A Method for Collecting Uniform Amount of Fresh Data from Areas with Varying Population Density

Citation:

Yuichi Kawamoto, Takayuki Nakazawa, Hiroki Nishiyama, Nei Kato, Yoshitaka Shimizu, and Tingting Jiang, "A Method for Collecting Uniform Amount of Fresh Data from Areas with Varying Population Density," IEEE International Conference on Communications (ICC 2015), London, UK, Jun. 2015.

A Method for Collecting Uniform Amount of Fresh Data from Areas with Varying Population Density

Yuichi Kawamoto†, Takayuki Nakazawa¶, Hiroki Nishiyama†, Nei Kato†,
Yoshitaka Shimizu‡, and Tingting Jiang§
†Graduate School of Information Sciences, Tohoku University, Sendai, Japan
¶Department of Information and Intelligent Systems, Tohoku University, Sendai, Japan
‡NTT Network Innovation Laboratories, NTT Corporation, Yokosuka, JAPAN
§Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
E-mails: † ¶{youpsan, takayuki, bigtree, kato}@it.ecei.tohoku.ac.jp
‡shimizu.yoshitaka@lab.ntt.co.jp, §virjtt03@vt.edu

***Abstract*—With the development of wireless communication technology, the utilization of the ambient information that the network users observe has attracted much attention. In this paper, we focus especially on the utilization of observed environmental signals for authentication systems. Ambient information taken as unique data at a particular time and place can be utilized to construct stronger authentication systems. However, since the environmental condition of the network is different for each location, which has a huge effect on the observed ambient information in the area, required data are also different for each place. Thus, in this paper, we propose an efficient data collection method which dynamically changes the way data is collected according to the requirements and the network condition. More specifically, our proposal aims to collect uniform amount of fresh data from areas with varying population density. Additionally, an algorithm to improve the efficiency of our proposal regarding the accommodation of the environmental condition of the network is introduced. Moreover, numerical results verify the effectiveness of our proposal.**

***Index Terms*—Data collection, Ambient information, and Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA).**

I. Introduction

Recently, wireless communication technologies have developed rapidly. In addition, the capability of the network devices has enhanced and a lot of the devices, not only PCs or smartphones, are now connected to the Internet [1]. In particular, the number of the devices for data gathering which are connected to a network wirelessly increases day by day [2], [3]. On the other hand, data collection with the participation of users has also attracted much attention. In this kind of data collection, each user with a wireless communication device collects required data from the environment and sends it to the central server via the wireless network. Data related to the users' environment such as temperature, humidity, precipitation, and communication signals, which are called "ambient information", are utilized in many network services. In this paper, we focus particularly on the utilization of environmental signals for security purpose.

Whether we are at home, at the office or at the station, network connection is available because some access points (APs) or base stations are deployed there. Also, some people also utilize mobile APs (such as smartphone tethering), which are not fixed on a single location. In this kind of situation, a list of information regarding environmental signals such as Service Set Identifier (SSID), and Radio Field Intensity (RFI), can be observed. Since information received from one device is different from what is received from other devices, these data can be considered unique to each device. Thus, it can be considered that it can help strengthening authentication systems.

Nowadays, many authentication systems use cryptographic algorithms. The algorithms can provide strong security but the strength depends on the created key. The more complex the key is, the more secure it becomes. Because the algorithm cannot insure sufficient strength of the security if the key is too simple, it requires strong key management mechanisms. Thus, authentication using communication signals as ambient information has attracted much attention as one of the novel way for some services. In [4], Y. Zheng et al. proposed to use the lists of communication signals to identify the geographical position of an user. Additionally, with the increasing use of various kinds of applications related to Social Network Service, authentication systems guaranteeing the users' correct positions are expected to be necessary to network services in the near future.

