Mining Social Interactions in Connection Traces of a Campus Wi-Fi Network

Eduardo Mañas-Martínez
Department of Signal Theory,
Networking and Communications,
University of Granada, Spain

Elena Cabrera
Department of Signal Theory,
Networking and Communications,
University of Granada, Spain

Katarzyna Wasielewska
Department of Signal Theory,
Networking and Communications,
University of Granada, Spain

David Kotz
Department of Computer Science,
Dartmouth College, USA

José Camacho
Department of Signal Theory,
Networking and Communications,
University of Granada, Spain

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; **Social aspects of security and privacy**; *Social aspects of security and privacy*; **Privacy protections**; **Social aspects of security and privacy**; • **Networks → Network privacy and anonymity**.

KEYWORDS

Privacy, Wi-Fi, connection logs, Dartmouth

ACM Reference Format:
Eduardo Mañas-Martínez, Elena Cabrera, Katarzyna Wasielewska, David Kotz, and José Camacho. 2021. Mining Social Interactions in Connection Traces of a Campus Wi-Fi Network. In *Proceedings of ACM Conference (SIGCOMM'21).* ACM, New York, NY, USA, 3 pages.

1 PROBLEM AND CONTRIBUTIONS

Wi-Fi technologies have become one of the most popular means for Internet access. As a result, the use of mobile devices has become ubiquitous and instrumental for society. A device can be identified through its MAC address within an autonomous system. Although some devices attempt to anonymize MAC addresses via randomization, these techniques are not used once the device is associated to the network [7]. As a result, device identification poses a privacy problem in large-scale (e.g., campus-wide) Wi-Fi deployments [5]: if the mobile device can be located, the user who carries that device can also be located. In turn, location information leads to the possibility to extract private knowledge from Wi-Fi users, like social interactions, movement habits, and so forth.

In this poster we report preliminary work in which we infer social interactions of individuals from Wi-Fi connection traces in the campus network at Dartmouth College [2]. We make the following contributions: (i) we propose several definitions of a *pseudo-correlation matrix* from Wi-Fi connection traces, which measure similarity between devices or users according to their temporal association profile to the Access Points (APs); (ii) we evaluate the accuracy of these pseudo-correlation variants in a simulation environment; and (iii) we contrast results with those found on a real trace.

2 APPROACH

To identify social interactions (such as connections to room-mates, co-workers, friends, and others), we define several mathematical descriptions of the *location coincidence* among devices. These descriptions leverage the temporal correlations in the devices' associations to APs. Thus, if two devices show a high correlation, we can infer that they may be related: the devices may be carried by the same individual, or they may be owned by room-mates or co-workers. In this work, we explore several different definitions of correlation perform. In future work, we plan to study temporal correlation patterns to infer social interactions.

We consider pseudo-correlation variants calculated using the following formula: $C(x,y) = \frac{S(x,y)}{N}$, where x and y are two devices, $S(x,y)$ refers to the number of sampling intervals when x and y are in a close location and N is the total of sampling intervals considered. In this work, sampling intervals are minutes and the total time is an hour, so that N is at most 60.

According to the previous definition, two devices increase their correlation when they are in a 'close' location at nearly the same time, and decrease it otherwise. We assume that two devices are not in a close location when they are associated to different APs, or when one of them is not associated to the network. Still, we need to take into account the case when both devices are not associated to the network during the same sampling time $\hat{t}$, and define three alternatives:

- $C_1(x,y)$: if $S(x,y)$ and N increase with $\hat{t}$, i.e., non-associated pairs of devices are considered to be together.
- $C_2(x,y)$: if only N increases with $\hat{t}$, i.e., non-associated pairs of devices are considered to be separated.
- $C_3(x,y)$: if neither $S(x,y)$ nor N increase with $\hat{t}$, i.e., the sampling times with non-associated pairs of devices are not taken into account at all.

