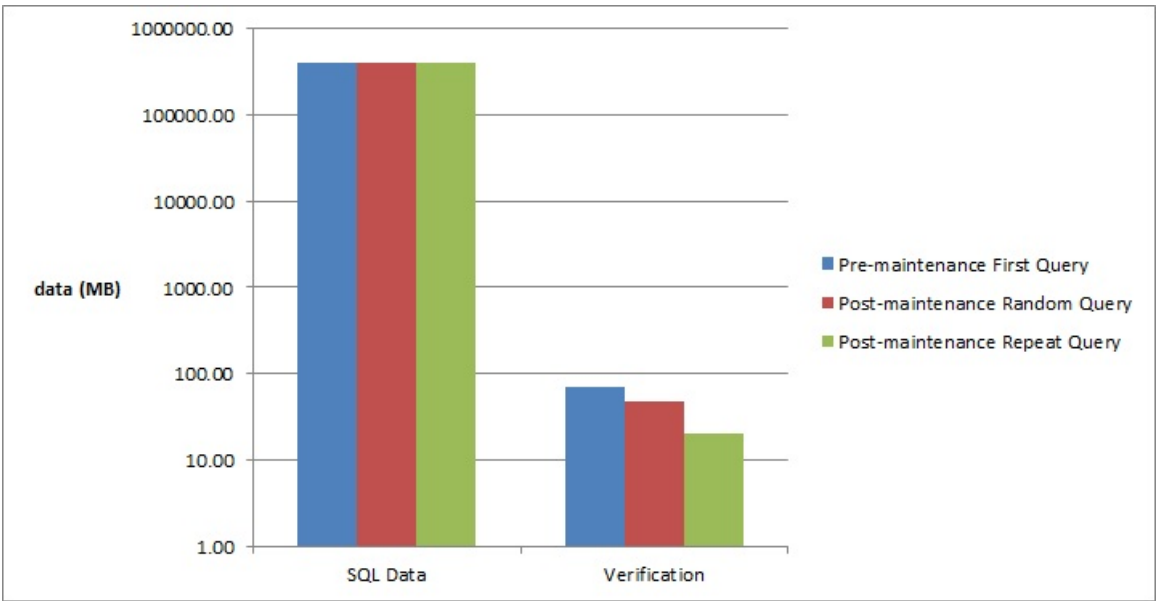


Fig. 4.8 SQL Data and Verification Overhead

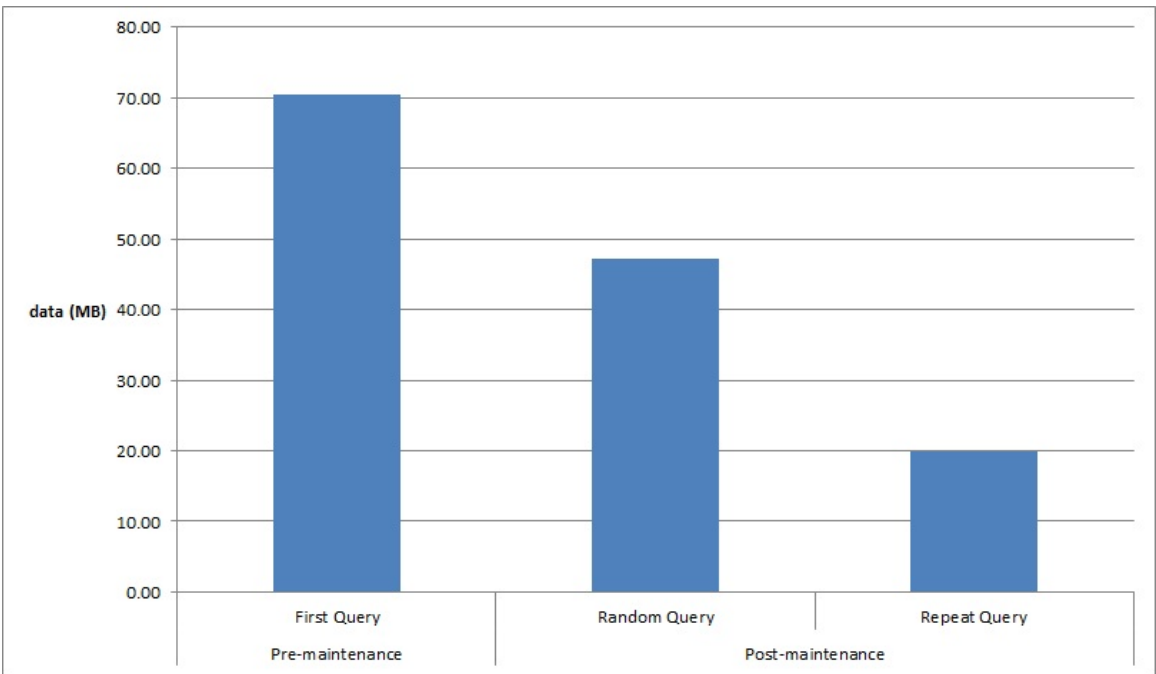


Fig. 4.9 SQL Verification Overhead Detail

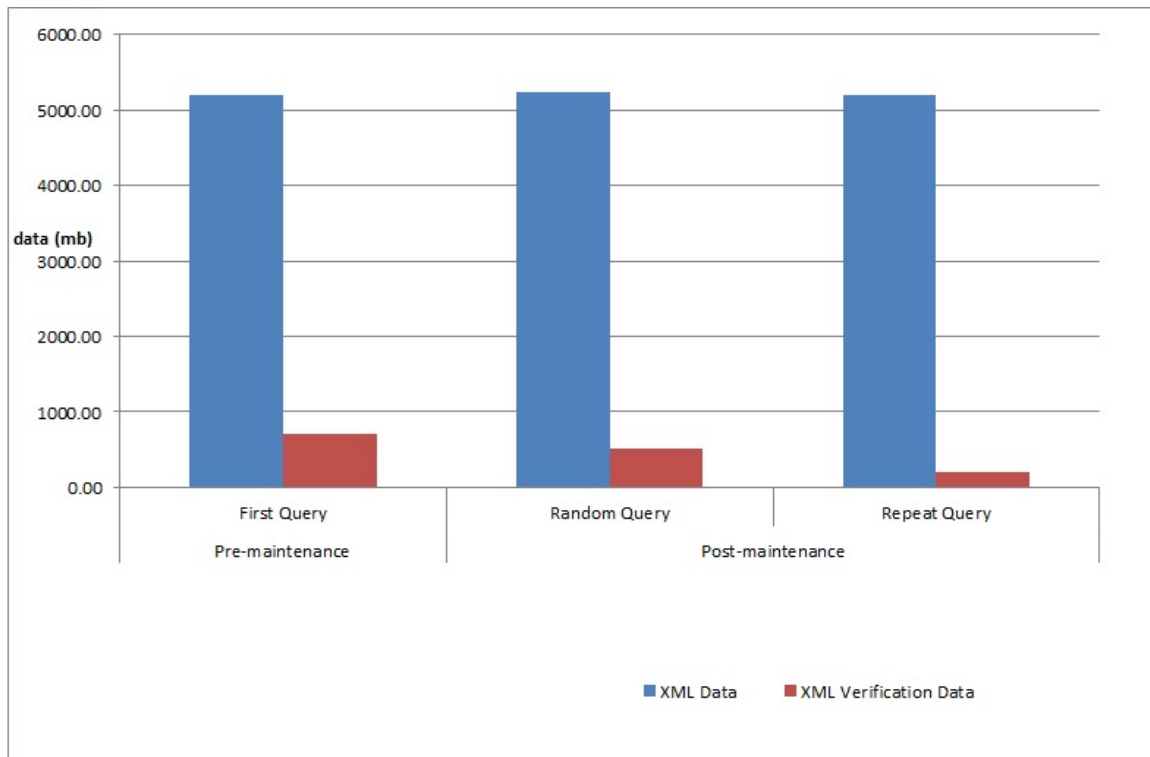


Fig. 4.10 XML Data and Verification Overhead

However, just comparing the verification data amount there was a significant decrease brought about by the maintenance process as shown in Fig. 4.9. Verification data was 72% of pre-maintenance levels even on the random query set, with the repeat query set resulting in 40% of previous values over 50000 queries performed in each set.

The XML data results were similarly positive as shown in Fig. 4.10. Verification data overhead was a much more significant component of overall data retrieved. With verification data accounting for 11.95% of pre-maintenance data, 8.89% of random and 3.75% repeat post-maintenance data. Verification data was 74% of pre-maintenance levels even on the random query set, with the repeat query set resulting in 42% of previous values over 50000 queries performed in each set.

Overall, there was a significant decrease in verification data required post-maintenance, with the most significant efficiencies gained where optimization can be based on repeating historic trends.

4.5.3 Results Discussion

Overall, the results show that the overhead cost of applying an authenticated query assurance model with searchable encryption to public health data structures is not detrimentally high.

The data sets used for the implementation were XML records and multimedia data. This provided a range of data types relevant to public health from the large file components (multimedia) to the smaller XML files that could be used for data submission/public health messaging dissemination. Additionally, our previous work [11] implemented a similar query assurance methodology over a greater range of record sizes, and in that instance data overhead was at acceptable levels and still significantly decreased by the maintenance process.

The results show that query verification can be feasibly applied to public health information data, with the actual time/data overhead being reasonable for querying against the data set used in our implementation. Further, the application of optimization to the verification tree can have positive effects to both computational time and data overhead, even where there is no query trend apparent.

In comparison, the insert time for verification nodes remains expensive and in some cases the verification tree may need to be optimized to reduce it. Increasing the number of child nodes per branch or requiring less strict tree balancing may have a positive outcome if this is required. Ultimately, due to the characteristics of this approach, its feasibility for the performance requirements of a particular system would need to be considered. However, we considered that it performs within acceptable levels of overheads in most cases.

Additionally, though we found that verification of larger multimedia files required additional verification time. However, compared to the additional time of data retrieval of large files the verification overhead was not overly significant.

We found the maintenance operation to be quite time consuming on a similar scale to the initial verification tree initialization, when applied to a dataset of this size with the

high number of queries performed against it. Any implementation of this type of approach would need to consider the improvement through optimization against the data-owners own computation time in performing the maintenance.

4.6 Conclusions and Future Work

This work explored the area of providing query assurance and security over encrypted public health information stored in shared or distributed databases. This will become an increasingly important capability for assured records as electronic health records constituted from multiple underlying repositories, including participant repositories in HPSNs become more extensively used and increasingly essential. We investigated whether a high level of query assurance could be feasibly provided in these conditions while maintaining reasonable overheads.

A test implementation was developed to measure the feasibility and efficiency of our approach, with specific consideration given to time and data overheads. We found that the overheads were not a major component of providing this type of implementation. Further, it was shown that our verification tree optimization technique was effective in reducing time overheads in our XML dataset and data verification overhead in both XML and multimedia datasets. However, some parts of the implementation continue to be moderately expensive: tree initialization, maintenance and insertion.

The use of a keyword trapdoor allowed the implementation to perform basic queries on the encrypted data, while preserving the confidentiality of both the keyword and stored data.

The implementation provides an authenticated level of query assurance. Additionally, as the records are encrypted and stored in an easily transmissible format, the approach is applicable to a range of different public health record types, such as data submissions or public health interventions.

REFERENCES

- [1] Andrew Clarke and Robert Steele. “Secure query assurance approach for distributed health records”. In: *Health Systems* 3.1 (Feb. 2014), pp. 60–73. ISSN: 2047-6965.
- [2] Reinhold Haux. “Health information systems—past, present, future”. In: *International journal of medical informatics* 75.3 (2006), pp. 268–281.
- [3] Tran Khanh Dang. “Security issues in outsourced xml databases”. In: *Open and Novel Issues in Xml Database Applications: Future Directions and Advanced Technologies* 1 (2009), pp. 231–261.
- [4] Andrew Clarke and Robert Steele. “Secure and Reliable Distributed Health Records: Achieving Query Assurance across Repositories of Encrypted Health Data”. In: *System Science (HICSS), 2012 45th Hawaii International Conference on*. Jan. 2012, pp. 3021–3029.
- [5] Feifei Li et al. “Dynamic Authenticated Index Structures for Outsourced Databases”. In: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. SIGMOD '06. Chicago, IL, USA: ACM, 2006, pp. 121–132. ISBN: 1-59593-434-0.
- [6] Pengtao Liu. “Query Assurance Verification for Outsourced Database”. English. In: *Advances in Future Computer and Control Systems*. Ed. by David Jin and Sally Lin. Vol. 160. Advances in Intelligent and Soft Computing. Springer Berlin Heidelberg, 2012, pp. 39–44. ISBN: 978-3-642-29389-4.

