Jestr

Journal of Engineering Science and Technology Review 15 (3) (2022) 147 - 152

JOURNAL OF
Engineering Science and Technology Review

Research Article

www.jestr.org

Evolution of the Use of Random MAC Addresses in Public Wi-Fi Networks

Carlos Andres Gomez[1,*], Laura Juliana Guerrero[2] and Luis Fernando Pedraza[3]

[1]*Universitaria Agustiniana, Faculty of Engineering, Telecommunications Engineering, Colombia.*
[2]*Datawifi SAS, Innovation Department, Colombia.*
[3]*Universidad Distrital Francisco José de Caldas, Technological Faculty, Telecommunications Engineering, Colombia.*

Received 18 September 2021; Accepted 9 June 2022

Abstract

The use of random MAC addresses is considered as an important tool to improve a user's privacy on a Wi-Fi network. Although this technique was proposed since 2014, its implementation was not effective immediately but had an uneven evolution over the following years. This paper studies the evolution of random MAC addresses based on the analysis of data collected in public Wi-Fi networks located in different areas of Latin American countries from 2016 to 2021. Statistical analysis is performed using the analysis of variance technique (ANOVA) that tests whether there are significant differences in the rate of MAC randomization over different periods of time and thereby identified an uneven evolution of the implementation of MAC address randomization measures in mobile devices with different operating systems, and also the massification of these measures from 2019. The development of the massification of randomized MAC addresses and the impact of the existence of centralized and widely controlled operating systems, such as iOS, is evident.

Keywords: MAC address, randomization, Android, iOS

1. Introduction

The development of data networks has created identifiers for the equipment involved in communications in order to meet the needs of information transmission; however, when the identifier can be associated with an end user, there is the possibility of using these identifiers to track individuals and thus violate their privacy. One of the most popular identifiers is the Media Access Control (MAC) address used by Wi-Fi and Bluetooth connections [1]. That is why the Institute of Electrical and Electronics Engineers (IEEE) created a methodology for randomizing these MAC addresses, applied especially in end-user devices such as cell phones and computers; although the adoption of these measures has taken many years and has depended on the efforts of several actors.

This paper analyzes the evolution of MAC address randomization in public Wi-Fi network connection scenarios in Latin American countries during the last years, using the proportion of randomized addresses and their operating system. Some previous studies have analyzed this process; however, the presence of randomized MACs was not of major relevance at the time of those studies [2], [3].

2. Literature Review

Different studies on the use of MAC addresses have focused on applications such as people identification, device tracking, travel time calculation in transportation systems, vehicle origin-destination estimation, etc. In [4], [5], [6], [7] and [8] the use of MAC addresses generated by devices associated to vehicles using Wi-Fi and Bluetooth technologies is studied to understand different variables such as travel times and origin-destination of vehicles. In [4] a system called Media Access Control Address Detection (MACAD) was developed that detects the signals of Bluetooth devices emitted by vehicles to identify them and perform an analysis on the travel time of each vehicle, in addition, different hardware alternatives were examined to improve the detection of the signals and the recognition of the MAC address, as a basis for the study of the travel time calculation.

Other researches [9] and [10] have used the mobilizing vehicles to scan MAC addresses through different routes, with which user location maps were developed. Advantages and challenges in using Bluetooth and Wi-Fi to collect MAC address data for crowd monitoring are presented in [11].

There have also been some studies on the implementation of MAC address randomization techniques in mobile devices and Wi-Fi networks, including the following. In [12], the need to protect the location privacy of clients in wireless networks is discussed. To solve these challenges, a location privacy protection package is proposed that includes a Dynamic MAC Address Assignment scheme and a shuffling scheme with a silence period. This strategy will hide the real identity of the client node in the WLAN and makes the involved nodes construct a mixing zone to prevent any adversary from tracking the client's movement in the WLANs. In [13], the first large-scale study on MAC address randomization is presented, detailing the implemented randomization policies, associated device models and device identification methods. In [14], MAC address randomization is analyzed by conducting a study with 160 cell phone models, determining whether randomization is used, if so, under what conditions it randomizes its MAC address, and whether known tracking vulnerabilities are mitigated.

