Characterizing Home Device Usage From Wireless Traffic Time Series

Katsiaryna Mirylenka, Vassilis Christophides, Themis Palpanas, Ioannis Pefkianakis, Martin May

Characterizing Home Device Usage From Wireless Traffic Time Series

Katsiaryna Mirylenka*
IBM Research - Zurich
Rüschlikon, Switzerland
kmi@zurich.ibm.com

Vassilis Christophides†
INRIA Paris -
Rocquencourt, France
vassilis.christophides@inria.fr

Themis Palpanas
Paris Descartes University,
France
themis@mi.parisdescartes.fr

Ioannis Pefkianakis†
Hewlett Packard Labs
Palo Alto, CA, USA
ioannis.pefkianakis@hpe.com

Martin May
Technicolor Research &
Innovation Center
Rennes, France
martin.may@technicolor.com

ABSTRACT

The analysis of temporal behavioral patterns of home network users can reveal important information to Internet Service Providers (ISPs) and help them to optimize their networks and offer new services (e.g., remote software upgrades, troubleshooting, energy savings). This study uses time series analysis of continuous traffic data from wireless home networks, to extract traffic patterns recurring within, or across homes, and assess the impact of different device types (fixed or portable) on home traffic. Traditional techniques for time series analysis are not suited in this respect, due to the limited stationary and evolving distribution properties of wireless home traffic data. We propose a novel framework that relies on a *correlation-based similarity* measure of time series, as well as a notion of *strong stationarity* to define *motifs* and *dominant devices*. Using this framework, we analyze the wireless traffic collected from 196 home gateways over two months. The proposed approach goes beyond existing application-specific analysis techniques, such as analysis of wireless traffic, which mainly rely on data aggregated across hundreds, or thousands of users. Our framework, enables the extraction of recurring patterns from traffic time series of individual homes, leading to a much more fine-grained analysis of the behavior patterns of the users. We also determine the best time aggregation policy w.r.t. to the number and statistical importance of the extracted motifs, as well as the device types dominating these motifs and the overall gateway traffic. Our results show that ISPs can exceed the simple observation of the aggregated gateway traffic and better understand their networks.

*Work done while visiting Technicolor R&I Center.
†Work done while at the Technicolor R&I Center.

Keywords

Wireless Traffic, Motif Extraction, Similarity Measurement

1. INTRODUCTION

The increasing diversity of home devices and network technologies have added layers of complexity to the connected home environment. Residential gateways (RGWs), Smart TVs, smartphones and tablets are just a few of the devices in today's connected home. Furthermore, several "over-the-top" services, such as video streaming (e.g., Netflix) and conferencing (e.g., Skype), are being delivered to a variety of devices and users using wireless home networks. Managing the connected home environment, and indeed optimizing the Quality of Experience (QoE) of residential users, emerges as a critical differentiator for Internet and Communication Service providers (ISPs and CSPs, respectively) and heavily relies on the analysis of home networks.

Although RGWs technology has been considerably improved to deliver IP-based, whole-home services as well as advanced WiFi capabilities as "Community WiFi" hotspots, ISPs still have little information about what lies beyond the subscribed RGW and how home network resources (e.g. bandwidth) are actually consumed by the underlying device ecosystem. For this reason, RGWs have been extended with continuous home network measurement capabilities under normal service operation and provide fine-grained connectivity, usage and performance data of home networks and devices[1]. Mining recurring patterns from these home traffic time series can enable a *data-driven* paradigm in network management [13] in order, for instance, to reduce the cost for *serving* and *diagnosing* remotely home networks [3, 22, 11], or even to improve residential QoE via home-specific *bandwidth sharing* [1] and *prioritization* [21] policies.

More precisely, home networks troubleshooting is almost always reactively initiated by residential users, requires a human intervention and as a consequence is a time consuming (e.g., 38 min. average time of a technical support call [26]) and not always feasible task (e.g., user's problem is solved in only 14% of technical support calls, partially solved in 24%

[1]Home network monitoring is of course subject of different low restrictions in different countries, and thus we are interested in non-intrusive passive probing techniques that guarantee anonymity

and not solved at all in 62% [26]). One of the reasons of low efficiency of remote technical support is that technicians cannot completely understand the problem and home network settings, mainly due to limited and sometimes inaccurate information that residential users provide. Extracting previously unknown recurring patterns (aka motifs [19, 7]) from residential traffic time series will bring strong evidence of regular user activity in homes that can be contrasted to the trouble description reported by users. In particular, traffic patterns enriched with detailed home device information is a valuable input for root cause diagnosis. Moreover, in their majority, ISPs typically broadcast firmware and software updates to all gateways at nights (some operators even on a daily basis). This may cause service outages, given that some gateways may exhibit an active network usage during night time. A fine-grained temporal characterization of residential bandwidth consumption will enable ISPs to differentiate RGWs firmware update policies according to the least cumbersome time window per home, thus, improving the overall QoE of residential users.

In addition, home network resources (bandwidth) are shared not only among residents using an increasing number of online applications (e.g., social networking, gaming, uploading/downloading, etc.) and real time services (TV on demand, teleconferencing), but also with guests, neighbors, or even the occasional passersby. Existing methods for bandwidth sharing and traffic prioritization are static and coarse. ISPs usually allocate a fixed percentage of home bandwidth to non-residential users, while traffic prioritization in commodity gateways is at best based on the network port on which traffic is sent or received. We believe that behavioral patterns extracted by gateway traffic time series can be used to support dynamic policies for sharing home bandwidth that consider the online habits of residential users. For example, in-home traffic congestion can be avoided by ordering the traffic patterns of different devices observed especially during afternoon and weekends. These patterns reveal the bandwidth consumption behavior of different groups of residential users (adults and children employ different devices during the same time-slots) while the comparison of traffic domination help us to distinguish between residents and guests (pattern-specific vs global traffic dominant devices).

To the best of our knowledge, this paper presents the first thorough analysis of traffic dynamics of heterogeneous wireless (WiFi) devices connected to 196 real RGWs, which are subscribers of a major European ISP. We focus on a time-oriented analysis of *continuous traffic data* to extract *previously unknown patterns* recurring of internet consumption that happen within, or across homes. We also assess the impact of different types of devices, such as laptops, desktops (classified as "fixed" devices), and tablets, smartphones (classified as "portables"), on these patterns. Unsupervised learning techniques are used for patterns discovery as the ground truth data regarding home activities are not available. Rather than partitioning homes or devices into distinct behavioral clusters, we are looking to extract *informative motifs of bandwidth consumption* within or across homes. Different from a previous analysis of the same dataset [23], which focused on coarsely aggregated gateway traffic, in this paper we conduct various types of time series analysis, including a per-device analysis, motif extraction, and search of dominant devices.

In our study, we develop a framework for analyzing the distribution properties of traffic data, the stationarity and predictability of usage patterns within and across homes, the similarity of device specific traffic to the overall home traffic, and others. We demonstrate that traditional techniques of time series analysis [20] are not suitable in our setting due to restricted stationarity of traffic time series. This is caused also by the fact that low-valued non-active traffic occupies the most of the probability mass of the traffic distribution, while the values of active traffic are detected as outliers. In contrast to Euclidean Distance and Dynamic Time Warping (DTW), the proposed approach better fits the requirements of our applications: (a) it correctly identifies similar trends, both when absolute values are important and when they are not; (b) it restricts the matches to time-aligned sequences; and (c) it provides a similarity measure, whose values (between -1 and 1) we know how to interpret based on the theory of statistics.

The main contributions of this paper are:

1) We propose a novel analysis framework for wireless home traffic data, namely: (a) a *correlation-based similarity* measure, which exploits the evolution characteristics, rather than the absolute traffic values, and is invariant to scaling; (b) a notion of *strong stationarity* that in addition to the similarity of data distributions imposes a correlation similarity across non-overlapping time windows; and (d) a definition of *dominant devices* based on the correlation similarity, that enables an intuitive and statistically grounded interpretation of the results.

