Extracting Usage Patterns of Home IoT Devices

Gevorg Poghosyan, Ioannis Pefkianakis, Pascal Le Guyadec, Vassilis Christophides

Extracting Usage Patterns of Home IoT Devices

Gevorg Poghosyan*, Ioannis Pefkianakis†, Pascal Le Guyadec‡ and Vassilis Christophides§
*Insight Centre for Data Analytics, University College Dublin, Ireland
†Hewlett Packard Labs, USA
‡Technicolor Research, France
§INRIA, France

***Abstract*—Ubiquitous connectivity and smart technologies gradually transform homes into *Intranet of Things*, where a multitude of connected, intelligent devices allow for novel home automation services. Providing new services for home users (e.g., energy saving automations) and Internet Service Providers (e.g., network management and troubleshooting) requires an in-depth analysis of various kinds of data (connectivity, performance, usage) collected from home networks. In this paper, we explore new Machine-to-Machine data analysis techniques that go beyond binary association rule mining for traditional market basket analysis considered by previous studies, to analyze individual device logs of home gateways. We introduce a multidimensional patterns mining framework, to extract complex device co-usage patterns of 201 residential broadband users of an ISP, subscribed to a triple-play service. Our results show that our analytics engine provides valuable insights for emerging use cases such as *monitoring for energy efficiency*, and *"things" recommendation*.**

***Keywords*-IoT; home networks; association rules**

I. Introduction

With the rapid growth of smart technologies, modern homes are gradually transformed into *Intranet of Things*. A variety of devices (smartphones, IPTVs etc.) get connected via wireless or wired home networks to offer multiple (not always integrated) services (e.g., home automation). Connectivity of things as well as residential broadband access is provided by home gateways capable of monitoring the operations and the performance of the connected devices.

In this paper, we are interested in analyzing device usage logs in order to support emerging use cases in smart homes such as *adaptive usage of home devices*and *"things" recommendation* [1]. Such use cases fall within the wider area of *human-cognizant* Machine-to-Machine communication aiming to predict user needs and complete tasks without users initiating the action or interfering with the service. While it is not a new concept, according to Gartner cognizant computing is a natural evolution of a world driven not by devices but collections of applications and services that span across multiple devices in which human intervention becomes as little as possible by analyzing past human habits.

To realize this vision, we are interested in co-usage patterns featuring spatio-temporal information regarding the context under which devices have been actually used in homes. For example, a network extender which is currently off, could be turned on at a certain time (e.g., evening) when it has been observed to be highly co-used with other devices (e.g., tablets). Alternatively, the identification of frequent co-usage of particular home devices (e.g., iPhone with media player), could be used by a "things" recommender to advertise the same set of devices at another home (say another iPhone user may be interested in a media player).

We advocate frequent pattern and association rule mining techniques since we believe that they are more easily understood by both end-users (for raising awareness regarding device energy or bandwidth consumption) and developers (for programming *if this then that* scripts of home automation), than the potentially more accurate but opaque Machine Learning techniques (e.g., classification). Traditional market basket analysis has been recently revised for extracting associations between users' interactions (e.g., communication and entertainment services) and context (e.g., time periods) captured by mobile devices [2], [3], frequent co-occurring mobile context events (e.g., a user listens to music during workdays, while driving) [4] or frequent co-usage patterns of different appliances under various contexts [5]. Unlike these works, we extract n-ary (vs. binary) patterns from device logs involving attributes of at least three distinct entities: $Device$, $Context$, and $Activity$. An extra $Gateway$ dimension is also considered when extracting recurring patterns across homes. Rather than decomposing our analysis problem into several binary ones ($Context \times Activity$, $Context \times Device$, etc.), we leverage recent advances in constraint-based algorithms [6], [7] for mining arbitrary n-ary relations. Our main contributions are:
1) We analyze a new dataset collected from home gateways, of subscribers of a large European ISP (Section II). Our dataset includes various information such as device connectivity, performance and usage data collected at a fine time granularity (per 30 seconds), under normal service operation, by an *important number* of gateways (201) on which a *large number* (2828) of fixed (e.g., desktops, laptops) and portable (e.g., tablets, smartphones) devices are connected, as well are also IPTVs and phones. To our knowledge, this is the largest scale study of triple-play home subscribers to date.
2) We introduce a discrete representation of gateway logs that is flexible enough to capture device activities spanning multiple contexts or vice versa (Section III). We enable an on-demand generation of device usage logs that combines

