The Open Data Set of FudanWiFi09

Yi-Qing Zhang, and Xiang Li*

Abstract

This article describes the open data set of FudanWiFi09 and how to use it. FudanWiFi09 contains four independent data files, i.e., access logs, location data, interaction data and sampled interaction data. The scholars who want to use these data files in their researches need to carefully read this article, acknowledge Fudan WiFi project and cite the appointed paper in IEEE Transactions on Systems, Man and Cybernetics: Systems, 2014.

Index Terms

Data Descriptions

I. Descriptions of the open data set

A. *How to use the data set*

This data set contains the following four independent data files and the detailed descriptions of each data file are in the next section.

1) access logs
2) location data
3) interaction data
4) sampled interaction data

This dataset was collected by Fudan WiFi project, with support from Fudan Univeristy. Collecting and making this data publicly available took a lot work.

IF YOU USE THIS DATA SET, CITE THE FOLLOWING PAPER IN YOUR PUBLICATIONS:

Y.-Q. Zhang, X. Li, J. Xu and A. V. Vasilakos, Human Interactive Patterns in Temporal Networks, IEEE Transcations on Systems, Man, and Cybernetics: Systems. vol. 45, pp.214-222, 2015.

B. *Data Disclaimer*

Data is provided without warranties as to performance or quality or any other warranties whether expressed or implied. By using this data set, you agree not to perform reverse engineering to extract users names, wireless access points locations, or any other information.

Y.-Q. Zhang and X. Li are with the Adaptive Networks and Control Lab, Department of Electronic Engineering, Fudan University, Shanghai 200433, PR China (E-mail:12110720032@fudan.edu.cn, lix@fudan.edu.cn).

∗ All correspondence should be addressed to X.Li.

C. Ethics Statement

The study was approved by the Informatization Office of Fudan University, which is responsible for deployment of the wireless network in Fudan University and data collection. Before the study, we sign a confidentiality agreement with the Informatization Office of Fudan University in conformity with the privacy regulations of the country laws for data download and analysis. The dataset used in this study is only involved de-identified information, and none privacy information of the WiFi users are available.

D. Licensing

II. Descriptions of four data files

In the data set, we define 0:00 in 18^{th} Oct, 2009 as the baseline time, zero. Any start times in data files are the differences between the physical time and the baseline time. The temporal resolutions of data files are all 1 minute.

A. Access logs

The file of access logs contains over 262104 online traces of 18718 campus members during 84 days (from 18^{th} Oct, 2009 to 9^{th} Jan, 2010) in 129 covered regions of wireless access points (WAPs) deployed in 6 public teaching buildings of one campus, Fudan University. This data is organized as a tab-separated list as shown in Table I.

Each 'userId' represents an anonymous user. The 'startTime' describes the starting time of the anonymous user getting access into the WiFi system. The 'duration' describes the consecutive times for the anonymous user keeping connecting with the WiFi system. The 'location' represents the covered region of one wireless access point.

B. Location Data

The file of location data contains two parts of information about 129 wireless access points and 6 teaching buildings. This data is organized as a tab-separated list as shown in Table II.

The 'location' represents the covered region of a WiFi access point. 'Floor No.' represents the WiFi access point is deployed in which floor and 'Bldg. No.' represents the WiFi access point is deployed in which teaching building. 'Bldg. longitude' and 'Bldg. latitude' represents the longitude and latitude of each teaching building.

TABLE I

ACCESS LOGS

userId	startTime	duration	location
1	414	333	514
2	424	96	315
...	...	...	...
...	...	...	...

TABLE II

LOCATION DATA

location	Floor No.	Bldg. No.
201	1	2
202	1	2
...	...	...
...	...	...
718	8	7
Bldg. No.	Bldg. longitude	Bldg. latitude
2	121.504348	31.297643
...	...	...
7	121.504502	31.299768

TABLE III

INTERACTION DATA

userId	userId	startTime	duration	location
1303	1366	1978	5	201
1329	1366	1973	10	201
...	...	...	...	...
...	...	...	...	...

C. Interaction Data

The file of interaction data contains 1214647 interactive records among 884800 pairs of users during 84 days (from 18^{th} Oct, 2009 to 9^{th} Jan, 2010) in 129 covered regions of wireless access points (WAPs) deployed in 6 public teaching buildings of one campus, Fudan University. This data is organized as a tab-separated list as shown in Table III.

This data file is generated from the data file of access logs based on the assumption of geographic coincidencesłco-

TABLE IV
SAMPLED INTERACTION DATA

userId	userId	startTime
4	6	444
8	11	469
...	...	...
...	...	...

locating in the same small region at the same time. To verify the assumption, we estimate the distances between any two users connecting to the same WiFi access point. The technical details are shown in the next section.

