Distributed Sensor Data Computing in Smart City Applications

Wei Wang*, Suparna De†, Yuchao Zhou†, Xin Huang*, and Klaus Moessner†
*Department of Computer Science and Software Engineering, Xi'an Jiaotong Liverpool University, China
†Institute for Communication Systems, Electrical and Electronic Engineering Department, University of Surrey, UK

***Abstract*—With technologies developed in the Internet of Things, embedded devices can be built into every fabric of urban environments and connected to each other; and data continuously produced by these devices can be processed, integrated at different levels, and made available in standard formats through open services. The data, obviously f a form of "big data", is now seen as the most valuable asset in developing intelligent applications. As the sizes of the IoT data continue to grow, it becomes inefficient to transfer all the raw data to a centralised, cloud-based data centre and to perform efficient analytics even with the state-of-the-art big data processing technologies. To address the problem, this article demonstrates the idea of "distributed intelligence" for sensor data computing, which disperses intelligent computation to the much smaller while autonomous units, e.g., sensor network gateways, smart phones or edge clouds in order to reduce data sizes and to provide high quality data for data centres. As these autonomous units are usually in close proximity to data consumers, they also provide potential for reduced latency and improved quality of services. We present our research on designing methods and apparatus for distributed computing on sensor data, e.g., acquisition, discovery, and estimation, and provide a case study on urban air pollution monitoring and visualisation.**

***Index Terms*—Internet of Things, Distributed Intelligence, Sensor Data, Sensor Services, Smart city**

I. Introduction

OUR understanding towards the Internet of Things (IoT) has been constantly evolving. Around ten years ago, the focus was mainly on Things' traceability and accessibility based on RFID tags. The IoT was described as a world-wide network of interconnected objects that are uniquely addressable. Later, as the number of heterogeneous objects became extraordinarily large, the research then focused on interoperability, representation, and abstraction of Things' capabilities from the "Semantics" and "Service" oriented perspectives [1], [2]. Over the years, we have witnessed the emergence of many IoT applications described as "smart" or "intelligent" (e.g., smart city, smart office, intelligent transportation). This was the case until recently, when researchers started rethinking about the question "what really makes an IoT application smart or intelligent?".

The answer is not surprising - "data", especially big data, which sweeps many of the research fields in recent days. The data-centric perspective views data as the most valuable asset in creating smart applications. One of the exciting research directions in this line aims to exploit insights from large amount of data through big data analytics. However, as the sizes of the IoT data continue to grow with increasing velocity, it becomes infeasible to transfer all the raw data to and process it at a centralised data centre. As an example, consider the case in which all the sensor data, which is potentially of low quality (e.g., noise or missing data), is transmitted to the data centre for processing. The big data processing platform at the centre needs to start many standard pre-processing tasks (e.g., data cleaning, integration and abstraction) before performing the analytics (e.g., map-reduce tasks and data mining algorithms). This process, especially the pre-processing, is time-consuming and costly, and usually results in high latency to service consumers. Wouldn't it be better if most of the pre-processing can be performed in a highly distributed way and in close proximity to service consumers?

The recent development on Fog Computing [3] and Mobile-Edge Computing (MEC) [4] enables cloud computing capabilities at the edge of a (mobile) network. These edge clouds or fogs have elastic resources for distributed data processing that do not suffer from the drawbacks of a traditional cloud architecture. The facilities provide the needed platform to perform intelligent computation in more efficient ways and enable applications and services with reduced latency and improved quality of services. It should be noted that the computation can also be performed on the other local autonomous units on the IoT, such as sensor network gateways or smart devices. In what follows, we present our recent research on designing methods and apparatus for distributed sensor data computation in the context of smart cities and demonstrate a case study on urban air pollution monitoring and visualisation.

II. Data Computing Architecture

In a nutshell, for sensor data consumers, there are two general ways to acquire data. The first is to discover the sensor services which can provide the needed functionalities and then to subscribe to those services. The second is to directly search for sensor data in streaming databases located in edge clouds. The two paradigms have their own respective advantages; while the first one allows the creation of loosely coupled IoT applications through service discovery and composition, the second is more suitable for applications that perform direct data processing and analytics.