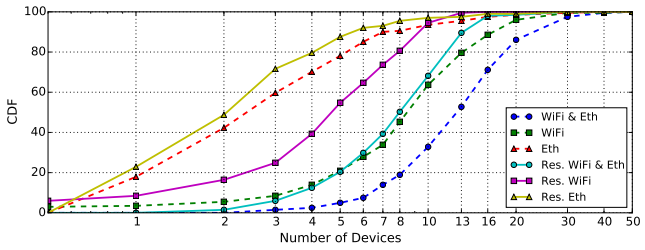


Figure 1. Number of devices per home.

usage evidence from multiple raw data logs (unlike unique transactions in market-basket analysis), while controls the spatial and temporal log resolution.

3) We extract frequent n-ary patterns and rules from device logs within or across houses (Section IV). These patterns uncover that devices are more frequently co-used at low traffic data rates, close to the gateway, during evenings. These co-usage patterns are significantly different across homes and independent of the number of devices.

4) We implement our data acquisition and analysis system, and discuss its performance requirements (Section V)[1].

II. Dataset Description

We analyze more than 21 million measurement reports collected over a 3-month period (February - April, 2014) from the home gateways of 201 residential broadband subscribers of a large European ISP, geographically distributed over 10 cities. The 2/3 of the gateways use fiber and the rest ADSL2+ to connect to Internet. Each gateway has 4 Ethernet ports, an 802.11b/g/n WiFi access point, FXO POTS ports to connect phones. An IPTV connects to the gateway through an Ethernet set-top box (STB) device.

The gateway reports each device's MAC address for each connected interface (Ethernet or WiFi), and its names given by users (e.g., "Kelitas-iPad"). It also reports every 30 seconds the device's *traffic data rate* in Kbps for both downlink (gateway to device) and uplink, and the WiFi *signal strength* (RSSI). The traffic for Ethernet devices is reported every 1 minute. Finally, the start time, duration, and direction (incoming or outgoing) of phone calls are reported, along with an indicator of which calls have been answered.

A device is defined by its MAC address - e.g., a laptop with WiFi and Ethernet interfaces appears as two devices. Out of the 2828 devices, 69.5% are WiFi and 30.5% are Ethernet. The number of WiFi and Ethernet devices per home varies from 3 to 50. Roughly 80% of the homes have more than 8 devices, as shown in Figure 1 (dashed lines). WiFi typically outnumber the Ethernet devices, and vary between 0 and 46 per home. Among the connected devices, there are also WiFi extenders connected to the gateway via Ethernet or WiFi. Although we can identify the devices behind an extender, we cannot specify their network interface; e.g., all the devices behind a WiFi extender which connects through Ethernet to the gateway, will appear as Ethernet devices. This justifies the high number of Ethernet devices (> 40) in some homes.

In our analysis we distinguish between *resident* (used regularly by residents), and *guest devices*, which are rarely used and typically belong to visitors. Devices which are connected for $\leqslant 7$ days are labelled as guest, or otherwise as resident. Figure 1 (solid lines) shows the distribution of resident devices at each gateway (varies from 2 to 28). In the rest of the paper we focus only on resident devices.

III. Core Entities in Home Intranet of Things

To represent device usage logs in residential Intranet of Things we rely on 4 types of entities shown in Table I. *Gateway*s are described by their identifier (Gid) and the broadband access type ($Access \in \{fiber, adsl\}$). *Device*s are described by their globally unique identifiers (Did), the gateway (Gid) and the physical interface ($Port \in \{wlan, eth, phone\}$) to which they are connected to, as well as, their MAC address (Mac). The need for globally unique Did stems from the fact that, devices may move across the homes, so the same MAC address may appear in more than one home. We also consider the device $Kind$ and $Vendor$ derived mainly by the MAC addresses. Table II describes the different device types observed in our dataset. Our gateways do not report any information regarding the device kinds. For all the devices with the exception of STBs, whose MAC addresses have been provided by our ISP, we have used a simple heuristic classification based on their MAC addresses and device names [9]. We have assessed the high accuracy of classification results against a ground truth collected by surveys from 49 homes of our deployment.

