PWS Cup 2019 Data Set

PWS Cup Organizing Committee

August 16, 2019

1 Introduction

An artificial data (PWSCup2019 artificial data) used in the contest is newly-generated on the basis of the SNS-based people flow data [1] that is an open data available to the public. This paper explains the characteristics of PWSCUP2019 artificial data generation. For the details on PWSCUP2019 content rules, refer PWS Cup 2019 contest rule document [2] and rule paper [3].

2 PWSCup2019 Artificial Data

2.1 Overview (Excerpt from Section 3.1 of the rule paper [3])

The SNS-based people flow data is a public data set that contains artificially generated traces covering Tokyo metropolitan area for six days (7/1, 7/7, 10/7, 10/13, 12/16, and 12/22 in 2013). The traces are generated on the basis of real trace data, but they do not contain any information that is easily collated with real users.

As shown in Fig. 1, we divide Tokyo metropolitan area (latitude from 35.65 to 35.75 and longitude from 139.68 to 139.8) equally into $32 \times 32 = 1024$ regions and assign region ID sequentially from lower-left to upper-right. The number of regions (location information) is $m = 1024$, and the size of each region is approximately 347m (height) $\times$ 341m (width) with an assumption that one degree longitude and latitude (in Tokyo area) correspond to 111km and 91km, respectively.

Next, we extract traces of all 10181 users who have more than 10 location information in Tokyo metropolitan area and use them as learning data to formulate a Markov model-based trace generator. Using the trace generator, we create a reference trace set $R^{(i,j)} = (r_1^{(i,j)}, \cdots, r_{2000}^{(i,j)})$ and an original trace set $O^{(i,j)} = (o_1^{(i,j)}, \cdots, o_{2000}^{(i,j)})$ for a user set $\mathcal{U}^{(i,j)} = \{u_1^{(i,j)}, \cdots, u_{2000}^{(i,j)}\}$ with $n = 2000$ users. These data sets are created independently per team $i \in [z]$ and per data set number $j \in \{1, 2\}$. The trace generator refers a feature of a user $u_k^{(i,j)}$ $(1 \leq k \leq 2000)$ when generating his reference trace $r_k^{(i,j)}$ and original trace $o_k^{(i,j)}$. $r_k^{(i,j)}$ and $o_k^{(i,j)}$ have high correlation, so teams can perform ID disclosure and trace inference attacks on public anonymized trace sets while referring to the corresponding reference trace sets.

The time is discretized by dividing the time from 8:00 to 17:59 in 30-minute interval. In the preliminary round, traces from the 1st and 2nd days are used for a reference trace, and traces from the 3rd and 4th

Table1　PWSCup2019 artificial data

Number of users	$n = 2000$
Target area	Latitude: 35.65 to 35.75, Longitude: 139.68 to 139.8
Number of location data	$m = 1024$ (divided into 32×32 regions)
Length of trace	$t = 40$ (from 8 : 00 to 17 : 59 for 2 days with time interval of 30 minutes)

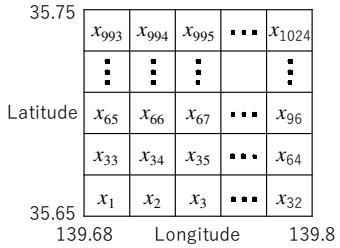

Figure1　Location information in the contest

days are used for an original trace. The length of each trace is thus $t = 40$, with time 1 representing 8:00 of the 1st day, time 40 representing 17:30 of the 2nd day, time 41 representing 8:00 of the 3rd day, and time 80 representing 17:30 of the 4th day.

Table 1 illustrates the overview of PWSCup2019 artificial data. Note that the trace length in the table is for the preliminary round; we might change the length (i.e., the days covered in the traces) in the final round.

We will not disclose the detailed specification of our trace generator. The generator creates PWS-Cup2019 artificial data in the following manners:

1. **Preserve population distribution:** The artificial data is generated in a way that the population distribution (i.e., the probability distribution on $m = 1024$ regions) per hour from 6 o'clock to 17 o'clock is close to the distribution shown in the original SNS-based people flow data.
2. **Preserve transition matrix:** The artificial data is generated in a way that the transition matrix of the Markov model (1024×1024 matrix) is closed to that of the original SNS-based transition flow data.
3. **Model user's home:** The artificial data is generated in a way that most users stay in regions where their home reside in the morning (about 95% at 6 to 7 o'clock, and about 30% at 8 o'clock). A home region is assigned randomly to each user while preserving the population distribution. Note that if we include regions at 6 to 7 o'clock in the reference/original traces, the adversary could know the home regions for almost all of the users. Since this setting is too favorable to the adversary and also unrealistic, we include regions from 8 o'clock in the reference/original traces in our contest.