2) We evaluate the effectiveness of the proposed framework using real data of wireless traffic observations and report the main findings: (a) there are many repetitive patterns within and across RGWs which describe the intrinsic user behavior of users and valuable to ISPs; (a) as networking time series are not stationary certain aggregation should be performed in order to find statistically significant patterns. The best time windows to aggregate home traffic data is found to be 8 hours for weekly patterns and 3 hours for daily patterns; (b) frequent weekly patterns correspond to heavy bandwidth usage both during weekdays and weekends, and frequent daily patterns correspond to (mostly) evening usage, (c) weekend usage tends to rely on portable devices, weekday usage relies more on fixed devices, while discontinuous usage within a day (mostly active in the evening or the morning) is still due to portable devices; and (d) almost every RGW involves a device that dominates its overall traffic, thus the behavior of this device should be mainly considered by ISPs while planning the updates.

The paper is organized as follows. Sections 2 and 3 describe the related work and our dataset. Section 4 shows the preliminary analysis of the wireless traffic. Section 5 describes the proposed methodology. Section 6 correlates the wireless traffic with home devices. Section 7 presents in detail the time aggregation and motifs analysis. Section 8 concludes the paper.

2. RELATED WORK

We consider several directions of related work that cover the specifics of analysis of wireless devices in home.

Wireless traffic analysis. The analysis of wireless traffic dynamics has been widely used to provide energy savings [12, 24], to build collaborative wireless networks [28], and to accommodate traffic offloading [16]. For example, recent proposals seek to power off idle Access Points (APs) [12]

or cellular base stations [24] to save energy. SEAR [12] forms clusters of WiFi APs and powers on/off APs of the same cluster based on the traffic demand that the cluster needs to serve. In a similar fashion, the system proposed in [24] powers off under-utilized cellular base stations when their traffic load is light, and power them on when the traffic load becomes heavy. Collaboration among APs has been explored in [28] to offer energy savings and load balancing in WiFi APs. The above designs though, are based on two key assumptions, which may not hold in RGWs. First, they are based on wireless traffic stationarity to predict idle times (e.g., the traffic volume is stable over short-term (2 hours), which can remain stable over several consecutive days [24]). Second, a set of devices (e.g., APs or base stations) show very similar temporal traffic characteristics. Our analysis shows that these assumptions do not hold in our case.

Human behavior. Multiple works demonstrate that the behavior of humans can be described through recurrent activity patterns. This is shown in the work [4] for social network data, in the work [32] for user behavior in microblog and in the work [10] for moving trajectories using mobile phone data. According to the results of work [10] the patterns of human mobility behavior are very regular. Home traffic data which we focus on is also determined by human behavior, but unlike previous studies it is not defined by a single individual but by a group of individuals who share one home. This leads to less repetability an large variety of possible activity patterns, making the analysis of RGWs more challenging.

Work [14] analyzes human correspondence behavior via mobile phone data. As in the other human activity studies [15, 9, 5], the data show high inhomogeneous in activities, meaning that periods of active events are much shorter than inactivity silence. This leads to the bursty time series with long tails in the probability density function. We observe this property for our data as well while pattern extraction is needed to take place on more regular data. Study [14] checks whether the inhomogeneous of correspondence behavior is due to daily or weekly periodical patterns or it is due to the nature of the behavior of human task execution. According to the results of this study, even after de-seasoning when daily and weekly periodical patterns are excluded, the data remains inhomogeneous, which means that the character of human activities is one of the main sources of bursty time series. Unlike mobile phone data, where long silence periods where observed in night time and during weekends, our networking data for certain homes have peak activities exactly in this "silence" periods. This means that in our case de-seasoning is not applicable as there are no strong daily and weekly silence period for all homes into consideration. Thus, we safely assume that the heterogeneity of our data is caused by human activities even to larger extend than for phone call data. In our work we concentrate on reducing inhomogeneous data characteristics by other means, such as by excluding background traffic from consideration and by specially devised distance measure for traffic time series. Then, we extract recurrent activity patterns of usage behavior using the technique of motifs.

Motifs. Mining of motifs or sequential patterns is an important task in time series analysis [19], [7]. As far as we know this kind of analysis has not been applied to the internet traffic data before as most of the studies use aggregated traffic values instead of time series. For this task, we considered several state-of-the-art tools for motif discovery, such as GrammarViz [17] and VizTree [18]. However, these tools are not suitable for our analysis for the following reasons. (a) These tools (and many other techniques available online) use Symbolic Aggregate approXimation (SAX) to represent time series, assuming that the distribution of time series values is normal [19]. This is not true for our traffic time series though, as their values follow the Zipf's law (we note that, contrary to the claims in [19], z-normalization does not lead to normal distribution if the initial distribution of the time series is not normal). At the same time we do not have ground truth data about the motifs in order to tune the alphabet size of SAX, which assigns more symbols near the value of zero, while in our case this region should have been coded with only one symbol. (b) GrammarViz seeks to discover motifs of different lengths, and exploits grammar distance for this. On the other hand, our data has clear time semantics, and we would like to discover motifs for week- and day shifting (i.e., non-overlapping) windows of fixed length, that is not enabled with GrammarViz.

3. DATASET DESCRIPTION

We analyze wireless traffic data collected from 196 residential gateways under normal service operation, involving subscribers of a large European ISP that are distributed over a large geographic area spanning 10 cities. The residential subscribers participate on a voluntary basis to our large scale data collection campaign. For privacy reasons, we do not collect data regarding running applications of home devices, demographics and activities users are engaged in.

The gateway platforms of our deployment have the following specifications: (i) ADSL2+ modem or fiber WAN access link, (ii) 4 ethernet ports, (iii) a WiFi access point enabled by a Broadcom 802.11b/g/n 2x2 radio. The 802.11 interface operates at the 2.4GHz band and supports PHY rates up to 300 Mbps. The most (67%) of our deployment's gateways are fiber (92% of the fiber plans provide 100/10 Mbps downstream/upstream speed, and for the rest it is 30/3 Mbps) and the rest are ADSL (with 24/1 Mbps downstream/upstream speeds). Each gateway logs the traffic counters at all network interfaces on the IP layer, and reports in bytes the cumulative outgoing (transmitted) and incoming (received) traffic of each connected device to the gateway. The focus of this work is the WiFi traffic. The gateway further reports the aggregated gateway traffic, which is the sum of the corresponding outgoing and incoming traffic of all its devices. These data measurements are automatically reported every minute by each gateway to a central server. Note that the wireless traffic reported by a gateway depends on the running applications' data rates and is bounded by the wireless effective throughput or the access link throughput (for traffic exiting/entering the home). A recent study though [23] has shown that wireless (and wired) network throughput is rarely the bottleneck.

Our dataset includes more than 20 million measurement reports collected over a 2-month period (starting from March 17, 2014). We were able to identify a total of 2147 distinct wireless devices (a device is defined by its MAC address). Using a heuristic-based algorithm [25], we were able to infer the type of a wireless device. The heuristic algorithm leverages the device MAC address (revealing manufacturer name) and the device names typically assigned by the user, which are reported by the gateway. For example, "Nintendo

Co., Ltd." is known to produce game consoles, "EPSON" – peripheral devices, while "Katy's-iPhone", indicates a smartphone manufactured by Apple. We have validated the effectiveness of the algorithm using ground truth data collected from surveys at 49 homes of our deployment. All 'light' devices such as smartphones, tablets and others are labeled as "portable", while laptops and desktops fall under the "fixed" category. There is also a category of "network equipment" that includes devices such as WiFi extenders, and additionally there is a small amount of "game consoles".

4. STANDARD DATA ANALYSIS

In this section, we study the main data characteristics (distribution and stationarity) of the traffic time series observed by the RGWs in our deployment, and discuss the challenges arising in this task when using traditional analytical techniques.

4.1 Traffic Data Distribution

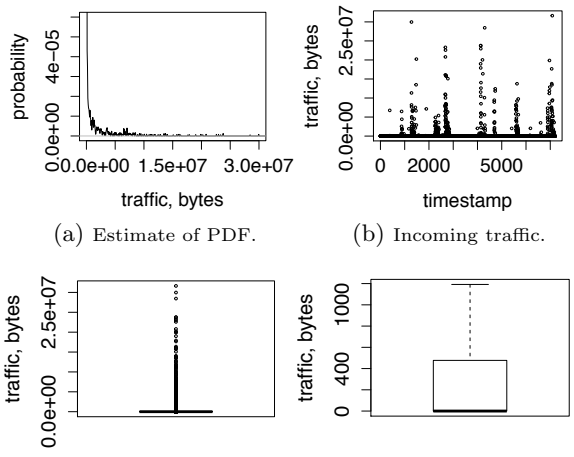
We first conduct a preliminary analysis using the wireless traffic data of the 10 most representative gateways with the highest number of observations for a single week period. Our analysis aims to answer the questions: (a) What are the main properties of the distribution of traffic values? Are probability density functions of traffic counters similar across gateways? (b) Which kind of traffic (outgoing, incoming, or overall) provides the most meaningful description of a gateway? To answer these questions, we exploit the following methods:

1) *boxplots* in order to visualize general probability distribution and outliers of time series, and

2) *estimation of probability density function (PDF)* using Kernel Density Estimators in order to assess and compare the probability distributions of time series.