A study on vulnerabilities in randomized MAC addresses is presented in [15]. In addition [16] indicates that the MAC address is not the only data that can be used for tracking, but also the content of the frames and their timing can be used to

*E-mail address: carlos.gomez@uniagustiniana.edu.co
ISSN: 1791-2377
doi:10.25103/jestr.153.15

achieve device tracking despite the randomization of the MAC address.

On the other hand [17] has studied the relationship in privacy management using MACs and routing layer in IoT applications. The author has proposed a data privacy preserving scheme using "ElGamal" cryptosystem after route selection in information routing of IoT systems. The author has succeeded in improving the effectiveness of secure data privacy preservation at the routing layer by experimental analysis of the routing layer

A limited number of studies have evaluated the real impact of random MAC addresses, largely due to the fact that their implementation in operating systems has been delayed, and as presented in this paper, only until 2020 did they become widespread. With respect to the above [2] analyzed several data sets of polling requests between the years 2013 and 2017, where a trend of overall increase in the use of absolute random addresses was observed. However of the total samples analyzed did not exceed 3% samples containing random MAC addresses. In [3] a study was developed in 2019 on randomization measures in the Wi-Fi connectivity market, finding that by that date MAC randomization functions had not yet caused serious problems in existing services.

3. MAC Addresses

The MAC address is a unique identifier used by a network interface (NIC) of most IEEE 802 network technologies, including Ethernet, Wi-Fi and Bluetooth. MAC addresses are primarily assigned by device manufacturers according to IEEE-controlled usage authorizations, so they are often referred to as an Ethernet hardware address or physical address. MAC addresses are formed according to the 6-byte principles used in two numbering spaces based on extended unique identifiers (EUI), the address usually includes an organization unique identifier (OUI) of the manufacturer represented in the first three bytes of this address. The last three bytes are the network interface controller (NIC) [2].

In order to generate MAC addresses randomly, the seventh bit of the first byte of the OUI is used as a flag: the local administrator bit (LA bit). The two least significant bits of the initial octet of the MAC address are used for special purposes. The least significant bit of octet 0 (the I/G bit) indicates an individual address (I/G=0) or a group address (I/G=1), and the second least significant bit of octet 0 (the U/L bit) indicates the universally administered address (U/L=0) or indicates a locally administered address (U/L=1) [18]. A universally managed address is understood as a globally unique address; while a locally managed address is used to generate random addresses, Fig. 1 shows the general scheme of a MAC address and the location of the LA bit.

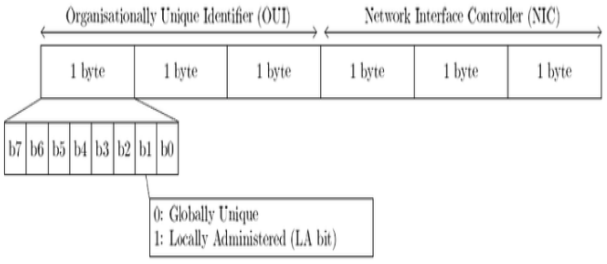

Fig. 1. MAC address structure. [2]

The use of the LA bit to indicate the existence of random MAC addresses has been the responsibility of the operating systems of end-user devices, especially mobile devices.

4. Material and methods

The massive application of randomized MAC addresses is measured in this work from samples of addresses of connections in public Wi-Fi zones from 2016 to May of 2021, in order to analyze the impact of these randomized addresses over time with respect to operating systems.