-
- [7] Tran Khanh Dang. “Ensuring Correctness, Completeness, and Freshness for Outsourced Tree-Indexed Data”. In: *Inf. Resour. Manage. J.* 21.1 (Jan. 2008), pp. 59–76. ISSN: 1040-1628.
 - [8] Haixun Wang et al. “Dual Encryption for Query Integrity Assurance”. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08*. Napa Valley, California, USA: ACM, 2008, pp. 863–872. ISBN: 978-1-59593-991-3.
 - [9] Einar Mykletun, Maithili Narasimha, and Gene Tsudik. “Authentication and Integrity in Outsourced Databases”. In: *Trans. Storage* 2.2 (May 2006), pp. 107–138. ISSN: 1553-3077.
 - [10] Viet Hung Nguyen and Tran Khanh Dang. “A Novel Solution to Query Assurance Verification for Dynamic Outsourced XML Databases”. In: *Journal of Software* 3.4 (2008).
 - [11] Andrew Clarke and Eric Pardede. “Outsourced XML Database: Query Assurance Optimization”. In: *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*. Apr. 2010, pp. 1181–1188.
 - [12] Maithili Narasimha and Gene Tsudik. “Authentication of Outsourced Databases Using Signature Aggregation and Chaining”. English. In: *Database Systems for Advanced Applications*. Ed. by Mong Li Lee, Kian-Lee Tan, and Vilas Wuwongse. Vol. 3882. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, pp. 420–436. ISBN: 978-3-540-33337-1.
 - [13] Min Xie et al. “Integrity Auditing of Outsourced Data”. In: *Proceedings of the 33rd International Conference on Very Large Data Bases. VLDB '07*. Vienna, Austria: VLDB Endowment, 2007, pp. 782–793. ISBN: 978-1-59593-649-3.

- [14] Premkumar T. Devanbu et al. “Authentic Third-party Data Publication”. In: *Proceedings of the IFIP TC11/WG11.3 Fourteenth Annual Working Conference on Database Security: Data and Application Security, Development and Directions*. Deventer, The Netherlands, The Netherlands: Kluwer, B.V., 2001, pp. 101–112. ISBN: 0-7923-7514-9.
- [15] Einar Mykletun, Maithili Narasimha, and Gene Tsudik. “Providing authentication and integrity in outsourced databases using Merkle hash trees”. In: *UCI-SCONCE Technical Report* (2003).
- [16] Kyriakos Mouratidis, Dimitris Sacharidis, and Hweehwa Pang. “Partially Materialized Digest Scheme: An Efficient Verification Method for Outsourced Databases”. In: *The VLDB Journal* 18.1 (Jan. 2009), pp. 363–381. ISSN: 1066-8888.
- [17] Ik Rae Jeong et al. “Searchable Encryption with *Keyword-Recoverability*”. In: *IEICE Transactions* 92-D.5 (2009), pp. 1200–1203.
- [18] Hyun Sook Rhee et al. “Trapdoor security in a searchable public-key encryption scheme with a designated tester”. In: *Journal of Systems and Software* 83.5 (2010), pp. 763–771.
- [19] Peter van Liesdonk et al. “Computationally Efficient Searchable Symmetric Encryption”. In: *Secure Data Management, 7th VLDB Workshop, SDM 2010, Singapore, September 17, 2010. Proceedings*. 2010, pp. 87–100.
- [20] Junji Shiraishi et al. “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules”. In: *American Journal of Roentgenology* 174.1 (2000), pp. 71–74.

**Faculty of Health Sciences
Author Contribution Statement**

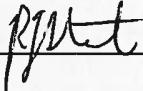
Candidate Name: Andrew Clarke

Degree Title: Doctor of Philosophy

Paper Title: Secure Query Assurance Approach for Distributed Health Records

As the corresponding author of the above paper, I confirm that the above candidate has made contributions to the following:

- Conception and design of the research
- Development of the prototype implementation
- Analysis and interpretation of the findings
- Writing the paper and critical appraisal of content

Signed  Name Prof Robert Steele Date 10th March 15

5 PRIVACY THRESHOLD APPROACH TO HPSN DATA AGGREGATION AND COLLECTION

Preamble

This chapter extends on a journal paper that was accepted in the Journal of the Association for Information Science and Technology [1]. It has been included as a chapter of this thesis with some additional detail added to subsection 5.5.1 and an additional subsection added 5.5.4 in addition to minor formatting changes to align with the thesis format. As part of preparing the paper for the thesis format the related work section has been moved to Chapter 2 . The previous chapter was focused on the implementation/evaluation of the distribution components of the system. This chapter is the second implementation/evaluation chapter and focuses on a local processing privacy threshold approach to public health data aggregation : expanding on the architecture introduced in chapter 3 detailing the major components of the systems. Further, additional detail on the types of sensors and sensor capabilities currently available/being developed that could be used as part of a health participatory sensing network are also discussed.

This chapter's technical and implementation/evaluation component is focused on a privacy threshold approach to public health data aggregation. This provides a specific approach to the requirement first detailed in 3.3.3. This approach uses a quasi-identifier score (QIS)

and multiple tiered thresholds to ascertain a specific level of data contribution and reduce privacy risks. This process is discussed at length with an algorithm provided. Additionally, further discussion of the types of data submission components and methods to ascertain their relative QIS value is presented.

These components are then utilized to build and evaluate a prototype implementation of a threshold privacy approach.

ABSTRACT

The pervasive availability and increasingly sophisticated functionalities of smartphones and their connected external sensors or wearable devices can provide a new data collection capability relevant to public health. Current research and commercial efforts have concentrated on sensor-based collection of health data for personal fitness and personal healthcare feedback purposes. However, to date there has not been a detailed investigation of how such smartphones and sensors can be utilized for public health data collection purposes.

Unlike most sensing applications, in the case of public health, capturing comprehensive and detailed data is not a necessity, but rather aggregate data alone is in many cases sufficient for public health purposes. As such, public health data has the characteristic of being capturable whilst still not infringing privacy, as the full detailed data of individuals that may allow re-identification is not needed, but rather only aggregate, de-identified and non-unique data for an individual. For example, rather than details of physical activity including specific route, just total caloric burn over a week or month could be submitted, thereby not identifying the individual.

In this chapter we introduce, prototype and evaluate a new type of public health information system to provide aggregate population health data capture and public health informational or behavioral intervention capabilities via utilizing smartphone and sensor capabilities, whilst fully maintaining the anonymity and privacy of each individual. We consider in particular the key aspects of privacy, anonymity and intervention capabilities of these emerging systems and carry out a detailed evaluation of anonymity preservation

characteristics.

5.1 Introduction

The recent rapid growth in both the capabilities and uptake of smartphones, suitable to act as health sensor platforms, has the potential to advance public health data collection and intervention in significant ways. Whilst increasingly research and development is concentrating on how mobile devices and sensors can be used as a tool for individual health data capture and feedback, this has not extended into investigation of how these devices can be used for public health data capture. Interestingly, the case for public health usage doesn't require the same level of precise data that would often be required in participatory sensing [2] applications in other domains. 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.

This chapter is a significantly extended version of a previous conference paper [3]. In particular this chapter differs in that it analyzes these novel smartphone-based public health information systems as a generic new type of system. It describes the results from building a significant prototype system and carries out a substantially more detailed privacy and anonymity analysis. We describe a class of smartphone-based information systems for anonymized public health data capture and intervention. Interventions [4], in this work are in the form of informational or behavior-related messages sent to an individual's smartphone, intended to create a health-related behavioral change, and are a key component of future Health Participatory Sensing Networks (HPSNs). In particular, as we later describe, a significant new capability enabled by these systems is that 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.

The introduced system eschews the need for a fully trusted central server, which might

prove impractical or a significant privacy risk on population-scale applications. Instead adopting an architecture which has a central aggregation server in communication with the end-user mobile devices, only via an intervening anonymizing layer, and uses local processing on each mobile device to ensure non-re-identifiability of the user from their submitted sensor data.

The anonymous communications layer could utilize onion routing [5] or mix networks [6] which are techniques for anonymous communication over a computer network using multiple intermediate nodes and encryption to protect privacy – these networks make it hard to trace the source and destination of an end-to-end communication. The system uses an anonymizing layer in combination with de-identification of data submitted, such that the content of the data submitted does not identify an individual, thus allowing anonymous submission/interaction between the participant and the HPSN. Beyond de-identification, the approach also addresses the risk of re-identification based on quasi-identifiers, such as information known about individuals outside the HPSN that could potentially be used to match with and re-identify the submitting individual. The conventional approach to addressing 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 [7] is assured for a data set, such that any individual is indiscernible from k other records based on quasi-identifiers.

However the type of public health information system introduced in this chapter, instead performs de-identification without a trusted aggregator or server, which significantly reduces privacy risks as there is no central point where sensitive information is stored that can itself pose a privacy threat to participants, or become a site for security lapses or target of malicious activity. Rather, anonymity and non-re-identifiability 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 desired public health data collection purposes. This is achieved in the system by utilizing quasi-identifier scores (QISs) as a quantified measurement of

approximate risk of possible re-identification and thereby enables a threshold approach to privacy limits. The threshold approach allows for automated, on-device calculation of the quantified privacy/re-identification risk of submitting various levels of detailed health sensor information in terms of QISs. The threshold approach supports the use of standard or default thresholds in terms of QISs as well as modification of thresholds on an individual's device to suit the preferences of a given individual. This allows the level of privacy disclosure an individual agrees to, to be managed without requiring a case-by-case approval.

5.2 Sensor Capabilities and Public Health Measures

In this section, we describe how data relevant to many public health measures can already be captured automatically via current or commercially available sensor capabilities. By measures we mean items that are indicators of health or healthy lifestyle or disease risk or disease. That is, the use of sensors for public health relevant data is not a speculative proposal, as many current commercial sensors already provide relevant functionalities.

We discuss this by first considering various current commercial sensor capabilities and then matching these to various accepted health risk factors, such as physical activity, blood pressure, blood glucose levels, weight etc.

5.2.1 Sensor Capabilities

The proliferation of commercial fitness and health sensors provides new mechanisms for population health data capture, even though these are currently targeted for use in relation to an individual's health and fitness. Commercially available sensors are also already able to capture many biomedical measures collected in public health data surveys. Such sensors include wearable patches, stretchable electronic tattoos, smartwatches, other wearables and implantable sensors along with the more widely deployed smartphones and connected sen-

sors. In addition, such public health data capture would have a number of characteristics quite distinct from traditional survey-based public health data capture approaches. These include:

- Being real-time/ near real-time
- Larger participant numbers/ proportion of population
- More detailed data
- Captured electronically
- Direct measurement, not human response
- Anonymized, as we discuss in this work

The area of personal health sensor and software development and commercialization [8] is currently a highly active area. This is possibly due to the relevance of these individual sensors to both the rapidly developing smartphone market and technologies, and the increasing interest to leverage such technologies for personal wellness, fitness, health and healthcare purposes [9, 10].

Fitness and Physical Activity Sensors

Commercial implementations such as Nike Fuel and Jawbone Up demonstrate the potential for, and achievability of continuous physical activity sensing. Jawbone Up extends beyond physical activity monitoring to include sleep patterns, sleep quality, and a nutritional diary. Other well-known examples of such wearable sensors include; FitBit, RunKeeper, myFitnessPal, Pebble Watch, the Basis Watch and Google Glass. Such fitness and health sensors are the most contemporarily available type of sensor that can be utilized for public health purposes, because such sensors are already achieving widespread interest and a great level of mass adoption.

Also of significant relevance is Google Now's, Activity Summary [11] which automatically provides a monthly estimate of how far an individual has walked and cycled, and comes as part of Google's Android mobile operating system – hence is already extremely widely deployed.

Vital Signs Sensors

Smartwatches such as the Mio Active are able to capture heart rate, the Amiigo wristband captures blood oxygen levels, Somaxis provides ECG and EMG sensors and the mc10 stretchable electronic tattoo can transmit heart rate and brain activity [8]. The capturing of vital signs is often more beneficial for individual health care, but it also adds new capabilities for public health data systems. Another example, the Sense A/S monitoring patch is able to measure blood pressure [8].

Blood Constituent Sensors

Increasingly, there are wireless-enabled patch technologies emerging that may be able to capture the levels of some blood constituents. Examples include the forthcoming Sano Intelligence [12] wearable patch which is touted to allow the capture of blood glucose and potassium levels, with further blood constituent capture planned for the future. Numerous continuous blood glucose monitoring systems are also currently available, mainly targeted for the management of diabetes.

Such sensor capabilities in a cheap and accurate form, have the potential to revolutionize individual health care, early detection and preventative health; and by extension also public health. That is, because such capabilities may be so beneficial in terms of individual health monitoring, health maintenance and early detection, that they could achieve wide adoption. If so, their possible role in public health data capture can also be proportionately significant.

