PMViewer: A Crowdsourcing Approach to Fine-Grained Urban PM2.5 Monitoring in China

ChengYi Zhang†‡, Yazhe Wang†‡‡, Peng Liu§, Tao Lin§, Lvgen Luo†‡, Ziqi Yu†‡, Xinwang Zhuo†‡
†State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
‡School of Cyber Security, University of Chinese Academy of Sciences
§College of Information Sciences and Technology, Pennsylvania State University, University Park
{zhangchengyi, wangyazhe, luolvgen, yuziqi,zhuoxinwang}@iie.ac.cn

Abstract—In recent years, the PM2.5 (particulate matter with a mean aerodynamic diameter of 2.5 micrometers or less) pollution has become a very serious problem in China. Currently, there are three types of monitoring approaches: government-led monitoring, Wireless Sensor Networks (WSN) approaches and Participatory Urban Sensing (PUS). There are three limitations in these state-of-the-art research approaches: a) the coarse-grained limitation of government-led monitoring, b) the deployment and maintenance cost of WSN approaches, c) the “Black Hole” problem and the “Black Time Window” problem of PMTI-based approaches in PUS. How to overcome these three main limitations is the biggest challenge. To address these limitations, we need a new way to collect PM2.5 data. Nowadays, IoT (Internet of Things) smart devices sold to various customers could steadily and directly collect the PM2.5 data in vast urban areas, but how to address these two challenges, how to address them is an open problem. In this paper, we propose PMViewer, a novel PUS approach to address these two challenges. PMViewer’s data are collected from tens of thousands of smart devices called AirBox through a crowdsourcing approach. We aim to offer a fine-grained spatial-temporal resolution for the public to monitor the urban PM2.5 pollution near their locations. PMViewer scrawls data from AirBox’s vendor server and parses the data to generate a map view to display real-time urban PM2.5 measurements. In this study, we design, implement and evaluate PMViewer. Evaluation results show that PMViewer efficiently and economically addressed these two challenges described above.

I. INTRODUCTION

In recent years, air pollution especially the PM2.5 pollution has reached alarming proportions in China. In Feb 2014, China banned outdoor school sports and cookouts as it grappled with a fourth straight day of thick, choking smog [1]. In Dec 2016, smog choked China so badly that planes in Beijing even could not land [2].

As a result, air quality monitoring especially PM2.5 has attracted great attention. Currently, there are three types of monitoring approaches: government-led monitoring, Wireless Sensor Networks (WSN) approaches and participatory urban sensing (PUS).

Government-led monitoring basically includes the satellite remote sensing and static measurement stations [14]. The results inferred from satellite images only reflect the air quality of the atmosphere rather than the ground air quality. The deployed static measurement stations are coarse-grained and require large and expensive facilities [13]. Wireless Sensor Networks (WSN) measure the air quality through numerous small and low-cost sensors. These approaches monitor emissions such as PM10 and NOX while finer detection like PM2.5 requires additional precise sensors, which incur extra deployment and maintenance cost.

Participatory urban sensing (PUS) [11] mainly include two categories: a) Collecting air quality data by utilizing participants’ mobile smartphones with additional sensors. AirCloud [13] uses front-ends sensor connected to the smartphone via Bluetooth 4.0. However, these approaches lack incentives because extra sensors are needed and they entirely rely on the participants to generate data. b) Utilizing the public mobile transportation infrastructures (PMTI) to detect urban air quality. Mosaic [14] and Srinivas et al. [12] use sensing devices deployed on buses to monitor the PM2.5 situation.

Generally, there are at least two inherent limitations of current PMTI-based approaches in PUS: i) The “Black Hole” problem: the public vehicles are running on limited and fixed transportation routes with constrained coverage. As each sensor on PMTI has a sensing range, each monitor area can be regarded as a cell. Along with the movement of PMTI, there are some zones which are covered by these cells. On the contrary, there are still some remaining zones not belonging to any cells. Thus, there are no measurement data in these zones which defined as “Black Hole”. ii) The “Black Time Window” problem: As PMTI are not fixed in locations, the time interval of data updating depends entirely on the next public vehicle coming to this location. It brings the “Black Time Window” problem when some time intervals do not receive any data.