The above methods lead to the following results: **(a)** The distribution of incoming and outgoing traffic of gateways follows Zipf's law (see Figure 1a), which means that the concentration of low traffic values is much larger than the amount of medium and high traffic values. The periods of really active traffic are very small, and thus detected as outliers in data distribution plots and boxplots. As an example, we show a typical time series in Figure 1b, an approximation of its PDF using Kernel Density Estimation zoomed around 0 of the y-axis in Figure 1a, and its boxplots with and without outliers in Figures 1c and 1d. This phenomena is also called inhomogeneity of data and as we mentioned in Section 2 it is typical for data describing human activities. **(b)** According to our results, there is a strong correlation between the incoming and outgoing traffic (mean = 0.92, median = 0.95, stddev = 0.08) of the gateways in our deployment. Since they are strongly correlated, we consider that the overall traffic of a gateway reflects the active behavior of a user without artifacts.

Summary. Since low traffic values account for most of the probability mass, the traffic values reported when devices are actually used are essentially detected as outliers in data distribution plots and boxplots considered by traditional time series analysis techniques. This motivates us to characterize the background traffic (Section 6.1) and remove it when looking for recurring patterns of active internet consumption. In this context, z-normalization alone does not help, as we want to consider also similarity of rankings of traffic values.



(a) Estimate of PDF. (b) Incoming traffic.

(c) Boxplot with outliers. (d) Boxplot without outliers.

Figure 1: Statistical analysis of a typical gateway.

4.2 Traffic Correlation and Stationarity

We now turn our attention to the following questions regarding the characteristics of the data: (a) Is the behavior of traffic counters similar among gateways? (b) Is there significant autocorrelation in the traffic of a gateway, or significant cross (lag-)correlation between gateways, or in general, what is the predictive power of the time series under examination? (c) Is the traffic of a gateway stationary in a short term period? (d) Is there a relationship between the number of connected devices and the traffic values? (e) Are home traffic time series sensitive to time aggregation? We use the following standard analysis techniques:

1) *correlation coefficients* – to measure similarity of RGWs;

2) *autocorrelation coefficients* – to evaluate how strong the connections between the values of a single time series are, and what their predictive power is;

3) *cross-correlation coefficients* – to measure how strong the connections between values of a pair of time series shifted in time are, and what their predictive power is;

4) *stationarity tests* (KPSS unit root test, Augmented Dickey-Fuller (ADF) test and others) – to check if time series are wide-sense stationary.

In all our experiments, when statistical tests are exploited, we use a significance level of $\alpha = 0.05$. For correlations tests, we use Pearson's, Spearman's, and Kendall's correlations, and interpret the strength of the correlation as follows: $[0.0; 0.1)$ →No Correlation, $[0.1; 0.3)$ →Low Correlation, $[0.3; 0.5)$ →Medium Correlation, $[0.5; 1]$ →Strong Correlation. This interpretation is widely accepted [2], [6], [30], though the borders may slightly vary depending on the application domain, for example, in medicine higher borders for strong correlation are usually required [29].

The above methods revealed the following results: **(a)** There are gateways, for which we can make predictions about their future behavior due to low, but statistically significant autocorrelations of their traffic time series. The example with the highest autocorrelation is shown in Figure 2(left). In this figure, the y-axis defines the value of Autocorrelation Function (ACF) that depends on the time lag between the time series values (x-axis). We note that no gateway exhibits a seasonal behavior. There is also some predictive power of one gateway given another, as some cross correlations with lags across gateways are significant. An example of a high cross-correlations is depicted in Figure 2(right). Even-though these observations suggest some predictive power, due to the significant amount of silence or

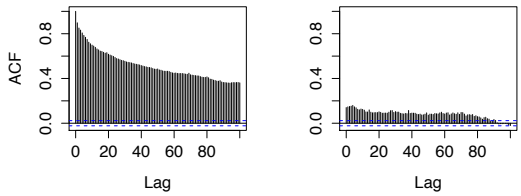


Figure 2: Autocorrelation (on the left) and cross-correlation (on the right) of gateways.

background traffic, ARIMA modeling for this time granularity cannot yield useful results, as it is not able to predict the rare bursts of the active traffic.

(b) Wireless traffic data is not stationary in the traditional sense, as all stationarity tests were rejected. This means that the data distribution characteristics change over time. For example, the covariance function of the time series is not constant in sliding window. We have also noticed that the Zipfian distribution of time series also evolves over time, meaning that the time series are also not stationary in the general sense.

(c) For all the gateways we checked, the correlation between the overall traffic time series and the number of connected devices time series was statistically significant, but low (mean = 0.37, median = 0.38, stddev = 0.21). This is an interesting result, indicating that traffic at a gateway depends more on the user behavior, rather than on the number of connected devices.

(d) For the time-aggregated time series with larger time binning, patterns become more visible as traffic peaks become more similar. At the same time, when excluding many points of low traffic, the essential information about the high traffic periods persists. Correlation and distribution heavily depend on time aggregation:

— The smaller the aggregation period is, the more different the data distribution within the week is. Almost all Kolmogorov-Smirnov tests were rejected for the smallest aggregations. For higher aggregation periods distributions become more similar.

— The smaller the aggregation period is, the lower the correlations between time series are. At the same time, for larger aggregation periods correlations either significantly grow, or disappear completely.

Summary. Our preliminary analysis of gateway traffic reveals that traffic time series are not stationary, neither in the general, nor in the wide sense: both the probability density function and the main time series characteristics (e.g., mean and covariance) change over time. Consequently, time series with current one minute binning are highly irregular, there are no stationary gateways, and similarity between different gateways of our deployment is very low. Hence, extracting meaningful patterns of bandwidth usage both across time and gateways requires to adapt new analytical methodology.

5. TRAFFIC ANALYSIS FRAMEWORK

We now define and describe the key concepts, on which we base our proposed analysis framework.

Similarity. The core issue when comparing traffic time series of residential gateways is to define a suitable similarity measure. As discussed earlier, absolute values of traffic volume are not helpful to understand seasonal usage of home devices within or across homes. Instead, we consider similarity in terms of correlation, which takes into account the monotonicity of traffic volume changes, rather than their absolute values, thus, providing invariance to scaling. We use three popular correlation coefficients as all kinds of dependencies between traffic time series described by these coefficients are important. *Linear* dependency is provided by Pearson's correlation, and *monotonicity* and *ranking-based* dependencies are provided by Spearman's and Kendall's coefficients. The correlation coefficients are not directly comparable but they have the same domain and semantic interpretation of strength, allowing us to use the highest possible value. More formally, the similarity measure is defined as follows:

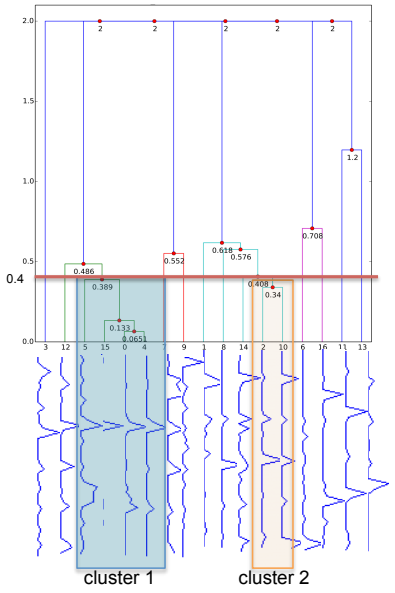


Figure 3: Hierarchical clustering of time series based on correlation similarity measure.