Refer Section 2.2. for more details.

2.2 Characteristics of the PWSCup2019 artificial data generation

2.2.1 Preserve population distribution and transition matrix

PWSCup2019 artificial data is generated in a way that the population distribution in every hour, from 6 o'clock to 17 o'clock, and the transition matrix are preserved. To quantitatively prove that these conditions hold true, we made the following experiment.

First, we calculated the population distribution and transition matrix in every hour from 6 o'clock to 17 o'clock using the traces of 10181 users in the SNS-based people flow data (refer Section 2.1). We define the population distribution of the SNS-based people data flow in time k ($6 \leq k \leq 17$) as a 1024-dimensional probability distribution $\mathbf{p}_k$ and denote its i-th element as $\mathbf{p}_k[i]$ ($1 \leq i \leq 1024$). We also define the transition matrix as a 1024×1024 matrix $\mathbf{Q}$ and denote its (i,j)-th element as $\mathbf{Q}[i,j]$. $\mathbf{p}_k[i]$ represents the probability of users being in region x_i at time k, and $\mathbf{Q}[i,j]$ represents the probability of users moving from region x_i to x_j in the next time interval. $\mathbf{p}_k$ and $\mathbf{Q}$ are calculated by counting the numbers of visits and transitions for all users and by normalizing the numbers to probability values.

Next, we followed the same procedures to calculate the the population distribution and transition matrix in every hour from 6 o'clock to 17 o'clock using the PWS artificial data (20 teams $\times$ 2 sets (ID disclosure and trace inference challenges) $\times$ 2 rounds (preliminary and final round) = 80 data sets). We define the population distribution of the PWSCup 2019 artificial data in time k ($6 \leq k \leq 17$) as a 1024-dimensional probability distribution $\tilde{\mathbf{p}}_k$ and denote its i-th element as $\tilde{\mathbf{p}}_k[i]$ ($1 \leq i \leq 1024$). We also define the transition matrix as a 1024×1024 matrix $\tilde{\mathbf{Q}}$ and denote its (i,j)-th element as $\tilde{\mathbf{Q}}[i,j]$. $\tilde{\mathbf{p}_k}$ and $\tilde{\mathbf{Q}}$ are calculated by counting the numbers of visits and transitions for all PWSCUP 2019 artificial data (i.e., all 80 data sets) and by normalizing the numbers to probability values.

Finally, we evaluated an error between the population distributions $\mathbf{p}_k$ and $\tilde{\mathbf{p}}_k$ ($6 \leq k \leq 17$) and an error between the transition matrices $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$. We used MAE (Mean Absolute Error) as an evaluation measure of the errors. For the population distribution, we calculated the average of MAE for 12 distributions from 6 o'clock to 17 o'clock.

The MAE of the the population distributions $\mathbf{p}_k$ and $\tilde{\mathbf{p}}_k$ is formulated as follows:

$$\frac{\sum_{k=6}^{17} \sum_{i=1}^{1024} |\mathbf{p}_k[i] - \tilde{\mathbf{p}}_k[i]|}{12 \times 1024}$$

The MAE of the transition matrix $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$ is formulated as follows:

$$\frac{\sum_{i=1}^{1024} \sum_{j=1}^{1024} \left| \mathbf{Q}[i,j] - \tilde{\mathbf{Q}}[i,j] \right|}{1024^2}$$

For the comparison, we also calculated the MAE of $\tilde{\mathbf{p}}_k$ and $\tilde{\mathbf{Q}}$ when each line of them exhibits uniform distribution (i.e., each element is 1/1024).

