Privacy Protection in WiFi-based Location Estimation

Zuotao Peng *, Katsuhiko Kaji *, Nobuo Kawaguchi * *Graduate School of Engineering, Nagoya University Fro-cho, Chikusa-ku, Nagoya, 464-8601, Japan, {peng, kaji, kawaguti}@ucl.nuee.nagoya-u.ac.jp

Abstract—In recent years, along with the spread of WiFi-based location estimation service, the privacy problem has been pointed out. For existing WiFi-based location estimation service, users' traces and position information can be obtained fraudulently. Personal privacy may be compromised. This research proposes an algorithm that determines the reliability of the user by considering the probability of both spatial and temporal to resolve the privacy problem in WiFi-based location estimation. Meanwhile, we are operating a portal Locky.jp on location estimation using WiFi. Based on the actual data of base stations stored in the Locky.jp database, we confirm this algorithm. As a result, the percentage of returning position for privacy invasion data and whole data (privacy invasion data is also included) are about 20% and 55% respectively. In summary, we know that this algorithm reduced the privacy invasion to about 20% in the WiFi-based location estimation system.

I. INTRODUCTION

Recently, with the development of WiFi-based location estimation technology, various WiFi-based location estimation systems and services have been proposed [1][2][3][4]. The WiFi-based location estimation system is based on the information such as signal strength WiFi base stations' positions which have been collected at various points in advance, and uses the base stations' information that user has observed, to estimate the user's position in some position estimation approach. With the spread of WiFi-based position estimation services, privacy problem has been pointed out. Nils showed that by using techniques such as impersonating the WiFi base station, the location of the base station's possessor could be identified [5]. In other words, it is possible for attacker to infringe the user's privacy (position and movement history) by impersonating the WiFi base station.

The privacy invasion in WiFi-based position estimation system is that the attacker obtains others' WiFi base stations' information or observation history in some way to steal their position information or movement history. As the methods of obtaining the information of others' WiFi base stations, sale or presentation of WiFi base stations with knowing the information of them, misusing of the base stations' information published on the internet can be considered.

For example, as shown in Figure 1, a large number of WiFi base stations' information has been collected in the database and the data are always being updated by location service provider. If a user sends the information of his observed WiFi A1 to the server, he is thought as being near WiFi A1, so the server will return the WiFi A1's position to him as his position. We assume that the attacker presents WiFi A8 to Tom and Tom sets the WiFi A8 at his home. Then the information of WiFi



Fig. 1. An example of privacy invasion

A8 and the position of Tom's home (lat8, lon8) is collected by the location service provider. For general users, without knowing who is the WiFi A8's holder, there is no privacy problem. But the attacker knows that WiFi A8's holder is Tom. If the attacker wants to know Tom's position, he can build a fake A8 and send the information to the server, the server will think that the attacker is near the A8 and return the real A8's position(lat8,lon8) to the attacker. In fact the attacker is not near WiFi A8 but he can get Tom's position. From above, we know that if the attacker knows someone's WiFi information, he can know his position by building a fake WiFi. So, keeping track of a particular base station becomes possible, and user's location can be identified, which may lead to criminal activities such as stalking.

This research proposes a reliability determination algorithm, considering both the spatial probability and the temporal probability, to solve the privacy problem in WiFi-based position estimation system as described above. Further, in order to understand the trade-off between the convenience and the privacy protection, we perform an evaluation experiment by using the database of Locky.jp [11], where a large number of WiFi observation data have been collected at various points. Further more, as the data of observation points are too large, it is difficult to use them practically. The database of this research only includes the data of WiFi base stations.

The remaining of the paper is organized as follows. Section II describes related work. In Section III the radio wave receiving probability model of WiFi base station is built. Section IV proposes the algorithm of determining the reliability of the user by considering the probability of both spatial and temporal. Next, Section V describes the evaluation experiments of this method. Finally, Section VI presents the main conclusions and the future works.

