Characterizing User Activity in WiFi Networks: University Campus and Urban Area Case Studies

Larissa Oliveira
Department of Computer Engineering
UC Santa Cruz
lmarinho@ucsc.edu

Katia Obraczka
Department of Computer Engineering
UC Santa Cruz
katia@soe.ucsc.edu

Abel Rodríguez
Applied Mathematics and Statistics
UC Santa Cruz
abel@soe.ucsc.edu

ABSTRACT

In this paper we investigate and characterize user activity in WiFi networks by analyzing and comparing the behavior of users that connect to two public WiFi networks, one of them deployed in a University campus and the other in a major urban area. We characterize WiFi network user activity based on two main features, namely: time users stay connected to Access Points and Access Point load. Overall, the main contributions of our work are as follows: (1) to the best of our knowledge, this is the first study comparing user activity in two different scenarions, i.e., a University campus WiFi network and an urban WiFi network; (2) our results validate previously observed characteristics of user behavior in WiFi networks, as well as unveil new behavior patterns, such as the fact that users on campus tend to stay connected to the network for longer periods of time when compared to users in an urban area; and (3) our work is the first study to formally test and validate the hypothesis that association times in WiFi networks follows a power law and to estimate the power-law's tail index.

Keywords

WiFi Networks; user Behavior; tail-distribution

1. INTRODUCTION

According to [1], mobile data traffic grew 74% worldwide in 2015 and is expected to increase almost eightfold by 2020. Additionally, by 2020, the total number of smartphones will be nearly 50% of the total number of devices and connections. This unprecedented growth demands a deeper understanding of how users move, connect, as well as generate and consume traffic. Understanding user activity in wireless access networks is essential to be able to scale and accommodate future connectivity- and traffic demand. Furthermore, better understanding mobile user behavior and activity can also greatly contribute to improve urban planning, including transit, transportation, and housing infrastructure, emergency response, as well as other services (e.g., food, shopping, entertainment, etc).

In this paper, our goal is to characterize user mobility- and usage patterns in WiFi networks based on data from real networks. In particular, we study traces from two different scenarios, namely a University campus and the downtown area of a major city. The former trace was obtained at the Dartmouth College campus during the 2005-06 academic year with a total of over 24,000 users and 3,300 access points. The latter trace was collected for 6 years (2004-2010) at the city of Montreal consisting of over 340 access points connecting 45,000 users. The main contributions of this study are as follows:

- To our knowledge, this is the first systematic study that compares user activity in a University campus and a major city.
- We have both validated results from previous related work and unveiled new user behavior. For example, we confirm that user connection times follow a power-law as has been observed previously; an example of a new result our study reveals is the fact that users tend to connect for longer periods of time in a university campus when compared to an urban network.
- Ours is the first study to apply statistical theory to test whether association times follow a power law and to estimate the tail index of their distribution.

2. RELATED WORK

In order to put our work in perspective, we propose a new taxonomy to classify existing studies on user mobility. We start by classifying related work into two main categories, namely studies based on descriptive- versus predictive analysis. In Table 1 we list relevant work on user activity in infrastructure-based WiFi networks and the features they investigate according to our taxonomy. Predictive studies on user mobility characterization in WiFi networks explore different features including: (1) **Association Time**: time interval during which the user stayed connected to an Access Point (AP) before moving to another AP or leaving the network; (2) **User Traffic**: amount of traffic (e.g., in number of bytes) users download and upload during the association time; (3) **Direction of Movement**: direction a user takes when moving between APs; (4) **AP Load**: access point usage, such as number of users connected to a given AP, total traffic handled by an AP, etc; (5) **Hotspots**: group users into communities according to social- and geographic features, such as popular hotspots. On the other hand, pre-

MSWiM '16, November 13-17, 2016, Malta, Malta

DOI: http://dx.doi.org/10.1145/2988287.2989172

Table 1: Relevant related work divided according to proposed taxonomy

Descriptive	Association Time	Traffic	Direction of Movement	AP load	Hotspots
Google WiFi [2]	√	√		√	
Dartmouth [8]	√		√		√
SF and NY [6]	√	√			
Predictive	Fractal Waypoint	Markov	Queuing Models	Clustering	SVM
MixedQueuing [4]			√		
MHMM [5]		√			
ToGo [12]					√
SAGA [13]				√	

dictive studies of user activity in WiFi networks can be classified according to the modeling approach adopted. Notable modeling approaches used in previous work include: (1) **Markov Models**; (2) **Fractal Waypoints**; (3) **Queuing Models**;(4) **Clustering**, and (5) **Support Vector Machines (SVMs)**.