D. *Sampled Interaction Data*

The file of interaction data contains 209137 interactive records among 187889 pairs of users during 84 days (from 18^{th} Oct, 2009 to 9^{th} Jan, 2010) in 129 covered regions of wireless access points (WAPs) deployed in 6 public teaching buildings of one campus, Fudan University. This data is organized as a tab-separated list as shown in Table IV.

This data file is sampled from the file of interaction data according from the temporal motif detection method, which restricts that at any time, one user can only interact with at most one neighbor.

III. DATA COLLECTIONS AND PREPARATIONS

A. *Project Setting*

The WiFi system involved in this study is deployed in the Handan campus of Fudan University in Shanghai, China. Fudan University has four different campuses, among which the Handan campus is the second largest one (228 acres). More than 75% students of the whole university study and live in this campus. The access points (APs) deployed in Handan Campus provide the dual-band(802.11g/n (2.4-GHz) and 802.11a (5-GHz)) free wireless accessing services to all the campus members (e.g., students, teachers, office staffs and visiting scholars).

In the WiFi Handan campus, all APs share the same network name (SSID), so the wireless clients (e.g., personal computer, video game console, smartphone and digital audio player) can seamlessly roam from one AP to another.(the roaming clients can be automatically detected by the APs without sensed by the WiFi users.) During every semester, there are more than 20,000 people working and living in the Handan campus. Each individual has a unique account provided by the Informatization Office of Fudan University to use the WiFi service for free.

We download the WiFi using data that records users' access-related events of the 2009-2010 fall semester (18/10/2009-9/1/2010) from the Informatization Office of Fudan University. When a client opens its WiFi service and successfully connects to an AP, its hardware information, such as the Media Access Control address (MAC address, the unique serial number of the device), CPU trademark and so on, is collected by the AP and transmitted

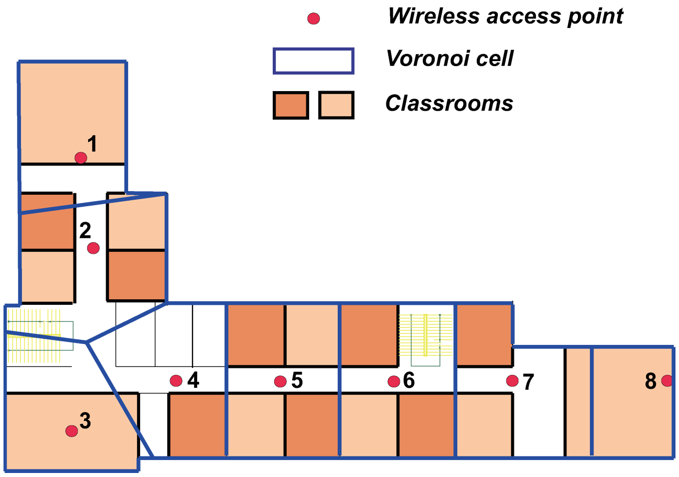

Fig. 1. **The 2^{nd} Teaching Building: Architectural Plan and WAPs Deployments.** The 2^{nd} teaching building has two-floors with the same architectural plan and WAPs deployments. The Figure displays the simplified architectural plan of the bottom floor. The division in terms of Voronoi cells defines the coverage of each WAP. The classrooms are colored by yellow. The situation that WiFi users use their devices outside the classroom(e.g., corridor, toilets, storage room) is not considered. Each classroom is divided into several irregular regions by the Voronoi cell based on the fact that each users' device is served by the *closet* WAP.

TABLE V

AVERAGE SPATIAL DISTANCES $< d >$ BETWEEN ANY TWO USERS SEEING THE SAME ACCESS POINTS UNDER POPULATIONS THE HETEROGENEOUS(HET) OR HOMOGENEOUS(HOM) MIXING IN EACH VORONOI CELL (WAP). THE AVERAGE SPATIAL DISTANCES ARE SHOWN AS $< d >$ WITH THE STANDARD VARIATION OF SPATIAL DISTANCES δ.

Teaching Building	Number of WAPs	$< d_{HET} >$(m)	δ_{HET}(m)	$< d_{HOM} >$(m)	δ_{HOM}(m)
the 2^{nd} bldg.	16	6.26	1.96	6.28	1.87
the 3^{rd} bldg.	21	7.14	0.96	7.23	1.01
the 4^{th} bldg.	15	7.86	1.71	7.97	1.87
the 5^{th} bldg.	28	6.98	1.37	7.12	1.43
the 6^{th} bldg.	33	7.11	1.59	7.09	1.60
the west bldg.	14	11.75	0.98	12.04	1.16

to the corresponding access controls(the central control components of the WiFI system) with a timestamp as its connecting time. When the wireless device leaves, the AP automatically collects the timestamp again as the disconnecting time of the client and transmits it to the corresponding access controls too. Therefore, a typical access-related trace contains online and offline time of the device as well as its hardware information and the connected AP.