Overall, there are three limitations in former PM2.5 monitoring mainly: a) the coarse-grained limitation of government-led monitoring, b) the deployment and maintenance cost of WSN approaches, c) the “Black Hole” problem and the “Black Time Window” problem of PMTI-based approaches in PUS. As a result, in the area of PM2.5 monitoring, the biggest challenge currently is described below: Challenge 1: How to overcome these limitations and offer a low-cost and fine-grained spatial-temporal resolution.

To address these limitations, we need a new way to collect
PM\textsubscript{2.5} data. Nowadays, with the development of IoT (Internet of Things), a plenty of IoT smart devices sold to various customers measure the air quality. These devices together could steadily and directly collect the PM\textsubscript{2.5} data in vast urban area. Whereas, it introduces new challenge as:

**Challenge 2**: How to obtain wide-spread real PM\textsubscript{2.5} data from tens of thousands of smart devices.

To the best of our knowledge, no existing work has addressed these two challenges described above, and how to address them is an open problem. In this paper, we propose PMViewer, an innovative crowdsourcing approach in PUS. Air quality data in PMViewer are collected from tens of thousands of smart devices called AirBox. Originally, AirBox was designed for individual use rather than for the public urban sensing. However, through our research on AirBox, we found a new insight that AirBox can be used for urban sensing. We found that AirBox leveraged the IoT technology to send data to the server of the vendor (we name it as H server) automatically. The H server can be queried by the public and we used this feature in our crowdsourcing approach. PMViewer scrawls data from the H server and parses the data and generate a map view to display real-time PM\textsubscript{2.5} measurements.

**Our contributions**:

- We propose a crowdsourcing approach to obtain urban PM\textsubscript{2.5} data from tens of thousands of smart devices.
- We design, implement and evaluate PMViewer, a low-cost and fine-grained approach for urban PM\textsubscript{2.5} monitoring. PMViewer can overcome the main limitations in former PM\textsubscript{2.5} monitoring approaches.
- We bridge the gaps between the original capabilities of the AirBox system and our goals. Meanwhile, anonymization on sensitive locations is conducted to protect the privacy of AirBox users.

We implement PMViewer as a public welfare website and it is available on http://www.ocaptains.com/AirBoxShow/. PMViewer is a non-profit platform for the environmentalists and the public who are concerned about the air quality.

**II. BACKGROUND AND PROBLEM STATEMENT**

AirBox is one of the smart devices designed to monitor household air quality including PM\textsubscript{2.5}. It is controlled by users with mobile app through the Internet. After the AirBox has been activated, it will always online and upload data to the H server regularly. Users can check their household real-time air quality on the app. In addition, the H server offers a function that users can scan and view another three AirBoxes’ data near their locations.

In this paper, we propose PMViewer in PUS to address these challenges above and our goals in PMViewer are described as follows: a) For everyone in China, he/she can easily query the PM\textsubscript{2.5} nearby their homes or offices, no matter if they bought AirBox or not. b) The cost for monitoring urban PM\textsubscript{2.5} should be as low as possible. c) PMViewer should have a high spatial and temporal resolution compared to government-led monitoring and PUS approaches.

**III. APPROACH OVERVIEW**

PMViewer is aimed at addressing the challenges and meeting the goals above. The architecture of PMViewer is shown in Fig. 1. Behind PMViewer, a server named as P server was built which interacts with the H server and offers a website to the public for monitoring urban PM\textsubscript{2.5} pollution.

There are three parts or steps inside the P server: First, the P server obtains data from the H server through a crowdsourcing approach. Second, it conducts anonymization on sensitive locations to protect user privacy. Third, cartographic visualization processing is designed for public welfare website. Along with the P server in PMViewer, Challenge 2 will be addressed in the first and the second step and Challenge 1 will be addressed after the third step. A brief overview of these steps is described below.