3. WIFI NETWORK TRACES

Our study explores user activity in two different WiFi network scenarios, namely a major urban center and a University campus. To this end, we study two WiFi traces which have been made publicly available from CRAWDAD. The first trace was collected over six years (2004-2010) from the WiFi network deployed in the city of Montreal, Quebec, Canada [11]. The second trace was collected at the Dartmouth College campus during the 2005-2006 academic Year [9]. In the remainder of this section, we describe the two traces in detail

3.1 Montreal Trace

Île Sans Fil (French for "Wireless Island", also known as ISF), is a non-profit organization that operates a network of WiFi hotspots in Montréal, Québec, Canada [11]. ISF provides free wireless network access to over 45,000 users and has a total of 346 unique access points (APs) deployed across Montreal's downtown area. All of the APs are located in publically accessible spaces, including cafes, restaurants, and bars, but also in libraries, funeral homes, doctors' offices, and Business Improvement Agencies (BIAs). They cover city parks and sections of popular commercial streets. While at of the end of the 6-year period, ISF had 346 APs deployed. The number of active APs per month varies, however it steadily increases during the trace collection period. At the last month of the trace, there were 185 active APs. Information available in the trace includes: users session (i.e., between login and logout) data such as account (user) id, MAC address, login- and logout times, AP id, and amount of data transferred (incoming and outgoing) for a period of six years from 2004 to 2010. The data has been sanitized in order to anonymize user-specifc information such as account id, connection id, user MAC address, and AP id.

3.2 Dartmouth Trace

The Dartmouth trace [9] was collected at the Dartmouth College campus during the 2005-2006 academic year. The campus occupies 200 acres with over 190 buildings, of which 188 had wireless coverage at the time the trace was collected. Over 3000 APs were deployed, providing WiFi coverage to the campus. Due to the compact nature of the campus, the APs installed in buildings are able to provide network coverage to most of the campus' outdoor areas. All APs share the same SSID, allowing wireless clients to roam seamlessly between APs. The 188 buildings with wireless coverage span 115 subnets, so clients roaming between buildings may be forced to obtain new IP addresses through DHCP (lease times were 6 or 12 hours at different points in the trace).

4. USER ACTIVITY PATTERNS

We analyze user activity for both the Montreal- and Dartmouth traces according to the Association Time and AP Load metrics as defined in Section 2.

4.1 AP Load

Montreal

Figure 1(left) shows the number of sessions (connections) for every user while Figure 1(right) shows the number of sessions for every unique AP for the Montreal trace. According to Figure 1(left), users have on average 14 sessions and the user that connects the most has over 13,000 sessions. The median number of sessions is 3 and the third quantile is 8. The tail index for the session distribution is 3.71, which indicates that, even though most users have very few sessions, approximately 23% of users have more than 8 sessions and 15% have more the 14 sessions. As shown in Figure 1(right), the average- and median number of sessions per AP is 692 and 218, respectively, while the maximum number of sessions per AP is over 10,000. Note that the least loaded APs have 57 sessions or less, which represents the first quantile.

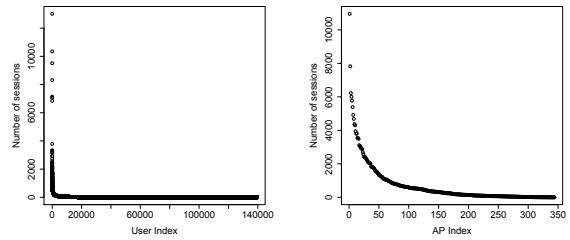

Figure 1: Montreal trace - Left: Number of sessions per user. Right: Number of sessions per AP.