Ambient sensors

Other initiatives such as Riderlog [13] and the Copenhagen Wheel [14] are moving towards capturing physical activity levels, and at the same time, additional contextual and environmental data. The Copenhagen wheel goes beyond physical activity sensing, to urban environment monitoring with air quality and noise sensors included in the implementation to provide additional data beyond just the activity of the individual.

5.2.2 Public Health Risk Factors

The various types of health data that can be collected via the above-mentioned sensors, already relate to a majority of public health measures:

- **Physical Activity Levels** – This is one of the most important lifestyle factors for chronic health conditions, other health risks and health in general [15]. This can now be quite accurately captured with already commercially available sensors and even via in-built smartphone capabilities alone [11].
- **Caloric Burn and Caloric Intake** – Caloric burn information can be captured by a range of activity sensors as described, and caloric intake can also be increasingly automatically captured [16].
- **Nutritional Data** – As mentioned, wearable patches have the ability to measure potassium levels, one of the markers of nutrition status [17].
- **Blood Pressure** – Blood pressure is a public health marker of cardiovascular disease [17] which is one of the most significant morbidity and mortality risks. As described, blood pressure can be captured via a wearable patch such as the Sense A/S amongst numerous others.

- Blood Glucose – A marker of diabetes [17] can be captured by wearable patches and other continuous glucose monitoring (CGM) devices. Recently the use of wirelessly connected contact lenses for measuring blood glucose levels from the surface of the eye has also been described [18].
- Body Mass Index (BMI) – Height is roughly invariant for adults and Bluetooth-enabled scales are increasingly available to capture weight.
- Body Fat Percentage and Lean Mass – Consumer grade scales and other measurement devices include body fat percentage and lean mass and wireless-enabled scales are increasingly available to capture these details.
- Sleep Pattern and Regularity – Sleep patterns are both an indicator and a preventative/risk factor for a number of conditions. Sleep quality can be captured by currently available commercial wristbands and other sensors.

5.3 Public Health Information System Architecture

The overall public health information system architecture (Figure 5.1) 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 (Public Health Groups) i.e. organizations involved in public health-related activities. The HPSN interfaces with Public Health Groups, which could include state or federal health departments, public health research institutions or other public health organizations.

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

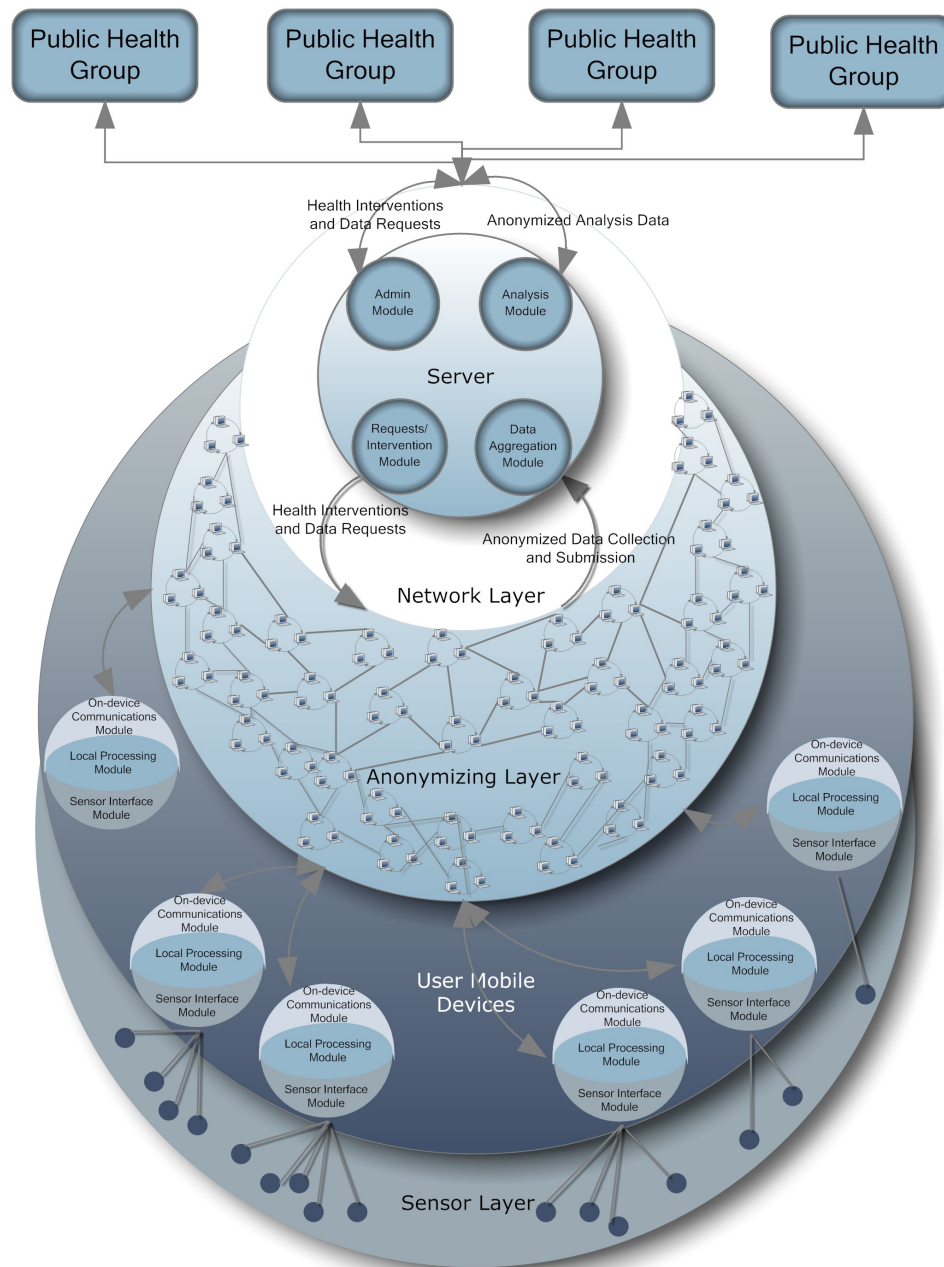


Fig. 5.1 Public Health Information System Architecture

data collection submissions are sent to the HPSS. The core functionality components of the HPSS are (1) Data Aggregation, (2) Analysis, (3) Intervention/Data Requests and (4) Administration.

The fundamental architecture can support different levels of both data collection and optionally 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.

5.3.1 Smartphone With or Without External Sensors

This is the base-case of a user utilizing a smartphone with or without additional external sensors, where the user is not required to take additional actions to participate in the public health data capture. This configuration has the advantage in that it has the greatest level of existing hardware deployment and ease of adoption – that is, smartphones without additional external sensors are currently the most prolific smartphone deployment case, though external sensors are increasing in popularity. Various types of data can prove to be important public health or epidemiological data sources. An example would be physical activity tracking [19] which has become increasingly popular in recent years, as well as its potential secondary usage for smart cities, including use for public health and population data capture and urban planning and environmental monitoring as discussed in our previous work [20].

5.3.2 Intervention Capabilities

This configuration additionally provides inputs to the individual to alter the actions they would have taken whilst 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’ participa-

tory 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-impacting behavior and hence have a component equating to a public health intervention. As such, the instigation to carry out ‘active’ sensing activities essentially constitutes a public health intervention input. Additionally for public health purposes, this can allow for immediate and continuous feedback of the effectiveness of campaigns upon population groups and sub-groups – a powerful new capability. This can contribute to the further understanding of the effect of informational inputs on such health-related behavior change as exercise behavior change [21] and for many other public health-related behavioral change campaigns.

5.3.3 Extension via Manual Input

This configuration combines the potential sensing capabilities of smartphones and external sensors with additional ‘human-sensing’ capabilities, allowing for larger volumes of sensor-based data to be complemented with subjective human-generated data and feedback. Further, this configuration can be implemented with or without intervention capabilities. Even without the benefits of interventions, the motivation for contributing data could be self-monitoring or altruistic/ citizen-scientist contribution, with the combination allowing the additional capability of providing human feedback in regards to interventions.

This is implemented through the addition of context-sensitive micro-surveys that are requested to be filled by users and attached to relevant collected sensor data. This allows for both data that is difficult to record through sensors alone, such as the context or purpose of physical activity (work, transport or recreation) and in some cases, data that may have been missed perhaps due to not wearing the sensors/mobile device for a period of time, to be added to the overall collection.

5.4 System and Prototype Components

The public health information system includes four major components: the HPSS, network layer, anonymizing layer and user mobile device. These architectural components and details of their prototype characteristics are described here.

5.4.1 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 on the HPSN server addresses the sending of data requests or interventions 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 the public health information system incorporates submissions of variable resolution (that is submissions for the same public health data collection 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 submissions 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, which can be optional that is not required for submission or mandatory components that must be part of the submission (see ‘Data submission policies’ subsection). Where a com-

ponent 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 Public Health Groups from the received sensing data.

5.4.2 Network and Anonymizing Layers

The network layer supports communication between the HPSS 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 ‘Intervention capabilities’ subsection).

The Anonymizing Layer consists of a mix network [6] or onion network [5], 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 allowing identification or re-identification.

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

The primary limitation of anonymous submission is 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.

5.4.3 User Mobile Device

Software incorporating the following modules is present on the end user's mobile device. The user's mobile device can operate according to the different levels of configuration identified in the 'Public health information system architecture' section. This would depend upon end user choices such 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 Public Health Group 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 health 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 policies 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 as well as 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, in relation to the local processing module, the section ‘Privacy threshold approach to public health data aggregation’, 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 data collection. This module can make use of existing communications standards such as the ISO/IEEE 11073 Personal Health Data standard to carry out standardized interfacing with external sensors where such standards are adopted.

5.4.4 Public Health Groups

Multiple public health groups are able to utilize the same HPSS and HPSN.

An end user might subscribe to more than one public health group’s public health data collection or intervention policy. For example, an individual could subscribe to data collection by two public health groups: the Department of Health and an Active Transport Initiative. They could for example, also subscribe to receipt of micro-surveys from the Active Transport Initiative and health interventions from the Department of Health.

The subscribed to policies, micro-surveys and health interventions would be updated periodically by these Public Health Groups. This is controlled by the client checking for updates, with the maximum valid period of a set of policies set by an expiring timestamp, one referring to the distribution process (maximum time before the data should no longer be distributed) and a second referring to the expected validity period (maximum period before

the policy needs to be updated).

Throughout an individual's daily schedule, their mobile device would automatically collect physical activity data, and data relevant to each organization is submitted intermittently throughout the day, utilizing the privacy-preserving mechanisms of the HPSN. The HPSN maintains user anonymity and does not allow re-identification – the submitted data just provides valuable input when combined with the mass of other individual's data, for aggregate population health measures.

For example, the Department of Health could be interested in overall physical activity in a day with age bracket and coarse location or location type information also submitted. Alternatively, the Active Transport Initiative might only be concerned with physical activity related to transportation (e.g. walking/cycling commuting) with the additional data of age bracket and start 'coarse location' and end 'coarse location' submitted. As another example, infrequently the end user might be prompted to complete a micro-survey related to active transport. For example, this could be a 30 second survey asking for a ranking of the five most significant factors as to whether on a specific day the individual would cycle/walk to work. This micro-survey would be presented to an end-user based on the locally stored data relating to that individual's travel habits – although it would have been sent to a much larger group. The end user's preferences also restrict how often micro-surveys from individual organizations and the overall HPSN system can request micro-surveys.

For example based on personal preferences and previous trend data, the end-user could also be prompted by a health intervention from the Department of Health, suggesting the health benefits of cycling.

The end user might also be interested to track their own health-related data and could find this a benefit that also assists their motivation to participate in the HPSN. The data displayed to them for this purpose is kept securely on their own device under their personal control (assuming that the mobile device includes its own security/anti-virus measures). The

data is never exposed via transmission to the HPSN, but this detailed data can separately be fed into such systems as part of their portable Personal Health Record [22] - for this reason this data for self-use can be more detailed and can be viewed without aggregation processing having first occurred if desired.

5.5 Privacy Threshold Approach to Public Health Data Aggregation

The public health information system, by applying granular and modular restrictions upon data collection controlled by the user, reduces real privacy risks through high levels of user control of contribution and restrictions on data potentially usable for re-identification. Additionally, the use of a local processing approach [23] to data submission and health intervention 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. This approach to privacy thresholds encourages the request of summary, calculated, classified and grouped data, rather than individually specific raw data, that would be likely to pose a privacy risk through potentially allowing re-identification.

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 against which to evaluate the requests.