The *Context* and *Activity* of Table I capture information related to the actual device usage. *Context* records the contiguous time intervals ($Begin$ and End timestamps) a particular device has been reported to be connected or disconnected in ($State$). Other contextual information could be considered depending on the device type and the scope of analysis. For example, for WiFi devices the $Quality$ of the received signal strength is an indicator of the device proximity to the gateway. In order to map RSSI to wireless link speed and quality, we use the thresholds presented in [10]. $Period \in \{night; morning; afternoon; evening\}$ or $Weekday \in \{workday; weekend\}$ represents useful temporal context that can be easily derived from the session timestamps and included to the device usage logs. Note that each device could operate only under one context within the corresponding time-interval recorded in the table *Context*.

Activity refers to the traffic rate of connected devices during contiguous time intervals ($Begin$ and End timestamps). We discretize the cumulative traffic rates (downlink and uplink) generated by a data device into different $Level$s capturing general classes of applications that could run on devices [11], as shown in Table III. Note that, although

[1]A preliminary version of our work appears in [8]. In this work, we present our complete set of experiments, and detail the extracted multidimensional patterns along with our proposed system architecture.

Table I
GATEWAYS, DEVICES, USAGE CONTEXT AND ACTIVITY SESSIONS.

(a) Gateway.

Gid	Access
g1	fiber
g2	adsl
g10	adsl

(b) Device.

Did	Gid	Mac	Port	Kind	Vendor
d1	g4	00:22:3a:*...*	eth	tv	Cisco
d3	g54		phone	phone	
d30	g2	9c:e6:35:*...*	wlan	portable	Apple

(c) Context.

Did	Begin	End	State	Quality
d30	5/3/2014, 23:30	6/3/2014, 00:10	connected	high
d30	6/3/2014, 00:10	6/3/2014, 00:20	connected	medium
d30	6/3/2014, 00:20	6/3/2014 00:40	connected	low
d30	6/3/2014, 00:40	6/3/2014, 07:00	disconnected	
d31	6/3/2014, 06:30	6/3/2014, 07:00	connected	

(d) Activity.

Did	Begin	End	Activity
d30	5/3/2014, 23:30	6/3/2014, 00:00	high
d30	6/3/2014, 00:00	6/3/2014, 00:30	medium
d30	6/3/2014, 00:30	6/3/2014, 00:40	idle
d31	6/3/2014, 06:30	6/3/2014, 07:00	low

the instantaneous peak traffic of the above applications may exceed their data rate bin, our gateways report an average rate over a 30-second or 1-minute period, which falls into the above bins. The activity levels for STBs and IP phones vary from the ones specified in Table III. We consider the STBs to be $idle$, when no content is being watched. During $idle$ activity, there still can be some traffic ($< 500\,kbps$) from STB firmware updates or from users browsing the menu. The activity level is $high$ when a user is watching TV (rate $\geqslant 500\,kbps$). Phone's activity level is considered $idle$ when there are no successful phone calls, and $high$ when there are active calls. Also note that each device could exhibit only one activity within the corresponding time-interval.

Table II
OVERVIEW OF HOME DEVICE KINDS.

Device Kind	Number of Devices	Device Sub-kind Examples
tv	328	set-top boxes, AppleTVs, chromecasts, media players
$phone$	135	IP phones, DECT phones
$portable$	931	tablets, smartphones
$fixed$	519	desktops, laptops, netbooks
$network\_equipment$	64	routers, WiFi extenders, media bridges, PLC modems
ip_camera	4	IP-cameras
$peripheral$	9	printers, scanners, projectors
$game\_console$	70	game consoles
nas	2	NAS
$other$	1	Raspberry Pis, Arduinos

Table III
ACTIVITY LEVELS AND CORRESPONDING APPLICATIONS.