II. RELATED WORK

In recent years, WiFi-based location estimation systems have been widespread. Later security analysis has shown these systems to be vulnerable to attack. Nils analyzed the security of public WLAN-based positioning systems [5]. They demonstrated that the system was vulnerable to location spoofing and location database manipulation attacks by using the Skyhook positioning system. Through these attacks, they showed the limitations of Skyhook and other similar public WLAN-based positioning systems, With knowing someone's personal WiFis, this problem would become privacy problem as said before. Furthermore, they discussed approaches for securing public WLAN positioning systems based on client-side integrity checks, secure data acquisition, and the mitigation of database poisoning. Specifically, for each localization request by the user, the current position is computed by the position estimation system. The resulting position is then compared to the latest stored position in the history record. Thus, we can get the user's moving distance, speed and so on. Client-side integrity checks include moving distance check, average speed check and trace check. Secure data acquisition is to check the user's reliability by signal fingerprints. The mitigation of database poisoning is to check user-supplied data before updating. However, they did not give the concrete methods and evaluation for the problem.

Google, who has collected amounts of WiFi data and is providing location-based service based on all of GPS, WiFi and IP address, announced a way for users to remove their WiFi information from Google's database in November 2011 [10]. If you do not want the location information of your WiFi to be utilized by Google, you can append "_nomap" to the name of the WiFi base station(SSID). For example, if the SSID is "Nuwnet", you need to change it to "Nuwnet_nomap". Then it is possible to opt it out from Google's database. But this method does not solve the privacy problem technically. In addition, users should be aware of their privacies and set the base stations by themselves.

From above, there is no effective methold to resolve the privacy problem until now. It is necessary and urgent for us to resolve this problem.

III. RADIO WAVE RECEIVING PROBABILITY MODEL

In this research, it is required to know which WiFi based stations should be observed at the user's estimated position. So it is necessary to determine the probability of receiving the radio wave of a WiFi base station at certain position.

The radio wave receiving probability model of WiFi base station is the relationship between the probability of receiving the radio wave of a WiFi base station and the distance from it. To get the model, We conducted an experiment at the Nagoya TOYOTA Auditorium. We set four WiFi base stations at first. Then the observation points were determined at every additional 10m while the distance from the WiFi base station was from 30m to 200m. We stopped for 1 minute and observed



Fig. 2. The relationship between the probability and the distance (the left graph shows it separately, the right graph shows it by average)



Fig. 3. The relationship between the probability and the signal strength (the left graph shows it separately, the right graph shows it by average)

the radio wave every 2 seconds at each observation point. In other words, we recorded the measurement 30 times at each observation point. For each observation point, the number of times that the WiFi base station was observed is defined as *count*, the probability P_{rev} of receiving the radio wave of the base station is defined as equation (1).

$$P_{rev} = count/30\tag{1}$$

As the results of the experiment, Figure 2 shows the relationship between the probability of receiving the radio wave and the distance (radio wave receiving probability model). Figure 3 shows the relationship between the probability of receiving the radio wave and the signal strength. From Figure 3, it can be seen that while the signal strength PL is stronger than approximately -80dBm, the probability of receiving the radio wave is about 1. While the signal strength PL is from -95dBm to -80dBm, the relationship between the probability of receiving the radio wave and the signal strength PL is from -95dBm to -80dBm, the relationship between the probability of receiving the radio wave and the signal strength can be considered as linear. Thus, the relationship is represented by the equation (2). As it is a Probabilistic Model, while $PL \leq -95$, instead of setting Prev(PL) = 0, it is better to give Prev(PL) a very small value.

$$P_{rev}(PL) = \begin{cases} 1 & (PL \ge -80) \\ \frac{PL}{15} + \frac{19}{3} & (PL \in (-80, -95)) \\ 10^{-3} & (PL \le -95) \end{cases}$$
(2)

In addition, the basic signal propagation model of the WiFi is known as equation (3) [6][7][8]. d is the distance from the base station. PL is signal strength. UL is the received signal strength in the reference distance 1m, n is the mean path loss exponent which is determined by the obstacles in the physical environment. In general, we set UL = -32dBm.

$$PL = UL + 10n\log(d) \tag{3}$$

Using equation (2) and (3), we can get the radio wave receiving probability model as equation (4). As mentioned



Fig. 4. The comparison of experimental data and the model

above, while $d \in [10^{\frac{63}{10n}}, 200]$, we give Prev(d) a very small value. Figure 4 shows the comparison of experimental data and this model with n = 2.88 which is determined by the environment of this experiment.

$$P_{rev}(d) = \begin{cases} 1 & (d \le 10^{\frac{24}{5n}}) \\ \frac{21}{5} - \frac{2}{3}n\log(d) & (d \in (10^{\frac{24}{5n}}, 10^{\frac{63}{10n}})) \\ 10^{-3} & (d \in [10^{\frac{63}{10n}}, 200]) \end{cases}$$
(4)

From the above, it is confirmed that the model is suitable to the experimental data.