Therefore, the randomization data are reviewed from a repeated measures approach, to analyze the increase in this phenomenon in the observation periods. For this purpose, the analysis of variance (ANOVA) technique is used to test if there are significant differences in the randomization rate of the MACs over different time periods. Likewise, the analysis is replicated including as an associated factor the operating system of the devices in order to verify if the evolution profile of the types of system presents a similar behavior in the observation periods.

4.1 Data samples

This paper analyzes the records of final user MAC addresses in public Wi-Fi zones, from 2016 to May 2021. These records are obtained by the company DataWifi SAS in Colombia, which provides data analytics services on Wi-Fi networks.

These data correspond to various public Wi-Fi zones located in Colombia (CO), Ecuador (EC), Peru (PE), Mexico (MX), Puerto Rico (PR), Dominican Republic (RD), Panama (PA), El Salvador (SV) and Argentina (AR), which represent different data observation points for this study.

Tab. 1 shows the consolidated data for each year of the study, differentiating the presence of data from Wi-Fi zones in different countries and observation points.

Table 1. Description of data

Year	Amount of data	Countries	Types of observation points
2016	21495	CO	Retail
2017	417613	CO, EC	Retail, Government, ISP
2018	1806237	CO,EC, PE	Retail, Government, ISP, Hotels, Banks
2019	11591024	CO,EC, PE, PA, SV	Retail, Government, ISP, Hotels, Banks, Restaurants, Hospitals
2020	10543398	C CO,EC, PE, PA, SV, MX, DR, PR	Retail, Government, ISP, Hotels, Banks, Restaurants, Hospitals, Transportation Systems

2021	8308144	C CO,EC, PE, PA, SV, MX, DR, PR,AR	Retail, Government, ISP, Hotels, Banks, Restaurants, Hospitals, Transportation Systems

The data collected has the following structure:

ID - Unique identifier of the sample
MAC user - MAC address used by the user device to manage its Wi-Fi connection.
MAC access point - MAC address of the Access Point device in the Wi-Fi zone. This data was not used in this study.
ID_zone - Wi-Fi zone identifier.
ID_Venue - Observation point identifier.

4.2 Data analysis methods

In this study ANOVA is used, which is based on partitioning the sources of variation in the data. Given that the total sum of variance of the observations y_{ij} s measured in terms of the deviations of each observation around the mean $\bar{y}$, without taking external factors into account, according to equation 1.

$$y_i - \bar{y} \tag{1}$$

When some external factor explaining changes in the response variable *y* is taken into account, the total variability can be decomposed into the variance between the response y_{ij} and the mean of the variable at the factor level $\bar{y}_{i.}$, and the variability between the means at the factor levels $\bar{y}_{i.}$ and the overall mean $\bar{y}$

From the above, the total variation can be broken down into two terms.

$$y_{ij} - \underline{y}_{..} = (\underline{y}_{i.} - \underline{y}_{..}) + (y_{ij} - \underline{y}_{i.}) \tag{2}$$

The first term of equation (2) corresponds to the sum of squares total(SST), which corresponds to the sum of squares regression or treatment (SSR) and the sum of squares error(SSE). [19].

$$SST = SSR + SSE \tag{3}$$

The sum of squares of the model describes the variability defined by the deviations between the mean estimates of the factors and the general mean, which in the concept of experimental design corresponds to observing and analyzing a response of two possible treatments, and thus estimating the mean response for each of the treatments and quantifying the deviations between these and the general mean. These deviations are expected to be maximum, since the treatment factor is the one that reflects the greatest variability in the responses. The sum of squares of the error measures the random variability that exists in the observations with respect to their factormean estimate.

Given r levels in the external factor, the ANOVA model can be described as:

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{4}$$

where $i = 1, \ldots, r$ and $j = 1, \ldots, n$

Where y_{ij} is the j-th observed response of the i-th level of the treatment factor, μ_i is the mean parameters of the treatments/factors and ϵ_{ij} are the random errors.