However, in order to develop authentication systems which utilize the list of ambient information, a network architecture capable of collecting the ambient information from many users is needed. Additionally, since the communication signals vary from time to time, data collection on a real-time basis is required to construct the list of fresh ambient information. Moreover, the changing speed of the communication signals is affected by the movement of people carrying mobile APs. For example, the changing speed is high in crowded places such as stations and shopping streets, but, it is low in quieter places. Thus, when collecting ambient information from users, it is also significant to consider the environment conditions such as whether the place is congested or quiet for efficient data collection. Furthermore, amount of data to construct the list is also an important aspect to consider in order to provide the authentication systems as many people as possible in a fair way. Thus, in this paper, we propose a method

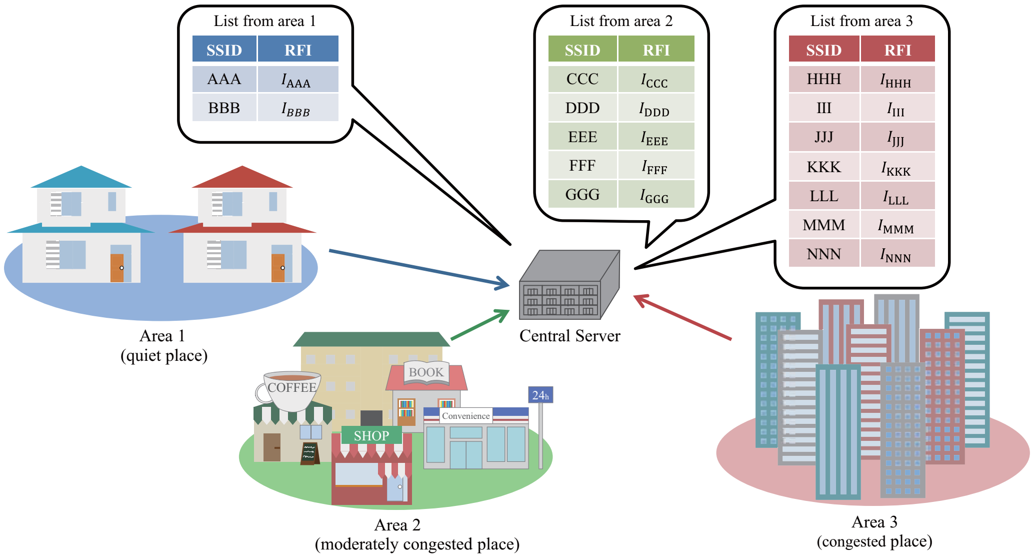


Fig. 1. An example of a system model.

for collecting uniform amount of fresh data from areas with varying population density.

The remainder of this paper is organized as follows. The assumed network model is presented in Section II. An experiment observing environmental signals, which is conducted in real field, is introduced in Section III. Section IV provides our proposed method for collecting uniform amount of fresh data from areas with varying population density. Section V shows the results of the numerical analysis. Finally, concluding remarks are provided in Section VI.

II. Assumed Network Model and Existing Data Collecting Methods

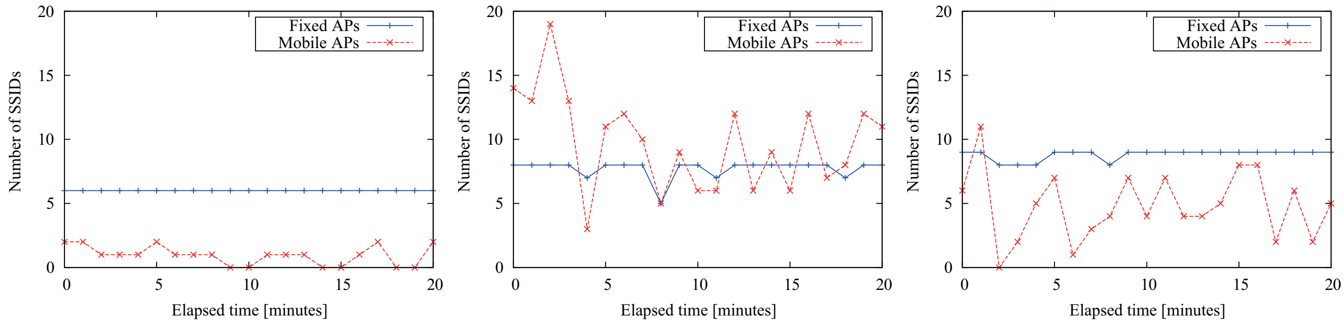
In this section, firstly, we present the assumed network model. Secondly, existing access control schemes are introduced as traditional data collection methods. Additionally, some shortcomings of these existing schemes are denoted with the introduction.

