MINING SOCIAL INTERACTIONS IN CONNECTION TRACES OF A CAMPUS WI-FI NETWORK

E. Mañas-Martínez[(1)], E. Cabrera[(1)], K. Wasielewska[(1)], D. Kotz[(2)], J. Camacho[(1)]

(1) Department of Signal Theory,Networking and Communications, University of Granada, Spain
(2) Department of Computer Science, Dartmouth College, USA

mmeduardoa@correo.ugr.es, elenacabrera@correo.ugr.es, k.wasielewska@ugr.es, david.f.kotz@dartmouth.edu, josecamacho@ugr.es

Topic: Device identification (MAC) poses a privacy problem in large-scale (e.g., campus-wide) Wi-Fi deployments: if the mobile device can be tracked, the user who carries that device can also be tracked. In turn, from location information we can extract private knowledge from Wi-Fi users, like social interactions, movement habits, and so forth.

Focus: We investigate methods to infer social interactions of individuals from Wi-Fi connection traces in the campus network at Dartmouth College.

Preliminary Results: Our approach to mine association traces to infer device/user interactions gives reasonable performance in simulated data, but other challenges need to be addressed for real data.

Mining Approach

Input Data: Wi-Fi association traces, with the Access Point (AP) ID, the User Device (UD) ID and the timestamp.

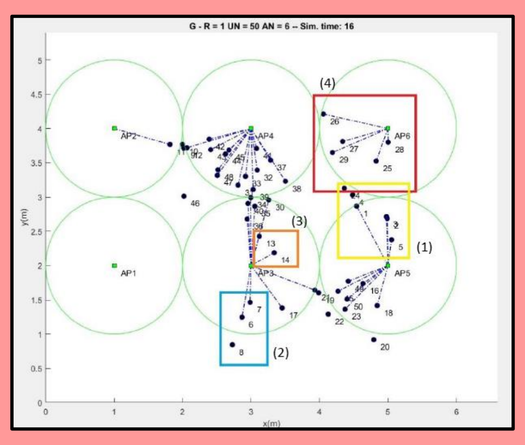
Preprocessing: For each UD and sampling time interval (e.g., one minute intervals), we compute the AP (if any) to which it is connected. The result is, for each UD, a time series of associations to APs

Infer social interactions: Pseudo-correlation Matrices leverage the temporal correlations in the devices' associations to APs. We consider the formula C(x,y)= $s(x,y)/N$, where x and y are two devices, S(x,y) refers to the number of sampling intervals when x and y are in the same AP, and N is the total of sampling intervals considered. We consider three variants of Pseudo-correlations

Usuarios		Usuarios									
		n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10
Usuarios	n=1	0.66	0.6	0.59	0.53	0.52	**0**	0.06	0.07	0	0
	n=2	0.6	0.83	0.69	0.56	0.75	0.13	0.16	0.18	0	0
	n=3	0.59	0.69	0.76	0.55	0.65	0	0.06	0.07	0	0
	n=4	0.53	0.56	0.55	0.66	0.54	0.02	0.06	0.07	0	0
	n=5	0.52	**0.75**	0.65	0.54	0.84	0.13	0.18	0.2	0	0
	n=6	0	0.13	0	0.02	0.13	0.71	0.51	0.28	0.05	0
	n=7	0.06	0.16	0.06	0.06	0.18	0.51	0.82	0.48	0.13	0
	n=8	0.07	0.18	0.07	0.07	0.2	0.28	0.48	0.5	0.09	0
	n=9	0	0	0	0	0	0.05	0.13	0.09	0.9	0.62
	n=10	0	0	0	0	0	0	0	0	0.62	0.7

Simulation Approach

We use **Bonnmotion v2.1.3** and **Matlab R2016a** to simulate an environment with a number of APs and UDs that move in groups



