34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Summarized Data to Achieve Population-wide Anonymized Wellness Measures

Andrew Clarke and Robert Steele

Abstract— **The growth in smartphone market share has seen the increasing emergence of individuals collecting quantitative wellness data. Beyond the potential health benefits for the individual in regards to managing their own health, the data is highly related to preventative and risk factors for a number of lifestyle related diseases. This data has often been a component of public health data collection and epidemiological studies due to its large impact on the health system with chronic and lifestyle diseases increasingly being a major burden for the health service. However, collection of this kind of information from large segments of the community in a usable fashion has not been specifically explored in previous work. In this paper we discuss some of the technologies that increase the ease and capability of gathering quantitative wellness data via smartphones, how specific and detailed this data needs to be for public health use and the challenges of such anonymized data collection for public health. Additionally, we propose a conceptual architecture that includes the necessary components to support this approach to data collection.**

I. Introduction

The growth in uptake of smartphone devices over recent years [1] has created an ideal environment for the growth of personal, participatory and opportunistic sensing as evidenced by the growth in both applications and research [2]–[4]. With the expected continued growth, it will soon be the case that the large majority of the population of many countries will be using smartphones or other mobile computing devices. The sensory data collection and usage possibilities are only just beginning to be explored.

However, it is generally assumed that collection of this data inherently implies the risk of privacy breaches or undue levels of individually-identifying data collection, prompting continuing research into protection of personal sensory information that is stored online through personal data vaults (PDV) [5] and strengthening access control rules and procedures.

Alternatively, we propose that for the purpose of collection of population-wide wellness data a more efficient approach is to limit the type of data disclosed from a mobile device to just anonymized summary or aggregate data, with consideration given to the potential for re-identification of an individual determined and minimized. In addition, the data can be submitted through an anonymous submission network, allowing for a high level of individual privacy without unnecessary detrimental restrictions to the type of data that could be collected. In this paper we provide further detail to this approach and present a conceptual architecture.

A. Clarke and R. Steele are with the Discipline of Health Informatics, University of Sydney, NSW, Australia, 2006 {andrew.clarke,robert.steele}@sydney.edu.au

II. Collection Mechanisms

There are two main forms of physical collection mechanisms based on mobile devices

1) Mobile device in-device sensors
2) Mobile device with additional external sensors

Though in many cases the in-device sensors of mobile devices are sufficient for data collection, complementation by external sensors is attractive to extend the range of sensing available such as through heart rate monitors (HRM) or air quality (indoor air quality was explored in our previous work [6]). In other cases complementation can overcome the current limitations of mobile devices in regards to battery life or need to wear the mobile device at all times. In this section we will detail some of the various approaches to data collection that are discussed in the current literature and in some cases already available as consumer products.

A. Physical Activity Data

Recent years have seen a significant increase in the detail and quantity of physical activity data tracked by individuals. This has coincided with the steep growth in smartphone uptake and its consequence of providing numerous in-device sensors suitable for this task. While there are a myriad of commercial applications, in this section we will just briefly detail some of the contemporary products that illustrate the potential move to more continuous and detailed data collection.

1) Jawbone Up: Jawbone Up [7] is a product focused around continuous activity monitoring based on a wristband that synchronizes with a mobile device. Its primary functions are to monitor physical activity, sleep patterns (additionally working as a bio-alarm clock) and a visual food journal. The combination of a mobile device and wristband avoids the issue of data collection having an impact on the battery life of already resource constrained mobile devices.

2) Nike+ Fuelband: Similar to Jawbone Up the Fuelband product is focused on all day physical activity sensing. However, it does not extend into sleep cycles and food journal functionality.

3) Headphone based HRM: In an attempt to make HRM more wearable in the everyday environment, there has been some recent research into combining HRM functionality into regular earbud headphones [8] with the potential outcome of increased HRM data collection.

B. Nutritional Data

1) Image Analysis: In recent research approaches, using image analysis via a smartphone's camera to identify the

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

components and size of food portions has been proposed [9]. While these types of approaches have a number of technical challenges regarding improving accuracy of matches and currently requiring user intervention to confirm the result, they are an examples of a shift to more effortless data collection in relation to nutritional data. If a similar trend to activity tracking is observed it could result in a large amount of quantitative data being collected by individuals for their own health goals.

