IC-ININFO

Local Processing to Achieve Anonymity in a Participatory Health e-Research System

Andrew Clarke*, Robert Steele

Discipline of Health Informatics, The University of Sydney, Sydney NSW 2006, Australia

Abstract

The use of participatory sensing in health e-Research applications is rapidly becoming a possibility due to the adoption of mobile computing technologies and sensing platforms. Such a change will have important benefits in the access to near real-time, large-scale up to population-wide data collection and analysis. However, there are numerous issues implied. Primarily of concern is how to ensure anonymity and privacy within these methodologies, and further the related issue of how to incentivize participants and remove barriers/concerns over participation. To address these concerns, in this paper we introduce a novel system to capture aggregate population health research data via utilizing smartphone capabilities while fully maintaining the anonymity and privacy of each individual contributing such data. A key and novel capability of this system is the support for customizable data collection; without the need to know specific details about an individual. The customized collection rules can be deployed on the local device based on detailed local data, and the resultant collection can be measured by the anonymous data collection network. In this paper we provide a conceptual architecture and describe a method for local processing of aggregate e-Research health data utilizing adaptive privacy thresholds to create a multi-party flexible approach to participatory data submission to support this novel health e-Research capability.

Keywords: e-Health, Privacy-preserving, Participatory Sensing, Mobile Health

1. Introduction

The recent growth in capabilities and uptake of mobile devices with sensors, or capable of acting as sensor plat-
forms has the potential to change elements of how health e-Research data is collected and interacted with. However, the need for further consideration of privacy and security issues and the limitations that these impose makes this new methodology challenging. While increasingly, smartphone or sensor-based data capture is used as a technique for individual health data capture or for small-scale deployments, this has not extended into a general health e-Research approach.

In our description of the capabilities of health participatory sensing networks (HPSNs) (Clarke & Steele, 2014) we have developed a broad classification that includes capabilities specific to the health domain, including such aspects as health interventions (Klasnja & Pratt, 2012), a key component of future HPSNs. Further, we have developed an approach that eschews the need for a fully trusted server-based approach that would likely prove impractical on population scale applications, in favour of a framework that utilizes an anonymous communications layer (MIX network (Sampigethaya & Poovendran, 2006) or Onion routing (Mauw et al., 2004)) in combination with de-identification of data submitted to provide anonymous submission/interaction. However, this alone would still incorporate the risk of re-identification based on quasi-identifiers in the form of information about individuals known outside the HPSN that could be used to match with and re-identify the submitted data. The most common approach to this type of risk is to use a trusted server or aggregation point to combine and obfuscate data to the point where k-anonymity (Kalnis & Ghinita, 2009) is achieved, such that any individual is indiscernible from k other records based on quasi-identifiers.

To provide an approach that doesn’t require a trusted server component we propose that a suitable level of anonymity can be provided by locally processing collected data into an aggregated generalized form that can still meet the need of health e-Research data collection. By utilizing quasi-identifier scores (QIS) and a threshold approach to privacy limits, the level of privacy disclosure an individual agrees to can be easily managed without requiring a case-by-case approval. Additionally, our approach involves the specification of and weighting of the data to be submitted to allow the local device to automatically calculate the resolution and breadth of data submitted so as to preserve privacy and anonymity whilst still submitting the data sufficient for health e-Research data purposes.

2. Related work

The interest in utilization of participatory sensing for a myriad of applications has spurred a number of different approaches to resolving or decreasing implicit security and privacy concerns. The more traditional approach would use a trusted server then k-anonymity (Kalnis & Ghinita, 2009) or a variant to anonymize the data before re-use for other purposes. Of course this approach suffers from the need for a fully trusted server as well as issues of a single point of failure in terms of privacy breaches. Other approaches combine a superficial anonymity (removal of identifiers and communications anonymity) with a central point of trust (Cornelius et al., 2008) to provide an anonymous approach. However, this alone is not well suited to a model where quasi-identifiers are a key submission component as de-identifier protection is still implemented at a central trusted point.