Based on this model we seek to test the following hypotheses:

$$H_0: \mu_1 = \ldots = \mu_r \tag{5}$$

$$H_a: not\ al\ \mu_i are\ equal$$

This hypothesis test can be tested using the sums of squares described above by means of the F statistic, described as:

$$F = \frac{MSTR}{MSE} = \frac{SSR/r-1}{SSE/n-r} \tag{6}$$

Where the F-statistic is calculated as the ratio of the mean square treatment (MSTR) to the mean square error (MSE). High values of the F-statistic support the hypothesis H_a since it implies that MSTR > MSE, while values close to 1 support H_0.

ANOVA on repeated measures is a way of analyzing longitudinal data, where time or order of measurement is used as the factor of r levels or repeated measures [20].

5. Results

Fig. 2 shows the evolution of the percentage of random MAC addresses over time. As of March 2020 there is an exponential increase in the randomness of MACs, which to date reaches over 45% of the observed MACs.

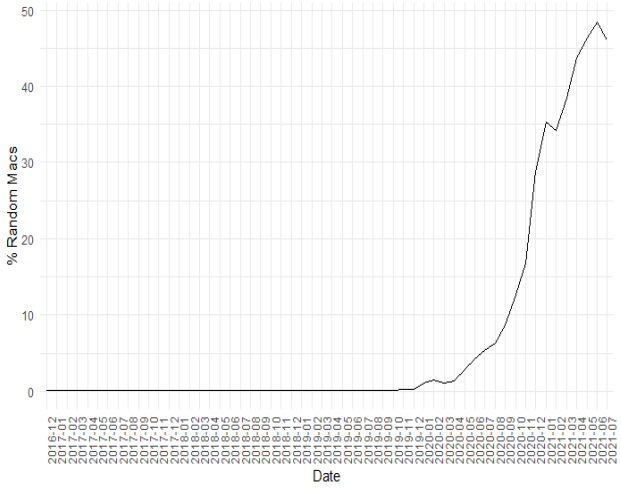

Fig. 2. Evolution of the percentage of random MAC addresses over time.

In the exploration of randomness, strong differences are observed between the proportions observed according to the type of operating system of the devices. Fig. 3 shows that while the percentage of MAC address randomness in devices with Android operating system is around 20-25% from the month of December 2020 to the month of May 2021, for devices with iOS operating system it is between 75-80%. However, in this particular sample the majority of data comes from Android devices, which makes the overall average randomness smaller between 25-50%.

Fig. 4 shows a positive linear relationship between the presence of iOS devices and the increase in MAC address randomness. In the case of Android devices, as shown in Fig. 5, the behavior is more random and no defined pattern is found. Additionally, it is corroborated in the data sample that most observation points have low concentrations of iOS devices and high concentrations of Android devices. Still, in both cases it is perceived that over time the proportionality of random MAC addresses has increased.

5.1. Time difference tests

The following is an analysis of the evolution of the proportion of random MAC addresses observed in device connections, for which two time comparison scenarios are evaluated. First, from July 2020 to May 2021 three observation periods with five months between them are tested. The second scenario corresponds monthly observation, from December 2020 to May 2021.

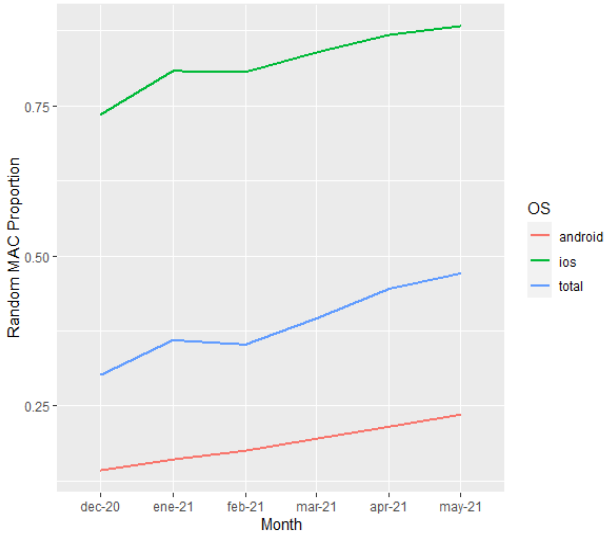

Fig. 3. Proportion of random MAC addresses by operating system.