IV. RELIABILITY DECISION ALGORITHM

The essence of the privacy problem in WiFi-based position estimation is that the attacker without going to the place near the base station can steal the position information of the base station. Therefore, the point to solve the problem is to determine the reliability of whether the user is in the estimated position.

For example, as shown in figure 1, in case that the attacker sends only information of the base station A8 to the server. If the base station A7, A9 exist very near to A8, there is a high possibility that A7, A9 are observed with A8 in the same time. So, there is doubt that whether the user is near A8 really. Furthermore, even if all of the base stations are sent, it is probable that the user can not be trusted. For example, in case that where the user was 10 seconds ago is known, because there is a limit to the moving speed, where the user exists now remains within a certain range. In this way, it can be determined the rough movement distance of the user by time. If the estimated distance is unrealistic, there is doubt that whether the user is in estimated location really.

From the above, this research uses both spatial and temporal information of the user in the estimated position to determine the reliability of the user. The flow of the algorithm is shown in Figure 5. First, we perform position estimation based on the observed base station information. Next, we calculate the spatial probability by using the radio wave receiving probability model and the observed base station information at the estimated position. At the same time, we calculate the temporal probability by using the history of a certain time. Finally, we consider both the spatial probability and the temporal probability, to calculate the reliability of the user.

A. The spatial probability

The spatial probability is the possibility that represents whether the user is really in the estimated position by con-



Fig. 5. The flow of this algorithm



Fig. 6. The flow of calculating the spatial probability

sidering whether the base stations which should be observed or should not be observed have been observed.

1) Premise: It is possible that new base stations are installed, so the base stations that do not exist in the WiFi database would not affect the reliability of user. Therefore, this study is only dealing with base stations that have been stored in the WiFi database.

2) Method of calculation: The base station information received by user is defined as set O. The information of each base station o_n includes BSSID b_n and signal strength r_n .

$$O = \{o_1, o_2, ..., o_n\}; o_n = (b_n, r_n)$$
(5)

First, we should estimate the user's position p(x, y) by using the information of equation (5) and a WiFi-based position estimation method [12].

Next, the following two types' base stations can be considered in the estimated position of the user. According to the flow that has been shown in Figure 6, we can calculate the spatial probability P_s of the user.

type1: observed base stations(set A).type2: the base stations that should be observed but have not been observed(set D).

For type1, we calculate the distance between the estimated position and each base station, and calculate the probability that these base stations can be observed in the estimated position by using the model shown in equation (4). The probability that base station b_n can be observed in estimated position p(x, y) is defined as P_{okn} .

$$P_{ok} = \{P_{ok1}, P_{ok2}, ..., P_{okn}\}$$
(6)

For type2, regard the base stations which are near than 200m from p(x, y) as the base station that should be observed. As said above, for these base stations, we can calculate the probability P_{ngm} that base station b_m can be observed in estimated position p(x, y).

$$P_{ng} = \{P_{ng1}, P_{ng2}, \dots, P_{ngm}\}$$
(7)

So, the probability $P(P_{ok}, P_{ng})$ that the user can obtained the base station information O in p(x, y) is as below.

$$P(P_{ok}, P_{ng}) = \prod_{i=1}^{n} P_{oki} \prod_{j=1}^{m} (1 - P_{ngj})$$
(8)

TABLE I. SEEK MAXIMUM PROBABILITY

	Algorithm MaxProb_Calculating(x,y)
1:	$P = 1$, $P_{all} = P_{ok} + P_{nq}$
2:	for every item $P_l \in P_{all}$ do
3:	if $P_l \geq 0.5$ do
4:	$P = PP_l$
5:	else
6:	$P = P(1 - P_l)$
7:	endif
8:	endfor
9:	return P

For every relevant base station, there are two states which are being observed and not being observed. So, if relevant base stations' number is (m + n), the user's possible observation patterns will be 2^{m+n} . Among these patterns, the spatial probability of the pattern with the maximum observed probability is considered to be 1. The algorithm to seek maximum probability in is shown in Table 1. For example, at one place, only base stations A1 and A2 can be observed. Furthermore, A1 and A2 are observed by 40% and 70% separately. The user may have 4 possible observation patterns. For every base station, if the observed probability is over 0.5, in the pattern of maximum probability, the base station would be observed or it would not be observed. Thus, the maximum probability here is (1-0.4)*0.7 (A1 is not observed and A2 is observed).