Activity Level	Application	Traffic Rate (kbps)
idle	mail client sync	< 7
$< 15\,kbps$	Skype/Viber text chat	< 4
low	Skype voice/video call	2×(24 to 300)
$[15, 500)\,kbps$	online radio	32 to 320
	YouTube (240p, 360p)	300 to 500
	web surfing (FB, news)	< 500
medium	YouTube (360p, 480p, 720p)	500 to 2000
$[500, 2000)\,kbps$	Skype video call HQ	2×(400 to 2000)
	file transfer, cloud sync	< 2000
high	YouTube HD (720p+)	2000 to 6000
$\geqslant 2000\,kbps$	Skype HD, conf. video call	2000 to 8000
	torrents	$\geqslant$2000

The adopted discretization aims to reduce the inherent data noise and scarceness in order to improve the statistical significance and interpretability of extracted patterns. The proposed data representation, can flexibly capture complex usage logs in Intranets of Things where a device activity may span multiple contexts or vice versa. For example, a user may switch from checking email to video watching (activity changes) on her tablet, while she is close to the gateway (context remains the same). Or, a user may move from living room to the garden (signal quality and the context change), while she is watching a video (activity remains the same). More precisely, the usage logs of a device are generated from the recorded Context and Activity sessions using the following join condition: $(C.Begin < A.End \wedge A.Begin < C.End)$. Since each device activity can be tracked only within a context, we need only to check for time intervals of contexts that overlap those of activities. The granularity of temporal aggregations clearly affects the density of generated usage logs, and is guided by the objectives of our analysis.

Table IV
USAGE LOG OF DEVICE D30.

Day	Begin	End	Quality	Activity
5/3/2014	23:00	00:00	high	high
6/3/2014	00:00	00:10	high	medium
6/3/2014	00:10	00:20	medium	medium
6/3/2014	00:20	00:30	low	medium
6/3/2014	00:30	00:40	low	idle

Table V
USAGE LOGS FOR GW $g10$.

Day	Did	Period	Quality	Activity
5/3/2014	d30	evening	high	high
6/3/2014	d30	night	high	medium
6/3/2014	d30	night	medium	medium
6/3/2014	d30	night	low	medium
6/3/2014	d30	night	low	idle
6/3/2014	d31	morning		low

IV. ANALYSIS OF DEVICE CO-USAGE

The analysis of usage logs in residential Intranet of Things involves different types of entities (e.g., $Gateways$, $Devices$, $Context$ and $Activity$), which may be described by several attributes depending on the scope of the analysis. The more we zoom into device usage scenarios, the more attributes need to be considered. Rather than reducing the complexity of the n-ary usage logs analysis into a frequent binary pattern and association rule mining problem by converting n-ary to binary logs (e.g., $Context \times Activity$, $Context \times Device$) [2], [4], [5], [12]–[14], we rely on recent advances in data mining over n-ary relations (e.g., $Device \times Context \times Activity$) [6], [7]. Our choice is motivated by the need to employ the same general purpose mining algorithms for serving different analytical use cases.

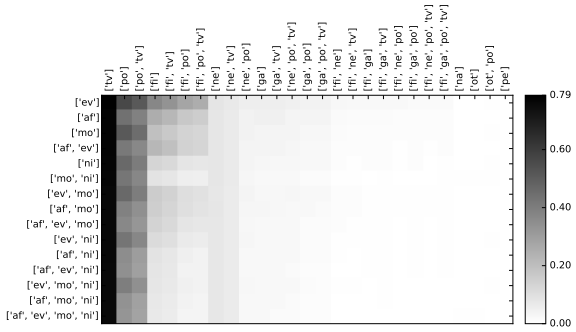
In general, n-ary usage logs are defined as subsets of the Cartesian Product of a number of finite and disjoint attribute domains: $U \subseteq D^1 \times ... \times D^n$. For example in the usage logs of Tables IV, V we have considered the domains Day ({(5/3/2014), (6/3/2014),...}), $Period$ ({morning, afternoon, evening, night}), $Quality$ ({high, medium, low}) and $Activity$ ({high, medium, low, idle}). Unlike extracting frequent co-occurring elements from the same domain (D^i), in our setting we are mining frequent sets of n-tuples ($D^1 \times ... \times D^n$) in U (called n-sets). For instance, we are interested in how regularly device $d30$ has been used close to the gateway with medium activity ([Quality:'high',Activity:'medium']) or what other

devices are exhibiting this usage pattern during evenings ([Period:'evening', Quality:'high', Activity:'medium']).