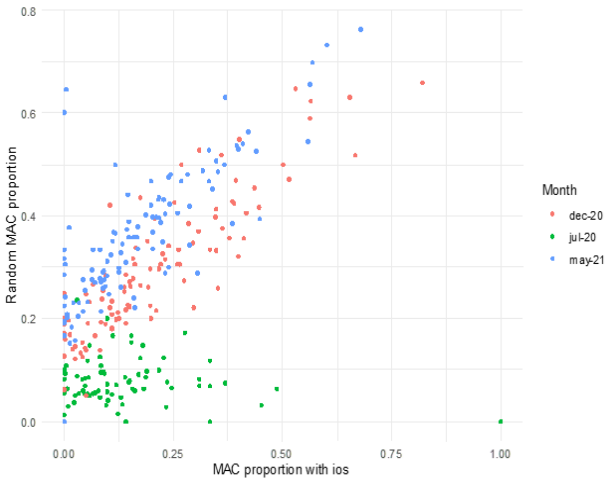

Fig. 4. Evolution of random MAC addresses on iOS.

Based on the devices connected to the available WiFi networks, the first scenario contains 78 observation points while the second has 109. From the MAC addresses observed during the time periods, the percentage of randomness at each point is calculated.

For the first scenario, Tab. 2 and Fig. 6 show how the average proportion of random MAC addresses has increased over time, from 7% to 35% in less than a year.

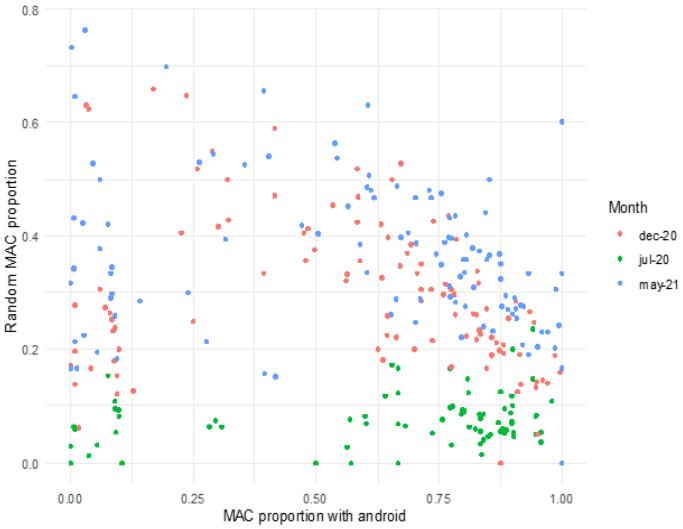

Fig. 5. Evolution of random MAC addresses in Android.

Table 2. Proportion of random MAC addresses in an interval of about one year.

Date	n	mean	Standard deviation
Jul-20	78	0.073	0.048
Dec-20	78	0.290	0.138
May-21	78	0.357	0.141

Fig. 6 shows that just as the number of random MACs has increased over time, so has the variability of this proportionality in connection locations.

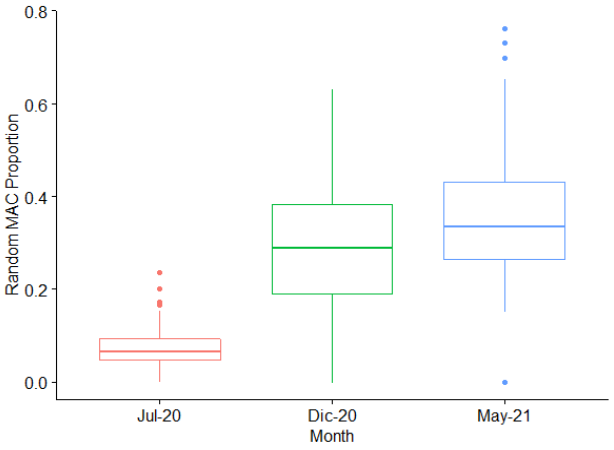

Fig. 6. Boxplot of random MAC addresses over an interval of about one year.