2) Electronic Tracking: Alternatively, approaches using electronic tracking via RFID tags and smart kitchen technology [10] or special food vessels [11] could be utilized by individuals. The previous approaches are limited as they require special equipment to be available at the point of food preparation. Ultimately, a more flexible mobile approach is needed while a combination of RFID, QR codes or other barcodes could be utilized to effectively quantify and track data about food items. Currently, this is still largely limited to barcode lookup or manual entry with only sporadic commercial application of any of these technologies for nutritional data collection.

III. Sufficient Data for Public Health Uses

It is generally assumed to collect useful information, a level of private data in the form of temporal and spatial fine grained data needs to be collected and transmitted to an external service for analysis and calculation of meaningful statistical information. However, we propose that some of the aggregation/calculation of wellness can be performed locally by the individual's mobile devices to decrease the sensitivity of transmitted and submitted data.

For the purpose of creating a population-wide wellness measure, highly detailed individual information is less critical, so a trade off to improve privacy is practical and favorable to adoption. The goal is to collect an overview of the preventative and risk factors affecting the community as a whole, rather than a particular individual. This would act as a complementary area of computational health to that of PHR (Personal Health Record) data collection and analysis [12] which is individual based. This can be achieved through the collection of aggregated totals for coarse location/time details and that goes well beyond what can be easily collected using contemporary survey methodologies. In addition, by keeping detailed data collection and processing limited to the mobile device, calculated fitness measures and trends can be generated without the need to keep a history or additional details about the individual (that could present both a perceived and real privacy risk) on the analysis server.

Examples of calculated measures that though based on private information, can be collectible without significant privacy risk, vary from well known measures such as calorie burn or intake and BMI, to more eclectic measures such as exercise intensity index (standardized index based on either subjective feedback or objective measurements such as HRM) and active transport percentage (proportion of a population that travels by a means that includes physical activity e.g. cycling or walking). More specifically, preventative and risk factors related to community health and wellness can be collected through this approach with protection from re-identification and privacy breaches.

The types of data that can be collected through this approach are quite varied, but some examples that are commonly of interest in previous large scale data collection efforts 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 wellness measure. In current considerations this can be split into the following three major groups:
  - Work Related Activity: The amount and intensity of physical activity completed during work.
  - Recreational Activity: Activity outside that associated with work or transportation.
  - Transportation Activity: Active transportation (walking, cycling or similar) as a form of physical activity is a current area of focus in many regions.
- Caloric Burn and Caloric Intake - This type of data could provide more detailed information as to the overall energy expenditure as related to nutritional intake.
- Nutritional Data - More detailed information could be provided regarding common nutritional issues within segments of the community.
- Body Mass Index (BMI) and change over time - This would allow for both the current snapshot of BMI as well as the individual trend without violating individual privacy.
- Sleep Patterns and Regularity - Sleep patterns are both an indicator and a preventative/risk factor of a number of conditions.

To reduce the risk of potential re-identification of collected data, the avoidance of individual submissions being detected to be from the same submitter needs to be assured. This would require the use of an anonymous submission network, in addition to consideration being given to the type of data requested. However, the collection of some demographic data would be extremely beneficial to the usefulness of this type of wellness measure, further complicating the process of anonymization.

IV. Aggregation of Population-wide Fitness Data

A. Challenges and feasibility for anonymized data collection

The challenges related to anonymized collection of personal wellness data require a threefold approach comprised of mobile device policies, secure auditing and an anonymous submission network. The mobile device policies center around restrictions on the level of detailed data that can be submitted, with only data that is of extremely minimal risk of re-identification based on its content being selected. The component of anonymous submission can be provided by a MIX network [13]. Additionally, the provision for secure auditing needs to be incorporated to assuage the user's privacy concerns. Due to the nature of the type of data that will be collected by these systems, the auditing log can quite reasonably be kept in a human readable format.

A non trivial issue with providing an anonymized data collection framework is detecting either incorrect or deliberately false data submissions. Due to the requirement of not tracking submissions from an individual over time, detecting a series of anomalous data submissions and then removing them is problematic. The most effective methodology is to maintain control over the mobile application by having the submissions authenticate that they are from a valid application without individually identifying the application.

Additionally, user perceptions of the data collection process are a key challenge as even where an extremely secure methodology of de-identification and secure submission is provided, if this is not transparent and communicated to the users, participation and motivation will likely be diminished. The submission of only aggregates in a human readable format will allow the user to see exactly what is being submitted in a understandable way, providing transparency. However, the use of the MIX network that provides submission without disclosing the individual's mobile device identification or other identifying detail related to the communication, would need to be concisely and prominently explained for the benefit of the users.