3 PRELIMINARY RESULTS

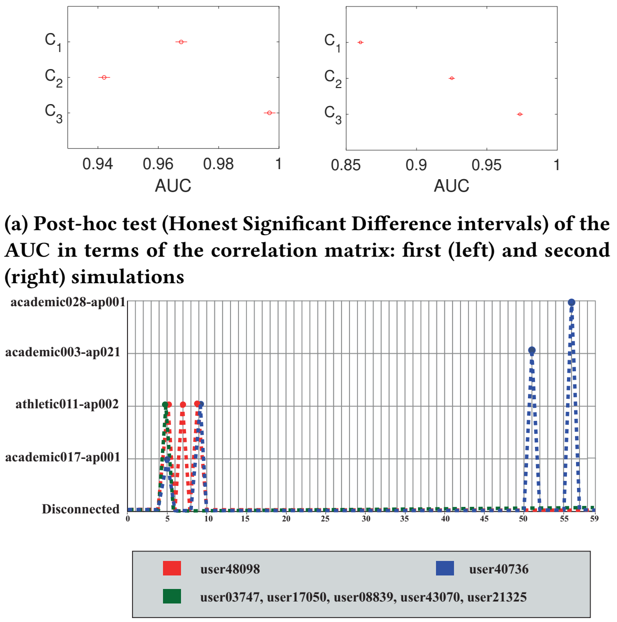
To understand the differences in accuracy of the alternative approaches, we developed a simulation engine [6], using BonnMotion [1] and the Matlab environment. In simulation, we can configure the movement of groups while controlling the devices in each group, the simulation time, the location of APs, and other parameters. The output of a simulation run is a connection log, with timestamps indicating when each device is connected to the network and through which AP. We assume that devices connect to the closest AP as long as its location is within the coverage radius of that AP. From the simulation data, we can evaluate the accuracy of a given correlation matrix in automatically identifying the devices in each group from the association logs. To measure accuracy, we use the *Area Under the ROC Curve* (AUC) metric [3].

We devised a full-factorial experimental design [8], considering 6 factors and 3 levels each, to understand how the accuracy changes with changes in the environment and the pseudo-correlation matrix chosen. The details are shown in Table 1. We consider 10 replicates for each experimental run. In total, we ran $10 \cdot (3^6) = 7,290$ experiments. We analyse the results with an analysis of variance (ANOVA) [8] and the box-cox transform [9] to meet normality assumptions. The ANOVA shows that all factors are significant (p-value « 0.01) and that the most relevant factors, according to the effect size, are the simulation area and the type of correlation. Figure 1a, left, shows the post-hoc test for the correlation matrix, which illustrates the superiority of C_3, with AUC close to 1.

To perform experiments on real data, we use the Wi-Fi trace of the Dartmouth College [4]. Dartmouth has a compact campus with over 200 buildings on 200 acres. At the end of 2018 the College had nearly 6,500 students, 3,300 staff, and 1,000 faculty, and the network included more than 3,000 APs. We have data about the users' network connections across seven years from 2012 to 2018 [2]. Although we aim to study all seven years, in this preliminary study we focus on one hour of data (11:00–12:00 local time, 1 Nov. 2018).

The original data [4] was collected under a protocol approved by Dartmouth's Institutional Review Board (IRB). We worked with a copy of the dataset that had been anonymized for use by researchers: each identifier (UserName, UserMAC, APName) had been replaced with a consistent, unique pseudonym of the same format. Ethical considerations regarding potential re-identification of users need to be explored in future work.