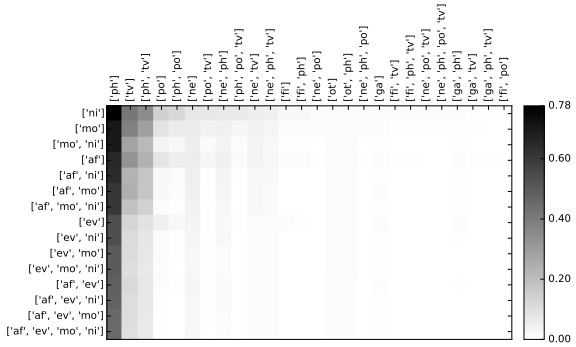
More formally, we are interested in extracting patterns under the form of *closed* n-sets [7]: $H = [X^1, \ldots, X^n]$ such that $X^i \subseteq D^i$. H is called a closed n-set iff (a) all elements of each set X^i are in relation with all the other elements of the other sets in U, and (b) X^i sets cannot be enlarged without violating condition (a). A notable characteristic of the Data-Peeler algorithm [7] we employ for extracting closed n-sets, is that it enables us to consider user-defined (anti-monotonic) constraints on the relevance of the mined n-sets. We may ask for patterns with a minimal number of elements in some domains (e.g. with at least 2 devices) and/or patterns covering at least a given number of tuples in U. N-ary association rules are based on the notion of *association* which boils down to closed n-sets defined on subsets of the original attribute domains. For example, $[Day{:}\{'06/03/2014'\}, Quality{:}\{'high','medium','low'\}, Activity{:}\{'medium','idle'\}]$ is an association on the support domain $Day \times Quality \times Activity$ while $[Day{:}\{'06/03/2014'\}, Activity{:}\{'medium'\}]$ is an association on $Day \times Activity$ support domain. The *frequency* (support) of an association is the subset of the support domain used to count the association occurrences on the remaining domains on which it is defined. Going back to our example, the association $[Quality{:}\{'high','medium','low'\}, Activity{:}\{'medium'\}]$ has as support relation Day and its frequency in the example usage log of device $d30$ is 1 (the three tuples only for $06/03/2014$). For formal definitions of a *rule* and *frequency* readers are referred to [6].

While frequency indicates the statistical significance of a rule (i.e., the joint probability $P(X, Y)$), its strength is measured by the confidence $\frac{|s(X \sqcup Y)|}{|s(X)|}$) (i.e. the conditional probability of $P(Y|X)$). In binary rules both $s(X \sqcup Y)$ and $s(X)$ are sets from the same domain. Since in an n-ary setting this is not always the case, in [6] authors introduced two alternative confidence metrics to cope with this issue. The form of n-ary association rules of our analysis does not fall into this tricky case and thus standard confidence metric is sufficient. We next detail the frequent n-ary patterns and rules we extract from the device usage logs either across or within the homes of our deployment.

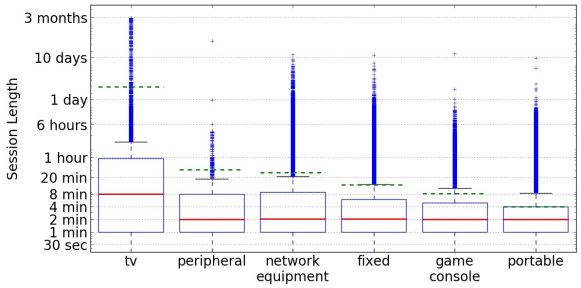
We first study device usage patterns across all the homes of our deployment. We seek to identify, *what types of devices are co-used more frequently* and *at what time periods in day?* Our analysis requires to extract n-ary associations from device usage logs defined over four domains: $Gateway$, Day, $Device$, $Period$. Since we are interested in frequent daily patterns across gateways, Gateway and Day are the support domains. For example, the support $s(\{'fi','po'\} \times \{'af','ev'\})$ reflects the number of days across all gateways, that fixed ('fi') and portable ('po') devices in a home were co-connected at both afternoon ('af') and evening ('ev').



(a) Co-connected device patterns in time.