A. Assumed network model

As mentioned in Section I, in this paper, we focus on the data collection of environmental signals as ambient information of users to construct the list of the ambient information, which is utilized for more secure authentication systems [5], [6]. In [5], [6], as the ambient information, some elements such as Service Set IDentifier (SSID), Received Signal Strength (RSS), Sequence Numbers (SN) and Media Access Control (MAC) addresses of the packets are introduced. As shown in Fig. 1 each list of the ambient information in each area is gathered in a central server. When users utilize the authentication system, they access the server and check if the list of ambient information that the user is observing matches the list which was uploaded to the server by other users. Since the ambient information of the users varies from time to time and place to place, it is unique data for that time and place. On the other hand, the coverage of each area is defined by the coverage of AP in each area. Each AP collects the list of ambient information from the users in its coverage area. Additionally, it is supposed that the density of users in each area is different. Depending on this condition, the changing speed of the observed list of the ambient information is subject to change. For example, in a quiet place like residential area or indoors cafe, the rate of fixed APs is high whereas the rate of mobile APs is low. Thus, the changing speed of the ambient information is low. On the other hand, in congested places like shopping streets there are usually many mobile APs. This causes the changing speed of the ambient information to be higher. Hence, in quieter places, it is possible to collect ambient information during a long timespan. On the other hand, in more congested places, it is required to collect data during a short period of time just in case the density of the point of ambient information collection becomes low, in order to keep the data in the central server fresh. Therefore, it is required to collect the list of ambient information according to the condition of each area to realize an efficient data collection.

B. Existing access control schemes of data collection methods

Access control schemes are categorized into two main types: fixed assignment schemes and contention-based schemes. As fixed assignment schemes, Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), and many other schemes have been developed. In these fixed assignment schemes, each user is assigned a certain amount of data transmission resource such as time, frequency, and so on. In such schemes, since everything is organized by a centralized control of APs or base station for the network, there are some advantages to collect data stably. However, sometimes wasted network resources assignment happens in



(a) Quiet place. (b) Congested place. (c) Moderate congestion.

Fig. 2. Measured number of SSIDs in different congested places.

the case where the change of network conditions is intense. For example, in the case where data from the users occur randomly, the performances of the fixed assignment schemes drastically decrease.

On the other hand, in the contention-based schemes, data transmissions occur distributedly based on the users' timings. As a famous contention-base scheme, Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) is often used in wireless networks, which has mechanisms to avoid collision of data sent by multiple users at the same time [7], [8]. In the process of CSMA/CA, when there is data to be sent at a user, the user checks the condition of the channel and starts transmission of the data if the user does not sense channel usage for a certain period of time. On the other hand, if the user senses the channel is being used, it stops the data transmission and waits for a random period time called " backoff time " . After waiting for the backoff time, the user starts retransmitting if the channel is free at that time. In the case where the channel is still being used by another user, the terminal selects a random backoff time again. At this time, the range of the chosen backoff time is increased in the process of CSMA/CA. By increasing the range in an exponential fashion each time the user senses the channel is busy, it decreases the probability that the data collision occurs again.

In the assumed network, it is considered that the number of users inside an AP varies from time to time. Thus, we decide to use a contention-based scheme in the assumed network in this research. Especially, we focus on CSMA/CA to collect the list of ambient information. To achieve efficient data collection, we propose a novel data collection method which is based on CSMA/CA in Section IV.

III. Actual Measurement of Environmental Signals in Real Field

In this section, we introduce the experiment that we conducted in real field to measure real environmental signals. In this experiment, we measured the list of SSID which is observed by a smartphone application as the unique ambient information of the observing user. By executing the measurement at three different places, we tried to confirm that the observed list of signals is unique and investigate the characteristic of the observed list in each place. At first, the experiment environment is introduced. Secondly, the experimental results are shown with some discussion.