Alternatives to these types of approaches include decentralized participatory sensing networks (Christin, 2010) using user interaction/awareness as part of the approach or keeping the data managed by the participant (Mun et al., 2010, Choi et al., 2011) and stringent user-definable access control mechanisms to manage sharing. While these approaches may be extensible to some requirements of HPSNs, the capabilities that are beneficial in other areas may make these approaches overly complex and unapproachable to individuals, limiting their feasibility for a large-scale implementation and deployment.

3. New model for e-Research data collection

The shift to a mobile participatory data collection methodology for research creates a number of possibilities in a variety of health-related research areas. These areas range from public health data collection and intervention (Clarke & Steele, 2014), environmental health (Outram et al., 2010), smart cities (Clarke & Steele, 2011), to biomedical (Steele & Clarke, 2013) and behavioral and social sciences research.

Our previous work (Clarke & Steele, 2014) provided a broad classification of health participatory sensing interaction models and their capabilities, specifically in the realm of collection of anonymized health data and interacting with individuals in the health context in a non-identifying manner. Both core functionalities have
impacts in the field of health e-Research. In this section we will discuss these models and the applicability to e-Research in detail.

3.1. Incidental participatory sensing

Incidental participatory sensing is the lowest tier of HPSN participation, characterised as sensing/participant data collection where that data would have already have been collected by individuals for their own personal use and analysis. As such, this requires a lower level of effort than more interactive forms of participation. This provides a growing and available large dataset for e-Research that in some cases is already publicly available. As the capabilities of sensors and the types of self-tracking increases the potential applications will further evolve.

3.2. Passive participatory sensing

Passive participatory sensing is defined as sensing that requires explicit additional effort to collect data that an individual would not have collected unless participating in a population e-Research health data capture task, but does not attempt to or require changes to the day-to-day behaviour of the individual. For example, this could include the individual using additional sensors that collect data relating to physical activity, diet, heart rate, sleep cycles, environment and biomedical measures specifically to contribute to participatory research data capture. This allows for potentially more complete data collection in comparison to that which is possible through purely incidental data collection.

3.3. Passive participatory sensing with subjective human sensing and feedback

This model combines the potential sensing capabilities of passive participatory sensing with human-sensing capabilities, allowing for large amounts of objective sensing data to be complemented with subjective human-generated data and feedback.

This could easily be implemented through the addition of context-sensitive micro-surveys that are requested and attached to relevant collected sensor data. This would allow for both data that is difficult to record through sensors alone and data that may have been missed to be added to the overall research data collection. In the e-Research field this provides an interesting methodology, especially in research fields that typically utilize survey-based methodologies, as it allows for potential minimization of known bias issues especially related to recall; selection and subjectivity issues by the combination with sensors; and the ability to request data at the most relevant moment.

Human-sensing has great potential as a complement to participatory health e-Research sensing, as it is able to bridge the gap created by limitations of sensing technology to reach the potential data collection goal. However, it would intrinsically increase the effort required to participate and would require a further assessment of motivation, incentives and potential participation levels.

3.4. Active participatory sensing

Active participatory sensing provides inputs to the individual to alter the actions they would have taken while involved in the participatory sensing network, in addition to the sensing capabilities of passive participatory sensing. Active participatory sensing in the health e-Research context has a somewhat different goal to that of many other contexts. While an active participatory model for typical sensing might focus on collecting a more complete data set in terms of spatial/temporal range, health and epidemiology-related active sensing would be more concerned with affecting a health-related action. The instigation to carry out ‘active’ sensing activities could essentially be considered to constitute a public health intervention input. As such, the behaviour change would be to attempt to improve the e-Research sensing data recorded in terms of research or health factors. It is assumed that active participatory sensing would have similar levels of technical sensor capabilities as for passive, with the focus shifted to the potential two-way communication that can be built on sensing data and an inherent feedback loop.
3.5. Active participatory sensing with subjective human sensing and feedback

The final classification comprises the combination of active participatory sensing with human-sensing and feedback. This creates a more complete level of data collection than individual components separately.