(b) Co-idle device patterns in time.



(c) Connectivity session duration distribution per device kind.

Figure 2. Device usage patterns across all homes of our deployment.

Any subset of this n-set (e.g., $s(\{'po'\} \times \{'af','ev'\})$) has support greater or equal than its superset.

We first observe that home residents rarely disconnect their TVs (STBs and other TV kinds) as indicated by the high support $s(\{'tv'\} \times \{'af','ev','mo','ni'\}) = 0.755$. A similar behavior has been also exhibited by phones that have been excluded from our subsequent analysis. Portable ($'po'$) are more often connected than fixed ($'fi'$) devices. Finally, network equipment ($'ne'$), game consoles ($'ga'$), nas ($'na'$), peripherals ($'pe'$) and other ($'ot'$) devices are sparsely used in our homes. The co-connectivity supports of such unpopular devices (cf. Table II) are naturally very low (< 0.08).

Figure 2(a) shows at evening and night all devices exhibit the maximum and minimum connectivity, respectively. TVs and network equipment are connected with almost the same frequency across all periods, as indicated by the similar supports of the columns $'tv'$ and $'ne'$. For the remaining device

kinds, the connectivity during the evening can be up to 3 times higher than that of the night. Depending on the device, connectivity can be more frequent in the afternoon than in the morning and vice versa. For example, fixed devices are more often connected in the afternoon than the morning ($s(\{'fi'\} \times \{'af'\}) = 0.25$ and $s(\{'fi'\} \times \{'mo'\}) = 0.19$, while the opposite holds for portable devices ($s(\{'po'\} \times \{'mo'\}) = 0.51$ and $s(\{'po'\} \times \{'af'\} = 0.44$).

We finally study device co-connectivity patterns in Figure 2(a). Obviously, all device kinds are co-connected with STBs since every home has at least one STB, which is connected almost every day. Portable and fixed devices are frequently co-connected as well, as the support of co-connected fixed devices ($s(\{'fi'\} \times \{any\_period\})$) is similar (varies between 0.03 and 0.09) to the one of co-connected fixed and portable devices ($s(\{'fi','po'\} \times \{any\_period\})$. Network equipment is mostly co-used with fixed and portable devices (and STBs), as shown in Figure 2(a). At particular time periods such as evening, network equipment is heavily co-connected with portable or fixed devices; this is because the support $s(\{'ne'\} \times \{'ev'\})$ and $(s(\{'ne','po'\} \times \{'ev'\}) + s(\{'ne','fi'\} \times \{'ev'\}) - s(\{'fi','ne','po'\} \times \{'ev'\}))$ is small (0.02). This is expected, since home devices are often connected through network extenders. Game consoles are also co-connected with portable and fixed devices.

To better ground our device co-connectivity patterns, we analyze the *duration of the connectivity sessions* of different devices. Figure 2(c) shows the box plot of all the session durations distribution in our dataset, where the dashed lines are the average durations. We omit from the plot the $other$ and nas devices, which have only one session during our monitoring period. As expected, the longest session belongs to a STB and lasts for the whole monitoring period (89 days). Interestingly, we observe TV sessions to last a few seconds, with the median to be $\sim$8 minutes. This is because the sessions under the TV category belong to either STBs or other TV devices (e.g., AppleTVs), which have shorter sessions. Specifically, the average session duration for STBs, and other TV types are 11.7 days and 1 hour, respectively. Despite the high connectivity session variations among the remaining devices, the median session durations are similar ($\sim$2 minutes) for all device kinds. This can be attributed to the fact that more than 99.5% of the connectivity sessions use WiFi, which gets disconnected after small idle periods.

We next investigate, *the activity level exhibited by connected devices*. Different from our analysis of co-connected devices, we now consider 68 out of the 201 gateways, where the activity of Ethernet devices can be reliably estimated (cf. Section II). Thus, the number of n-tuples determining the frequency of extracted patterns is 68×89. We first differentiate idle, from low, medium, high activity levels. Then we construct the usage logs including all instances when devices in a home are idle during a certain time period. A connected device is considered to be idle during a time period, if all the gateway reports during this period show the device to be in idle activity level. Two (or more) devices are co-idle at a certain time period if the above condition holds, and they overlap with the period under analysis. Figure 2(b) illustrates the heat map of the support of *co-idle* devices at different periods, e.g., the support $s(\{'fi','po'\} \times \{'af','ev'\})$ is the number of days across all the gateways, that fixed and portable devices were idle, at both afternoon and evening. For illustration purposes, we sort the heat map cells based on decreasing support order and omit the patterns with very low support ($\leqslant 0.003$).