A. Experiment environment

The measurements are conducted at three different places. First measurement was done in quiet place A (e.g. cafe along the street with little pedestrian traffic). Second measurement was conducted in crowded place B (e.g. the premises of a large station). At last, we took the third measurement in the middle congested place C. Here, the congestion means simply amount of people, i.e. quiet means there are few people and crowded means that there are a lot of people. In the measurement, a Samsung Galaxy S III, in which Android OS is installed, was utilized to observe the environmental signals. We measured SSID as environmental signals every minute for 20 minutes at each place.

B. Results of the experiment

In this experiment, we measured the list of SSIDs observed in different places for 20 minutes. From the observed list, it is understood that the SSID is classified into two types, which are fixed APs and mobile APs, which includes that of smartphones using tethering. To clearly classify the SSIDs into the two groups, we need a definition for separation. In this analysis, the APs' SSIDs observed more than 80% of the measurement period are defined as fixed APs. Since sometimes the signals of fixed APs are not observed due to signal fluctuation, we do not set the rate as 100%. All other APs are defined as mobile APs.

The measurement results in the three places (a quiet place, a congested place, and a middle ground between the both) are shown in Fig. 2. The red line shows the number of observed mobile APs' SSIDs and the blue line denotes that of fixed APs' SSID. From these results, it is clearly shown that the number of mobile APs' SSID is very few in the quiet place by contrast to the large amount in the congested place. Thus, it is understood that the number of mobile APs tends to increase with the number of the users in the place. Additionally, changes to the number of mobile APs are also growing in intensity proportionally to the number of the users. On the other hand, the number of observed fixed APs' SSID stays pretty much the same in every place.

From the results, we understood that the difference in the density of users causes the difference of the number of observed ambient information and its changing speed caused by mobile APs. Thus, an efficient data collecting method considering the condition of the coverage area is needed.

IV. Proposed data collection method

In this section, we propose a method for collecting uniform amount of fresh data from areas with varying population density. In this proposal, by considering the network condition and the requirements from the system, we aim to collect same amount of data from both areas where many users are and fewer users are with keeping the highest throughput.

A. Algorithm of proposed data collection method

As mentioned above, the density of users is different depending on the place. Additionally, the users inside the coverage always change from time to time. Thus, we propose a data collection method which dynamically adjusts to the network and users' conditions based on the CSMA/CA scheme. In our proposal, each AP which collects the list of ambient information decides the amount of users which will execute the data transmission and the time period during which the users are allowed to transmit according to the requirements of the density of the collected ambient information. To decide the number of users which will execute the data transmission during the determined time period, we set the rate of the number of users sending data in the group of all users inside the coverage of the AP as α. Additionally, the time period is denoted as T. By controlling the values of α and T dynamically, our proposal achieves to collect uniform amount of data from areas with varying population density and high throughput which results into a fresh ambient information collection and also satisfies the requirements of the density of the collected ambient information.

α and T are calculated by the AP at the end of each time period and broadcasted to the users. Each user receiving the message including these parameters from the APs decides whether or not they send the list of ambient information during the time period randomly according to the value of α. Next, the users which decide to send the data choose an amount of time to wait before starting the data transmission randomly between 0 and T. By choosing the waiting time randomly, wasteful data collision is avoided. The value of T is calculated to optimize the expected throughput with the expected amount of users transmitting data, which depends on the value of α. However, since it is difficult for each AP to determine how many users are inside its coverage, it is impossible to determine the value of α adequately. Thus, in our proposal, each AP sets the value of α according to Algorithm 1. In this algorithm, the AP dynamically changes the value of α after the end of each time period according to the throughput and number of collisions occurred during that time period, n_c. In our consideration, APs can identify data collision by observing changes to the signal level. So, when the throughput is lower than the expected value, the AP estimates whether the value