5.5.1 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 a specific zip code or 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 health-related value to be collected. A data collection can have multiple measures for comparison. Examples of possible population-wide anonymized health or wellness measures are discussed in our previous work [24] 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 contemporary sensing capabilities. Emerging wearable patches that may be able to capture some blood constituent information [25], 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. Although fine-grained location information would not be generally required for public health, some examples include places physical activity occurred as a location type (e.g. work, home, gym or parks), 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 time or interval of time in which a measure occurred. Rather than submitting the specific time of a measure, a time period in which the measured value occurred can be submitted, lowering the potential risk of re-identification. Additionally, 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.

5.5.2 Data Submission Policies

Data submission policies will have:

1. Mandatory data requirements – Typically a Measure value and high priority demographic dimensions. If this is not submittable without breaching an individual's privacy threshold the submission is not made for that individual.
2. Optional data requirements – Additional data components that can be submitted alongside the mandatory data requirements. To allow for the calculation of the highest level of data that can be submitted without breaching the threshold, the optional data components will be weighted by importance and whether a less specific data submission is acceptable for a data component as a secondary weighting.

An algorithm (see 'Algorithm for data collection policy processing' subsection) 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 to also be adjusted. 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.

Additionally, as shown in Figure 5.2 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

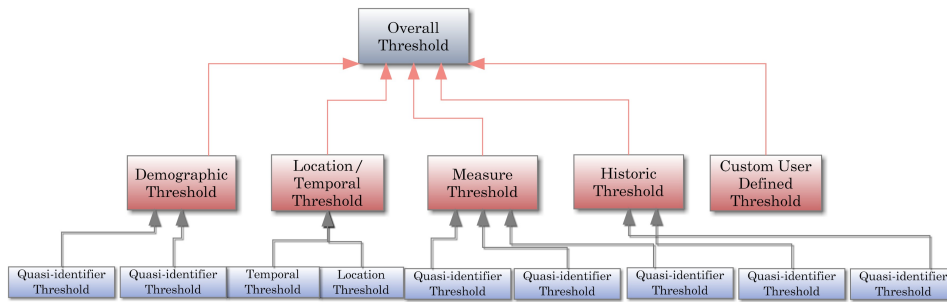


Fig. 5.2 Algorithm for Data Collection Policy Processing

‘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, such as not reporting on measures in certain locations, or time periods or combinations of data components that they would like to restrict in addition to the standard thresholds.

The data request is processed by the local processing module by adapting the data submission request to the anonymous submission settings on the local device. Firstly (Figure 5.3), 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.

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 5.3.

This provides for a personalized adjustment of the submission requirement to meet the

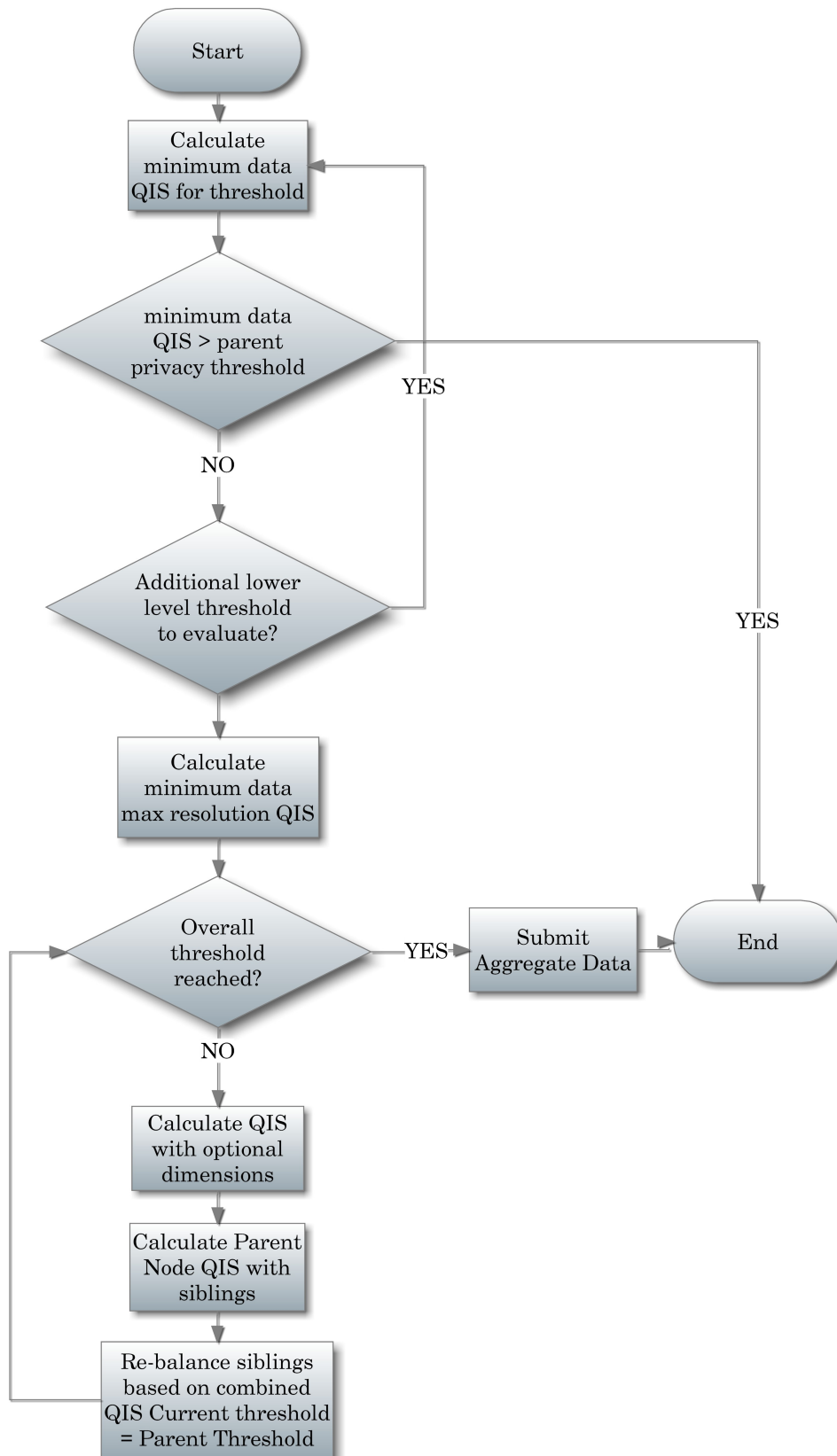


Fig. 5.3 Data Collection Rule Processing Algorithm

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.

5.6 Privacy

The goals of a HPSN are of a different nature than those of other such participatory sensing networks that for example provide noise maps [26] or air quality data [27]. Rather, the goal is to collect aggregate population-scale health data and deliver public health interventions.

5.6.1 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 rather than the distance and locations (or alternatively the distance in combination with conditions i.e. elevation, wind, pace intensity etc. could be submitted). This is because it is overall physical activity levels or alternately sedentary behavior levels that are of interest in relation to public health. By submitting only the caloric burn arising from 50km of cycling, the re-identification level can be shown to be close to zero. This is because the measure (caloric burn of 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) [28] use a bicycle for transport or recreation. Though this is more or less common based on particular demographic groups, with

that 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 all individuals, such as rarer demographics (see analysis in Section ‘Evaluation’).

An additional way that location data could be used in public health data collection is through location classification. That is, rather than submitting the coarse or fine grained location data, the type of location is instead submitted. This could incorporate reporting that a measure is linked to a work or home location (without revealing the location of either), or for example, for physical activity that it is linked to a park, gym, trail/track, urban street etc. Similar classifications would be possible for many types of measures collected for public health, and the privacy advantage of processing that classification locally rather than submitting sensitive data is significant. In the case of location classification, individual classifications are unlikely to be used to re-identify an individual. However, it would be reasonable to apply a threshold on the number of locations submitted to avoid potential exploitation. An example of such exploitation might be where an individual could be potentially re-identified due to having a unique set of locations for a given demographic set that can be correlated with known external data, perhaps from social media.

An example of adaptive local processing of location privacy is demonstrated in previous work [29]. However, as public health data submission does not require location data at all for many submissions, this means the 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. A simple calculation such as below could be used.

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

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

5.6.2 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.

Knowing more specific details of the time in which a measure occurred 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 (L) – 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 (G) – Is it possible to break the total period (and the associated measure) into its component parts and at what resolution.
- Start time (S) – 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 (E) – 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:

$$T_{QIS} = T_{calc}(S, E) + L / (1 - G) \quad (2)$$

In the above formula T_{calc} evaluates the start and end time of the data submission request, where the start and end date are related to common time periods e.g. day, week, month, quarter etc. a set value is used relative to the broad or narrow nature of the period. Where the start or end time is more targeted, an additional weighting is added to the most closely matching set value for a common period.

5.6.3 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 as 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 giving us good baseline data for demographic thresholds. 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 [28]. 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 μ is the individual demographic details.

$$D_{QIS} = 1 - Pr(\mu_1, \mu_2, \mu_3, \dots, \mu_n) \quad (3)$$

5.6.4 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, though unlikely 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.

$$M_{QIS} = \omega_A A_1 + \omega_B B_2 + \omega_C C_3 + \dots + \omega_D D_n \quad (4)$$

In the above formula A , B , C and D are individual measures and ω is the resolution for the measure. This reduces the above identified risk by limiting the number and detail of additional measures on a submission.

5.6.5 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, that is, assigning specific sensing ‘tasks’ to individuals. 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 program refinement. 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.

5.6.6 Overall Threshold

The overall threshold is calculated by combining the L_{QIS} , T_{QIS} , D_{QIS} and M_{QIS} in two stages, where there is a first stage threshold over L_{QIS} and T_{QIS} as there are close connections between location and temporal privacy and a second stage over all QIS values.

Stage 1:

$$\theta_{L/T} > \gamma_L L_{QIS} + \gamma_T T_{QIS} \quad (5)$$

Stage 2:

$$\theta_{LTDM} > \gamma_{LT} \theta_{L/T} + \gamma_D D_{QIS} + \gamma_M M_{QIS} \quad (6)$$

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

5.7 Evaluation

To demonstrate the operation of this approach we constructed a prototype system focusing on the local processing submission components. To achieve this, the prototype system generates a set of clients each with randomized demographics, measures, location and time records, and in the case of this evaluation 100,000 clients were generated. These clients then process a set of data submission requests which are submitted to the prototype server and evaluated for privacy considerations.

The prototype evaluation used population distributions from the Greater Sydney Metropolitan area to generate the individual client’s demographics including: age, gender, ethnicity,

income and education. The prototype client and server were 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 evaluation measures the approach's effectiveness at specific privacy levels in the following areas:

- Reduction of potentially re-identifiable unique demographic combinations.
- The proportion of submissions that met maximum submission detail including all optional dimensions.
- The proportion of dimensions/measures not submitted or submitted at a diminished precision.

The submission rule approach, allows for a great deal of specification in the types of flexibility the client has available, to adjust the data before submission. To demonstrate this our test submission includes all identified data components, including five demographic dimensions, three measure dimensions, temporal dimension and a location dimension using location type classification.

Thresholds are set for each of the data components, then executed on the local device based on the specific dataset. The dimensions are grouped into mandatory and optional components of the submission. As such, two demographic, one measure, temporal at lowest resolution and location (min level of granularity) dimensions are set as mandatory. A further three demographics, two measures, temporal at highest resolution and location (max level of granularity) dimensions are set as optional.