So the spatial probability P_s is defined by the following equation.

$$P_s = \frac{P(P_{ok}, P_{ng})}{\mathbf{MaxProb_Calculating}(\mathbf{x}, \mathbf{y})}$$
(9)

B. The temporal probability

The temporal probability is the possibility of representing the rationality of the estimated position based on the average acceleration and the limit of speed by considering the speed of the past inferred from position information history to some extent.

1) Premise: Since it is intended to represent the rationality of the movement range, We consider the temporal probability is not changed if the user does not move.

2) The speed of the past: It is believed that the speed of the user during a certain time period is stable. So the speed of the past can be regarded as one criterion for the estimated speed of the user now. In this study we calculate the speed of the past v by using the history of a predetermined time T. The time



Fig. 7. The flow of calculating the temporal probability

interval of the position estimation is $\triangle t$. The average speed for each time interval in the history is calculated. In order of close to the present time, they are defined as $\{v_1, v_2, ..., v_N\}$. The speed of the past v is defined by the equation below [9].

$$v = r \cdot v_1 + \frac{r}{2} \cdot v_2 + \dots + \frac{r}{2^{N-1}} \cdot v_N$$
$$(r = \frac{2^{N-1}}{2^N - 1})$$
(10)

 $\frac{r}{2^{N-1}}$ is discount rate. With focusing more on the recent history, it also reflects over all of a fixed period of time.

3) Average acceleration: The difference between the estimated average speed and the speed of the past is defined as the average acceleration Δv .

4) Method of calculation: We set the limit of user's moving speed is V_{max} . The flow of calculating the temporal probability is shown in figure 7. First, we determine the speed of the past v by the method in equation(10).

Next, we calculate the current moving speed v' and the average acceleration Δv . We set that loc(t) is the function of position estimation and the distance calculation function is dis(p1, p2). The user's estimated move distance d_w , v' and Δv are shown in equations below. v' is calculated in the same method with equation (10).

$$d_w = dis(loc(t_i), loc(t_{i-1}))$$
(11)

$$v' = \frac{d_w}{\Delta t} \cdot r + \frac{v}{2} - \frac{v_N}{2^N} \cdot r \tag{12}$$

$$\Delta v = \frac{v' - v}{\Delta t} \tag{13}$$

We know that $\triangle v$ satisfies normal distribution $N(0, \sigma^2)$. Variance σ^2 differs by the moving means [14][15].

$$f(\triangle v) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\triangle v^2}{2\sigma^2})$$
(14)

We can think that the temporal probability is 1 when $\triangle v$ is 0, furthermore it falls when the absolute value of $\triangle v$ increases. Thus, the temporal probability P_t is defined by the equation(15), while u(x) is the unit step function.

$$P_t = u(V_{max} - v')\frac{f(\Delta v)}{f(0)}$$
(15)

C. The reliability of user

In this research, considering both the probability of spatial P_s and temporal P_t , we determine the reliability P of the user as the equation below.

$$P = P_s \cdot P_t \tag{16}$$

V. EXPERIMENT

To protect privacy, it is necessary to sacrifice the convenience. In order to understand the trade-off between convenience and privacy protection in WiFi-based position estimation system, we perform the evaluation experiment based on the database of Locky.jp.

A. Experiment settings

1) Parameter settings: As shown in Table II, we set all parameters in this paper. This experiment uses only the history of 300 seconds or less. Set the interval 2 seconds as the same with Locky.jp. Then, because the vehicle has maximum acceleration generally, we set σ^2 by considering the acceleration model of vehicle. The max speed is set as the speed of SHIKANSEN. Set the threshold 0.5 about whether trust the user.