Our analysis shows the devices with the longest idle times to be the IP phones and TVs, whose maximum support is 0.78 and 0.41, respectively. This behavior implies that TVs and phones are frequently connected, but rarely used. Portables (support ranges from 0.004 and 0.1) show longer idle times than fixed devices (support ranges from almost zero to 0.02). Network equipment shows relatively long idle times with supports between 0.02 and 0.08. Finally, game consoles are rarely observed idle (supports from 0.003 to 0.008) compared with the other device kinds. Note that peripherals devices are absent from Figure 2(b). This can be attributed to the fact that residents often connect peripherals (such as printers) to use them directly, so we do not observe periods where such devices are connected but idle. There were no nas devices in the 68 homes of our study. Figure 2(b) shows that for all devices apart from game consoles, the highest supports in decreasing order appear at night, morning, afternoon and evening. On the contrary, the game consoles present the highest supports in reverse order (evening, afternoon, morning, night). This suggests that when residents in our deployment homes play video games, they usually don't use other devices. As expected, frequently connected but idle devices, are often co-idle with other devices. For example, the support of idle phones $s(\{'ph'\})$ is similar to the support of co-idle phones and portable devices $s(\{'ph','po'\})$, particularly during night. Other device kinds appear to be co-idle as well. For example, network equipment is co-idle with portable devices, which is expected since the traffic typically routes through the networking device to the gateway.

The $Quality$ dimension, which solely applies to wireless devices, was not considered so far, because 91% of our gateway reports for wireless devices are classified in the high quality category. The remaining 8.5% and 0.5% of records are classified in the medium and low categories, respectively.

Different from the ISP-wide analytics presented so far, we next perform an in-depth analysis of individual homes aiming to support home-automation use cases. We seek to identify, *what types of devices are co-used more frequently*, *at what time periods in day* and *what is their activity level?* We select as case study a home with 17 resident and 10 guest devices of various kinds and vendors (see Table VI) that in their majority use WiFi to connect to the gateway.

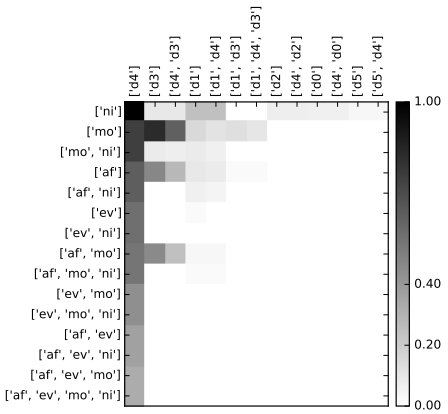


Figure 3. Co-idle device temporal patterns for our case study home.

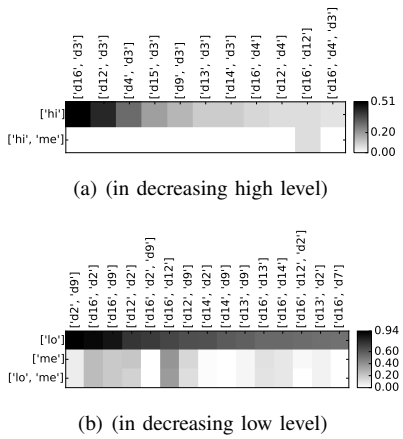


Figure 4. Co-active patterns during evening for case study home.

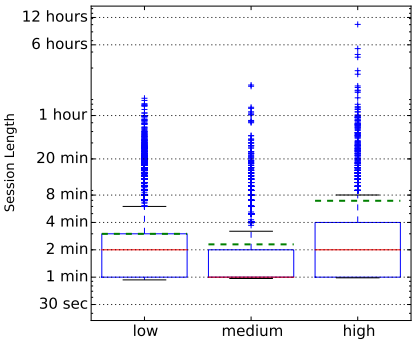


Figure 5. Activity session durations per activity level.