Simulation example with different groups of UDs

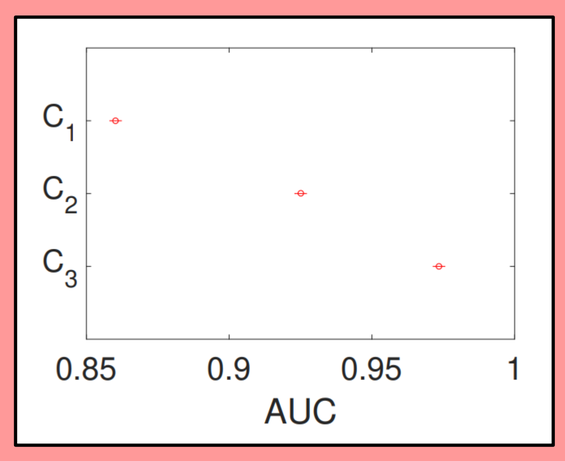
Simulation experiment: **Full factorial design** from **AUC**

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

Results by **ANOVA**

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	0.17375	2	0.08688	943.6	0
X2	0.15246	2	0.07623	827.98	0
X3	0.00432	2	0.00216	23.46	0
X4	0.42961	2	0.21481	2333.08	0
X5	0.00419	2	0.0021	22.76	0
X6	0.14947	2	0.07474	811.75	0
Error	0.66999	7277	0.00009		
Total	1.5838	7289			

Honest Significant Difference Intervals of the different pseudo-correlation variants for **AUC**

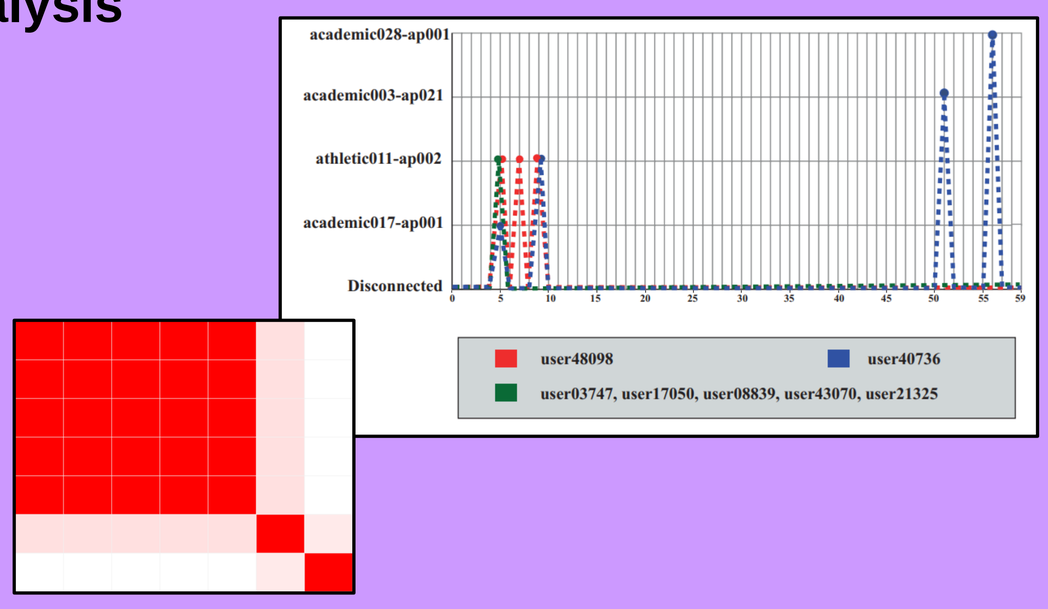


Variant **C3** with optimal **AUC**

Real Data Analysis

To perform experiments on real data, we use the Wi-Fi trace of the **Dartmouth College** [1]**.** The dataset has been anonymized: each identifier (UserName, UserMAC, APName) had been replaced with a consistent, unique pseudonym of the same format. In this preliminary study, we focus on one hour of data (11:00–12:00 local time, 1 Nov. 2018).

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 C3, includes 7 devices not associated to any AP for most of the hour, and only coincide in a single AP during a single minute



Future work will focus on analyzing the 7 years of the Dartmouth College Wi-Fi trace, re-evaluating the assumption of "proximity" between two users, exploring uncertainty measures to improve accuracy and assessing related privacy concerns.

[1] 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.

[**ACKNOWLEDGEMENTS:** This work was supported by Dartmouth College, by ACM SIGMO-BILE, 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-Curiegrant agreement No 893146.