Additionally, this model allows for the interaction of the two components adding the capability to give human-sensing or feedback related to a specific intervention. This creates a higher level of capability than available in any of the other models and allows for more complete and useful anonymized research information to be collected.

Fig. 1. Participatory Health e-Research System Architecture

4. Privacy threshold approach to data aggregation

Previous approaches to mobile device data collection have often operated on a trusted server or trusted aggregator model where potentially private details are submitted and privacy/security assured by the server. This can result in considerations of perceived or real privacy risks that can affect levels of participation and actual privacy. Alternatively, other approaches remove extensively all potential quasi-identifiers or other semi-private information to preserve privacy. However, this significantly reduces the types of applications that can be supported, and would be of limited usefulness for e-Research health data collection due to these restrictions. Our
approach proposes that by applying granular and modular restrictions on data collection controlled by the user, perceived and real privacy concerns can be alleviated. To further allow individual control of privacy, the individual also has granular controls of the researchers or public health groups they interact with and in what manner. Additionally, the use of a standardized format for the data requests, health interventions and privacy threshold policies allows the on-device adaptation of an automatically calculated ‘compromise’ data submission level that matches the data request as closely as possible without breaching variable user-defined privacy thresholds.

In this section we will define the overall conceptual model for such an anonymizing health e-Research system, the core categories of data submission components, a format for data requests that allows prioritization of measures/dimensions for submission and a privacy threshold structure to evaluate the requests.

4.1. Conceptual model

The conceptual model for the proposed anonymizing health e-Research system involves one or many central health e-Research servers that are utilized by a variety of e-Researchers and public health groups which 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 as in Figure 1.

There are two primary data transmissions from and to the health participatory sensing server respectively: (i) e-Research data requests are distributed from the server, and (ii) anonymized data collection submissions are sent to the server. The core functionality components of the health participatory sensing server are Data Aggregation, Analysis and Intervention/Data Requests.

4.2. Data submission components/dimensions

The core concept of local processing (on the 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 country of submission rather than postcode, the QIS decreases to reflect the increased generality. The approach also takes into account when multiple quasi-identifiers are submitted together, as such a set of quasi-identifiers have a combined QIS value that is assessed against privacy limitations. The four core data components in determining the combined QIS are: Measures, Location, Temporal and Demographic and are described below.

4.2.1. Measures

Measures are aggregate or calculated values that refer to a specific value to be collected. A data collection can have multiple measures for comparison. Examples of possible population wide anonymized wellness measures are discussed in our previous work (Clarke & Steele, 2012, Steele & Clarke, 2013) and include values such as physical activity patterns and intensity, caloric burn and caloric intake, nutritional data, BMI, sleep regularity and patterns, and biomedical measures. However this is not an exhaustive list and rather just representative of current sensor capabilities on mobile devices. QIS for this component would vary but not be overly high as its potential to allow re-identification would be limited without significant other components.

4.2.2. Location

Location is a useful component where the place a measure occurred is of meaningful use. Examples include places physical activity occurred, and active transport data (where physical activity is combined with commuting/transportation) etc. A fine granularity location would have a high QIS score, while a more general location would be lower. There are many cases of public health data capture where precise location can be sacrificed and the location information used just to establish town or suburb-based variation in particular calculated public health measures.
4.2.3. Temporal

Temporal is the component indicating in which period of time a measure occurred. Often the overall temporal range would be set by the data request, however, to keep the QIS value low keeping the temporal value of the returned result less precise is preferred. For example, a data request could specify all activity for a measure in a month period with temporal values being which day/week the activity occurred.

4.2.4. Demographic

Demographic component includes all the other data in relation to the individual that may be additionally submitted for research purposes for example gender, age, ethnicity etc.

4.3. Data request format

Data submission policies will have two main components. Firstly, core data requirements - typically a particular measure that is being collected. If this is not submittable without breaching a privacy threshold the submission is not possible – but measure values on their own will almost never act to allow re-identification of an individual. Secondly, supplementary data requirements - the additional dimensions which can be submitted alongside the core requirements. To allow for the calculation of the highest level of data that can be submitted without breaching the threshold, the additional dimensions will be weighted by importance and whether less specific data submission is acceptable for a dimension/measure.