Algorithm 1 Proposed data collection algorithm

AP broadcasts the set of α and T to all of users
Each user decides whether to send data or not randomly according to α
AP receives data during T and counts the number of data collisions
/* After the time period T */
if $n_\mathrm{c} > \phi$ **then**
    $\alpha = \alpha - \Delta$
else
    $\alpha = \alpha + \Delta$
end if
Recalculate the optimal value of T according to the new α

of α was set lower or higher than the optimal value through the number of collisions. Since it can be calculated with the behavior of data transmission in CSMA/CA and observation of actual network, we skip the introduction of the number of occurred data collisions when the throughput is maximized, namely ϕ, due to the limitation of space. In the case where the value of α was set too low, there was some period during which no data was received, which results in no data collision and n_c becoming lower than ϕ. On the other hand, in the case where a high value of α is set, data collision occurs many times, which results in the lower throughput and n_c becoming higher than ϕ. Therefore, each AP determines whether the value of α should be changed to higher or lower than the value which was used at the previous interval. Here, we define the amount of increment / decrement as Δ. Actually, the value of Δ has an effect on the capability of staying close to the optimal value in the proposal, but, we use it as a constant value in this paper. On the other hand, the value of T can be optimized to maximize the throughput during the time period after the value of α is fixed as expressed in the following subsection. In this way, the proposed method controls the number of users sending data during a interval and the length of the interval, which causes efficient data collecting based on CSMA/CA.

B. Optimization of the value of T

In our proposal, CSMA/CA is supposed to be used as the access control scheme. Additionally, each user chooses a random amount of time to wait before starting the data transmission during the time period of T. Thus, the throughput is calculated by the existing analysis of CSMA/CA. Therefore, we use this analysis to optimize the value of T in this paper. In this analysis, we assume that all of the data transmitted from the users is that collects ambient information.

From the research in [9], the probability that data transmission succeeds in a network using CSMA/CA is expressed as follows:

$$p_\mathrm{s} = \frac{e^{-aG}}{G \cdot (1+2a) + e^{-aG}}, \tag{1}$$

where a denotes the ratio of propagation delay to packet transmission time. Here, we suppose that each user sends a packet to AP when data occur at the user. Additionally, the traffic occurring rate inside the coverage of the AP, G (packets/slot), is assumed to be in a Poisson process. Moreover, in the assumed network that the proposed method is adopted, the value of G is expressed as follows:

$$G = \frac{N \cdot \alpha \cdot t}{T}, \tag{2}$$

where N and t denote the total number of users in the coverage of the AP and the length of a time-slot, respectively. At this time, the throughput, S (packets/slot), is expressed as follows:

$$S = G \cdot p_{\mathrm{s}} = \frac{G \cdot e^{-aG}}{G \cdot (1 + 2a) + e^{-aG}}. \tag{3}$$

It is presented that there is a value of G which maximizes S in researches [7], [8]. Therefore, we define the optimal value of T, T_{opt}, as the value that achieves the maximum throughput. The value of T_{opt} is expressed as follows:

$$T_{\mathrm{opt}} = \arg\max_{T} S. \tag{4}$$

C. Analysis on the expected performance of the proposal

At last, we analyze that how much the requirement of data collection density is satisfied with the proposed data collection method. Here, we represent the value of p_{s} in the case where the proposed method is adopted in the assumed network as $p_{\mathrm{s}}^{\mathrm{prop}}$. Then, the probability that the data transmission from the user which decides to execute the data sensing according to the value of α succeeds with the proposal during the time period, P, is expressed as follows:

$$P = 1 - (1 - p_{\mathrm{s}}^{\mathrm{prop}})^{M+1}, \tag{5}$$

where M denotes the maximum number of attempted retransmissions during one time period. In the process of CSMA/CA, contention window, CW, is used to determine the backoff time. A random number is chosen from the range between 0 and CW, and the product between this random number and the length of a time slot is set as the backoff time. Additionally, the value for the contention window stops increasing when it exceeds the value of $CW_{\max}$. After that, the random value is chosen from the range between 0 and $CW_{\max}$ repeatedly. Moreover, the value of the contention window increases along with the number of retransmissions according to the following expression:

$$CW = (CW_{\min} + 1) \cdot 2^m - 1. \tag{6}$$

TABLE I
PARAMETER SETTINGS

Radius of the coverage of AP	30m
Maximum value of contention window ($CW_{\max}$)	15
Minimum value of contention window ($CW_{\min}$)	1023
Length of a time-slot (t)	20μs
Ratio of propagation delay to packet transmission time (a)	0.5