Figure 2(left) shows a scatterplot of the logarithm of the total number of sessions per AP versus the logarithm of the number of users in each AP. Note that the number of users per AP represents a lower bound for the number of sessions, so all points in the graphs must lie above the 45 degree line. The graph indicates that the relationship is roughly linear (in the log-log plot), with limited- and roughly constant variability. On the other hand, in Figure 2(right), we show a scatterplot of the total number of sessions per user versus the number of APs to which a user connects. Interestingly, the relationship between these two variables is generally not linear in this case. Indeed, we see evidence of at least three clusters. One of them corresponds to what could be considered "static users", i.e., users who frequently connect to a relatively small number of APs (data points located along the vertical axis). Another cluster consists of the points located along the horizontal axis which correspond to users that connect very sporadically, but they do so to a large number of different APs. Note that the latter type of users appears to be relatively more frequent that the former. Finally, we observe the cluster represented by the points in the

bottom left corner of the graph, indicating that most users connect to under 40 APs over less that 4,000 sessions.

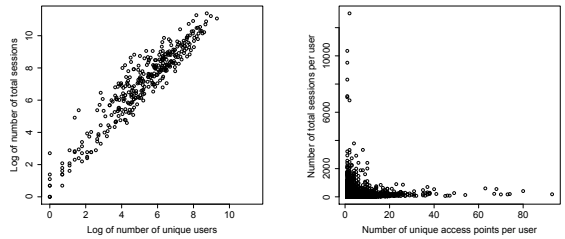

Figure 2: Montreal trace: Left: Logarithm of the total number of sessions per AP versus the logarithm of the number of users in each AP. Right: Total number of sessions per user versus number of APs a user connects to.

Dartmouth

We also analyzed AP Load for the Dartmouth trace. Again, first we study the usage patterns for users and APs separately which is plotted in Figure 3. In Figure 3(left), the number of sessions per user is shown and we can see that, on average, users have approximately 160 sessions, whereas very few users have more than 10,000 sessions. In Figure 3(right), we show the number of sessions per AP. The average load per AP for Dartmouth is 158 sessions and the highest load is 4000 sessions. When we compare Montreal and Dartmouth in terms of the number of sessions per AP, that is AP Load, we observe that access points in Montreal have on average, and overall, a higher load then Dartmouth access points. This result is to be expected since the sample population in Montreal is at least ten times higher and the number of access points in Montreal is approximately ten times lower. However, we note that users have on average a higher number of sessions in Dartmouth, a university campus, than in Montreal. The same observation also holds when comparing the number of user sessions at the Dartmouth campus to the results obtained in [2] for Google's Mountain View WiFi network.

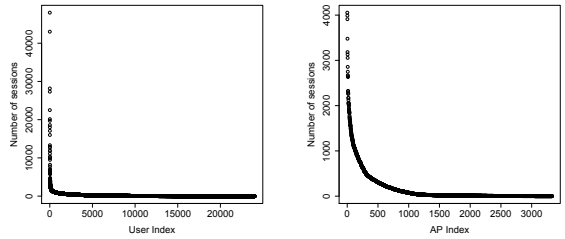

Figure 3: Dartmouth trace: Left: Number of sessions per user. Right: Number of sessions per AP.

Next, we study AP usage patterns for Dartmouth. Figure 4(left) plots the log of the total number of sessions per AP versus the log of the number of users in each AP. Similarly to the Montreal trace, we observe that the relationship is also almost linear (in the log-log plot) but exhibits much higher variability. In Figure 4(right), we plot the total number of sessions per user versus the number of APs to which users connect. Like in the Montreal case, we observe a cluster of users in the bottom left corner of the graph who connect to less than 400 APs over less that 5,000 sessions. However, unlike Montreal, there are no other easily identifiable clusters.

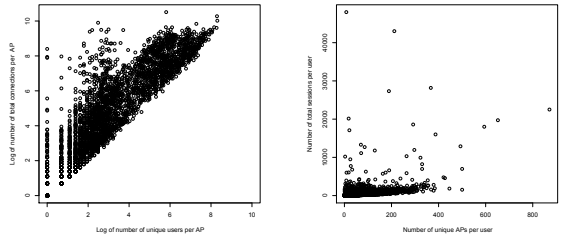

Figure 4: Dartmouth trace: Left: Logarithm of the total number of sessions per AP versus the logarithm of the number of users in each AP. Right: Total number of sessions per user versus number of APs a user connects to.