![Fig. 2. Privacy threshold structure](image)

The “inclusion weighting” (see sample data request rule) will be used to decide which additional dimensions can be submitted alongside the core data. Where there are required dimensions that cannot be met without reaching a QIS that would breach a privacy threshold, the submission does not occur. In other cases our algorithm will calculate the inclusion of dimensions versus the resolution of data to create the most useful data submission as established by the researchers (based on weightings) that can be achieved. This will determine beyond the inclusion decision, the level of detail that is submitted e.g. for temporal data, reducing the resolution down to a larger time period rather than an exact time could avoid breaching a location/time threshold limit as well as lowering the submission QIS for the overall threshold allowing for more detailed data of other dimensions/components.
Sample data request rule:

```xml
<SubmissionRule>
  <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>
</SubmissionRule>
```

Additionally, shown in Figure 2 is the tree based approach to the privacy threshold structure, where all lower level thresholds as well as the overall threshold cannot be exceeded by a data submission QIS. Apart from the thresholds related to the data components we identified in the previous subsection, there are the additional thresholds of ‘Historic’ and ‘Custom User Defined’. Historic relates to a limitation as to the frequency 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 requestor or to all requesters generally. Finally, the user-defined threshold allows for the limitation of certain contexts or combinations of dimensions that they would like to restrict in addition to the standard thresholds.

4.4. Algorithm for data collection rule processing

On processing of the rule, the data request is adapted to the anonymous submission settings on the local device, as illustrated in Fig. 3. Firstly, a check is done as to whether the required minimum data can be submitted (dimensions with an inclusion weighting greater than 0.5), at the minimal level of precision. Then the level of precision is increased based on the resolution rating until maximum precision or the privacy threshold is met. If there is an additional margin until the threshold at this point, optional dimensions are included based on the formula for optional dimension inclusions that are calculated based on the inclusion weighting and precision weighting giving an optimal inclusion structure. This algorithm is performed for all the lower level thresholds individually then adjusted to meet and balance at the parent node threshold, then adjusted to meet the root threshold and re-balanced (see Fig. 3).

A similar approach is used to evaluate intervention/micro survey targeting weighted features to meet a minimum threshold.
This allows for the previously described privacy rules to be dynamically applied to individual privacy thresholds for specific data or dimensions/measures. This allows an easy to manage system of user-level privacy control that does not remove the usefulness of data for public health research and health e-Research purposes.

4.5. Summary

Through this approach and its anonymizing capabilities a novel e-Research capability and methodology is achieved. This allows raw detailed data to be collected and made available from mobile devices - potentially to be used for a myriad of research purposes. With de-identification and anonymization performed for each request, privacy can be provided without strictly limiting the content and applications of data or requiring a trusted server or data collector.

The potential benefit of such an approach is to avoid the common issues associated with duplication of data collection, data siloing, limited data sets and limited re-usability of collected data.
Additionally, the incorporation of public health interventions (that could also be used in e-Research) in addition to data collection allows for the scale of the functionality and the types of stakeholders in the system to be broadened. Intrinsically, a key novel capability provided is the ability to target a specific intervention or other communication to an individual based on detailed, locally collected data without infringing on individual privacy or in any way identifying the individual to the HPSN.

5. Conclusion

This paper presents a health e-Research system and approach based on local processing of aggregate population health data that utilizes privacy thresholds and a compromise approach to data submission that supports the data collection model and builds upon HPSN capabilities. To this end we have discussed the application of health participatory sensing models to provide novel e-Research capabilities, an approach to submission rules/health intervention rules that allows a compromise between individual privacy and e-Research application requirements and an algorithmic approach to computing QIS to compare to privacy threshold values. Additionally, we have identified the core components of HPSN data submission and described an overall conceptual model for this submission.

References