Then, M is expressed by Eq. 7. In Eq. 7, k is the maximum value which satisfies the following expression:

$$t \cdot \sum_{m=0}^{k} \{(CW_{\min} + 1) \cdot 2^m - 1\} \leq T, \tag{8}$$

where $CW_{\max}$, $CW_{\min}$, and ω denote maximum and minimum value for the contention window and the number of retransmissions when the value of contention window reaches $CW_{\max}$, respectively. Furthermore, by using Eq. 6, the value of ω can be calculated as follows:

$$\omega = \log_2 \left(\frac{CW_{\max} + 1}{CW_{\min} + 1} \right). \tag{9}$$

V. PERFORMANCE EVALUATION

A. Parameter settings

Table I shows the parameter settings. In this performance evaluation, the radius of the AP which collects the ambient information from the users is set to 30m. Additionally, we suppose that the users are deployed randomly in the coverage of the AP and that CSMA/CA is used as the based access control scheme. The maximum and minimum values for the contention window are set to 15 and 1023, respectively, as the settings used in the process of CSMA/CA. Moreover, a time-slot length and the ratio of propagation delay to packet transmission time are set to 20μs and 0.5, respectively. To make evaluations in different situations, the required density (number of users per square meter) of ambient information, D, is set to 0.004, 0.008, 0.011, respectively. To simplify the analysis, it is supposed that the data collection in the proposal is executed after the value of α is set to an adequate value. It means that the data collection begins after the value of n_{c} becomes ϕ.

$$M = \begin{cases} \dfrac{T - \sum_{m=0}^{k} \{(CW_{\min}+1) \cdot 2^m - 1\} \cdot t}{\{(CW_{\min}+1) \cdot 2^{k+1} - 1\} \cdot t} + k & \left(CW_{\min} \cdot t < T < t \cdot \sum_{m=0}^{\omega} \{(CW_{\min}+1) \cdot 2^m - 1\} \right) \\[2ex] \dfrac{T - \sum_{m=0}^{\omega} \{(CW_{\min}+1) \cdot 2^m - 1\} \cdot t}{CW_{\max} \cdot t} + \omega & \left(t \cdot \sum_{m=0}^{\omega} \{(CW_{\min}+1) \cdot 2^m - 1\} \leq T \right) \end{cases}. \tag{7}$$

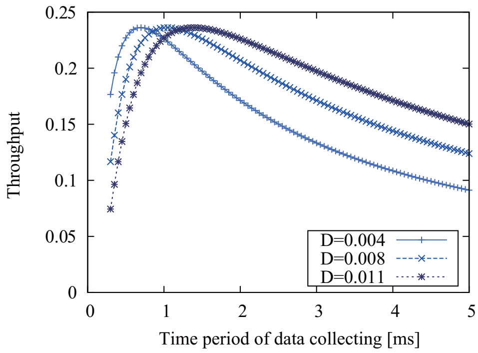


Fig. 3. Throughput vs. Time period of data collecting.

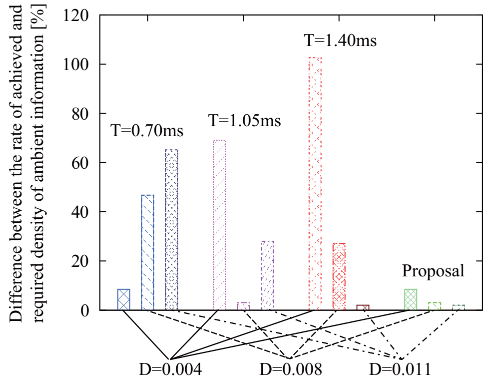


Fig. 4. Difference between the rates of achieved and required density of ambient information with or without our proposal.

B. Numerical results

At first, we investigate the relationship between the throughput and the length of the time period to confirm the existence of an optimal value for the time period. Fig. 3 shows the results of throughput with the change of the time period. From these results, it can be seen that an optimal value of the time period which maximizes the throughput exists for each required density of ambient information. Thus, in our proposal, by calculating the throughput with the required density, the optimal value of T is set and broadcasted repeatedly.