**Step 1: Crowdsourcing Approach on Data Acquisition.** As mentioned in Section II, the H server offered a function that users could view nearby AirBoxes’ data. Our breakthrough of obtaining data from the H server was derived from this function. Specifically, thousands of ‘virtual AirBoxes’ were generated throughout the nation and each virtual AirBox scans actual AirBoxes nearby as much as possible. Based on this step, PMViewer breaks through the boundaries of an individual AirBox and obtains nationwide urban AirBoxes’ data.

**Step 2: Anonymization on Sensitive Locations.** Data obtained from the H server contain locations (latitude and longitude coordinates) of the AirBoxes’ users are greatly related to user privacy. The P server implements it through a two-phase location offset process and visualization protection.

**Step 3: Cartographic Visualization Processing.** It could be convenient to the public if they can browse urban PM\textsubscript{2.5} situations online. We realized cartographic visualization on Baidu Map [10]. Locations of AirBoxes are marked on the map and citizens can view this map and zoom in on selected areas including their own locations. After this step, anyone having access to the Internet can view the PM\textsubscript{2.5} situation nearby their homes or offices.

**IV. CROWDSOURCING APPROACH ON DATA ACQUISITION.**

Based on AirBox, users can scan and view data of nearby AirBoxes. We name this function as function-scan. It is based on HTTP requests and responses between the app and the H server. Basically, each response is limited to three other AirBoxes’ data. Our study found that carefully constructed HTTP requests can break these limitations. We analyzed the
contents of HTTP requests and found that every request contains four parameters mainly: deviceId, lat, lng and count. In these parameters, deviceId refers to AirBox identification, lat and lng together refer to the geographical location of AirBox, count refers to the request number of other AirBoxes nearby. If we want to leverage the function-scan to acquire data of urban AirBoxes, we need to buy tens of thousands of Air Boxes and place them in different urban areas throughout the nation. It is highly impractical, however, we discovered that deviceId, lat and lng are the meta data to differentiate different AirBoxes in the H server. If we change these parameters’ value, tens of thousands of ‘virtual AirBoxes’ can be created. Meanwhile, we found that the value of count is the constant 3 which limited the amount of returned data of AirBoxes. If we design the function-scan with virtual AirBoxes’ meta data value and the value of count carefully, a crowdsourcing approach could obtain numerous AirBoxes’ data.

The overview of crowdsourcing approach is shown in Fig. 2. We intend to generate thousands of nationwide distributed ‘virtual AirBoxes’ in different urban district centers. Then, the function-scan is executed with the value of count at a big number which leads to a large scan range. Due to the large scan range, duplicated data exist and deduplication is needed.

V. ANONYMIZATION ON SENSITIVE LOCATIONS

To protect the user privacy, we offset the actual locations. We name the original actual location as actual-location, and name the location after offset as offset-location. They are both in the form of latitude and longitude coordinates, and offset-location shares the same PM2.5 value with the corresponding actual-location. We propose two basic rules: (i) The offset-location is within a certain distance away from the actual-location. (ii) The offset-location still represents the PM2.5 situation close to the actual-location.

In the P server, we generate the offset-location with the leverage of Geocoding WebService [4]. The overview of actual location offset processing is shown in Fig. 3. This Geocoding WebService offers two basic functions: geocoding and reverse-geocoding. Geocoding means converting the detailed addresses into latitude and longitude coordinates. On the contrary, reverse-geocoding means converting latitude and longitude coordinates into the detailed addresses. We execute the offset processing through the following two steps:

(i) Utilizing reverse-geocoding: The P server converts the actual-location into the detailed address. The requests in reverse-geocoding are actual-location and the responses are the detailed addresses. The detailed address consists of the address_component and the semantic_description. The address_component is the basic address including country, province, city, district, street and street number. The semantic_description is the description of the POI. POI (Points-of-Interest) are some specific points that someone may find useful or interesting on the electronic map around the user’s location. After this step, the P server can get the detailed address of the actual-location, even the sensitive street number is available. We reconstruct the detailed address with the street number removed and the semantic_description reserved.

(ii) Utilizing geocoding: The P server converts the reconstructed addresses into latitude and longitude coordinates (the offset-location). The requests in geocoding are the reconstructed addresses. The responses contain latitude and longitude coordinates which are the offset-location. Until now, the P server obtained the offset-location. This location differs from the actual-location because the street number have removed. In addition, when geocoding is execute in the second step, the semantic_description is included to ensure that the offset-location is not too far from the actual-location.