DEFINITION 1. **Correlation similarity measure** *$cor(X,Y)$ between two time series $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^n$ is a maximum correlation coefficient among statistically significant Person's $r(X,Y)$, Spearman's $\rho(X,Y)$ and Kendall's $\tau X,Y$ correlation coefficients:*

$$cor(X,Y) = max[r_{p_v<\alpha}(X,Y), \rho_{p_v<\alpha}(X,Y), \tau_{p_v<\alpha}(X,Y)].$$

If none of the three correlation coefficients are statistically significant $cor(X,Y) = 0$.

The significance of the corresponding correlation coefficient test is defined w.r.t. the zero-hypothesis $H0$ is that there is no correlation, or the coefficient is equal to zero. The significance level α is (as before) set to 0.05.

A correlation-based similarity measure also leads to meaningful threshold values for time-series similarity. For example, in the hierarchical clustering of traffic time series illustrated in Figure 3, the distance measure is set to $1-cor(\cdot,\cdot)$. As the correlation ≥ 0.6 is considered to be high, the corresponding threshold on a distance measure 0.4 is used to detect similarity clusters.

Spearman's and Kendall's correlation coefficients are insensitive to z-normalization of the data, while the Pearson correlation coefficient is normalization dependent. As we will see in Section 6.2, our correlation-based similarity measure allows to better grasp the actual device usage compared to similarity measures based on absolute values, such as the popular Euclidean distance. Euclidean and DTW distance also do not meet the needs of similarity measurement in the work [31]. This work extracts the patterns of human behavior but in terms of time series of item popularity over

online media and also proposes a distance measure invariant to scaling. But, unlike our case, they also consider the patterns to be similar if the peaks of activities are shifted in time. In case of behavioral patterns that are valuable to ISPs it is important that the traffic is active simultaneously or within the same aggregated time periods.

Stationarity. Since our goal is to extract time-evolving patterns and characterize these patterns across different dimensions of interest, we define a new notion of stationarity adapted to the peculiarities of our traffic time series. Our traffic data has very strong time semantics — traffic depends highly on the day of the week and on time of usage, so we cannot expect that the data distributions during the weekend and working days are the same. Thus, we are interested in time-framed patterns (from one day to another, or from week to another) and consider regularity of behavior in terms of non-overlapping time windows. In this respect, we measure the correlation similarity of the time series with itself, comparing each window of the chosen period with each other, in order to measure the entire stationarity of a period of interest. We also check whether the traffic data distribution changes significantly from one period to another using a non-parametric comparison test for arbitrary probability distributions (i.e. Kolmogorov-Smirnov test). More formally:

DEFINITION 2. *A time series of a gateway is* **strongly stationary** *for a particular window size if:*
• *it has a correlation similarity measure* > 0.6 *among all non-overlapping windows in consideration;*
• *the Kolmogorov-Smirnov test (that checks if the distributions of two time series is the same) is not rejected for all possible window pairs.*

The main difference between our 'strong stationarity' notion and the classical stationarity is that instead of using sliding windows, we use non-overlapping windows. Furthermore, apart from the similarity of the distributions, we also check the correlation similarity between the windows, which makes this a 'strong' notion of stationarity. Asserting that the time series of a gateway is strongly stationary, ensures that the underlying bandwidth usage is regular and can be described by a repetitive usage pattern.

Time aggregation. We use the notions of correlation similarity measure (Definition 1) and strong stationarity (Definition 2) in order to formulate optimization criteria for choosing the best time periods for aggregating traffic values, or the best binning of time series. More formally the problem is defined as follows:

DEFINITION 3. *Given a mapping function W of nonoverlapping time windows of length g defined over a set of times series U, the* **best aggregation granularity** $g_{best} \in G$ *is defined as*

$$g_{best} = \arg\max_{g \in G} E[cor(x(g), y(g)],$$

where $x(g), y(g) \in S$, a set of time series S defined from U through W: $S = W(U)$, $x(g)$ and $y(g)$ are aggregated traffic volume values according to the time binning g.

Mean is used as an unbiased estimate of $E[\cdot]$.

As we will see in Section 7.2, deciding which is the best traffic aggregation binning is crucial for extracting unknown meaningful recurring patterns of medium-term (week) and short-term (day) usage behaviors of a residential gateway. These patterns are called **motifs**.

Dominant devices. We also need to detect dominant devices per gateway, that is, the devices that have traffic time series very similar to the overall traffic of a gateway. We define a dominant device as follows.

DEFINITION 4. *Device d is* **ϕ-dominant** *per a gateway if the correlation similarity between its traffic and gateway traffic is larger than a threshold ϕ.*

Besides determining dominant devices of the traffic reported by the gateways of our deployment (Section 6.2), this definition will enable us to better characterize the motifs in terms of types of devices that contribute to the traffic of the motif the most(Section 7.2).

6. GATEWAY DEVICE ANALYSIS

In this section, we are interested in identifying the active usage traffic generated when residents actually run online applications, as well as in detecting the devices that contribute the most to the overall traffic.

6.1 Active Device Traffic

As we have seen in Section 4.1, the majority of the traffic volume values reported by the gateways are rather low. This background traffic is essentially attributed either to the control traffic generated by the operating systems of devices (e.g., during sleep mode or application software updates), or to low traffic generated by light applications running in the background (e.g., when a mail server checks for new emails, or when twitter updates the message list). Background traffic has its own patterns and fluctuations that influence our time-series analytics given that the majority of wireless devices in our data, such as tablets and smartphones, are frequently in the idle state and use their wireless radio rarely in order to increase the battery life.

In order to extract recurring patterns from active usage traffic generated when residents actually run online applications, we need to exclude the background traffic. Background traffic can be separated from active traffic by setting a threshold τ on the number of bytes in traffic time series. To obtain active traffic time series, all the values which are lower than τ are set to zero. Deciding on an appropriate threshold τ for background traffic is far from trivial, given the lack of a ground truth (both on the operating systems for particular sleep policies and the applications running on devices). Instead, we can exploit a general statistical technique based on the probability distribution of the traffic time series, as the boxplots described in Section 4.1. Given that the interval in which most of data values of time series belong falls between the whiskers of the plot, we use the upper whisker of a boxplot in order to define τ. This is supported by the fact that background traffic values are the most frequent in our data, while active traffic is sparse.

As this threshold is device specific, we estimate it for each device, per outgoing and incoming traffic separately. We study the background traffic for four weeks of data. For this period, we observed the traffic for 934 devices connected to user gateways. According to the histograms for outgoing and incoming traffic (refer to Figure 4), the background threshold for most of the devices is below 5000 bytes per minute (i.e., less than 1 Kbps), while for almost all the devices τ is lower than $40,000$ bytes for both outgoing and incoming traffic, which corresponds to a rate of 5.3 Kbps. There are

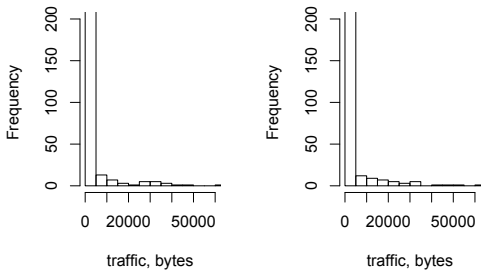


Figure 4: Distribution of τ of outgoing (on the left) and incoming (on the right) traffic.
24 devices with $\tau > 40000$ bytes for incoming traffic, and 15 devices with $\tau > 40000$ bytes for outgoing traffic.

We checked if there is a dependency on the type of device and the distribution of traffic it produces. We grouped the devices by τ as follows. 'Small' group corresponds to $\tau <= 5000$, 'medium' to $\tau \in (5000, 40000]$, and 'large' to $\tau > 40000$. We use the types of devices defined in Section 3. In the small and medium groups portable devices dominate, while in the last group fixed devices are the most popular as on PCs and laptops much more applications can run simultaneously in non-active mode. Thus, background traffic can be a significant feature for device type classification.

In summary, to exclude background traffic from consideration, we use a threshold per device of $\tau_{back} = min(\tau, 5000)$. This value is an upper border of the background traffic for the majority of the devices, as illustrated in Figure 4. Our threshold of 5000 bytes per minute for background traffic is also consistent with the previous works [25], [23], which set it to 1 kbps, thus making our threshold more tight.