A one-way ANOVA is now fitted to test whether there are significant differences between the proportion of randomness observed in the periods. Tab. 3 shows there is a significant increase in the randomness over time. Tab. 4 shows all the pair comparisons are found to be significantly different, thus concluding that the increase in randomness in the year of observation had a significant increase. From Tab. 4 it can be concluded that the difference between July and December 2020 is much larger than that observed between December 2020 and May 2021.

For the second scenario, the evolution of randomness in MAC addresses is analyzed for six continuous months, in order to verify whether the change remains significant. As shown in Tab. 5 and Fig. 7, the randomness continues to increase over time; although the change is not as abrupt as that observed from July to December 2020 or May 2021. Additionally, the variability of randomness at the observed locations remains more stable.

Table 3. ANOVA for three measurements in the months of Jul-20, Dec-20 and May-21.

Component	Degrees of freedom	Sum Square	Mean Square	F value	P-value
time	2	3.442	17.208	124.3	<2e-16***
Error	231	3.197	0.0138		

Table 4. Pairwise t-test between measurements of jul-20, dic-20 and may-21

Time 1	Time 2	statistic	P value
dic-20	jul-20	14.0	2.26e-22***
dic-20	may-21	-7.44	3.69e-10***
jul-20	may-21	-17.8	1.34e-28***

Table 5. Proportion of random MAC addresses for 6 continuous months.

Time	n	mean	Standard deviation
Dec-20	109	0.288	0.14
Jan-21	109	0.321	0.144
Feb-21	109	0.326	0.142
Mar-21	109	0.338	0.129
Apr-21	109	0.351	0.134
May-21	109	0.361	0.135

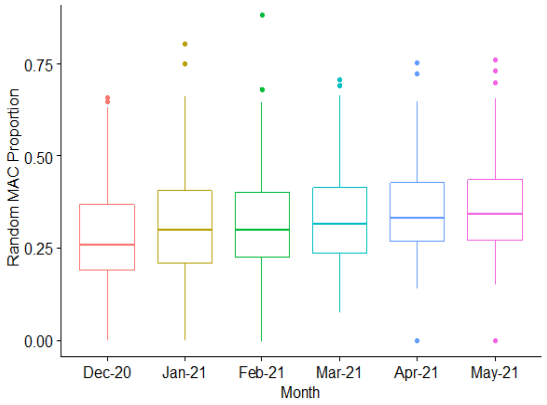

Fig. 7. Behavior of randomness over time.

The test of differences by means of ANOVA is presented in Tab. 6. Once again, it is observed that there is at least one significant difference between the proportionality of randomization in the observed months.

Table 6. ANOVA for 6 continuous months.

Component	Degrees of freedom	Sum Square	Mean Square	F value	P-value
time	5	0.369	0.07375	3.909	0.00169**
Error	648	12.224	0.01886		

Tab. 7 shows the pairwise test to identify between which months there is a significant difference. The pairs turn out to be different when there is more than one month difference between them, except for the change between December 2020 and January 2021 where the increase in randomness is significant.