4.2 Association Time

Montreal

Figure 5 shows the histogram of the logarithm of association times for the whole Montreal trace. We observe that a little less than 30% of users spend approximately 50 min connected, while a little over 20% remain connected for approximately 20 min and another 20% remain connected for up to 2h. We also observe that a small percentage of users connect over very long or very short times. To better understand the behavior of these users, we analyze in Section 5 the tail of the association time distribution and test the hypothesis that it follows a power-law. This power-law behavior of the association times has has been observed in related work [10], [8], [4]. However, to our knowledge, our work is the first to conduct a formal analysis to test the power-law hypothesis for association times in WiFi networks, where we estimate power-law parameters such as the tail index.

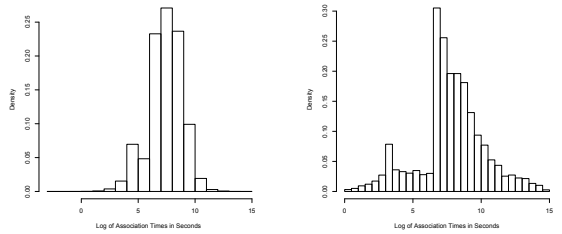

Figure 5: Left - Montreal trace: Histogram of the log of association times (in seconds). Right - Dartmouth trace: Histogram of the log of the association times (in seconds).

In our study, we also try to uncover temporal patterns in the association times. Figure 6(left) shows the distribution of the log of the median association time per day of the week for the entire trace. We divided each day of the week in 8 periods, namely: from 0am to 3am, from 3am to 6am, etc. in order to explore finer-grained temporal patterns for the time users remain connected to an AP.

Using the average median association time of 29 minutes (red horizontal line in Figure 6(left)) as baseline, we observe different behavior depending not only on the day of the week

but also on time of the day. As expected, on Monday, we notice higher association times during business hours (i.e., 9am-6pm), whereas during early morning and late night, association times are lower. The median and average association times during business hours are 30- and 75 minutes, respectively. Tuesday and Wednesday show similar behavior, i.e., higher association times during business hours, however the period extends from 9am-7pm. We notice a somewhat different behavior on Thursdays and Fridays: while the median association time is still around 29min, the average association time is 60min between 9am and 3pm. Interestingly, on weekends, we observe longer average association times, which happen during the period between 6am and 3pm. We also notice that the median and average association times for the weekend in on average higher than for weekdays, namely 36 and 80 minutes, respectively. In order to explore coarser grain patterns, we also investigated the distribution of the median association times on a monthly basis throughout the year. Figure 6(center) shows a boxplot of the log of the median association time per month for the entire trace period. Here we can see that the average median association time is approximately 30min (red horizontal line in the plot). However, we observe that in the first six months after the network was deployed, i.e., from August 2004 to February 2005, the average median association time is approximately 18min. We believe this lower association time (relative to the overall average median association time of 30min) is due to the fact that the Ile San Fils network had just been deployed and was still growing.

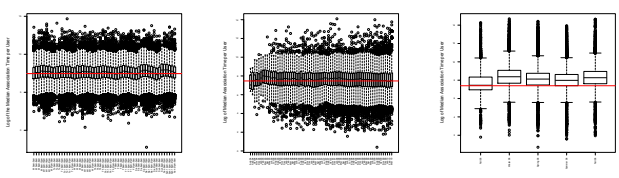

Figure 6: Montreal trace: Left: Boxplot of the log of the median association time per day of the week, and (center) per month. Dartmouth trace: Boxplot of the log of the median association time per quarter (in seconds) for the 2005-06 academic year.