TABLE II.	PARAMETER SETTINGS	
-----------	--------------------	--

parameter	Т	Δt	σ^2	Vmax	threshold
value	300 s	2 s	2	100 m/s	0.5

2) Experimental data: In this experiment, as shown in Table III, we divide the database of Locky.jp into two parts, which are training data and evaluation data.

TABLE III. EXPERIMENTAL DATA

	observation info	base stations	period
Training Data	10,471,524	756,415	2005/7/6 - 2010/3/3
	observation info	observation points	period
Evaluation Data	6,958	2,917	2010/3/3 - 2010/4/9

3) Position estimation: Privacy problem only exists where position can be estimated. So location estimation is required in advance. We estimate the position of each observation point by proximity method [12]. In 2917 observation points, 1259 points' positions can be estimated. Therefore, we will only consider these 1259 points.

B. Result of experiment

1) Result: Using the algorithm in this paper, we calculate the spatial probability, temporal probability, reliability of each observation point. The results are shown in figure 8.

2) Percentage of returning the position: We use the threshold in Table II. If the result is bigger than the threshold, the position will be returned. Therefore, the percentage of returning the position by considering the spatial probability, temporal probability, reliability of each observation point respectively is shown in table IV. The percentage of returning position by considering the reliability is 55.1%. Because there are many observation points that do not move in evaluation data, so the temporal probability and the reliability become small.



Fig. 8. The result of whole data (1259 observation points), upper left is spatial probability, upper right is temporal probability, and the lower is reliability

TABLE IV. PERCENTAGE OF RETURNING POSITION FOR WHOLE DATA



Fig. 9. Stimulate the privacy invasion data using moved base stations

C. Verification experiment

Moved base stations adversely affect the accuracy of location estimation. We consider the base stations which are relocated by the holder, the base stations in Shinkansen, and mobile WiFi as moved base stations. We delete the moved base stations for the training data in this experiment [13]. However, in the evaluation data, there are moved base stations.

We use the data of moved base stations to simulate the situation of privacy invasion in this paper. For example, as shown in figure 9, before moving, base station A1 and A2 were in Tokyo and base station A3 were in Nagoya. After moving, base station A1, A2, A3 were all in Tokyo. When some user went from loc1 to loc3 through loc2 in Tokyo, at loc2, although the real position is Tokyo, the estimated position is Nagoya. Because the base station A3 was in Nagoya in the training data(before moving).

As situations of privacy invasion, if attacker knows the information of the base station A3(such as BSSID) which is in Nagoya, he can build an fake base station A3 in the loc2(Tokyo) in some way. So the attacker can get the position of A3(Nagoya) in loc2(Tokyo) through the WiFi-based location estimation system.

From the above, it can be said that the observation points which include moved base stations have the same situation with privacy invasion. In this experiment, we use the observation points(83 points) whose strongest base station is moved over 200m as privacy invasion data, to verify the privacy protection methods of this study.



Fig. 10. The result of the privacy invasion data (83 privacy invasion observation points), upper left is spatial probability, upper right is temporal probability, and the lower is reliability

1) Extract the result of privacy invasion data: We extract the result of the spatial probability, temporal probability, reliability of privacy invasion data(83 observation points). The results are shown in figure 10.

2) Percentage of returning position for the privacy invasion data: Using the threshold in Table II, the percentage of returning the position by considering the spatial probability, temporal probability, reliability of the privacy invasion data(83 observation points) respectively is shown in table V. The percentage of returning position by considering the spatial probability is over 85%. Because most observation points in privacy invasion data have only one base station, only considering the spatial probability can not prevent privacy invasion. The percentage of returning position by considering the reliability is only about 20%.

TABLE V. PERCENTAGE OF RETURNING POSITION FOR THE PRIVACY INVASION DATA

	the spatial probability	the temporal probability	reliability
percentage	85.5%	21.7%	20.5%

D. Consideration

From the evaluation experiment, we know that the percentage of returning position for the privacy invasion data is only about 20%, In other words, this algorithm reduced the privacy invasion to about 20%. However, the percentage of returning position for the whole data is about 55%. There are three reasons. First, the privacy data is also included in the whole data. Second, if the user do not move, the temporal probability will not increase. So, the reliability is low. Third, to protect the privacy, we have to sacrifice convenience to some extent. From above, we should grasp the trade-off between convenience and privacy protection in WiFi-based position estimation system.