5.7.1 Results

Our approach to local processing for re-identification protection, is based on the premise of a trade-off between the amount of individually specific data requested and the level of

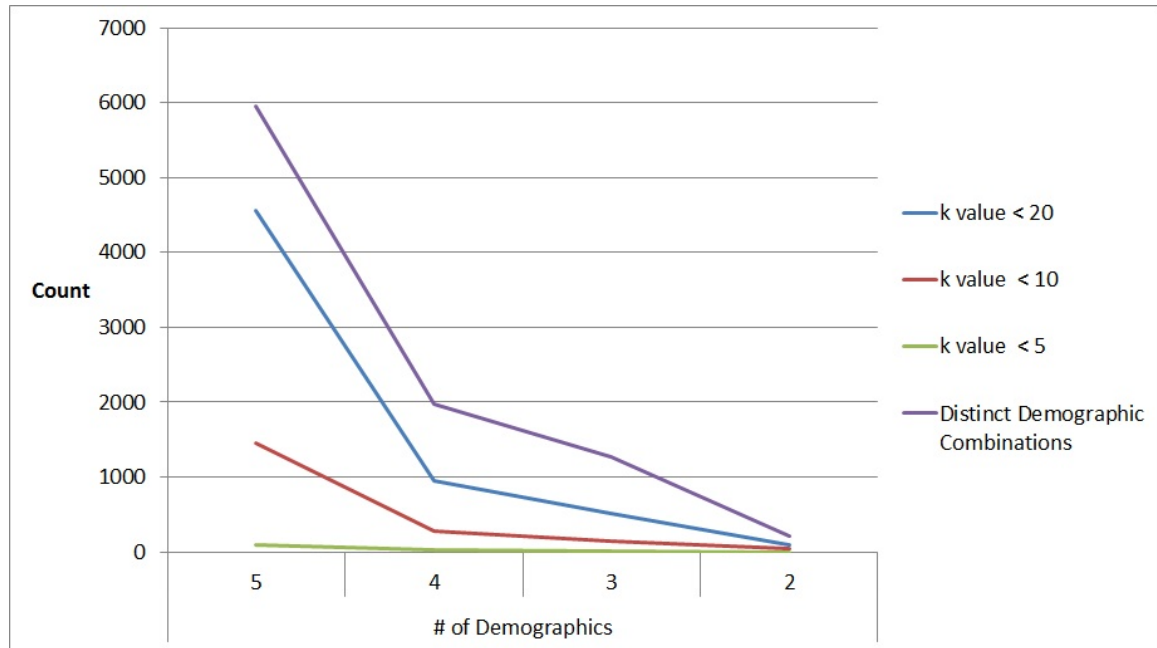


Fig. 5.4 Demographics Removal and Impact on k -anonymity Value

submission and privacy thresholds a system will have. As such, a comparison against the more typical approach is relevant. In a more typical case, rather than making the decision on the local device, the number of demographics required would be adjusted at a system-wide level. As such, the reduction of the number of demographics reduces the overall unique combinations and hence the chance of re-identification. The results of this type of approach, for a sample of 100,000 submissions are displayed in Figure 5.4 where there is a clear benefit of decreasing the number of demographics required for submission if keeping a low k -anonymity value is a priority.

However, in a more adaptive approach, such as using local processing to make decisions based on greater knowledge of potential re-identification, this can be improved by only removing demographic combinations at higher risk of re-identification due to lower representation and therefore k -anonymity values. In our approach, this is evaluated based on a set threshold and the individual's demographic probability, based on population averages. In Figure 5.5 we show how this approach on the same underlying data can support having

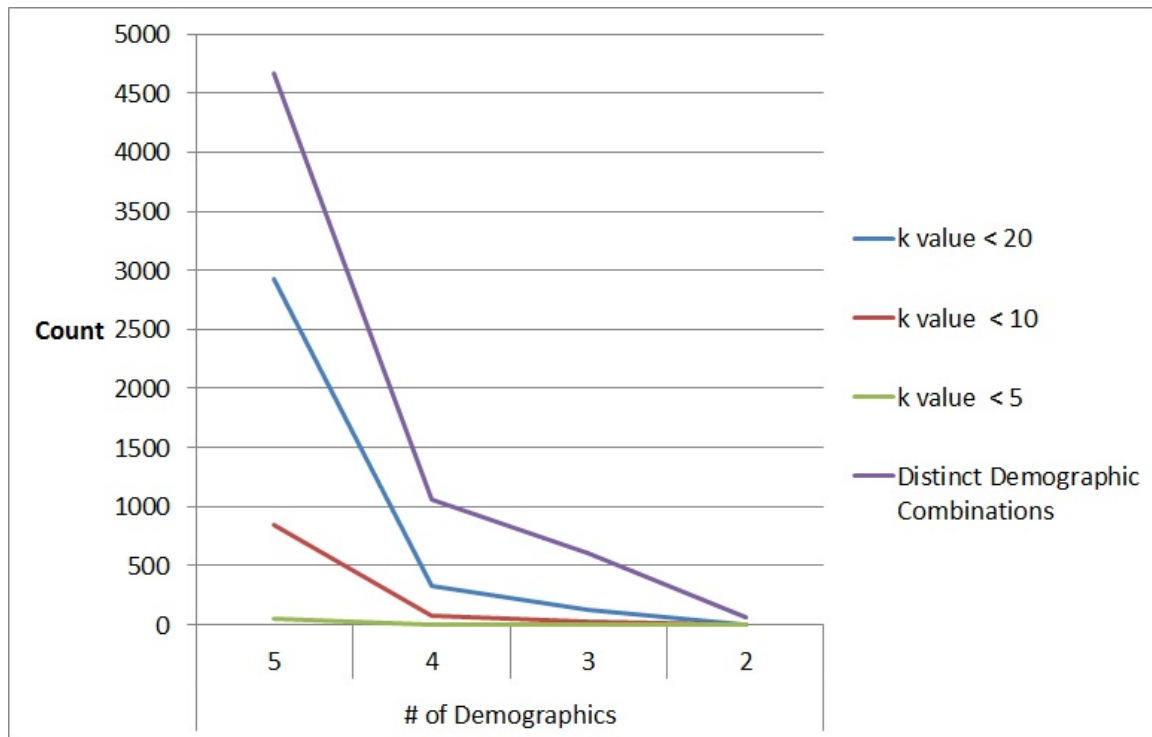


Fig. 5.5 Local Processing Impact of k -anonymity Value

greater demographic requests in a data submission while alleviating privacy risk.

As figure 5.5 demonstrates, the local processing approach provided a steep decline in the number of distinct demographic combinations that had a k value lower than x . In fact, for the most critical level of k value, which is 5, a suitable k -anonymity level was reached at 4 demographics (for which only 3 distinct demographic combinations had $k < 5$), while the approach in Figure 5.4 required reduction to 2 for a comparable k -anonymity value. That is, in Figure 5.4, 4 demographics included led to 23 distinct demographic combinations with $k < 5$, 3 demographics included led to 13 combinations with $k < 5$, and only with 2 demographics included was there 0 combinations with $k < 5$. The number of demographics included in a submission, relate to the demographic threshold and overall threshold limits, where a lower threshold would result in a limitation of the number of demographics or the detail of a demographic. Alternatively, the number of demographics with low k value could also be decreased by increasing the overall number of submissions, however, this is somewhat out-

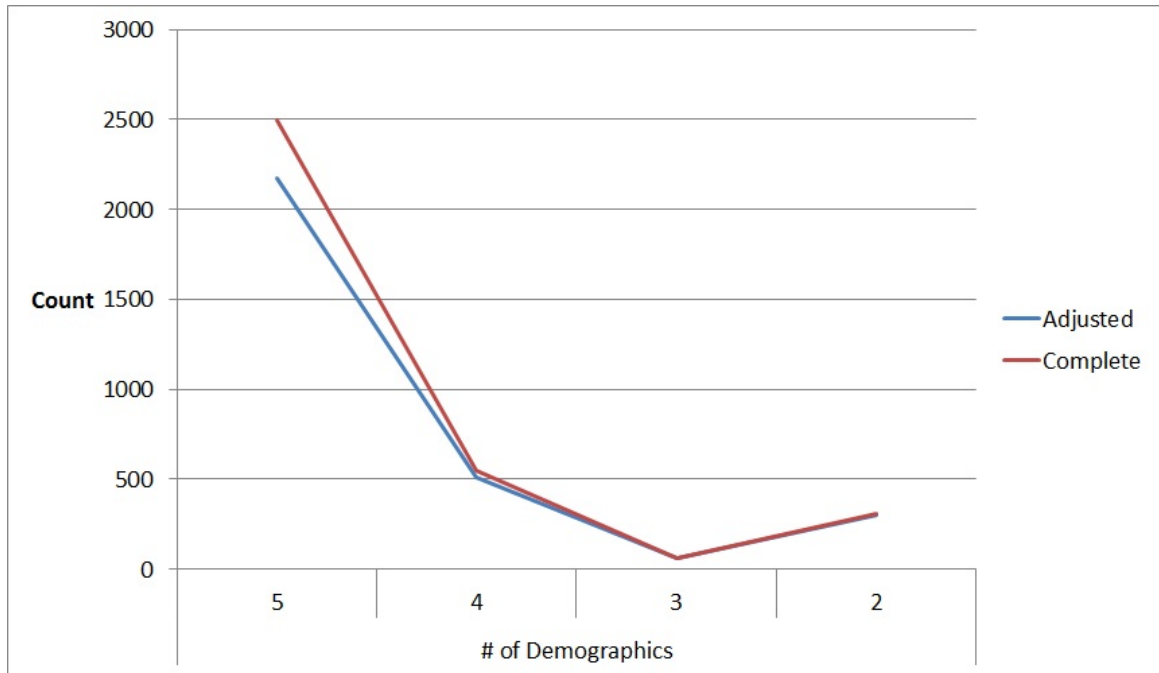


Fig. 5.6 Comparison of Adjusted and Complete Demographic Combinations

side the control of the system. This assumption could be taken into account in our threshold approach to provide lower thresholds, where larger numbers of submissions are expected to provide further improvements over a system level demographic setting approach.

In Figure 5.6 we compare the number of demographic combinations that are adjusted by excluding one or more optional demographics to the number of complete demographic combinations. As shown, our implementation returned the majority of demographics combinations in a complete manner, at the varying threshold levels. Additionally, as it is the least represented demographic groups in terms of population numbers that were adjusted, a larger majority of overall submissions were complete (59%).

Our analysis of the use of location and temporal components in data submissions considers that in many cases location types or very coarse locations and broad time periods will not be sensitive for re-identification purposes (see Privacy section). However, even if utilized for such a purpose, it could be mitigated by a threshold approach and we can look at this in terms of a similar k -anonymity analysis. In Figure 5.7, we display the k -anonymity

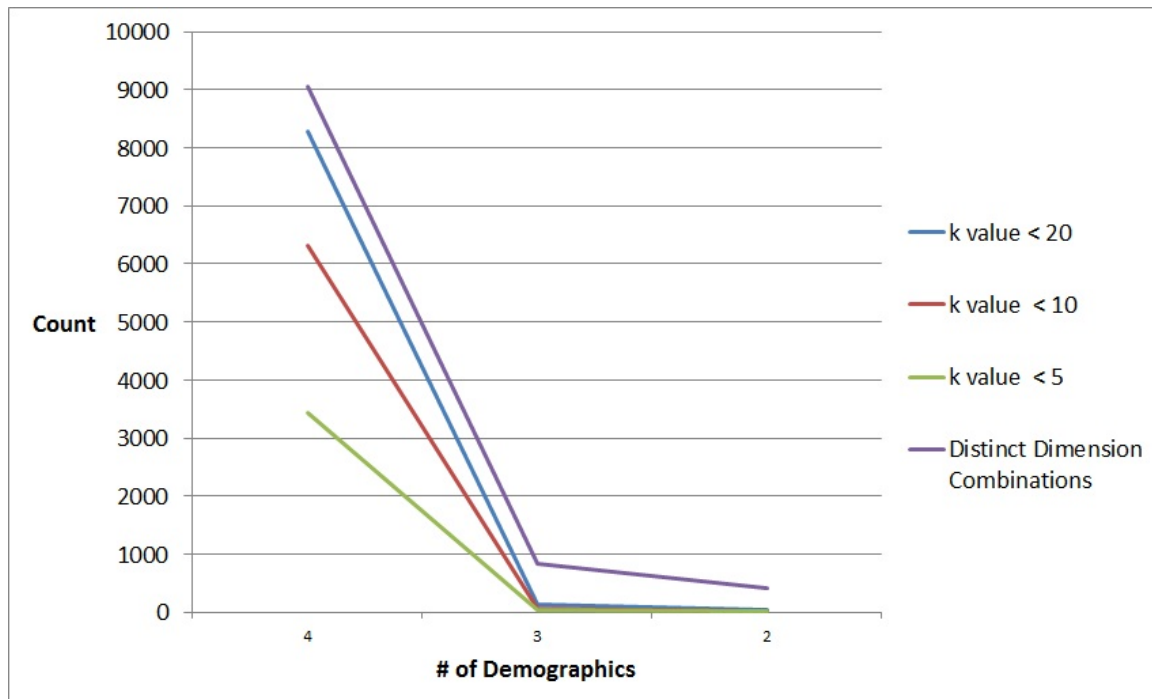


Fig. 5.7 Local Processing Impact of k -anonymity Value with Location Types

values for a mix of a location dimension and multiple demographic dimensions, which has a typical graph as the number of demographics are decreased. To keep the location type dimension from further creating low k value distinct dimension combinations, our threshold approach took the n (where n is a randomized variable $4 < n < 10$) most common locations types for each individual and submitted all other location types as undisclosed to the HPSS. This is potentially a key consideration, as even with four demographics, the addition of the location types we included would have extended the number of distinct dimension combinations out to a possible 22176, making it quite likely that almost all combinations would have had a k -value lower than 5.