As we will see in Section 7, background traffic removal reveals more regularity in traffic time series. The ability to automatically detect the background traffic of a device will also help ISPs to improve the energy saving policies *without* using data aggregation from multiple homes as has been proposed in the literature [8].

6.2 Dominant Devices

A device is considered to be dominant if it characterizes the general behavior of a gateway over time with respect to the overall traffic. As before, we only consider wireless traffic and wireless devices.

Definitions 1 and 4 are used in order to detect ϕ-dominant devices. As we are interested in high time series similarity we have chosen ϕ threshold to be 0.6, as before only statistically significant correlations were reported. We perform the search of dominant devices for all the gateways that have at least one observation per week for each week of consideration, we have observed 153 such gateways. The data contains the time series of all devices that were connected to a gateway after March 17, 2014 together with its overall traffic. If there are several dominant devices detected, we rank them in descending order of their correlation similarity.

According to the results, 7 gateways had 3 dominant devices, 43 gateways had 2 dominant devices, 99 gateways had 1 dominant device, and only 4 gateways did not have any dominant device. In most of the cases, there is at least one dominant device per gateway, meaning that the bandwidth consumption of the gateway is determined by the usage of the device. There might be several dominant devices, which can indicate that the number of residents regularly using network is higher than one. There are at most 3 dominant devices per gateway, we ranked them in the descending order of correlation similarity value, hence, first dominant devices has traffic time series that is the most similar to the overall traffic time series of a gateway.

We also checked what type of devices are dominant per gateway. Overall, among dominant devices we detected 74 fixed, 67 portable, 53 unlabeled, 9 network equipment devices, and 3 game_consoles. The distribution of the different device types, depending on the ranking of dominance, is shown in Figure 5. The plot shows that there are many gateways, for which the dominant devices for all the ranks are fixed devices. This is attributed to the fact that fixed devices produce in general more traffic and are usually connected for longer periods to the gateway. Still among dominant devices there is a significant number of portable devices, which are increasingly being used nowadays.

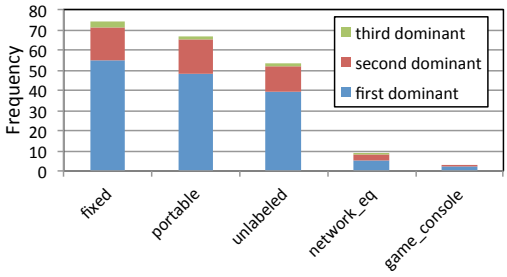


Figure 5: Distribution of types of dominant devices given their ranking.

The knowledge of dominant devices over long periods (one month in our case) is of great importance to ISPs, as it can provide a high level profiling of gateways.

Using Other Distance Metrics. For the sake of completeness we have compared the dominant devices obtained using our correlation-based similarity measure with those obtained when using the Euclidean distance or simply the absolute traffic volume used in work [23]. For the Euclidean distance computation, we consider the time series of a gateway $X = \{x_i\}_{i=1}^n$ and time series of a device $Y = \{y_i\}_{i=1}^n$, where n is the number of observations for four weeks of data. The Euclidean distance is computed using the formula: $dist_{Eucl} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$. Alternatively, the devices which produce the highest volume of traffic are considered to be dominant traffic-wise.

As there are no clear thresholds for Euclidean and traffic-based dominances, we compare the results of our correlation-based dominant devices, where we used meaningful threshold, with the devices that are ranked using the Euclidean distance(ascending order) and the traffic volume(descending order). Using the correlation dominance we have detected 206 dominant devices. For some gateways there can be two or three dominant devices. For each gateway we detect its correlation-based dominant devices and obtain a ranking of dominant devices based on the other two measures. Then if the devices in the ranking are the same, meaning that the first device in one ranking is also the first in the second ranking and so on, we say that the devices are detected equally using the two measures.

Among the 206 dominant devices, 182 (88%) are ranked the same as Euclidean-based dominant devices, and 151 (73%) are ranked the same as traffic-based dominant devices. Nevertheless, there are many cases, where dominant devices have lower overall traffic (around 15%), even though they closely follow the traffic time series of the gateway, with an exception of a few bursts (3 or 4). Our similarity measure is able to detect these devices, which cannot be detected

using the Euclidean and traffic-based distances.

We have also tried more strict ϕ threshold for dominance, $\phi = 0.8$. Even with very tight constraint on dominance, there is still large amount of gateways (67%) that have at least one dominant device and the ratio of fixed devices among the dominant devices is even larger.

Dominant Devices and Number of Residents. Having the results of a recent user survey over a subset of 49 gateways in our deployment, which contained information on the number of users per gateway, we checked if the number of dominant devices is correlated with the number of residents. The result of this analysis showed that there is no evidence of significant correlation. This may be due to the fact that different users are active during different periods of time, and in case of multiple users (and therefore multiple devices) the number of overall dominant devices is lower.

On the other hand, in the gateways with one user, there is always one dominant device detected. In the case of two users (9 gateways), 2 dominant devices are detected in 5 gateways (56%) and 1 dominant device is detected in 4 gateways (44%) and there are no three dominant devices detected. We have calculated the correlation coefficient between the number of dominant devices and the number of users only for gateways with 1 and 2 users, and we obtained a statistically significant correlation value = 0.53. In the case where the number of users > 3, the effect of multiple devices, discussed in the previous paragraph, is present again, and only 1 or 2 dominant devices are detected.

Summary. The most interesting findings of our analysis of dominant devices per home are:

1) Each gateway has at least one dominant device.

2) The majority of dominant devices are fixed devices.

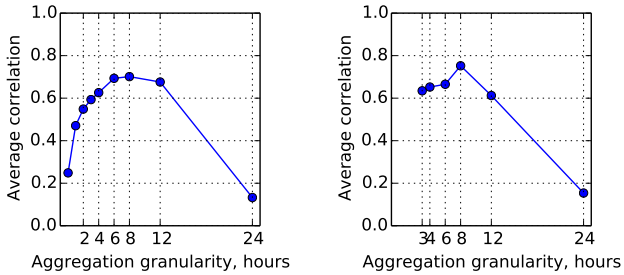
3) Correlation based dominance is beneficial because it identifies dominant devices that are missed by Euclidean, or traffic-based distance notions.

4) The number of dominant devices provides a lower bound for the number of residents per gateway.

7. TIME AGGREGATIONS AND MOTIFS

In this section, we are interested in investigating which time aggregations of gateway traffic better reveal regular user behavior in houses. For example, whether the produced traffic is more significant in the morning rather than in the evening, during the weekdays rather than the weekends, etc. By checking various meaningful time aggregations ranging from 30 minutes to 12 hours, we have experimentally verified that the larger the aggregations are, the more correlations are observed within or across gateways. This is due to the fact that the periods of non-active traffic become smaller, while traffic peaks become more similar. If at the same time we exclude the points of background traffic, the actual usage patterns of home gateways become more visible.

Since there is no golden standard as to which aggregation should be used, the choice is usually application driven. In our work we rely on the notion of strong stationarity (see Definition 2) capturing repetitive usage patterns of gateways to systematically determine the right aggregation level. If a time series is strongly stationary, then the patterns found for this time series are stable and we can generalize the results of the analysis over this time series. This is not always truth if standard bining such as morning, working hours, late afternoon, evening, night is used, because usually the borders of this binning, number of bins and their length is not experimentally verified but just arbitrary set.



(a) Average correlation for stationary gateways starting at midnight

(b) Average correlation for stationary gateways starting at 2am

Figure 6: Aggregation curves for weekly patterns.

We consider two kinds of windows: daily-windows starting from midnight to reveal short-term patterns of usage behavior and weekly windows staring from Mondays to reveal medium-term patterns. For four weeks of data and a weekly period with 3 hours aggregation 7 % of gateways appeared as stationary. Thus, though there are strongly stationary gateways, still most of them change their behavior from week to week. Finally after removing the background traffic (see Section 6.1) 11% of gateways were detected as stationary. In the next section we choose the best aggregation period according to maximization of time series correlation.

7.1 Best Aggregation Period

In this section, we discuss how to choose the best aggregation period, according to the maximization of the time series correlations across time. The problem is formally stated in Definition 3. Background traffic is removed from all time series, as described in Section 6.1.