Dartmouth

We also analyze the distribution of the association time for the Dartmouth trace. Dartmouth campus' WiFi network has been studied in previous work, for example [3] [13] [5] [8] [4]. To our knowledge, we are the first ones to study the most recent version of the trace available from [9]. We find that, differently from traces collected in major cities (e.g., Montreal, San Francisco and New York [6]), Dartmouth campus users tend to remain connected for longer periods of time. More specifically, on average, users spend more than 1h connected to campus APs. Figure 5 (right) shows the histogram of the log of the association time for the entire trace. Similarly to Montreal, we observe a bimodal distribution. In the first cluster, users remain connected for up to 11 min, which encompass approximately 20% of users. The second cluster encompass approximately 62% of users that remain connected for more than 20 min. Within the first cluster, we observe that 10% of users have an association time of 6 min. On the other hand, approximately 15% of users have an association time of over 2h. Similarly to Montreal, we can identify a small percentage of users (7%) that either remain connected for very long or very short periods of time. We hypothesize that some of these longer connections times, e.g., in the order of days, might be attributed to desktop- or even laptop computers in dorms or offices which remain connected for long periods of time. To better understand the behavior of these extreme users, we analyze in Section 5 the tail of the association time distribution and test that it follows a power-law. Our results align with results from studies of the Dartmouth campus network for previous academic years [8] [2] [5].

We also investigated the distribution of the median association time during every quarter of the academic year. Here, we are interested in capturing temporal usage patterns over a shorter time scale. Figure 6 (right) shows a boxplot of the log of the median association time per quarter for the 2005-2006 academic year. Using the average median association time of 38 min (red horizontal line in Figure 6) as baseline, we can observe some variation among quarters. For the Fall quarter of 2005, we observer lower median association times, on average 28 min, whereas during Winter quarter of 2006 we observe a higher median association time 72 min. Even though the average median association time from Spring 2006 to Summer 2006 decreased (3h to 48 min), we can see they still have a very similar distribution. This is an interesting result, since we could expect that most students wouldn't remain on campus during summer vacations. Finally, we notice an increase in the median association time during the beginning of the next academic year, Fall 2006. We also observe that the average median association time for Fall 2006 and Winter 2006 have similar distributions.

5. HEAVY-TAILED ASSOCIATION TIMES

In this section, we investigate the hypothesis that the heavy-tail distribution of user association times follows a power-law as suggested by our results reported in Section 4.2 and by relevant related work [10], [8], [4]. Previous work suggests that many of the reported power-law distributions in the literature have not been rigorously validated. This is especially true when the power-law assumption is made based on log-log plots. Instead, in this paper we aim to formally test the hypothesis that association times follow a power law. We also estimate the distribution's tail index, which is key to estimate future association times.

The probability density function of the power-law distribution (also known as the Pareto distribution) is defined as:

$$f(x) = \frac{\alpha - 1}{x_{min}} (\frac{x_{min}}{x})^{\alpha} \quad (1)$$

for $\alpha > 1$ and $x_{min} > 0$. We do not attempt to use the power-law distribution to model the full distribution of the association times. Instead, it is used to model its right tail, i.e., the distribution of large association times. The value x_{min} represents the minimum value for which the distribution follows a power-law. Note that as the value of x_{min} increases, the amount of information available about the behavior of the tail decreases. The most common approach to estimate x_{min}, also known as the threshold, is by inspecting the log-log plot of the data. However, this method is very subjective and error prone. Instead, we use a Kolmogorov-Smirnov (K-S) test, which looks at the maximum distance between the data and the cumulative density function of

the power-law distribution whose parameters are estimated using maximum likelihood. This test is implemented using the `R` `poweRlaw` [7] package. The uncertainty associated with the parameters are estimated via a bootstrap procedure also available in the `poweRlaw` package. Figure 7(left) shows the power-law fitted to the heavy tail of the association times for Montreal while Figure 7(right) shows the same graph for Dartmouth. The tail parameters estimated for Montreal were $\alpha = 3.68$ $sd = 0.22$ and for Dartmouth $\alpha = 2.79$ $sd = 0.004$, where sd is the standard deviation. The $p - value$ associated with the K-S test was 0.06 for Montreal and 0.10 for Dartmouth.