VI. EVALUATION

A. How about the Spatial Resolution of PMViewer?

1) Comparing with Government-led Monitoring: PMViewer offered more monitor nodes than the government-led monitoring approaches. For example, in Fig. 4, the number of active AirBoxes of PMViewer in Beijing are nearly 10 times larger than that of government-led monitoring [6]. To quantify the granularity, we use Average Nearest Neighbor (ANN) as the metric. There are three parameters: ANN Ratio, observed mean distance (DO) and expected mean distance (DE) [3].

$$\text{ANN Ratio} = \frac{D_O}{D_E} = \frac{\sum_{i=1}^{n}d_i}{n} \quad D_E = \frac{0.5}{\sqrt{n/A}} \quad (1)$$

In this equation, di represents the distance between a point i and its nearest neighboring point, n is the total number of points, A is the area of the minimum enclosing rectangle around all points. If the ANN Ratio is less than 1, the pattern exhibits clustering, otherwise the trend is toward dispersion. We analyzed ANN Ratio between the PMViewer and government monitors [7] in six cities. The comparison results in Fig. 4 shows that the ANN ratio of PMViewer is basically less than 1 compared to the government-led monitoring.
2) Comparing with Other PUS Approaches: As a new approach in PUS, PMViewer overcomes the Black Hole problem of PMTI-based approaches in PUS. We evaluate PMViewer through the comparison with Mosaic [14], which is a typical PMTI-based approach. To clearly compare the spatial resolution between PMViewer and Mosaic, we chose the area (2.9*3.1 km$^2$) covered by Mosaic in Hangzhou, China. In this area, there are 8 Mosaic nodes installed on buses while there are 17 PMViewer AirBoxes in this area. Although Mosaic plans to deploy more nodes in public infrastructures [14], it will incur more hardware cost and maintenance cost.

B. How about the Temporal Resolution of PMViewer?

The temporal resolution of PMViewer mainly depended on the real-time ability of the P server, referred to $T_{ability}$:

$$T_{ability} = T_{dataObtain} + T_{dataAnonymization} + T_{show}$$  \(2\)

$T_{dataObtain}$ refers to the time that the P server crawls data from the H server. It is the most time-consuming part with dozens of minutes to pass. $T_{dataAnonymization}$ refers to the time that the P server needs to finish the process of location anonymity. It can be quickly completed by background programs. $T_{show}$ is the time used for visualization. It is almost real time because the dynamic website can finish it in a few milliseconds. Through the analysis above, we found that $T_{dataObtain} \gg T_{dataAnonymization} + T_{show}$, which implies that the measurement of $T_{ability}$ could be represented by $T_{dataObtain}$. To evaluate the actual time interval, we conducted several tests in crawling data from the H server and the result is shown in Fig. 5. We can evaluate that urban $PM_{2.5}$ data can be updated up to every 20 minutes. We evaluate the temporal resolution of PMViewer in two aspects: Firstly, in government-led monitoring, the data updating period is at least an hour [7], while in PMViewer, it needs no longer than 20 minutes. Secondly, as PMViewer aimed to address the “Black Time Window” problems of PMTI-based approaches in PUS, we compare the temporal resolution with PMTI-based approaches. We choose Mosaic [14] as the representative of PMTI-based approaches. Mosaic utilizes buses to monitor the urban $PM_{2.5}$ situation. The temporal resolution relies on the interval between buses. As the buses with Mosaic node are few (only 8 in Hangzhou), the time interval for data updating at one location is changeable and would take half an hour or more. In comparison, PMViewer offers a much better resolution with fewer time intervals and more steady results.

C. How about the Data Anonymization?

In Section V, we proposed two basic rules for offset-location. To ensure the rule (i), after the first step of location offset processing, we removed the street number in the original address. As a result, the offset location generated in the second step would be different from the actual location. To ensure the rule (ii), we reconstructed the address which combined with the semantic description. We evaluate whether the offset addresses are still near the original address. We randomly chose 30 AirBoxes and calculated the distances between their original locations and the offset. The result is shown in Fig. 6 and show that offsets are ranging from 10 to 180 meters which are acceptable for the rule (ii).