7.1.1 Weekly Patterns

First, we consider the best aggregation for medium-term patters of weekly behavior. As traffic depends heavily on the time of the day from day to day we considered all time aggregations starting at midnight that are factors of 24 hours, namely 1, 2, 3, 4, 6, 8, 12, 24 hours and additionally we considered initial time series aggregation, which is 1 minute. We also try aggregation granularities which are larger then 2 hours, starting from 2am and 3am. For the analysis, we consider all the user gateways that have at least one traffic observation every week during the 4 weeks of interest. The total number of such gateways is 153.

An aggregation period is considered to be the best if it reveals the highest correlation of traffic time series of a gateway values from one week to another.

In order to compare aggregations, we calculate the average correlation among all the week-week pairs separately for each gateway, and for strongly stationary gateways (Definition 2). The plots of the average correlation values are shown in Figure 6.

The maximum points for strongly stationary gateways are reached at granularity periods of 6, 8 and 12 hours for aggregations from midnight (Figure 6a) and 8 hours for aggregations starting from 2am (Figure 6b). When considering all the gateways, large correlation points are reached for 3, 4 and 6 hours of aggregations starting at midnight, while 8 hour aggregations starting at 2am is still a maximum point. Since the 8 hours aggregation period starting at 2am is an absolute winner for weekly patterns, we use this aggregation for our further analysis. Note that this aggre-

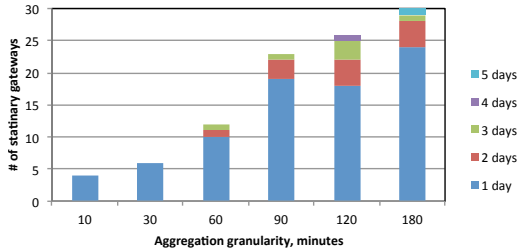


Figure 7: Stationary gateways per aggregation window.

gation has a meaningful semantic interpretation: each day is divided into 3 periods, namely, "morning" between 2am-10am, "working hours" between 10am-6pm, and "evening" between 6pm-2am. The traffic behavior of gateways for this aggregation is also of interest to ISPs.

7.1.2 *Daily Patterns*

As the initial observations are received at a rate equal to 1 minute, and for daily patterns we need to have a reasonable amount of points, we try the following aggregation periods: 1, 5, 10, 30, 60, 90, 120 and 180 minutes. We do not consider aggregations larger than 180 minutes as it is not desirable to have less than 8 data points per pattern. All the binning values are factors of 1440 minutes, which constitute a day.

As before we define the aggregation period to be the best if it reveals the highest correlation of traffic time series of a gateway values for daily patterns. Unlike weekly patterns, we do not require every day to be similar to each other, but we expect that Mondays should be highly correlated with Mondays, Tuesdays with Tuesdays, etc. For the analysis, we consider all the user gateways that have at least 1 traffic observation every day during the 4 weeks of interest. Their number is 100.

For daily patterns we also studied strongly stationary gateways, where the behavior of the same days of the week is stationary (in the sense of Definition 2). Note that for stationarity we require not only highly correlated observations among the corresponding days of the week (e.g., all Mondays should be correlated with each other), but we additionally require that the probability distributions of all instances of that weekday should be similar. The number of stationary gateways per aggregation granularity is shown in Figure 7. We also decompose the total number of stationary gateways to the number of gateways which have one stationary day of the week, two stationary days, and so on.

Figure 7 shows that the number of stationary gateways grows with the aggregation granularity. Additionally, more days are stationary within the same gateways if the granularity is larger.

In order to compare the aggregation granularities we calculate the average correlation among all the pairs of the same day of the week separately for all gateways and for gateways that appeared to be strongly stationary. The plots of average correlation are shown in Figure 8.

The results show that small aggregation periods correspond to low regularity in the data; moreover, there are no gateways with at least 1 stationary day of week for aggregation granularities of 1 and 5 minutes. The correlation value for all the gateways grows significantly up to 1 hour aggregation, then it becomes stable up to the level of 180 minutes or 3 hours. At the same time, the average correlation of the stationary gateways keeps growing with the larger aggregation granularity, and the highest value is reached for the 3 hours aggregation granularity, which is the aggregation period we use for daily patterns in the rest of this study.

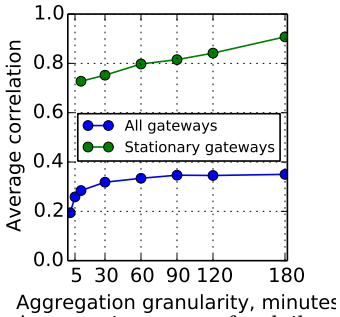


Figure 8: Aggregation curves for daily patterns.

7.2 Motif Discovery

As mentioned in Section 5, for motif extraction we only consider patterns that correspond to medium-term (week) and short-term (day) usage behaviors across time and across gateways. Patterns within a particular gateway only or of a longer period can also be identified following the proposed methodology.

The following definition of a motif is used in our work:

DEFINITION 5. *Given a set of times series U, a mapping function W of non-overlapping windows, which extracts periods of interest S from U according to a window length and its starting point synchronizes with the corresponding timestamp of U (beginning of a day or a week), $S = W(U)$,* **motif** *is a set $M \subseteq S$, $M = (m_i)_{i=1}^{k}$, where k is a support of a motif. M has the following properties:*

1. **individual similarity***: $\forall i \in \{1, 2, ..., k\}$, $\exists j \in \{1, 2..., k\}$: $cor(m_i, m_j) \geq \phi$,*
2. **group similarity***: $\forall i, j \in \{1, 2, ..., k\}$, $i \neq j$: $cor(m_i, m_j) \geq 3/4\phi$,*

Thus, when a new subsequence is included in a motif it is very similar to at least one existing subsequence in S and it is reasonably similar to all the rest s_i, $i = 1, ..., k$. In our case $\phi = 0.8$.

In other words, two time series are considered to constitute a motif if the correlation distance between them is very high. The threshold we have chosen is 0.8. Several motifs can be combined if all the time series that comprise the motifs have high correlations with each other. In this case, the correlations should be ≥ 0.6.

Motif extraction enables us to enrich traditional analysis of the aggregated traffic reported by a gateway. As a matter of fact, we can detect detailed behaviors within the same house that can be attributed to different residents (e.g., adults or children), or to different habits (i.e., daily or weekly patterns). Identification of diverse behaviors inside a single house goes beyond the current state of the art. Furthermore, to provide additional information about the obtained motifs, we analyze them across the following dimensions:

1. How many dominant devices per gateway contributed to the motif? In this case we consider dominance for the corresponding time period of time series that formed the motif.
2. How do the dominant devices per motif and gateway relate to the overall dominant devices of a gateway detected for a period longer than 4 weeks?
3. What is the distribution of the dominant devices per motif? We consider portable, fixed devices, network equipment and others as discussed in Section 6.1.
4. Are there daily motifs, which are more common among weekends than working days and vice versa?

5. What gateways contribute the most to the motifs?
For the analysis we concentrate on significant motifs with high support values, so called motifs of interest.

7.2.1 *Weekly Motifs*

In order to find weekly motifs we use 8 hour aggregation period starting at 2am, as it is the best time aggregation according to the experimental results in Section 7.1.1. The motif search was done on user gateways that have at least one observation for each week, out of the six weeks starting from March 17th. The number of such gateways is 147.

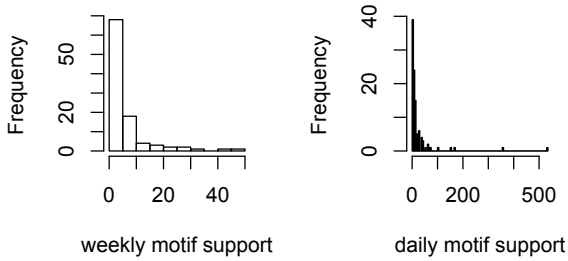


Figure 9: Distributions of support values for weekly (left) and daily (right) motifs.

As a result, 101 motifs are discovered from 882 (147*6) weeks of observations. Out of those, 14 motifs have support $\geq$ 10. The distribution of the support values is shown in Figure 9. For weekly motifs, the participation of gateways in motifs is rather low, but at least one week time series per gateway contributed to the motif construction. The top gateways among the gateways with the highest contribution provided at least 6 time series for weekly motifs while on average number of the distinct motifs per gateway is 2.76. The distribution of the number of distinct weekly motifs per gateway can be seen on Figure 10.