To accomplish this, users will be explicitly informed on their device that it is not possible or allowable for any of the submitted information to be associated with them personally once submitted.

B. Conceptual Architecture and Framework

The conceptual architecture is comprised of four layers: sensors, mobile application, communications and analysis server as shown in Fig 1. The sensor layer is composed of internal and external sensors that are available to the mobile device. This could include HRM, accelerometer, gyroscope and GPS etc.

The mobile application layer contains the software components required for the functionality of anonymous wellness data collection. The wellness application provides the user interface and experience, and shares information with the secure auditing and reporting policy and anonymization components. The reporting policy and anonymization component aggregates data provided from the wellness application according to data rules and user options, preparing the data in a format that can be submitted with potential re-identification avoided.

The reporting policy and anonymization component sends its processed data through a MIX network [13] to anonymize the reporter's details. The submitted data is then stored and analyzed at the analysis server that has received no identifying data or communications.

Due to its inherent privacy support and guarantees, this system is not designed to support personal intervention. The capability is one for population health data collection only. That is not to say other separate technology-based capabilities may not support an individual's health and healthcare supported via health sensors.

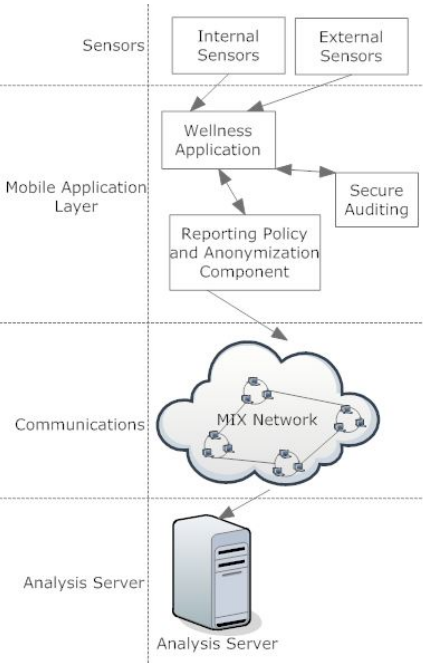

Fig. 1. Anonymized Wellness Data Collection Conceptual Architecture

C. Case examples of Data Collection

In many cases of data collection, individual's anonymity can be assured through k-anonymity [14] or similar approaches, which assume that if an individual can not be identified from a k sized group of submissions, anonymity is assured. However, it typically implies that detailed information be submitted to the trusted server or intermediary and mixed with other data before being used. However, the approach we have suggested in this paper is that through only submitting aggregated and calculated measures of detailed data, the submission can already be at the point of assumed anonymity due to the weakness of quasi-identifiers provided to the third party and the anonymous submission network (in this case a MIX network and the lack of submission history data). The risk analysis of quasi-identifiers remains an ongoing research effort [15].

A prominent example would be Body Mass Index (BMI), as an indicator for a number of chronic diseases. It has potential importance in large scale data collection. The submission of just the BMI calculation (based on mass and height) provides a level of ambiguity of the component values that could more easily be potential quasi-identifiers usable in re-identifying a submitter when taken into account with other demographic details such as coarse location or gender.

Physical activity data is another area where detailed data can be aggregated or calculated to provide meaningful values while also decreasing the sensitivity of the data. While the data would initially be collected for the usage of the individual with detailed time, location and intensity values, through processing a summarization of this data can be submitted. An example would be submitting the total activity recorded over the last week with breakdown into categories of active transportation, work physical activity and recreational physical activity, with perhaps some further detail in relation to the number of days each category was active. This provides potentially statistically important information in regards to preventative and risk factors for the community without collecting data linked to an individual or which can be used to re-identify the user without already having the detailed knowledge of the individual beyond what was submitted. Splitting the different aggregates into multiple submissions and limiting demographic information that can be included can further decrease the chance of re-identification.

Additionally, if technological methods for collecting nutritional data become more mainstream the potential for anonymous submission becomes more attractive. In this case kJ (kilojoule) intake from food could be compared to physical activity kJ values. Or specific nutritional data rules could be focused on, where occurrences of vitamin/mineral deficiencies in dietary intake could be reported on. Additionally, overall trends such as high fat intake, low calcium intake or any other area of interest could be identified. This, along with the ability to track how these values change over time with regards to external influence, would be of great interest. As with other data collections, to re-identify a submission would require detailed information about the individual, and since multiple submissions from the same user cannot be linked, there is no benefit to re-identifying an individual submission.