Overall, the threshold approaches and local processing modification of public health relevant data collection as shown in these results can be effective in improving the k -anonymity value of demographic groups and hence a reduction is potential re-identification and misuse of collected data. In terms of demographics, the types of demographics with less even distributions create more modifications. The cost of this approach in terms of decreased

information collection as compared to just excluding entire dimensions is also an improvement. The approach could also be further improved through calibration of the HPSN over time.

5.8 Discussion and Future Work

Systems that collect public health related data have significant implications in terms of privacy, anonymity, ethical considerations and technical challenges that need to be considered in development of a public health information systems approach.

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 [30]. Additionally, the growth in smart device ownership and personal health tracking and quantification will continue to drive the potential of health participatory sensing.

The proposed smartphone-based public health information system focuses 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 manner similar to portions of a private electronic health record. These considerations could potentially also be resolved within the bounds of the existing described approach.

However, privacy and public perceptions of such participatory sensing approaches need

to be further researched. As such, future work could include studies of perceived privacy of participatory sensing applications specific to the health domain. A useful extension in this regard would be to consider incentivization, adoption and health organization acceptance of such approaches.

5.9 Conclusion

This chapter describes smartphone-based public health information systems for population-scale anonymous capture of public health data and intervention. The type of system described also has 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. Additionally we have considered the privacy, anonymity and intervention properties and implications of such systems.

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 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. We provided a detailed evaluation of the privacy preserving characteristics of such systems at the level of large user numbers.

REFERENCES

- [1] Andrew Clarke and Robert Steele. “Smartphone-based Public Health Information Systems: Anonymity, Privacy and Intervention”. In: *Journal of the Association for Information Science and Technology* (2015). ISSN: 2330-1643.
- [2] Jeff Burke et al. “Participatory sensing”. In: *In: Workshop on World-Sensor-Web (WSW’06): Mobile Device Centric Sensor Networks and Applications*. 2006, pp. 117–134.
- [3] Andrew Clarke and Robert 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*. Jan. 2014, pp. 2908–2917.
- [4] Predrag Klasnja and Wanda Pratt. “Healthcare in the pocket: Mapping the space of mobile-phone health interventions”. In: *Biomedical Informatics* 45.1 (2012), pp. 184–198. ISSN: 1532-0464.
- [5] Sjouke Mauw, J. Verschuren, and Eric de Vink. “A Formalization of Anonymity and Onion Routing”. In: *Computer Security - ESORICS 2004*. Ed. by Pierangela Samarati et al. Vol. 3193. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2004, pp. 109–124. ISBN: 978-3-540-22987-2.

- [6] Krishna Sampigethaya and Radha Poovendran. “A Survey on Mix Networks and Their Secure Applications”. In: *Proceedings of the IEEE* 94.12 (Dec. 2006), pp. 2142–2181. ISSN: 0018-9219.
- [7] Panos Kalnis and Gabriel Ghinita. “Spatial k-Anonymity”. In: *Encyclopedia of Database Systems*. Ed. by Ling Liu and M. Tamer Özsu. Springer US, 2009, p. 2714. ISBN: 978-0-387-35544-3, 978-0-387-39940-9.
- [8] Melanie Swan. “Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0”. In: *Journal of Sensor and Actuator Networks* 1.3 (2012), pp. 217–253. ISSN: 2224-2708.
- [9] Robert Steele. “Social media, mobile devices and sensors: Categorizing new techniques for health communication”. In: *Sensing Technology (ICST), 2011 Fifth International Conference on*. Nov. 2011, pp. 187–192.
- [10] Robert Steele et al. “Elderly persons’ perception and acceptance of using wireless sensor networks to assist healthcare”. In: *International Journal of Medical Informatics* 78.12 (2009). Mining of Clinical and Biomedical Text and Data Special Issue, pp. 788–801. ISSN: 1386-5056.
- [11] MobiHealthNews. *Google adds activity tracking to Android app*. 2012. URL: <http://mobihealthnews.com/19551/google-adds-activity-tracking-to-android-app/>.
- [12] Sano Intelligence. 2012. URL: <http://rockhealth.com/accelerator/portfolio-companies/sano-intelligence/>.
- [13] Bicycle Network. *Riderlog*. 2012. URL: <http://www.bv.com.au/general/ride-to-work/91481/>.
- [14] Christine Outram, Carlo Ratti, and Assaf Biderman. “The Copenhagen Wheel: An innovative electric bicycle system that harnesses the power of real-time information

- and crowd sourcing”. In: *EVER Monaco International Exhibition & Conference on Ecologic Vehicles & Renewable Energies*. 2010.
- [15] Darren E.R. Warburton, Crystal Whitney Nicol, and Shannon S.D. Bredin. “Health benefits of physical activity: the evidence”. In: *Canadian Medical Association Journal* 174.6 (2006), pp. 801–809.
- [16] Robert Steele. “An Overview of the State of the Art of Automated Capture of Dietary Intake Information”. In: *Critical Reviews in Food Science and Nutrition* (2013).
- [17] AIHW. “Biomedical Component of the Australian Health Survey: Public Health Objectives.” In: (2011).
- [18] Samuel Gibbs. *Sweet solution? Google tests smart contact lens for diabetics, the-guardian*. 2014. URL: <http://www.theguardian.com/technology/2014/jan/17/google-tests-smart-contact-lens-diabetics>.
- [19] Predrag Klasnja et al. “Using Mobile & Personal Sensing Technologies to Support Health Behavior Change in Everyday Life: Lessons Learned”. In: *AMIA Annual Symposium Proceedings*. 2009, pp. 55–59.
- [20] Andrew Clarke and Robert Steele. “How personal fitness data can be re-used by smart cities”. In: *Intelligent Sensors, Sensor Networks and Information Processing (ISS-NIP), 2011 Seventh International Conference on*. Dec. 2011, pp. 395–400.
- [21] Noora Hirvonen et al. “Information behavior in stages of exercise behavior change”. In: *Journal of the American Society for Information Science and Technology* 63.9 (2012), pp. 1804–1819. ISSN: 1532-2890.
- [22] Robert Steele, Kyongho Min, and Amanda Lo. “Personal health record architectures: Technology infrastructure implications and dependencies”. In: *Journal of the American Society for Information Science and Technology* 63.6 (2012), pp. 1079–1091. ISSN: 1532-2890.

- [23] Andrew Clarke and Robert 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*. Jan. 2014, pp. 2908–2917.
- [24] Andrew Clarke and Robert 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*. 2012, pp. 2158–2161.
- [25] Ariel Schwartz. *No More Needles: A Crazy New Patch Will Constantly Monitor Your Blood*. 2012. URL: <http://www.fastcoexist.com/1680025/no-more-needles-a-crazy-newpatch-will-constantly-monitor-your-blood>.
- [26] Mariusz Wisniewski et al. “NoizCrowd: A Crowd-Based Data Gathering and Management System for Noise Level Data”. In: *Mobile Web Information Systems*. Ed. by Florian Daniel, GeorgeA. Papadopoulos, and Philippe Thiran. Vol. 8093. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 172–186. ISBN: 978-3-642-40275-3.
- [27] Bratislav Predic et al. “ExposureSense: Integrating daily activities with air quality using mobile participatory sensing”. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*. Mar. 2013, pp. 303–305.
- [28] Austroads. *Australian Cycling Participation : Reporting for the National Cycling Strategy 2011-2016*. 2011. URL: <http://www.austroads.com.au/abc/images/pdf/AP-C91-11.pdf>.
- [29] Berker Agir et al. *Adaptive Personalized Privacy in Participatory Sensing*. Tech. rep. 2012.

-
- [30] Robert Steele and Andrew Clarke. “The Internet of Things and Next-generation Public Health Information Systems”. In: *Communications and Network* (2013).

**Faculty of Health Sciences
Author Contribution Statement**

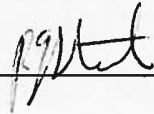
Candidate Name: Andrew Clarke

Degree Title: Doctor of Philosophy

Paper Title: Smartphone-based Public Health Information Systems: Anonymity, Privacy and Intervention

As the corresponding author of the above paper, I confirm that the above candidate has made contributions to the following:

- Conception and design of the research
- Development of the prototype implementation
- Analysis and interpretation of the findings
- Writing the paper and critical appraisal of content

Signed  Name Prof Robert Steele Date 10th March 15

6 TARGETED AND ANONYMIZED HPSN PUBLIC HEALTH INTERVENTIONS

Preamble

This chapter is based on a conference paper, that was published in the Annual International Conference of the Engineering in Medicine and Biology Society (EMBC) [1]. It has been included as a chapter of this thesis with only minor formatting changes to align with the thesis format.

This chapter forms the final chapter of the implementation/evaluation section. The preceding chapters 4 and 5 dealt with the distribution of data collection policies and the de-identification of data before submission. This final chapter deals specifically with the distribution, application and reporting collection of the use of public health interventions within a health participatory sensing network. Though these topics were touched on briefly in chapters 4 and 5, they are covered in greater detail in this section with an appropriate prototype implementation and results. Additionally, much of the methodology for this final chapter was already introduced in previous chapters, and as such, references to previous papers (including papers used as chapters in this thesis) are used extensively in this chapter.

This chapter relates back to subsection 3.3.4 in the initial foundation chapter and expands and details the approach utilized.

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 chapter, 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 chapter we describe such a platform, its place in the HPSN and demonstrate its feasibility through an implementation.

6.1 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 tar-

getting. Additionally, health participatory sensing systems such as HPSNs [2] 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 [3] which deal with identified individuals and their personal health data and possible communication such as appointment reminders or medication adherence. Instead, HPSN focus 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 a solution to these problems; which is an extension and combination of our prior work in relation to HPSNs [2, 4] and query assurance [5]. 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.

6.2 Related Work

The rich capabilities of participatory sensing have garnered interest in its usage for a range of quite disparate areas. This includes air quality and pollution sensing [6], urban area noise level data collection [7] and public health data collection [2] 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 [8] 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 [9] to provide an anonymous approach. While this is quite effective when the participatory sensing network 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 [10] using user interaction/awareness as part of the approach or keeping the data managed by the participant [11, 12] and stringent user-definable access control mechanisms to manage sharing. The limitation of these approaches when considering HPSNs is 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.

6.3 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 [13] 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 [2] 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 [14]. 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 [2], 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.

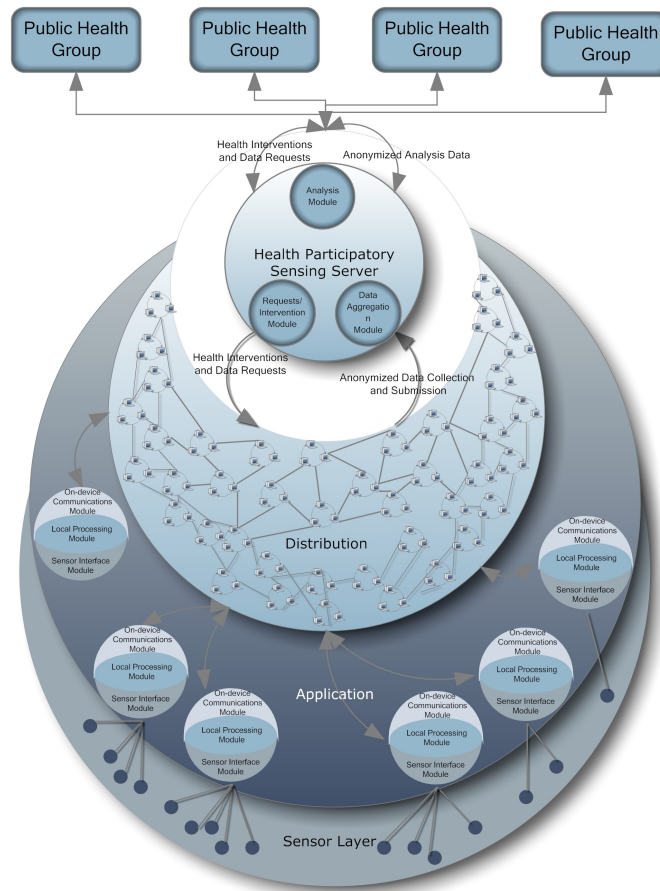


Fig. 6.1 Public Health Information System Architecture (see section 5.4)

6.4 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 public health data collection system, the intervention approach 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 6.3 and our previous work [2].

The platform components and their inter-relationships are illustrated in Fig 6.1, 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.

6.4.1 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 [15] or onion network [16], 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 [5] 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 timestamp 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, as well as keeping verification data overhead size and processing time to acceptable levels.

6.4.2 Application

As public health interventions are performed on the local device, the decision as to the intervention to perform, must also 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.