As shown in Figure 1, the two types of data acquisition services can be seamlessly integrated into an edge cloud, which can be used as the fundamental platform to implement

978-1-5386-2723-5/17/$31.00 2017 IEEE

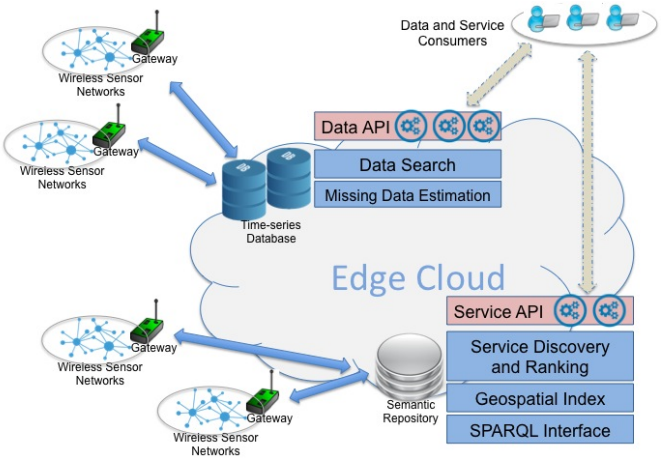

Fig. 1. Integrating data and services in an edge cloud infrastructure

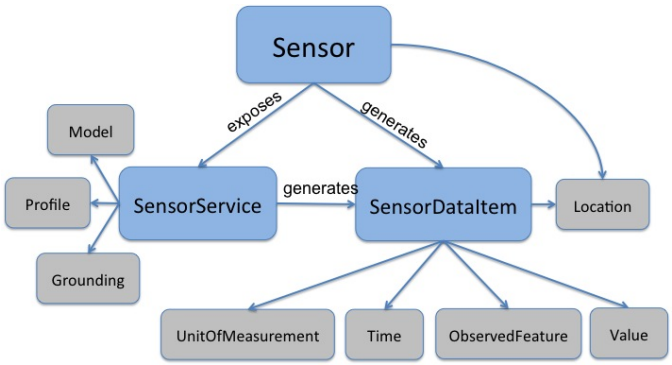

Fig. 2. Simplified semantic model for sensor and sensor Data

data services and distributed intelligence. Storage is one of the core functionalities offered by the edge clouds, for example, to store metadata or semantic descriptions of the sensor services as well as the observation and measurement data collected from sensors. The edge clouds can provide sensor service discovery functionalities if sensors are exposed as services (e.g., conventional Web services or REST services). Streaming data from mobiles sensors or smart phones can be directly stored in a storage facility. The edge clouds also can provide search or discovery functionalities according to the data consumers' requests. Quality of the sensor data (e.g., missing values) is a prominent issue to be considered for data consumers. Section VI shows how intelligent processing methods can be implemented within the distributed framework to ensure data quality for further analytics.

III. Semantic Modelling and Representation

The importance of using semantic technologies and service-oriented architecture for IoT has been well recognised by the IoT community. Given the fact that the number of highly distributed and heterogenous "Things" connected to the Internet increases rapidly each year, providing interoperability and scalability becomes a fundamental requirement to support object representation, discovery, data integration, storage, and analytics. On one hand, semantic modelling provides rich metadata for "Things" and a common basis for interoperability among different "Things"; on the other, abstracting "Things" functionalities as Web services offers homogeneous and scalable ways to access their functionalities.

We have developed ontological models for sensor services and sensor data in the EU FP7 IoT.est project (http://ict-iotest.eu/iotest/), which allow to generate fine-grained semantic annotations, and to create meaningful linked sensor data for service and data discovery. A simplified semantic model is shown in Figure 2, the objects (e.g., Sensor, SensorService or SensorDataItem) can be described not only with datatype properties and literal values (e.g., ID, name, isMobile and description), but also with object type properties by linking to other objects (e.g., Location, and SensorType). A "Sensor" object can be described using the existing Semantic Sensor Network Ontology (http://purl.oclc.org/NET/ssnx/ssn) and exposes its functionalities as a "SensorService". A SensorService is an abstract representation of the sensor capabilities and can be accessed using standard Web service interfaces. Both "Sensor" and "SensorService" can generate "SensorDataItem", which in turn can be semantically annotated according to the ontological model, by using Location, Time, UnitOfMeasurement, value and a semantic reference to the observation feature, e.g., Temperature defined in the Climate and Features ontology (https://www.w3.org/2005/Incubator/ssn/ssnx/cf/cf-feature). The lightweight semantic model for sensor data is particularly suitable for sensor data from mobile sources whose spatial and temporal properties keep changing constantly.