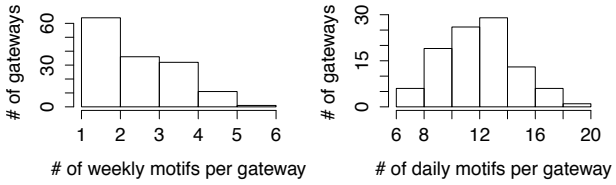


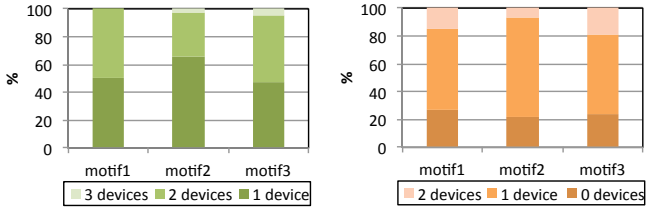
Figure 10: Distributions of the number of motifs, where a gateway participated in.

Some of the motifs of interest are shown in Figure 11.

In general, all the detected weekly motifs correspond either to the patterns of everyday evening usage, or to the patterns where certain days are the most influential.

We further elaborate on weekly motifs of interest, as shown in Figure 11. We label the motif in Figure 11a - motif1, Figure 11b - motif2, and Figure 11c - motif3. The distribution of the number of dominant devices for these motifs is shown in Figure 12a. We observe that these motifs have one or two dominant devices, and most of them correspond to the overall dominant devices of a gateway according to Figure 12b. Nevertheless, there are some devices, which are dominant per motif time slot, but not dominant for a gateway overall. The number of these devices is at least 2.5 times smaller than the number of common dominant devices.

We notice that motif1 and motif3 are mainly observed for portables (Figure 13), while motif2 is observed for fixed devices. This can be attributed to portable devices being used in the evenings. But, more regular users, like the ones contributing to motif2, tend to use fixed devices.



(a) Distribution of the number of dominant devices.

(b) Distribution of intersections with overall dominant.

Figure 12: Dominant devices for weekly motifs.

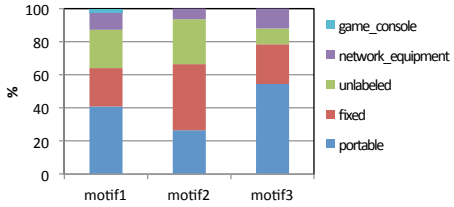


Figure 13: Distribution of types of dominant devices for weekly motifs.

7.2.2 *Daily Motifs*

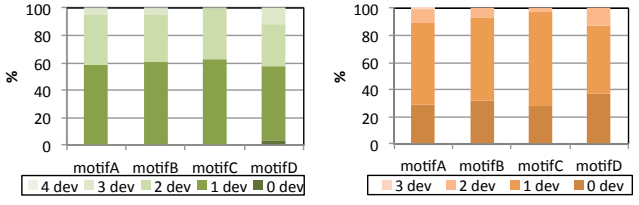
We have extracted daily motifs for the aggregation period that is considered to be the best in terms of the highest number of correlated patterns, as discussed in Section 7.1.2. This is the three hours aggregation which gives 8 points for each daily time series.

The daily motifs were extracted from time series of 100 user gateways (out of 196 gateways), which have at least one observation per day in raw time series for the observation period of four weeks. In total 112 motifs were extracted, with 48 motifs having support larger than 10. The distribution of the support values is shown in Figure 9. The most popular (with the highest support) daily motifs are connected to various evening usages, while there are still motifs with daily behaviors and mixed morning and evening behaviors. Surprisingly, for daily motifs each gateway contributed with at least 16 time series. The top gateways (among the gateways with the highest contribution) provided at least 28 time series for daily motifs. At the same time the average number of district motifs per gateway is 12.5.

The distribution of the number of distinct daily motifs per gateway is shown in Figure 10. The support of the daily motifs is more repetitive between the same homes than in the case of the weekly motifs. This can be attributed to the fact that more days per gateway were considered. 28 data windows are used for daily motifs, while 6 data windows are used for weekly motifs.

Analysis. We analyze in detail 4 representative daily motifs shown in Figure 14. The distribution of the number dominant devices is illustrated by Figure 15a.

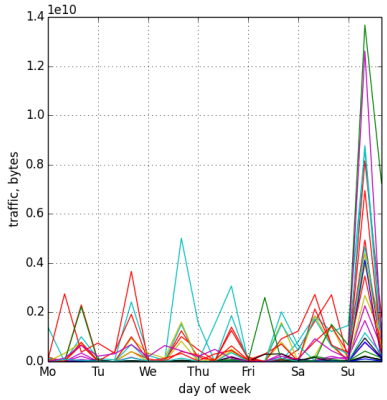
We observe that motifs usually have one or two dominant devices. Unlike the weekly motifs, many of them do not correspond to the overall gateway's dominant devices (Figure 15b). However, the majority still coincides with the overall dominant. This higher ratio of new dominant devices



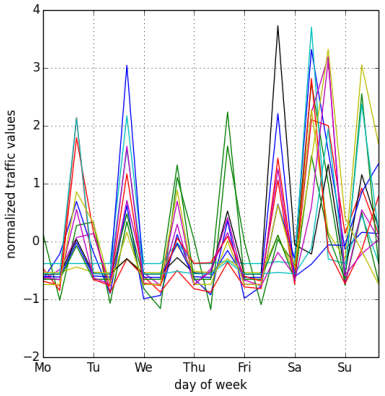
(a) Distribution of number of dominant devices.

(b) Distribution of intersections with overall dominant.

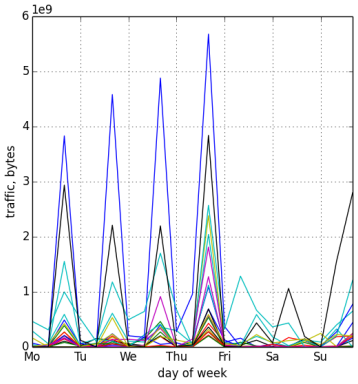
Figure 15: Dominant devices for daily motifs.



(a) Heavy Weekend Users: Support = 26 time series (23 % occur during different weeks in the same gateways).

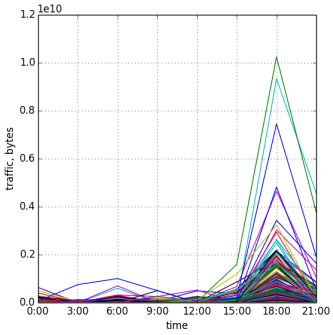


(b) Everyday Users: Support = 13 time series (15 % occur during different weeks in the same gateways).

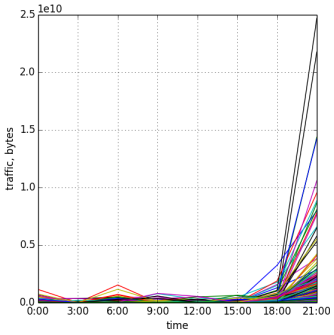


(c) Workdays Users: Support = 21 time series (29 % occur during different weeks in the same gateways).

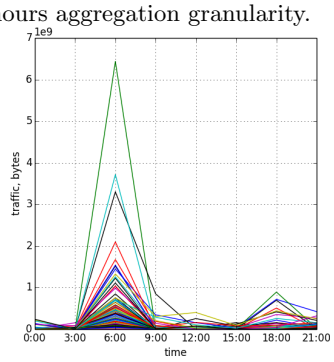
Figure 11: Weekly motifs for 8 hours aggregation granularity.



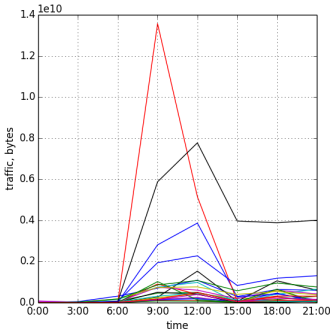
(a) Afternoon users: Support = 356 time series (95% occur within the same gateways).



(b) Late evening users: Support = 534 time series (98% occur within the same gateways).

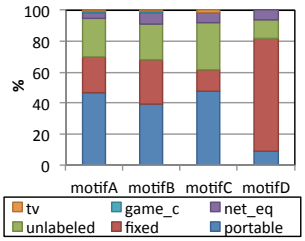


(c) Morning and evening users: Support = 105 time series (83% occur within the same gateways).