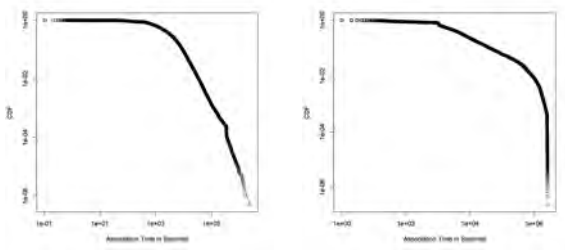

Figure 7: Montreal trace: (left) Power law fit for the tail of association times for Montreal and (right) for Dartmouth

Both K-S goodness-of-fit ($p - values$) for Montreal's and Dartmouth's association time distributions indicate that they can indeed be modeled by a power-law. We further observe that the tail of the distribution of Dartmouth's association times is heavier than Montreal's. This can be observed by inspecting Figure 7(left) and (right) as well as by their tail parameter α. Dartmouth's α is lower that Montreal's indicating a heavier tail since the lower α is, the slower the distribution decays and therefore, the heavier the distribution's tail. Consequently, large association times tend to be relatively more common for Dartmouth than Montreal. Indeed, since for the power-law distribution only moments of order $\lfloor \alpha \rfloor - 1$ exist, the tail of Dartmouth is heavy enough that the conditional variance is infinite.

6. SUMMARY AND DISCUSSION

In this section, we summarize the main findings of our study:

- `AP Load`: We find that for the Montreal trace, the relationship between number of sessions and number of users is roughly linear, whereas the relationship between number of sessions and number of APs is not. We also observed similar trends for Dartmouth. However, as expected, the AP load for Montreal is on average and overall higher than Dartmouth's. For the Montreal network, we can identify 3 different user clusters based on the number of user sessions: there are the so-called static users, i.e., users that mostly connect to a few APs (i.e., less than 5), users that are quite mobile, i.e., connect to a large number of APs (i.e., more than 50), and finally most users connect to less than 40 APs. However, for Dartmouth, we can only identify one clear cluster, which are users that connect to less than 400 APs.
- `Association Time' Heavy-Tailedness`: We test the hyphotesis that the association times for both Montreal and Dartmouth follow a power-law and confirm that both distributions can indeed be modeled as power-laws. Dartmouth has a heavier-tail when compared to Montreal, therefore large association times tend to be more common for Dartmouth.

7. ACKNOWLEDGMENTS

Financial support was granted by the CAPES Foundation Grant ID:13428-13-6 and by NSF under project CNS 1321151.

8. REFERENCES

[1] Global mobile data traffic forecast update. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[2] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren. Usage patterns in an urban wifi network. *IEEE/ACM Trans. Netw.*, 18(5):1359–1372, Oct. 2010.

[3] W. Bao and B. Liang. Insensitivity of user distribution in multicell networks under general mobility and session patterns. *IEEE Transactions on Wireless Communications*, 12(12):6244–6254, December 2013.

[4] Y.-C. Chen, J. Kurose, and D. Towsley. A mixed queueing network model of mobility in a campus wireless network. In *INFOCOM, 2012 Proceedings IEEE*, pages 2656–2660, March 2012.

[5] M. Gani, H. Sarwar, and C. M. Rahman. Prediction of state of wireless network using markov and hidden markov model. *Journal of Networks*, 4(10):976–984, 2009.

[6] A. Ghosh, R. Jana, V. Ramaswami, J. Rowland, and N. K. Shankaranarayanan. Modeling and characterization of large-scale wi-fi traffic in public hot-spots. In *INFOCOM, 2011 Proceedings IEEE*, pages 2921–2929, April 2011.

[7] C. S. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015.

[8] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *INFOCOM*, volume 6, pages 1–13, 2006.

[9] D. Kotz, T. Henderson, I. Abyzov, and J. Yeo. CRAWDAD dataset dartmouth(v.2009-09-09). http://crawdad.org/dartmouth/campus/20090909, Sept. 2009.

[10] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863, April 2009.

[11] M. Lenczner, B. Grégoire, and F. Proulx. CRAWDAD dataset ilesansfil(v.2007-08-27). http://crawdad.org/ilesansfil/wifidog/20070827, Aug. 2007.

[12] J. Manweiler, N. Santhapuri, R. Choudhury, and S. Nelakuditi. Predicting length of stay at wifi hotspots. In *INFOCOM, 2013 Proceedings IEEE*, pages 3102–3110, April 2013.

[13] B. Nunes, K. Obraczka, and A. Rodrigues. Saga: Socially- and geography-aware mobility modeling framework. In *Proceedings of the 15th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '12, pages 367–376, New York, NY, USA, 2012. ACM.