IV. Service Discovery and Ranking

Service discovery is a powerful apparatus for data consumers to obtain sensor data. Adding semantic annotations to sensor services and publishing them as linked data enable structured queries and semantic reasoning. However, this also introduces challenges to the discovery process. For example, the semantic description data should be stored in distributed repositories (preferably in close proximity to the sensors) and the discovery mechanism should be able to find sensor services efficiently on an extremely large scale. This is fundamentally different from the discovery of Web services which is mostly performed in a centralised fashion. In fact, implementing a centralised storage with large-capacity for semantic description data is trivial with the current technologies; nevertheless, maintenance of the data and handling frequent updates are inefficient, time-consuming and error-prone as the operating environments of Wireless Sensor Networks (WSNs) are highly dynamic and the sensors themselves are constrained in processing capabilities and energy.

With these challenges in mind, we have developed a semantic sensor service discovery platform based on spatial indexing and semantic search [5]. The rationale behind this is that sensor discovery in typical IoT scenarios is location dependent and the geographical information can be exploited

TABLE I
COST COMPUTATION FOR DIFFERENT SENSOR SERVICES

Service ID	service_103	service_134	service_52	service_49
Sensor ID/name	55/G15	4/E17	26/U42	21/U46
Importance	2.1065	0.9387	0.9188	2.5582
Energy	0.56	0.96	0.1	0.18
Link quality summary to gateway	0.1681	0.3151	0.2536	0.2091
Link quality summary from gateway	0.1607	0.3647	0.2357	0.1973
COST	30.0283	**8.7136**	72.0554	39.7263

to effectively reduce the search space. The spatial index also allows for approximate search, i.e., find services nearby that provide same or similar functionalities as required by the query. This is particularly suitable for sensor service discovery as it is unlikely that service consumers are able to describe the search needs in exact ways.

The platform can be seamlessly integrated into an edge cloud as shown in Figure 1, in which the discovery process is depicted as several functional components. Semantic description data about all the sensor services is stored in a semantic repository, which can be queried with the SPARQL language (https://www.w3.org/TR/rdf-sparql-query/), a special purpose language designed for the semantic Web and the Linked Data. Instead of indexing the geographical information of each individual sensor service, our method indexes the gateways of WSNs. The advantage is that frequent changes about the sensors can be constrained locally and do not propagate to upper levels in the index. The discovery procedure takes a SPARQL query and first searches against the spatial index. The index returns one or more WSN gateways with which the sensors are associated. The discovery process then queries the semantic repository and retrieves a list of sensor services. Finally, the addresses (e.g., URLs) of the services are returned to the requesters, who will be able to directly subscribe to the services. As the search space is significantly reduced with the spatial index, the semantic query can be performed in efficient ways.

Frequently, instead of returning all the discovered service URLs to the consumers, we need to choose one or a few "best" sensor services from the discovery results in many applications, such as automated service composition, runtime adaptation or data integration. Existing methods exploit either the semantic descriptions or quality of service information for ranking; nevertheless, this information might not be always available. We have developed a novel ranking method by looking into the WSNs and estimating the cost of accessing sensor services [6]. The computation is performed on the gateway of a WSN. In general, requesting a sensor service involves data communications with several other sensor nodes in the network (e.g., those serve as the relay nodes). The process generates cost, measured in the unit of energy consumed by all nodes in the communication loop.

The ranking is performed using the contextual information of sensor nodes extracted from the WSN, for example, energy level, importance and link quality summary. Not all nodes in a WSN are equally important and the term "importance" in this context measures how important a node is to the WSN as a whole. For example, an important node might frequently act as the relay node for communications between the gateway and other nodes. Low energy of an important node may signify that the WSN is in a dangerous state of breaking down. The gateway continuously observes the communication patterns within the WSN and infers an overlay topology for the WSN based on its best knowledge. The overlay topology is then used to infer the importance based on the random walk model. The link quality summary measures the quality of the communication paths between the gateway and a node, and is used to calculate the most probable communication path. The cost is the sum of cost incurred at all nodes in a service access loop. Low cost implies less energy consumption for the WSN and would be able to prolong the WSN lifetime. In Table I, a query for temperature at a specific location returns four sensor services, $Service\_103$, 134, 52 and 49. The method infers that querying $Service\_134$ would incur the least cost, 8.7136, and would be ranked as the best in this case. The evaluation shows that the method has great potential in preserving the energy of the WSN.