Tabla 7. Pairwise t-test para 6 meses continuos

Time 1	Time 2	statistic	p-value
Apr-21	Dec-20	6.73	1.28e- 8
Apr-21	Jan-21	3.35	1.60e- 2
Dec-20	Feb-21	-3.99	2.00e- 3
Dec-20	Jan-21	-4.60	1.70e- 4
Dec-20	Mar-21	-7.70	1.04e-10
Dec-20	May-21	-9.73	2.85e-15
Feb-21	May-21	-3.73	5.00e- 3
Jan-21	May-21	-4.77	8.67e- 5
Mar-21	May-21	-3.65	6.00e- 3

6. Discussion

The results obtained show an evolution in MAC address randomization techniques. As evidenced in Figure 2, as of September 2020, a proportion of random MAC addresses higher than 10% of the total addresses detected in the analyzed public Wi-Fi network datasets is detected.

This allows us to compare the findings with similar works [2] and [13], through the Table 8, where the results of similar Wi-Fi zones have been taken: public and open Wi-Fi zones, as the case of Sapienza dataset of [2], since private or experimental Wi-Fi zones are hardly comparable to the scenario studied in this work.

Table 8. Comparison of results with similar studies.

	Year	Total MAC Addresses	Total random MAC	% Random MAC
[2]	2013	160.000	320	0,2%
[13]	2017	2.604.901	1.388.565	53,3%
This work	2017	417.613	405	0,1%
	2018	1.806.237	1.815	0,1%
	2019	11.591.024	16.860	0,1%

	2020	10.543.398	980.939	9,3%
	2021	8.308.144	3.471.709	41,8%

7. Conclusions

In this paper, the evolution of MAC address randomization was studied over the last few years in public Wi-Fi hotspots. From this, it became evident that randomized MAC addresses had not had a real impact on wireless local area networks (WLANs) until 2019, despite the fact that randomization measures began to be implemented since 2014. After 2019, there is an exponential increase in the use of randomized MAC addresses, which is related to the default implementation by Android 10 and iOS 14 mobile operating systems and later versions. Therefore, the use of random MAC addresses has a direct relationship with the mobile operating systems, in particular a linear behavior over time of the implementation of this measure was evidenced in smartphones with iOS operating system, which is a unified and controlled operating system; while the use of random MAC addresses in the Android operating system was observed irregular and with little control, which may be due to the absence of centralized control.

The increase in randomness showed an almost exponential growthsince the beginning of 2019, and it was observed that during the last year and the last six months of this study this trend continues to increase, although not at the same rate. The differences in the randomness of MAC addresses between the months of July and December 2020, and between December 2020 and May 2021 turn out to be notorious. Since then the rate of increase of this randomness has been reduced but has not finished stabilizing and continues to grow. During the last months of the study, an increase rate of approximately one percentage point for each month was observed.

In addition, further research is needed on the impact of MAC address randomization on data network management, as it has been shown that the MAC address is not the only tool for maintaining the link to the unique identification of a device.