Above are just a few specific examples. The potential data for numerous different purposes that could be collected is vast.

D. Public Health Interventions

This approach could allow for more sophisticated and targeted public health campaigns to be created, via traditional media or emerging communications [16]. It would not be possible for there to be individually targeted public health interventions as per the privacy guarantees. However, there could be for example targeted public health interventions per suburb or per lifestyle or nutritional area of need, with the further advantage of receiving timely feedback on the effectiveness of the campaign based on continual data collection.

V. Discussion and Conclusion

The key benefits of this approach are the extension of epidemiological and population health data capture to the scale of millions, or potentially population-wide, and for this to be continual and real-time. In addition this is to be implemented in a way, using aggregate measures per person, that does not allow individual identification and hence does not breach privacy or allow individual monitoring. The disadvantages are that in-device and external smartphone sensors will continue to have limits to the scope of health data they can capture for the foreseeable future.

In this paper, we proposed a framework and architecture to facilitate the collection of non-identifying, population-wide wellness measures. This approach increases the ease and capability of gathering quantitative wellness data via smartphones, while achieving a high level of privacy. The core concept is based on increased local processing so that only the required information is submitted with consideration given to avoiding the risk of re-identification. In addition to this, the utilization of an anonymous submission network removes the potential for re-identification through the communication layer.

References

[1] Gartner. (2011) Gartner says worldwide smartphone sales soared in fourth quarter of 2011 with 47 percent growth. [Online]. Available: http://www.gartner.com/it/page.jsp?id=1924314 [Accessed: Apr. 20, 2012]

[2] R. Kwok. (2009) Personal technology: Phoning in data. [Online]. Available: http://www.nature.com/news/2009/090422/full/458959a.html [Accessed: Apr. 20, 2012]

[3] L. Kazemi and C. Shahabi, "Towards preserving privacy in participatory sensing," in *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2011, pp. 328–331.

[4] A. Clarke and R. Steele, "How personal fitness data can be re-used by smart cities," in *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on*, Dec. 2011, pp. 395 –400.

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

[6] R. Steele and A. Clarke, "A real-time, composite healthy building measure drawing upon occupant smartphone-collected data," *in 10th International Healthy Buildings Conference*, July 2012.

[7] Jawbone. (2011) Up by jawbone with motionx technology empowers you to live a healthier life. [Online]. Available: http://content.jawbone.com/static/www/pdf/press-releases/up-press-release-110311.pdf [Accessed: Apr. 20, 2012]

[8] M.-Z. Poh, K. Kim, A. Goessling, N. Swenson, and R. Picard, "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.

[9] G. Villalobos, R. Almaghrabi, B. Hariri, and S. Shirmohammadi, "A personal assistive system for nutrient intake monitoring," in *Proceedings of the 2011 international ACM workshop on Ubiquitous meta user interfaces*, ser. Ubi-MUI '11. New York, NY, USA: ACM, 2011, pp. 17–22.

[10] P.-Y. Chi, J.-H. Chen, H.-H. Chu, and J.-L. Lo, "Enabling calorie-aware cooking in a smart kitchen," in *Persuasive Technology*, ser. Lecture Notes in Computer Science, H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segersthl, and P. hrstrm, Eds. Springer Berlin / Heidelberg, 2008, vol. 5033, pp. 116–127.

[11] J. Lester, D. Tan, S. Patel, and A. Brush, "Automatic classification of daily fluid intake," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on*, March 2010, pp. 1 –8.

[12] R. Steele and A. Lo, "Future personal health records as a foundation for computational health," in *ICCSA (2)*, ser. Lecture Notes in Computer Science, O. Gervasi, D. Taniar, B. Murgante, A. Laganà, Y. Mun, and M. L. Gavrilova, Eds., vol. 5593. Springer, 2009, pp. 719–733.

[13] K. Sampigethaya and R. Poovendran, "A survey on mix networks and their secure applications," *Proceedings of the IEEE*, vol. 94, no. 12, pp. 2142 –2181, Dec. 2006.

[14] "Spatial k-anonymity," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Springer US, 2009, p. 2714.

[15] K. El Emam, "Risk-based de-identification of health data," *Security Privacy, IEEE*, vol. 8, no. 3, pp. 64 –67, May-June 2010.

[16] R. Steele, "Social media, mobile devices and sensors: Categorizing new techniques for health communication," in *Sensing Technology (ICST), 2011 Fifth International Conference on*, 28 2011-Dec. 1 2011, pp. 187 –192.