D. How about the Hardware Cost of PMViewer? Can All Citizens Get the Real-time $PM_{2.5}$ Situation?

If person A bought an AirBox and placed it at home, he or she gets the real-time $PM_{2.5}$ situation. Meanwhile, person A can also access to the PMViewer to check other $PM_{2.5}$ situation. But for person B without any AirBoxes, he or she can also have access to PMViewer for $PM_{2.5}$ situation nearby their location. Person B and person A can get nearly the same result, which inferred that all citizens can get the real-time $PM_{2.5}$ situation with no extra hardware costs.

E. AirBox is mainly Placed inside Homes or Offices. Does It Reflect the Actual Air Quality Situation Outdoor?

As the indoor air quality is normally better than the outdoor, we should discuss whether the result from PMViewer reflects
the actual outdoor air quality or not. We select a government-led monitor node deployed in Shijingshan district of Beijing, and a PMViewer node near it. We compared the monitor data of these two nodes for a whole day on October 5th, 2016. The results shown in Fig. 7 implied that PMViewer can properly reflect the actual air quality outdoor.

VII. RELATED WORK

A. Government-led Monitoring

Government-led monitoring basically includes satellite remote sensing and static measurement stations [14]. Satellite remote sensing directly measures the concentration of certain air pollutants. However, the results inferred from satellite images only reflect the air quality of atmosphere rather than the ground air quality. Static measurement stations require large and expensive facilities (on the order of 50K-100K dollars) [13] in fixed stations. These facilities are coarse-grained placed where the PM2.5 monitors are sparse and far from each other.

B. Wireless Sensor Networks Approaches

WSN approaches have been studied and deployed extensively in the environmental monitoring field due to their broad applicability and application prospects. WAPMS [9] is deployed to monitor air pollution in Mauritius. Yajie Ma et al present a distributed infrastructure based on WSN for urban air pollution monitoring in London [15]. Generally, these approaches mainly monitor emissions such as PM10 and NOX, while finer detection like PM2.5 is not included as additional precise and costly sensors are required.

C. Participatory Urban Sensing

The PUS approach mainly includes two categories [11]: a) Utilizing the participants’ mobile smartphones with additional sensors to collect and aggregate air quality data. Generally, these devices need additional hardware if there are not essential sensors in the mobile phones. GasMobile [8] uses a small USB hardware device directly connected to the smartphone to monitor the air quality. AirCloud [13] consists of custom-designed sensor front-ends and air quality analytic engine in the cloud. Generally, PUS based on mobile smartphones lacks incentives for extra sensors needed and they entirely rely on the participants to generate data. b) Utilizing the public mobile transportation infrastructures (PMTI) to sense urban air quality. Chaosheng Xiang et al [5] present a mobile sensing system using a public bicycle system to monitor urban air quality. Srinivas et al [12] and Mosaic [14] present sensing devices deployed on public buses to sense urban air quality.

VIII. CONCLUSION

In this paper, we propose PMViewer, a novel PUS approach for the public to monitor the urban PM2.5 situation nearby their locations. To offer a low-cost and fine-grained spatial-temporal resolution in PM2.5 monitoring with smart devices, there are two challenges: How to overcome limitations of former PM2.5 monitoring approaches and how to obtain wide-spread real PM2.5 data from tens of thousands of smart devices. PMViewer aims to address these two challenges as its data are collected from tens of thousands of smart devices called AirBox through a crowdsourcing approach. PMViewer scrabbles data from AirBox’s vendor server and parses the data to generate a map view to display real-time urban PM2.5 measurements. We design, implement and evaluate PMViewer. Evaluation results show that PMViewer efficiently and economically addressed these two challenges described above.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their insight and detailed feedback. Our work was supported by The National Key Research and Development Program of China NO.2017YFB0801900 and Youth Innovation Promotion Association of CAS. Peng Liu was supported by NSF CNS-1505664 and NSF SBE-1422215.

REFERENCES