V. DATA QUERY AND SEARCH

Compared to service discovery and subscription, searching in time-series databases can directly retrieve sensor data. It is suitable for data from both sensors installed in fixed locations and mobile sensors (e.g., smartphones and sensors installed on public transportation systems for opportunistic sensing). Sensor data has been described as frequently updated, timestamped and structured data (FUTS) [7], which change rapidly along not only the temporal dimension, but also the spatial dimension.

The FUTS data can be described according to the lightweight semantic model shown in Figure 2 and stored in a time-series database in an edge cloud. Historical and near real-time data can be queried by calling the data retrieval APIs. We have designed such a data search method based on a number of searching criteria, e.g., location of interest, observed features, and spatial extent can be specified to perform various queries within a desired time window [8]. Spatial extent can be specified in terms of geographical regions of interest or distances from a specified point. Aggregation operations (e.g., minimum, maximum or average) on the data values are also supported.

VI. IMPROVING DATA QUALITY USING REGRESSION ANALYSIS

Data quality is always an issue in sensor data or big data. For example, we found that the sensor data on pollutants collected from the London Air Quality Network project (http://www.londonair.org.uk/LondonAir/Default.aspx) over a 1-year period contains many missing and incorrect values. These missing or incorrect values obviously cannot be used for data analytics in smart city applications. Therefore, we need to estimate the missing and incorrect values in order to

TABLE II
INFORMATION ABOUT THE SENSING SITES AND AIR QUALITY MEASUREMENTS

Sensing Sites	Sensing Area	Measurements
Islington-Arsenal	Urban background	NO, NO2, NOx, PM10
Islington-Holloway Road	Roadside	NO, NO2, NOx, PM10
Haringey-Priory Park South	Urban background	NO, NO2, NOx, O3

improve its quality. We apply a well-known machine learning technique, Support Vector Regression (SVR) to predict the missing values and compare its performance in terms of accuracy to the state-of-the-art techniques that employ Locally Weighted Regression (LWR) [9]. The data discovery mechanism described earlier is applied to select the data points for training. LWR calculates the Euclidean Distances between records in the training set and the query (i.e., inputs to predict a missing data point). Instead of training a regression model using the entire training set, it selects the k-nearest data points for training. SVR selects the training data in a similar way and applies a kernel function to map inputs of the training set into a new space.

Experiments are performed based on the air quality data collected from three sensing sites from the London Air Quality Network. The details of three sensing sites, areas and the air quality measurements are shown in Table II.

The measurements include Nitric Oxide (NO), Nitrogen Dioxide (NO2), Oxides of Nitrogen (NOx), Ozone (O3), and PM10 Particulate (PM10) in different places. All of them are measured in the unit of ug/m3. Data points are daily mean values in the year of 2015. By listing each measurement in each sensing site as a column, a dataset is created with 365 rows and 12 columns (in total 4,380 cells), which contains 179 missing data points originally. Ten test datasets are generated by randomly creating up to 30% missing data points from the original dataset.

In the preliminary study, NO2 in Arsenal is chosen as the output of the regression model and the rest of the columns are used as input. Mean Squared Errors are calculated between actual and predicted values. The results are plotted in Figure 3, in which SVR shows an overall better predicted accuracy. Its performance is also more stable compared to LWR, which implies that the SVR can better handle overfitting.

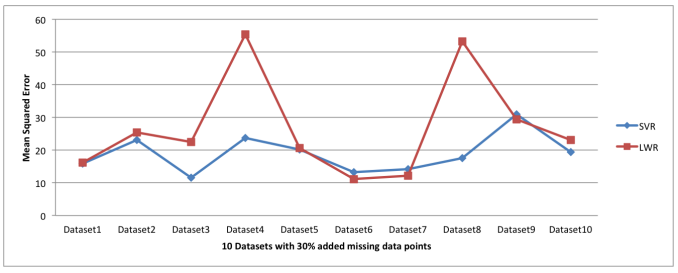

Fig. 3. Comparison of prediction performance between Support Vector Regression and Locally Weighted Regression

VII. CASE STUDY: URBAN AIR POLLUTION

We apply the aforementioned techniques for distributed intelligence in a case study on urban air pollution. Many urban areas have built air quality tracking stations that monitor pollutant levels. Exposure to large quantities of these pollutants introduces severe health threats. According to the report of the World Health Organisation, 2.4 million deaths annually are directly attributable to air pollution [10]. It has become imperative to effectively monitor and predict pollution levels to allow citizens to better plan their daily activities and for city authorities to take rapid remedial actions.