Table 2 shows the MAE of the population distribution and transition matrix. For both cases, we can see that the PWS Cup 2019 artificial data exhibits much smaller MAE than that of the uniform

Table2　MAE of population distribution and transition matrix

	PWS Cup 2019	Uniform distribution
MAE of population distribution	3.70×10^{-4}	9.22×10^{-4}
MAE of transition matrix	9.89×10^{-4}	1.87×10^{-3}

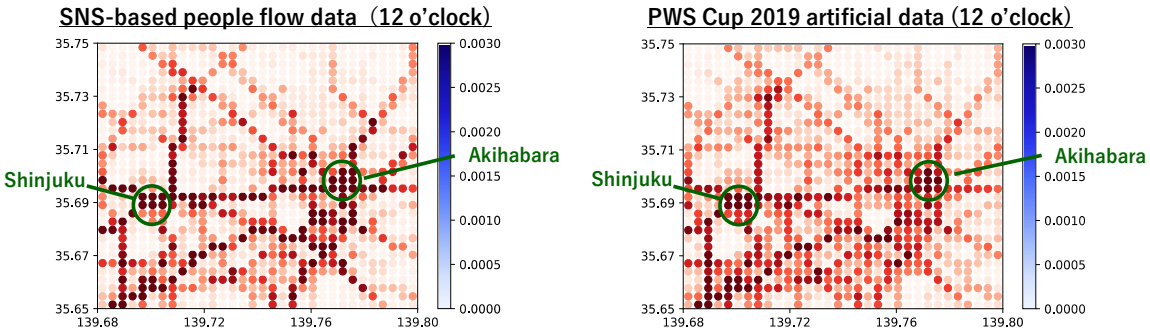

Figure2　Population distribution at 12 o'clock: $\mathbf{p}_{12}$ (left) and $\tilde{\mathbf{p}}_{12}$ (right)

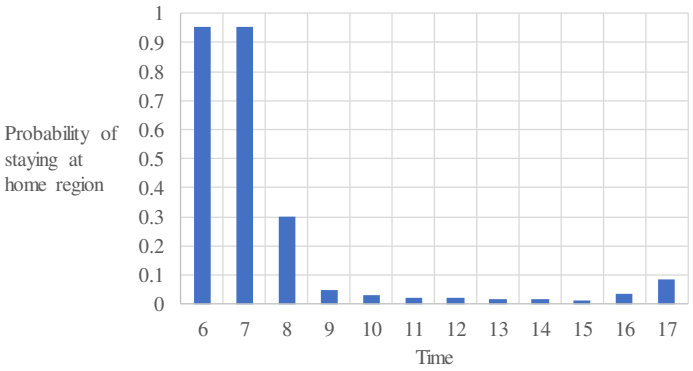

Figure3　Probability of users staying at home

distribution. Figure 2 visualizes the population distribution $\mathbf{p}_{12}$ and $\tilde{\mathbf{p}}_{12}$ at 12 o'clock. We can see that $\mathbf{p}_{12}$, the population distribution of the SNS-based people flow data at 12 o'clock, is concentrated around Akihabara and Shinjuku area. We can also see that the PWSCup 2019 artificial data exhibits similar population distribution. In this experiment, we evaluated the population distributions and transition matrices exhibited by the whole PWSCup2019 artificial data (i.e., 80 data sets). We also confirmed that the population ditribution and transition matrix is preserved in each data set.

2.2.2　Model user's home

The PWSCup2019 artificial data is generated in a way that most users stay in regions where their home reside at 6 o'clock and 7 o'clock.

Figure 3 illustrates the probability of users staying in regions where their homes reside in each time period (from 6 o'clock to 17 o'clock). The probability is calculated by by counting the numbers of visits at the home region for all PWSCUP 2019 artificial data (i.e., all 80 data sets) and by normalizing the

numbers to probability values. We can see that the probability of staying at home region is about 95% in 6 and 7 o'clock and about 30% in 8 o'clock (note: we only use the location data after 8 o'clock in the contest). We can also see that the probability gradually increases from 16 o'clock; the probability is about 8% at 17 o'clock.

References

[1] Nightley and Center for Spatial Information Science (CSIS), SNS-based People Flow Data, [online] Available: `http://nightley.jp/archives/1954`.

[2] PWS Cup 2019 Contest Rule: `https://www.iwsec.org/pws/2019/cup19-rule-e.pdf`

[3] Takao Murakami, Hiromi Arai, Makoto Iguchi, Hidenobu Oguri, Hiroaki Kikuchi, Atsushi Kuromasa, Hiroshi Nakagawa, Yuichi Nakamura, Kenshiro Nishiyama, Ryo Nojima, Takuma Hatano, Koki Hamada, Yuji Yamaoka, Takayasu Yamaguchi, Akira Yamada, Chiemi Watanabe, "PWS Cup 2019: Location Data AnonymizationCompetition," CSS2019.