Secondly, we evaluate the efficiency of the proposed method by contrast with the case where a fixed value for the data collecting time period is adopted. Fig. 4 shows the difference between the rate of achieved and required density of ambient information in each case where the required density is set to 0.004, 0.008, 0.011, respectively. For comparison, the time period is set to 0.70, 1.05, and 1.40, respectively. The reason why these values were picked up is that they are the same value of the optimal value of T in each case where the required density is set to 0.004, 0.008, 0.011, respectively. From the results, it is shown that the proposed method achieved the minimum value of the evaluation index. By comparing with the case where a fixed value for the time period is adopted, it is understood that the proposed method always uses the optimal value of T, even if the required density is different. Therefore, it is confirmed that the proposal achieves an effective data collection according to the requirement from the system.

VI. Conclusion

In this paper, we proposed a a method for collecting uniform amount of fresh data from areas with varying population density. Along with the increase in attention to utilizing observed users' ambient information for security purpose, a data collection from many users in real-time basis is required. To achieve an efficient collection of the ambient information from the users, in our proposal, the difference of the network condition between different places is considered. Additionally, we propose an algorithm to accommodate the change of the network condition dynamically. Numerical results represent the effect on the data collection from the users on the proposed method. From the results, we confirmed that the proposed method achieves an efficient data collection that satisfies the required density of the ambient information collection.

Acknowledgement

This research was partially funded by the project, "Cognitive Security: A New Approach to Securing Future Large Scale and Distributed Mobile Applications, " of Japan-US Network Opportunity: R&D for "Beyond Trillions of Objects " supported by the National institute of Information and Communications Technology (NICT), Japan. Additionally, this research was supported in part by US NSF grant CNS-1405747.

References

[1] Y. Kawamoto, H. Nishiyama, N. Kato, N. Yoshimura, and S. Yamamoto, "Internet of Things (IoT): Present State and Future Prospects" *IEICE Transactions on Information and Systems,* vol. E97-D, no. 10, pp. 2568-2575, Oct. 2014. (Invited Paper)

[2] L. He, Z. Yang, J. Pan, L. Cai, and J. Xu, "Evaluating Service Disciplines for On-Demand Mobile Data Collection in Sensor Networks," *IEEE Trans. on Mobile Computing,* vol. 13, no. 4, pp. 797-810, Apr. 2014.

[3] S. He, J. Chen, D. K. Y. Yau, Y. Sun, Cross-layer Optimization of Correlated Data Gathering in Wireless Sensor Networks, *IEEE Transactions on Mobile Computing,* vol. 11, no. 11, pp. 1678-1691, Nov. 2012.

[4] Y. Zheng, M. Li, W. Lou, and Y. T. Hou, "Sharp: Private proximity test and secure handshake with cheat-proof location tags," *17th European Symposium on Research in Computer Security (ESORICS 2012),* Pisa, Italy, Sep. 2012.

[5] J. D. Nielsen, J. I. Pagter, M. B. Stausholm, "Location privacy via actively secure private proximity testing," *2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops),* pp. 381-386, 19-23 Mar. 2012.

[6] L. Xiao, Q. Yan, W. Lou, G. Chen, Y. T. Hou, "Proximity-Based Security Techniques for Mobile Users in Wireless Networks," *IEEE Transactions on Information Forensics and Security,* vol. 8, no. 12, pp. 2089-2100, Dec. 2013.

[7] F. Wang, D. Li, Y. Zhao, "Analysis and Compare of Slotted and Unslotted CSMA in IEEE 802.15.4," *5th International Conference on Wireless Communications, Networking and Mobile Computing, 2009. WiCom '09,* pp. 1-5, 24-26, Sep. 2009.

[8] Y. Cheng, H. Li, P. J. Wan, X. Wang, "Wireless Mesh Network Capacity Achievable Over the CSMA/CA MAC," *IEEE Transactions on Vehicular Technology,* vol. 61, no. 7, pp. 3151-3165, Sep. 2012.

[9] Y. Yang, T. S. P. Yum, "Delay distributions of slotted ALOHA and CSMA," *IEEE Transactions on Communications,* vol. 51, no. 11, pp. 1846-1857, Nov. 2003.