In addition, the spatial probability of the whole data and the privacy invasion data are almost the same (87.2% and 85.5%). That is because that many observation points have only one base station in the evaluation data. Using only spatial probability can not judge these observation points whether they are privacy invasion data.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented an algorithm that determined the reliability of the user by considering the probability of both spatial and temporal to resolve the privacy problem in WiFibased location estimation. In addition, we also built the radio wave receiving probability model of WiFi base station.

To evaluate this algorithm, we performed the evaluation experiment based on the database of Locky.jp. At last we knew that this algorithm reduced the privacy invasion to about 20%.

In future, we will apply this algorithm to large-scale data. Furthermore, in order to respond to various situations, attacking experiments are also considered. Finally, we hope it can be used in practice.

REFERENCES

- Anthony LaMarca, Jeffrey Hightower, Ian Smith, Sunny Consolvo: Self-Mapping in 802.11 Location Systems, In Proceedings of the Seventh International Conference on Ubiqutous Computing, pp.87-104(2005).
- [2] Julian Lategahn, Frank Kuenemund, Christof Roehrig: Mobile Robot Localization Using WLAN,Odometry and Gyroscope Data, International Journal of Computing, Vol.9,Issue 1,pp.22-30(2010)
- [3] Nobuo Kawaguchi : Locky.jp : Locky.jp: Wireless LAN Position Estimation and Its Application, The Transactions of Human Interface Society, Vol.10,No.1,pp.15-20(2008).
- [4] Katsuhiko Kaji, Nobuo Kawaguchi: indoor.Locky: indoor.Locky: Wireless LAN Indoor Location Platform Using UGC, Journal of Information Processing of Japan, Vol.52,No.12(2011) V1-230(2010).
- [5] Nils Ole Tippenhauer, Kasper Bonne Rasmussen, Christina Popper, and Srdjan Capkun: Attacks on Public WLAN-based Positioning Systems, Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services (MobiSys)(2009)
- [6] Varela, F., Sebastiao, P., Correia, A., Cercas, F., Velez, F.J., Robalo, D.and Rodrigues, A.: Unified Propagation Model for Wi-Fi, UMTS and WiMAX Planning in Mixed Scenarios, in Proc. of PIMRC '10-21st IEEE International Symposium on Personal Indoor, and Mobile Radio Communications, Istanbul, Turkey (2010).
- [7] Varela, F., Sebastiao, P., Correia, A., Cercas, F., Velez, F.J., Robalo, D. and Rodrigues, A.: Validation of the Unified Propagation Model for Wi-Fi, UMTS and WiMAX Planning, in Proc. of PIMRC '10-21st IEEE International Symposium on Personal Indoor, and Mobile Radio Communications, Istanbul, Turkey(2010).
- [8] Muzaiyanah Hidayab, Abdul Halim Ali, Khairul Bariah Abas Azmi: Wifi Signal Propagation at 2.4 GHz, Microwave Conference, APMC 2009. Asia Pacific(2009).
- [9] Sebastian Thrun, Wolfram Burgard, Dieter Fox: *Probabilistic Robotics*, The MIT Press(2005).
- [10] Google : Greater choice for wireless access point owners, http://googleblog.blogspot.jp/2011/11/greater-choice-for-wirelessaccess.html (2012).
- [11] Locky.jp : Portal Website for WLAN-based Location Estimation, http://locky.jp/ (2013)
- [12] Yuusuke Nitta, Shigeyoshi Ohno: A study on Methods for Wireless LAN Based Location Estimation, Bull. Polytechnic University, No41-A(2012)
- [13] He Tao, Kaji Katsuhiko, Kawaguchi Nobuo : Adaptation of Integrity Maintenance Method for Location Estimation to Large Scale Wireless LAN Observation Database, MoMuC2011-29(2011)
- [14] Investigation of barrier on bicycle-pedestrian track with probe bicycle, Japan Society of Civil Engineers, Vol.22(2005)
- [15] Takashi Yonekawa : Development of a driving simulator that aims for reality of city driving http://www.jari.or.jp/resource/pdf/H20ITS/11-4.pdf, JARI ITS seminar(2009)