An urban air quality monitoring application for mobile phones has been developed based on the methods for sensor data acquisition, discovery and missing value estimation. The pollutant data is collected over a period of 1 year (January to December 2015) from 3 sensing sites in central London (as described in Section VI). The application provides a near real-time visual clue to users about the air quality in different urban areas. In a larger-scale environment monitoring system, local and regional data can be aggregated and mined to provide timely feedback especially in case of an emergency such as a toxic pollution alert.

The Air Quality Index (AQI) is used to rescale the pollutant concentrations from 1 (low-risk) to 10 (hazardous), as it is not easy to assess health risks by comparing pollutants directly. We calculate the AQI values based on the model designed by the Department of Environmental and Food Regulatory Agency, UK. The results together with the corresponding monitoring site?s geographical information are visualised as a heat map, as shown in Figure 4. Pollution levels are depicted with coloured circles, with larger radius denoting higher pollution level.

VIII. CONCLUSION AND FUTURE RESEARCH

Given the unprecedented, ever-increasing amount of data on the future IoT, even the largest data centres with the state-of-the-art big data processing techniques will face serious challenges in terms of processing efficiency. The idea of distributing intelligent computation to local resources has the potential to alleviate these challenges and provides a foundation for application-independent data processing. The computation inside an edge cloud or "fog" helps reduce the data sizes and provides high quality data for further data analytics. The work presented in this paper primarily focuses on sensor data collected from WSNs; however, the paradigm is also applicable to many other types of data in smart city applications. The intelligent processing at the local resources is not limited to those discussed here. Other data integration and abstraction methods can also be implemented to further reduce the sizes of the raw data and to produce useful information granules at various levels of resolution, e.g., events, irregularities, anomalies and patterns, which can be used in high level analytical applications.

Following the vision of truly smart cities, we envisage three important future research directions. The first is to develop knowledge representation method for data and information granules. Semantic technologies has been shown successful in