B. The Voronoi Decomposition

In the campus WiFi system of Fudan University, there are overlapped regions between adjacent WAPs in the same floor. However, the device will connect to the WAP which has the highest strength of signal in order to keep data exchanging smoothly. In physics, wireless signals as electromagnetic wave signal is followed by the inverse-square law, indicating the signal is doubly attenuated with the growth of the spatial distance. Therefore, the overlapped region of adjacent WAPs will not influence the devices automatically connecting to the *closest* WAP. Only when the WAP is overloaded, the device will be switched to another adjacent WAP. In this WiFi system, each WAP can serve around 50 users. From our statistics, no WAP is overloaded. Besides, the WAPs in adjacent floors have none overlapped covered regions because the building materials can dramatically attenuate the wireless signals. Therefore, each floor can be decomposed by the WAPs into the corresponding Voronoi tessellations.

C. Preliminary for Estimating Spatial Distances between two users

It is impossible to precisely identify the position of a user within a Voronoi cell, and the spatial distance between any two users inside one Voronoi cell can not be directly calculated. Therefore, we propose an approximative method to estimate the average spatial distance of any two users in Voronoi cells as follows.

The approximative method need the empirical data of population numbers in each Voronoi cell as the preliminary to estimate the average spatial distance of any two users. The empirical total number of population in different classrooms can be analyzed from ‘ Fudan University Curriculum Schedule of the 1^{st} Semester in 2009-2010’ [1]. However, the population’s number of each Voronoi cell can not be directly analyzed because the number of users in a part of the classroom is unknown. Although we do not know the spatial population distribution in each classroom, without loss of generality, we can assume that the population is homogeneously mixing in each classroom. Therefore, the number of users in a part of the classroom can be estimated in proportion to the corresponding area, which is analyzed from the blueprint of all the teaching buildings. Finally, it is possible to show that the population in some cells(e.g., ‘WAP 4’ and ‘WAP 5’ in Figure 1) covering several classrooms is heterogeneously mixing, and in other cells(e.g., ‘WAP 3’ and ‘WAP 8’) covering a part of one classroom is homogeneously mixing.

As shown in Figure S5, the bottom floor of the 2^{nd} teaching building is decomposed into 8 Voronoi cells. The blue line presents the corresponding Voronoi cells. The regular rooms in yellow are classrooms, where WiFi users may interact with each other(activity regions). Other places(e.g.,corridor, toilets, storage room) in white are not considered for WiFi users interactions in this paper. Most of cells cover multi-classrooms(activity regions), thus the average spatial distances which is estimated using empirical data in the preliminary is the result under heterogeneous(HET) clustering. Furthermore, we also work the average spatial distances under homogeneous(HOM) mixing by conserving the total number of population in each cell but ignoring the number distribution in different covered classrooms (i.e., in the cell of ‘WAP 4’, the total number of population is equal to the empirical data, but these population are homogeneous distributed in each classroom.).

[1] http://www.jwc.fudan.edu.cn/s/67/t/179/3e/0d/info15885.htm.

D. The Results of spatial distances between any two users

Given the number of population in each activity region of Voronoi cells, the cells can be mapped to a cartesian coordinate system, and any dot inside the cell can be described by a cartesian coordinate. We randomly locate users to each activity region of Voronoi cell with the corresponding users' number to get the cartesian coordinate of each user, which is used to calculate the Euclidean distance between any two users. For instance, if there are N individuals in 'WAP 1', $N(N-1)/2$ Euclidean distances will be generated, the average Euclidean distance represents the average spatial distance between any two users 'seeing' this WAP. In the process on locating users to the cells, random fluctuation is introduced. Thus we apply Monte Carlo Method, repeating the whole locating process *M* times, and calculate the average value to reduce the fluctuation. The average spatial distance between any two users in each building is the average result of all average spatial distance of deployed WAPs. Table V shows the average spatial distances between any two users in each building under population heterogeneous(HET) mixing and homogeneous(HOM) mixing, which are both about 6-7 meters (excluding the *west* teaching building which has a large fluctuation on different floors, not shown in this paper), indicating that there is no cluster effect in each cell to estimate the average spatial distances.