An initial analysis showed that the simulation was not modelling the real data with fidelity. The most correlated group of devices during the hour analyzed, according to C_3, includes 7 devices. They are not associated to any AP for most of the hour, and only coincide in a single AP during a single minute (see temporal pattern of associations in Figure 1b). To understand why C_3 was performing so differently in the simulation and real data, we devised the *mean association rate (MAR)* statistic, defined as the average percentage of time a device is associated to the network. We found that our simulation has an average MAR=42.8% while the real trace, during the hour considered, has an average MAR=8.0%. To reduce the MAR, we repeated the simulation considering that any device would be switched off, on average, 4 out of 5 sampling times, yielding a MAR=8.6%. Repeating the analysis, we found a generalised reduction of accuracy (Figure 1a, right), specially in C_1, and that C_3 was still outperforming the other matrices.



(a) Post-hoc test (Honest Significant Difference intervals) of the AUC in terms of the correlation matrix: first (left) and second (right) simulations

(b) Time profile of selected users

Figure 1: Simulation and experimental results

Table 1: Factors and levels of the experiment

Factors	Level 1	Level 2	Level 3
Type of correlation	C_1	C_2	C_3
Number of APs	20	60	120
Devices per group	1	2	4
Simulation area	50x50	100x100	200x200
Space between group members	0.1	0.5	1
AP coverage radio	2	4	8

4 CONCLUSION AND FUTURE WORK

Our approach to mine association traces to infer device/user connections gives reasonable performance in simulated data, but other challenges need to be addressed for real data. Future work will focus on:

- Exploring uncertainty measures connected to a correlation score.
- Extending the analysis to the seven-year trace, and exploring temporal correlation patterns.
- Re-assessing the assumption of 'closeness' for two devices connected to the same AP using geographical information of the APs.
- Assessing related privacy concerns, in particular, the possibility to re-identify users in the anonymized data.

ACKNOWLEDGEMENTS

This work was supported by Dartmouth College, by ACM SIGMOBILE, and (earlier) by US NSF award number 0454062. We appreciate the support of colleagues in the collection and anonymization of Wi-Fi data: most notably Tristan Henderson, Chris McDonald, and Nathan Schneider. This work was also supported by the Agencia Estatal de Investigación in Spain, grant No PID2020-113462RB-I00, and the European Union's Horizon 2020 Marie Skłodowska-Curie grant agreement No 893146.

REFERENCES

[1] Nils Aschenbruck, Raphael Ernst, Elmar Gerhards-Padilla, and Matthias Schwamborn. 2010. BonnMotion: A Mobility Scenario Generation and Analysis Tool. In *Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques (SIMUTools '10)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, Article 51, 10 pages. https://doi.org/10.4108/ICST.SIMUTOOLS2010.8684

[2] José Camacho, Chris McDonald, Ron Peterson, Xia Zhou, and David Kotz. 2020. Longitudinal analysis of a campus Wi-Fi network. *Computer Networks* 170 (2020), 107103. https://doi.org/10.1016/j.comnet.2020.107103

[3] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[4] David Kotz and Kobby Essien. 2005. Analysis of a Campus-wide Wireless Network. *Wireless Networks* 11, 1–2 (Jan. 2005), 115–133. https://doi.org/10.1007/s11276-004-4750-0

[5] Feng Lyu, Ju Ren, Nan Cheng, Peng Yang, Minglu Li, Yaoxue Zhang, and Xuemin Shen. 2019. Big Data Analytics for User Association Characterization in Large-Scale WiFi System. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*. IEEE, New York, USA, 1–6. https://doi.org/10.1109/ICC.2019.8761511

[6] Roberto Magán-Carrión. 2016. *Supervivencia en redes ad hoc. Mecanismos de tolerancia y reacción frente amenazas de seguridad.* Ph.D. Dissertation. Universidad de Granada.

[7] Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C. Rye, and Dane Brown. 2017. A Study of MAC Address Randomization in Mobile Devices and When it Fails. *Proceedings on Privacy Enhancing Technologies* 4 (2017), 268–286.

[8] Douglas C Montgomery. 2017. *Design and analysis of experiments*. John Wiley & Sons, New Jersey, USA.

[9] Jason Osborne. 2010. Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation* 15, 1 (2010), 12.