6.4.3 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 [4]. 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 [4], 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 [4].

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.

6.4.4 Analysis

The analysis of public health participatory sensing data relies on collection of sufficient data for public health uses [17], which differs from what would be required in most other participatory sensing systems. As such, generally aggregate non-specific demographic level data is needed, as well as the measured values and the types of interventions performed.

6.5 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 broad demographics targets, 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 [5].

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 [5] (see Section 6.4.1). 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, the participant device 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 [5]. As such, we perform our implementation pre and post optimization incorporating 5000 requests for each. The results of the verification overhead are displayed in figure 6.2.

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 [4], extended and modified for public health interventions, which uses a threshold and priority approach to decide what information is reported for analysis before locally pro-

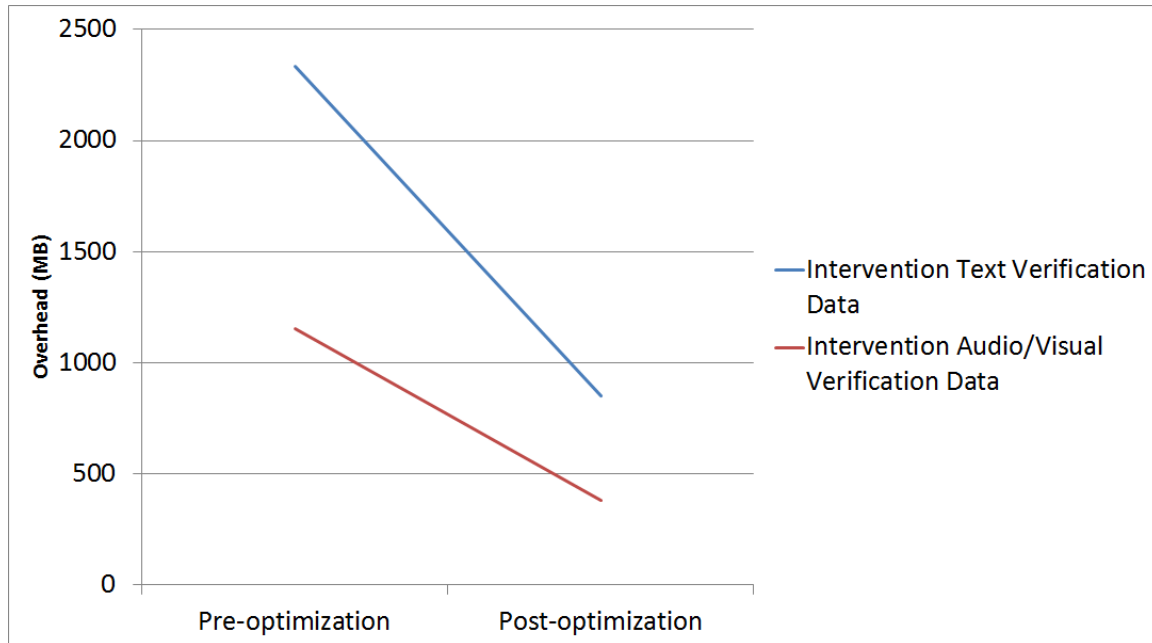


Fig. 6.2 Public Health Intervention Distribution Verification Overhead

cessing the result and submitting through an anonymous communications network [4] (see Section 6.4.3 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 Metropolitan [18] area to generate the individual client's demographics including age, gender, ethnicity, income and education. The prototype client and server are both developed in Java (1.6), the client uses

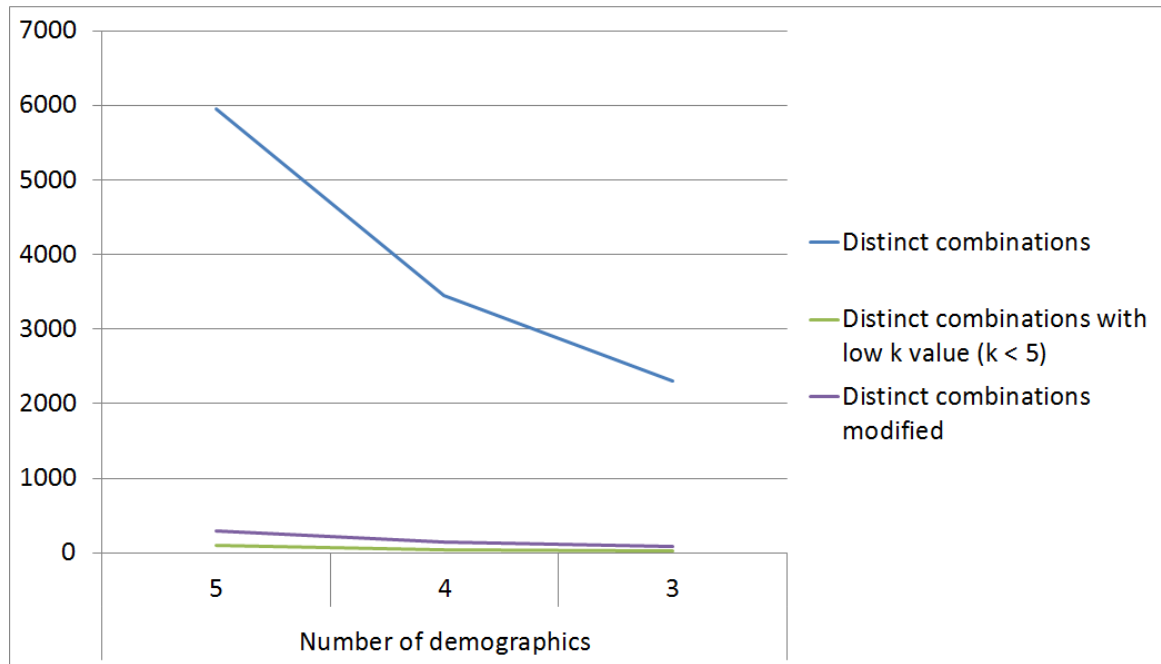


Fig. 6.3 Public Health Intervention Distinct Demographic Combinations to Low k Value Combinations

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 6.3, 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.

6.6 Conclusion

This chapter 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 via utilization of 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] Andrew Clarke and Robert Steele. “Targeted and anonymized smartphone-based public health interventions in a participatory sensing system”. In: *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. Aug. 2014, pp. 3678–3682.
- [2] Andrew Clarke and Robert Steele. “Health Participatory Sensing Networks”. In: *Mobile Information Systems* 10 (3 2014), pp. 229–242.
- [3] Robert Steele, Kyongho Min, and Amanda Lo. “Personal health record architectures: Technology infrastructure implications and dependencies”. In: *Journal of the American Society for Information Science and Technology* 63.6 (2012), pp. 1079–1091. ISSN: 1532-2890.
- [4] Andrew Clarke and Robert 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*. Jan. 2014, pp. 2908–2917.
- [5] Andrew Clarke and Robert Steele. “Secure query assurance approach for distributed health records”. In: *Health Systems* 3.1 (Feb. 2014), pp. 60–73. ISSN: 2047-6965.
- [6] Bratislav Predic et al. “ExposureSense: Integrating daily activities with air quality using mobile participatory sensing”. In: *Pervasive Computing and Communications*

- Workshops (PERCOM Workshops), 2013 IEEE International Conference on.* Mar. 2013, pp. 303–305.
- [7] Mariusz Wisniewski et al. “NoizCrowd: A Crowd-Based Data Gathering and Management System for Noise Level Data”. In: *Mobile Web Information Systems*. Ed. by Florian Daniel, George A. Papadopoulos, and Philippe Thiran. Vol. 8093. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 172–186. ISBN: 978-3-642-40275-3.
- [8] Panos Kalnis and Gabriel Ghinita. “Spatial k-Anonymity”. In: *Encyclopedia of Database Systems*. Ed. by Ling Liu and M. Tamer Özsu. Springer US, 2009, p. 2714. ISBN: 978-0-387-35544-3, 978-0-387-39940-9.
- [9] Cory Cornelius et al. “Anonymsense: privacy-aware people-centric sensing”. In: *Proceedings of the 6th international conference on Mobile systems, applications, and services*. MobiSys '08. Breckenridge, CO, USA: ACM, 2008, pp. 211–224. ISBN: 978-1-60558-139-2.
- [10] Delphine 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.* 2010, pp. 847–848.
- [11] Min Mun et al. “Personal data vaults: a locus of control for personal data streams”. In: *Proceedings of the 6th International Conference*. Co-NEXT '10. Philadelphia, Pennsylvania: ACM, 2010, 17:1–17:12. ISBN: 978-1-4503-0448-1.
- [12] Haksoo Choi et al. “SensorSafe: A Framework for Privacy-Preserving Management of Personal Sensory Information”. In: *Secure Data Management*. Ed. by Willem Jonker and Milan Petkovic. Vol. 6933. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 85–100. ISBN: 978-3-642-23555-9.

- [13] Gartner. *Gartner Says Worldwide Smartphone Sales Soared in Fourth Quarter of 2011 With 47 Percent Growth*. 2011. URL: <http://www.gartner.com/it/page.jsp?id=1924314>.
- [14] John Rula and Fabián E. Bustamante. “Crowd (Soft) Control: Moving Beyond the Opportunistic”. In: *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*. HotMobile '12. San Diego, California: ACM, 2012, 3:1–3:6. ISBN: 978-1-4503-1207-3.
- [15] Krishna Sampigethaya and Radha Poovendran. “A Survey on Mix Networks and Their Secure Applications”. In: *Proceedings of the IEEE* 94.12 (Dec. 2006), pp. 2142–2181. ISSN: 0018-9219.
- [16] Sjouke Mauw, J. Verschuren, and Eric de Vink. “A Formalization of Anonymity and Onion Routing”. In: *Computer Security - ESORICS 2004*. Ed. by Pierangela Samarati et al. Vol. 3193. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2004, pp. 109–124. ISBN: 978-3-540-22987-2.
- [17] Andrew Clarke and Robert 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*. 2012, pp. 2158–2161.
- [18] Australian Bureau of Statistics. *Census Community Profiles Greater Sydney*. 2011. URL: http://www.censusdata.abs.gov.au/census%5C_services/getproduct/census/2011/communityprofile/1GSYD.

**Faculty of Health Sciences
Author Contribution Statement**

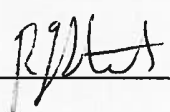
Candidate Name: Andrew Clarke

Degree Title: Doctor of Philosophy

Paper Title: Targeted and Anonymized Smartphone-based Public Health Intervention in a Participatory Sensing System

As the corresponding author of the above paper, I confirm that the above candidate has made contributions to the following:

- Conception and design of the research
- Development of the prototype implementation
- Analysis and interpretation of the findings
- Writing the paper and critical appraisal of content

Signed  Name Prof Robert Steele Date 10th March 15

7 DISCUSSION AND FUTURE WORK

Systems that collect public health related data have significant implications in terms of privacy, anonymity, ethical considerations and technical challenges that need to be considered in development of a public health information systems approach.

However, HPSNs show great promise in their capability for both collecting relevant, timely and useful public health information that makes the implications well worth investigating with the development of strategies and approaches to mitigating the negative implications will be key to potential usage.

The use of mobile device technology for public health data collection and intervention is largely still in its infancy, as such, the on-going development of new sensing technologies and experience with usage will continue to influence the types of participatory sensing possible and the potential in the public health context. These collection and intervention related requirements are introduced in Chapter 3 and the discussion and future work of this work in presented in Section 7.2 below.

This thesis focused on some of the technical and privacy issues related to the distribution of collection rules and interventions, collection of data and intervention results and the submission of data. Additionally, it also dealt with efficiencies of distribution and a multi-data owner system. Finally, as this was a largely undefined area of participatory sensing, a certain amount of classification and description of interaction and capabilities was required.

The distribution component provided a suitable approach for distribution of data collection policies and public health interventions. This was the most explored area in previous

work with many alternate approaches available with the contribution of this thesis focusing on gaining some additional efficiencies through routine maintenance of the verification objects to reduce security overhead. The second requirement was that the approach had to be compatible with a multiple data-owner environment which is addressed in the implementation. The implementation was used to evaluate the efficiencies of the query assurance approach and provided some improvements over prior work. These distribution related requirements are introduced in Chapter 4 and the discussion and future work of this work is presented in Section 7.1 below.

7.1 Query Assurance

There is a trend towards development of more complex and integrated future public health information systems, with data held between multiple organizations and even potentially components held with the consumer [6]. The query assurance and distribution approaches described in this thesis aim to specifically target the issues of public health data stored over a distributed architecture with multiple data owners, such as a HPSN. In Chapter 4 we proposed a query assurance method matched to this challenge, as the approach primarily provides query assurance even in cases where the data is stored on hardware not controlled by the data owner and secondarily provides encryption as part of the approach.