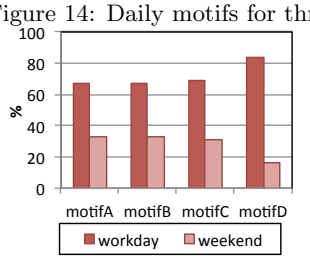


(d) All day users: Support = 24 time series (67% occur within the same gateways).

Figure 14: Daily motifs for three hours aggregation granularity.



(a) Distribution of the types of dominant devices.



(b) Distribution of working days and weekends.

Figure 16: Dominant devices and days. Daily motifs.

is attributed to the small time intervals (i.e., a single day), while overall dominance is determined for 4 weeks of data.

The distribution of dominant device types is shown in Figure 16a. Motifs A, B and C which correspond to high usage behavior in the morning or/and in the evening are mainly created by portable devices, while motif D, with all day active usage(and similar daily motifs we detected), is more generated by fixed devices. In general, daily patterns have higher percentage of portable devices as dominant devices, especially in the cases of not continuous usage behaviors.

Most of the motifs correspond to both working days and weekends, but all day usage (motifD) contains more working days (Figure 16b). The relative support of working-day motifs is larger than that of weekends, as working days are more frequent in the training set.

Summary.

1) Portable devices are sources of short-term morning or evening activities, while fixed devices generate day and night long-term patterns.

2) Gateways participate in several motifs. This supports our hypothesis that a gateway involves multiple distinct behavioral patterns. Our approach is able to reveal those, thus leading to an accurate characterization of the gateways, which in turn provides valuable insights to ISPs.

We observe that the discovered motifs reveal fine-grained regular behaviors that can be exploited in order to better manage residential networks. For example, our analysis identifies groups of users, irrespective of their location and demographics, that share similar time periods of low activity (in different parts of the day, or night). These can be used by ISPs and CSPs to schedule maintenance processes in a way that minimally interferes with the user activities.

8. CONCLUSION

In this work, we analyze the wireless traffic time series of 196 home gateways. We describe a similarity measure suitable for capturing the characteristics of traffic *evolution* within and across gateways. We propose a notion of stationarity that, in addition to the similarity of data distributions, also imposes a correlation similarity across non-overlapping time windows. This work is a first step towards understanding fine-grained regularities on residential traffic consumption. ISPs and CSPs could leverage such analytics to enable remote maintenance services such as: (i) troubleshooting and firmware/software updates of RGWs, and (ii) hotspot resource management and collaborative networks, which require *fine time-scale identification* of gateways' and home devices' active and idle times. Existing methods rely heavily on wireless traffic stationarity for such predictions, which

do not hold in home networks (cf. Section 2). Our analytics framework uncovers the best aggregation of home traffic values, in order to identify motifs and to detect gateways with similar/different traffic patterns. We are currently working towards integrating our time series correlation and motif extraction, in a streaming big data analytics platform, such as Apache Storm or Amazon Kinesis.

Acknowledgments

We would like to thank Gevorg Poghosyan, Augustin Soule, Pascal Le Guyadec, Henrik Lundgren, Jaideep Chandrashekar and Christophe Diot for their precious help in making sense of wireless home traffic data. This work was partially funded by the European ICT FP7 User Centric Networking project (grant no. 611001).

References

[1] M. Chetty, R. Banks, R. Harper, T. Regan, A. Sellen, C. Gkantsidis, T. Karagiannis, and P. Key. Who's hogging the bandwidth?: The consequences of revealing the invisible in the home. In *CHI 2010*. Association for Computing Machinery, Inc., April 2010.

[2] G. W. Corder and D. I. Foreman. *Nonparametric statistics for non-statisticians: a step-by-step approach*. 2009.

[3] L. DiCioccio, R. Teixeira, and C. Rosenberg. Measuring home networks with homenet profiler. In *Passive and Active Measurement*, 2013.

[4] N. Eagle and A. S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behav Ecol Sociobiol*, 2009.

[5] J.-P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *National Academy of Sciences*, 101(40):14333–14337, 2004.

[6] S. B. Eom, M. A. Ketcherside, H.-H. Lee, M. L. Rodgers, and D. Starrett. The determinants of web-based instructional systems' outcome and satisfaction: An empirical investigation. *Instructional techn.: Cognitive asp. of online programs*, 2004.

[7] P. Ferreira and P. Azevedo. Evaluating deterministic motif significance measures in protein databases. *ALMOB*, 2(1), 2007.

[8] E. Goma, M. Canini, A. Lopez Toledo, N. Laoutaris, D. Kostić, P. Rodriguez, R. Stanojević, and P. Yagüe Valentin. Insomnia in the access: Or how to curb access network related energy consumption. *SIGCOMM Comput. Commun. Rev.*, 41(4):338–349, 2011.

[9] B. Gonçalves and J. J. Ramasco. Human dynamics revealed through web analytics. *Phys. Rev. E*, 78:026123, Aug 2008.

[10] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 2008.

[11] S. Grover, M. S. Park, S. Sundaresan, S. Burnett, H. Kim, B. Ravi, and N. Feamster. Peeking behind the nat: an empirical study of home networks. In *IMC Conference*, 2013.

[12] A. P. Jardosh, K. Papagiannaki, E. M. Belding, K. C. Almeroth, G. Iannaccone, and B. Vinnakota. Green wlans: On-demand wlan infrastructures. *Mob. Netw. Appl.*, 14(6):798–814, 2009.

[13] Y. Jennifer. The data-driven approach to network management: Innovation delivered, 2010.

[14] H.-H. Jo, M. Karsai, J. KertÂl'sz, and K. Kaski. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055, 2012.

[15] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E*, 83:025102, Feb 2011.

[16] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong. Mobile data offloading: How much can wifi deliver? *ACM CoNEXT*, 2010.

[17] Y. Li, J. L. 0001, and T. Oates. Visualizing variable-length time series motifs. In *SDM*, 2012.

[18] J. Lin, E. J. Keogh, and S. Lonardi. Visualizing and discovering non-trivial patterns in large time series databases. *IVI*, 4(2), 2005.

[19] J. Lin, E. J. Keogh, L. Wei, and S. Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Min Knowl Discov*, 15(2), 2007.

[20] S. MAKRIDAKIS and M. HIBON. Arma models and the box–jenkins methodology. *Journal of Forecasting*, 16(3), 1997.

[21] J. Martin and N. Feamster. User-driven dynamic traffic prioritization for home networks. In *ACM SIGCOMM Workshop on Measurements Up the Stack*, W-MUST '12, 2012.

[22] A. Patro, S. Govindan, and S. Banerjee. Observing home wireless experience through wifi aps. In *ACM MobiCom '13*.

[23] I. Pefkianakis, H. Lundgren, A. Soule, J. Chandrashekar, P. Le Guyadec, C. Diot, M. May, K. Van Doorselaer, and K. Van Oost. Characterizing home wireless performance: The gateway view. In *accepted for IEEE INFOCOM 2015*.

[24] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li. Traffic-driven power saving in operational 3g cellular networks. In *MOBICOM*, pages 121–132, 2011.

[25] G. Poghosyan. Device analytics in home networks. Master's thesis, EPFL, August, 2014.

[26] E. S. Poole, W. K. Edwards, and L. Jarvis. The home network as a socio-technical system: Understanding the challenges of remote home network problem diagnosis. *Comput. Supported Coop. Work*, 18(2-3):277–299, June 2009.

[27] K. Poularakis, I. Pefkianakis, J. Chandrashekar, and L. Tassiula. Pricing the last mile: Data capping for residential broadband. *ACM CoNEXT*, 2014.

[28] C. Rossi, C. Borgiattino, C. Casetti, and C. F. Chiasserini. Energy-efficient wi-fi gateways for federated residential networks. In *IEEE WoWMoM'13*.

[29] R. Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1), 1990.

[30] L. Waluyan, S. Sasipan, S. Noguera, and T. Asai. Analysis of potential problems in people management concerning information security in cross-cultural environment -in the case of malaysia-. In *HAISA*, 2009.

[31] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.

[32] C. Zhang, Y. He, and Y. Ji. Temporal pattern of user behavior in micro-blog. *Journal of Software*, 8(7), 2013.