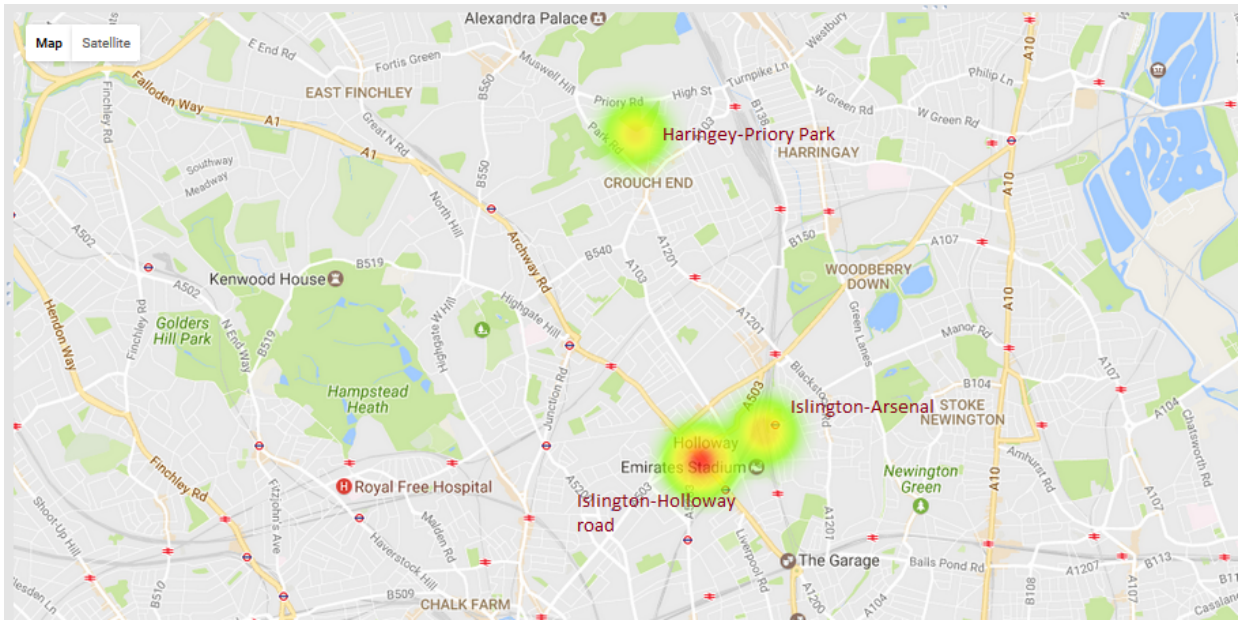

Fig. 4. Map visualisation of different values of Air Quality Index in London

knowledge representation for the Internet of Things. However, knowledge representation for different information objects in smart city application is much more complex and needs to capture the variety of data types, formats, resolution of information granules. The second is develop knowledge discovery techniques for big smart city data analytics. Current methods for knowledge discovery and data mining in smart cities mostly focus on data from either the physical world, social world, or individual domains, however, a smart city should be viewed as an inseparable organism and data correlations among different city domains need to be discovered based on the data collected from the cyber, physical and social worlds. This has the potential to uncover the concealed insight beneath the data, and provide useful knowledge for human understanding in order to better monitor, plan and regulate our city lives. The third is how to evaluate the trustworthiness of derived insights in smart cities through continuous analytics. Given the fact that smart city data is usually noisy, dynamic and of low quality, credibility of the discovered knowledge needs to be evaluated, by exploiting the different data sources from the cyber, physical and social worlds. Temporal factors need to be taken into consideration to perform continuous analytics in order to refine the developed trust model.

ACKNOWLEDGMENT

This work is supported by the collaborative European Union and Ministry of Internal Affairs and Communication (MIC), Japan, Research and Innovation action, “iKaaS” under EU Grant number 643262.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” *Computer Networks*, vol. 54, pp. 2787–2805, 2010.

[2] D. Guinard, V. Trifa, S. Karnouskos, P. Spiess, and D. Savio, “Interacting with the SOA-Based Internet of Things: Discovery, Query, Selection, and On-Demand Provisioning of Web Services,” *IEEE Trans. Serv. Comput.*, vol. 3, no. 3, pp. 223–235, Jul. 2010.

[3] Datta, Soumya Kanti and Bonnet, Christian and Härri, Jérôme, “Fog computing architecture to enable consumer centric Internet of Things services,” in *19th IEEE International Symposium on Consumer Electronics, ISCE 2015*, 06 2015, pp. 1–2.

[4] ETSI, contributing organisations: Huawei, IBM, Intel, Nokia Networks, NTT Docomo, Vodafone, “Mobile-Edge Computing - Introductory Technical White Paper,” 2014.

[5] W. Wang, S. De, G. Cassar, and K. Moessner, “An experimental study on geospatial indexing for sensor service discovery,” *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3528–3538, 2015.

[6] W. Wang, F. Yao, S. De, K. Moessner, and Z. Sun, “A Ranking Method for Sensor Services Based on Estimation of Service Access Cost,” *Information Sciences, Elsevier*, vol. 319, pp. 1–17, 2015.

[7] Y. Qin, Q. Z. Sheng, N. J. Falkner, S. Dustdar, H. Wang, and A. V. Vasilakos, “When things matter: A survey on data-centric internet of things,” *Journal of Network and Computer Applications*, vol. 64, pp. 137 – 153, 2016.

[8] Y. Zhou, S. De, W. Wang, and K. Moessner, “Enabling query of frequently updated data from mobile sensing sources,” in *The 13th IEEE International Conferences on Ubiquitous Computing and Communications (IUCC2014)*. IEEE, 2014, pp. 946 – 952.

[9] H. Kurasawa, H. Sato, A. Yamamoto, H. Kawasaki, M. Nakamura, Y. Fujii, and H. Matsumura, “Missing sensor value estimation method for participatory sensing environment,” in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*. IEEE, 2014, pp. 103–111.

[10] World Health Organisation, “Global Health Risks: Mortality and burden of disease attributable to selected major risks,” 2009.