The types of data likely to be stored in public health databases by health organizations and providers are more fully understood and of a traditional nature, leading to a level of certainty as to what will tend to be stored in the foreseeable future. However, the types of data collected by health consumers are developing rapidly and it is not inconceivable that they may be the larger data sources in future health information systems.

This would include the collection of subjective data, as well as increasingly objective data collected by consumer grade mobile sensors [6] and other consumer health communications technologies [7]. A recent report suggested that one in four adult Internet users

track their own health data online [8], suggesting the possible growing importance of this source of public health data. The types of data collected already are potentially important for consideration of risk and preventative health factors when taken in combination with the individual's complete health record. The proliferation of personal fitness data systems may, in many cases lead to the storage of such data in separate or proprietary repositories, as is seen in current practice. This anticipates future health information system architectures where some parts of health data, for example, some of the arguably less sensitive data such as fitness data, are held within different proprietary repositories, but public health groups are still able to reliably integrate and access this data as if it is a single record. This is an important functionality for realizing computational health capabilities, such as data mining to support individual care, and population health analysis [6]. The numerous organizations holding public health-related data and the lack of complete trust between these organizations, underscores the need for a generic query assurance approach, extensible to multiple repositories, owners and data types as described in this chapter. The query assurance techniques described can be applied to repositories of health or fitness data collected by mobile or pervasive sensors.

In summary, emerging health information systems have great potential in the realm of more complete data and communications capabilities, but will have to address the challenges of such systems, including the greater challenges to query assurance and security. As such, we consider the following would be important to further development of secure query assurance over encrypted public health information:

Combining more complex forms of searchable encryption with query assurance could be attempted, with the goal to increase query efficiency or allow more complex queries to be performed against the encrypted data set.

In another research path, possible improvements to the initialization and maintenance portion of the approach could be considered. Further, extension of this implementation to

provide secure auditing would be a useful approach.

7.2 Public health data collection and intervention

The threshold approach to de-identification of individual data aimed to meet two goals. Firstly, fill the existing gap in the prior work by providing a suitable low overhead approach that removes the need for a trusted third party, and secondly, provide an approach more suitable to public health data - the types of data collected likely to be very dissimilar to other participatory sensing applications where location is a key component. Hence the existing privacy approaches have limited utility when applied to the public health context. The evaluation of the approach indicated that while it is not a zero-risk system, privacy risks can be significantly reduced through minor reductions of the detail of data submitted in line with local threshold limits and demographic and distribution analysis.

Lastly, an implementation of the key new functionality of HPSNs over traditional public health data collection methodologies was provided, that is, the capability to utilize public health interventions on the same platform. The distribution and reporting threshold limitations of these interventions was evaluated, providing a key initial work in this area.

The proposed smartphone-based public health information system focuses 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 could potentially also be resolved within the bounds of the existing described approach.

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

8 CONCLUSIONS

This thesis describes Health Participatory Sensing Networks, their capabilities, challenges and technical requirements for public health data collection and intervention. Specifically, this thesis defined the capabilities, interaction types and incentivization approaches for these participatory systems.

Further, this work provided technical detail and prototype implementations and evaluations of some of the key challenges of HPSNs. Namely, the distribution of HPSN data, the anonymous collection of public health data and the usage of public health interventions.

The smartphone-based public health information systems include an approach based on local processing of aggregate data for public health use 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, this thesis 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. Further, a detailed evaluation of the privacy preserving characteristics of such systems at the level of large user numbers was provided.

The evaluation of each section found that privacy and security could be substantially improved with minor overheads and data restrictions.

Future work providing a much deeper analysis of incentivization approaches for this specific participatory sensing model could be beneficial. Additionally, the further evaluation

of this approach utilizing different demographic data to check or additionally fine tune the privacy thresholds could improve the overall privacy for individuals.

REFERENCES

- [1] Apple Inc. *Health*. URL: <https://www.apple.com/ios/whats-new/health/>.
- [2] Dan Graziano. *The complete guide to Google Fit*. 2014. URL: <http://www.cnet.com/how-to/the-complete-guide-to-google-fit/>.
- [3] Samsung Electronics Co. Ltd. *S Health*. URL: <http://content.samsung.com/us/contents/aboutn/sHealthIntro.do>.
- [4] Jeff Burke et al. “Participatory sensing”. In: *Workshop on World-Sensor-Web (WSW’06): Mobile Device Centric Sensor Networks and Applications*. 2006, pp. 117–134.
- [5] Predrag Klasnja and Wanda Pratt. “Healthcare in the pocket: Mapping the space of mobile-phone health interventions”. In: *Biomedical Informatics* 45.1 (2012), pp. 184–198. ISSN: 1532-0464.
- [6] Robert Steele and Amanda Lo. “Future Personal Health Records as a Foundation for Computational Health”. In: *ICCSA* (2). 2009, pp. 719–733.
- [7] Robert Steele. “Social media, mobile devices and sensors: Categorizing new techniques for health communication”. In: *Sensing Technology (ICST), 2011 Fifth International Conference on*. Nov. 2011, pp. 187–192.
- [8] Susannah Fox and Sydney Jones. “The social life of health information”. In: *Washington, DC: Pew Internet & American Life Project* (2009), pp. 2009–12.

A IMPLEMENTATION DETAILS AND PSEUDOCODE

A.1 Query Assurance Algorithms

Algorithm 1 DB Initialization

```
1: Read data from file
2: Send data to DB server
3: while hash depth not reached do
4:   Hash element
5:   Send to Path tree on verification server
6:   Re-sign root node
7:   if index required then
8:     Send index to Index tree on verification server
9:     Re-sign index tree
10:  end if
11: end while
```

Algorithm 2 Client driven data accessing

```

1: query database
2: query Verification server
3: receive query result
4: receive verification object
5: hash query result
6: if hash (query result) exists in verification object then
7:   while signature not reached do
8:     hash (lowerHash)
9:   end while
10:  verify signature with highest level hash
11:  if verification passes then
12:    query result is valid
13:  else query result is invalid
14:  end if
15: end if

```

Algorithm 3 Server driven Data accessing

```

1: query database
2: Server translates query to verification Tree
3: receive query result + verification object
4: hash (query result + path)
5: if hash exists in verification object then
6:   while while signature not reached do
7:     hash(lowerHash)
8:   end while
9:   verify signature with highest level hash
10:  if verification passes then
11:    query result is valid
12:  else query result is invalid
13:  end if
14: end if

```

Algorithm 4 Refreshing verification tree time stamps

```

1: Server maintains list of timestamped nodes
2: Server sends time stamps nodes to Data Owner
3: Data owner verifies signatures+ time stamp authenticity
4: Data owner resign the nodes and returns to Server

```

Algorithm 5 Maintenance

```

1: for  $i = 0$  to indexNo do
2:   leaf = firstLeaf
3:   for  $j = 0$  to leafCount do
4:     if leaf.readCount > threshold then
5:       temp = leaf
6:       while nextLeaf.readCount > threshold do
7:         leaf = nextLeaf
8:       end while
9:     end if
10:    if temp != null then
11:      find common ancestor(temp,leaf)
12:      apply time stamp + signature to ancestor
13:      temp = null
14:    end if
15:  end for
16: end for
17: for  $j = 0$  to pathTreeNo do
18:   check readRecords
19:   if readRecords > granThreshold then
20:     re-hash data to finer granularity
21:   end if
22:   set time stamp duration based on modIficationRate
23: end for

```

A.2 Local Processing Algorithms

Algorithm 6 Test Implementation

```
1: for  $j = 0$  to 10000 do  
2:   Mobile Device Individual Init()  
3:   Mobile Device Local Processing()  
4: end for
```

Algorithm 7 Mobile Device Database Clear

```
1: Drop Table Demographic  
2: Drop Table Individual  
3: Drop Table IndividualToDemographic  
4: Drop Table Measure  
5: Drop Table Location  
6: Drop Table Location_Type
```

Algorithm 8 Mobile Device Database Init

- 1: Mobile Device Database Clear()
 - 2: Create Table Demographic
 - 3: Create Table Individual
 - 4: Create Table IndividualToDemographic
 - 5: Create Measure Table Measure
 - 6: Create Table Location
 - 7: Create Table Location_Type
 - 8: Insert Greater Sydney area ancestry data into Demographics Table (Demographic_ID, Description, Type_ID, Probability)
 - 9: Insert Greater Sydney area age group data into Demographics Table (Demographic_ID, Description, Type_ID, Probability)
 - 10: Insert Greater Sydney area gender data into Demographics Table (Demographic_ID, Description, Type_ID, Probability)
 - 11: Insert Greater Sydney area income group data into Demographics Table (Demographic_ID, Description, Type_ID, Probability)
 - 12: Insert Greater Sydney area education level data into Demographics Table (Demographic_ID, Description, Type_ID, Probability)
 - 13: Insert location type data into Location_Type Table (Location_Type_ID, Description, Probability)
-

Algorithm 9 Mobile Device Individual Init

- 1: Generate random ancestry, age group, gender, income group and education level.
 - 2: Insert ancestry into IndividualToDemographic Table
 - 3: Insert age group into IndividualToDemographic Table
 - 4: Insert gender into IndividualToDemographic Table
 - 5: Insert income group into IndividualToDemographic Table
 - 6: Insert education level into IndividualToDemographic Table
 - 7: Randomize number of workouts x
 - 8: **for** $j = 0$ to x **do**
 - 9: Generate randomized location_type
 - 10: Generate randomized latitude
 - 11: Generate randomized longitude
 - 12: Generate randomized measure/s
 - 13: Generate randomized datetime
 - 14: Insert lat, long and location_type into Location Table
 - 15: Insert individualID, measure_type, value, description, time, locationID into Measure Table
 - 16: **end for**
-

Algorithm 10 Mobile Device Local Processing

```
1: Set Overall Threshold
2: Set Demographic Threshold
3: Set Temporal Threshold
4: Set Measure Threshold
5: Retrieve data request
6: Set optionalDemographics
7: Set optionalNoLocations
8: Set optionalMeasures
9: Set mandatoryDemographics
10: Set mandatoryNoLocations
11: Set mandatoryNoMeasures
12: Set mandatoryTemporalDetail
13: Set datetime Period
14: Calculate OverallQIS for mandatory details
15: while OverallQIS < Overall Threshold and additional optional dimensions available do
16:     add optional dimension
17:     for each sub threshold do
18:         if subQIS > sub threshold then
19:             remove optional dimension
20:         end if
21:     end for
22:     re-balance optional data based on available threshold.
23: end while
24: if OverallQIS < Overall Threshold then
25:     Submit Aggregate Data
26: end if
```

A.3 Local Processing Sample File

Sample data request rule:

```
<Submission_Rule>
    <Name>PhyActDuration_Healthy_Bodies</Name>
    <core_data>
        <measure>Physical Activity Duration (Time)</measure>
        <min_detail>HH</min_detail>
        <max_detail>HH:MM:SS</max_detail>
        <resolution_weighting>.65</resolution_weighting>
    </core_data>
    <optional_dimension>
        <value>Start Time</value>
        <min_detail>DD/MM/YYYY</min_detail>
        <max_detail>DD/MM/YYYY HH:MM:SS</max_detail>
        <resolution_weighting>.3</resolution_weighting>
        <inclusion_weighting>.9</inclusion_weighting>
    </optional_dimension>
    <optional_dimension>
        <value>Activity Type</value>
        <min_detail>N/A</min_detail>
        <max_detail>N/A</max_detail>
        <resolution_weighting>N/A</resolution_weighting>
        <inclusion_weighting>.7</inclusion_weighting>
    </optional_dimension>
</Submission_Rule>
```

A.4 Local Processing Results

# of Demographics				
	5	4	3	2
k value < 20	4555	947	507	92
k value < 10	1457	274	146	35
k value < 5	97	23	13	0
Distinct Demo-graphic Combina-tions	5952	1979	1275	208

Table A.1 Demographics Removal and Impact on k -anonymity Value

# of Demographics				
	5	4	3	2
k value < 20	2930	334	127	2
k value < 10	840	81	24	0
k value < 5	50	3	0	0
Distinct Demo-graphic Combina-tions	4668	1055	605	64

Table A.2 Local Processing Impact of k -anonymity value

# of Demographics				
	5	4	3	2
Adjusted	2175	509	62	300
Complete	2493	546	64	305

Table A.3 Comparison of Adjusted and Complete Demographic Combination

# of Demographics			
	4	3	2
k value < 20	8276	136	36
k value < 10	6322	65	22
k value < 5	3443	40	11
Distinct De- mo- graphic Com- bina- tions	9048	832	418

Table A.4 Local Processing Impact of k -anonymity Value with Location Types