References

1. M. Cunche. (2013). I know your MAC Address: Targeted tracking of individual using Wi-Fi. Journal of Computer Virology and Hacking Techniques 10(4):219-227. DOI: 10.1007/s11416-013-0196-1
2. C. Matte, M. Cunche. (2018). Spread of MAC address randomization studied using locally administered MAC addresses use historic. https://hal.inria.fr/hal-01682363 ; Research Report. RR-9142, Inria Grenoble Rhône-Alpes. 2018.
3. C. Ansley. (2019). MAC Randomization in Mobile Devices. SCTE•ISBE expo 2019 and NCTA Fall Technical Forum 2019
4. Y. Wang, Y. Malinovskiy, Y. Wu, U. Lee. (2011). Error modeling and analysis for travel time data obtained from bluetooth MAC address matching. University of Washington, Department of Civil and Environmental Engineering. Research Report Agreement T4118 Task 46 Bluetooth Time Data TNW 2011-01 TransNow Budget 61-8390
5. C. Bakula, W. Schneider, J. Roth. (2012). Probabilistic Model Based on the Effective Range and Vehicle Speed to Determine Bluetooth MAC Address Matches from Roadside Traffic Monitoring. Journal of Transportation Engineering. Volume 138 Issue 1 - January 2012
6. T. Tsubota, T. Yoshii. (2017). An Analysis of the Detection Probability of MAC Address from a Takahiro Tsubota and Toshio Yoshii Moving Bluetooth Device. Transportation Research Procedia, Volume 21, Pages 251-256, ISSN 2352-1465, https://doi.org/10.1016/j.trpro.2017.03.094.
7. S.M. Remias, A. M. Hainen, M. Jijo, V. Lelitha, S. Anuj, D. Bullock. (2017). Travel Time Observations Using Bluetooth MAC Address Matching: A Case Study on the Rajiv Gandhi Roadway: Chennai, India. Purdue University, West Lafayette, Indiana, 2017. https://doi.org/10.5703/1288284316505
8. M. Blogg, C. Semler, M. Hingorani, R. J. Troutbeck. (2010). Travel Time and Origin-Destination Data Collection using Bluetooth MAC Address Readers . Transport Research Forum 2010 Proceedings 29 September – 1 October 2010, Canberra, Australia.
9. A. Hidayat, S. Terabe, H. Yaginuma. (2017). Mapping of MAC Address with Moving WiFi Scanner. International Journal Of Artificial Intelligence Research. DOI: https://doi.org/10.29099/ijair.v1i2.27
10. A. Hidayat, S. Terabe, H. Yaginuma. (2018). International review for spatial planning and sustainable development A: Planning Strategies and Design Concept, Vol.6 No.3 (2018), 154-167. ISSN: 2187-3666 (online). DOI: http://dx.doi.org/10.14246/irspsda.6.3_154
11. N. Abedi, A. Bhaskar, E. Chung. (2013). Bluetooth and Wi-Fi MAC Address Based Crowd Data Collection and Monitoring: Benefits, Challenges and Enhancement. Australasian Transport Research Forum 2013 Proceedings
12. M. Lei, X. Hong, S. Vrbsky. (2007). Protecting Location Privacy with Dynamic Mac Address Exchanging in Wireless Networks. IEEE GLOBECOM 2007 proceedings
13. J. Martin, T. Mayberry, C. Donahue, L. Foppe, L. Brown, C. Riggins. E. Rye, D. Brown. (2017). A Study of MAC Address Randomization in Mobile Devices and When it Fails. Proceedings on Privacy Enhancing Technologies,2017(4) 365-383. https://doi.org/10.1515/popets-2017-0054
14. E. Fenske, D. Brown, J. Martin, T. Mayberry, P. Ryan, E. C. Rye. (2021). Three Years Later: A Study of MAC Address Randomization In Mobile Devices And When It Succeeds. Proceedings on Privacy Enhancing Technologies, 2021, 164 - 181.
15. M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, F. Piessens. (2016). Why MAC Address Randomization is not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms. In Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (ASIA CCS '16). Association for Computing Machinery, New York, NY, USA, 413–424. DOI:https://doi.org/10.1145/2897845.2897883
16. C. Matte, M. Cunche. (2016). On wi-fi tracking and the pitfalls of mac address randomization. In New Internet Object Challenges: Human- Machine Interaction and Human Factors, 2016.
17. G. Kalyani, S. Chaudhari. (2021). Enhanced Privacy Preservation in the Routing layer with Variable-length packet data for Attack Free IoT Sector. Journal of Engineering Science and Technology Review. 14. 95-99. 10.25103/jestr.141.10.
18. IEEE. (2008). Guidelines for Use of Extended Unique Identifier (EUI), Organizationally Unique Identifier (OUI), and Company ID (CID). IEEE Standards Association. IEEE. Retrieved 5 August 2018.
19. M. H. Kutner, C. Nachtsheim, J. Neter, W. Li. (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill.
20. P. J.Diggle, P. Heagerty, K. Liang, S. L. Zeger. (2002). *Analysis of longitudinal data* (1st ed.). Oxford University Press