Table VI
OVERVIEW OF DEVICES IN OUR CASE STUDY HOME.

Device ID	Interface	Device Kind	Vendor	Number of Days Connected
$d0$	WiFi	portable	AMPAK Technology	69
$d1$	Ethernet	peripheral	Fujitsu	46
$d2$	WiFi	portable (iPhone)	Apple	89
$d3$	Ethernet	IPTV		87
$d4$		IP phone		89
$d5$	WiFi	data	Unknown	32
$d6$	WiFi	portable (iPhone)	Apple	68
$d7$	WiFi	portable	Samsung	74
$d8$	WiFi	portable	Samsung	56
$d9$	WiFi	portable	Sony	89
$d10$	WiFi	data	Azurewave Technologies	15
$d11$	WiFi	portable	Samsung	50
$d12$	WiFi	fixed	Hon Hai Precision	88
$d13$	WiFi	fixed	Intel	81
$d14$	WiFi	fixed (laptop)	GVC Corporation	68
$d15$	WiFi	fixed	Hon Hai Precision	46
$d16$	WiFi	media player	Liteon Technology	89

The number of days that a device appears to be connected varies from 15 to the total 89 days of interest.

Figure 3 shows the heat map plot of the support of *co-idle* devices at different time periods, where the support domain is the Day. As expected, Phone ($d4$) and STB ($d3$) have the longest idle times. Besides these devices, the peripheral $d1$ and the data device $d5$ exhibit the longest and the shortest idle times, respectively. Interestingly, we do not observe any strong correlation between the number of connected days and the idle times. For example, the peripheral $d1$, is connected less days than the portables $d0$ and $d2$. For all the devices, the highest supports in decreasing order appear at night, morning, afternoon and evening, (cf. Figure 3).

Two (or more) devices are considered to be co-active with activity level x at a certain time period, if they have overlapping activity intervals at level x, during this period. In the heat map of Figure 4(a) we present the co-active devices along with their activity levels, in decreasing support of $'high'$ activity level. The support domain is the Day, and we filter out co-active devices with less than 4 days. In the sequel, we focus on high, medium activity, in the evenings. As we can observe in Figure 4(a), almost all co-active device sets include STB ($d3$) and phone ($d4$). The highest support is $s(\{'d16','d3'\} \times \{'hi'\}) = 0.51$. This means that, the STB and the media player show high co-activity level, for 45 out of the 89 days of monitoring period. The large co-activity supports for phone and STB is justified from the fact that, we define only idle and high activity levels for those devices. Apart from these frequently used devices, a media player and a fixed device ([’d16’, ’d12’]), exhibit a high activity level with 4 days of support.

Figure 4(b) depicts device co-usage at *low* activity, in decreasing support (the support domain is again the Day). Due to space restrictions, we present only the 15 highest support device sets. We observe two key differences compared with the high activity level patterns of Figure 4(a). First the supports are overall higher, with the peak support $s(\{'d2','d9'\} \times \{'lo'\}) = 0.94$ (83 days). This implies that co-used devices operate more at low than high activity level, in our case study home. Second, the device sets do not include STB and phone because low activity does not apply to them. Device sets include portable, fixed, media player devices and their combinations. We finally observe co-active devices at different activity levels with the peak support to be 65 days ($s(\{'d16','d12'\} \times \{'low','medium'\}) = 0.73$). We observe the activity with the peak supports to be 0.9, 0.67, 0.51, 0.4 for evening, afternoon, morning and night. Although, there are common co-activity patterns across periods, residents co-use different devices at different times, which calls for different home profiling among time periods.

Figure 5 shows the box plot of activity session lengths distribution (the dashed line is the average length). We have excluded phone and STB devices which have only idle and high activity levels. We observe short sessions with median values of 1 minute for medium and 2 minutes for low and high activity levels, respectively. The highest average session duration never exceeds 8 minutes and is observed for high activity level. Short activity sessions imply longer idle times, which can be exploited in home automation scenarios. However, we still observe variations in session durations with the peak of over 10 hours for high activity level. During these long sessions, the residents could be constantly downloading data (e.g., Torrents).

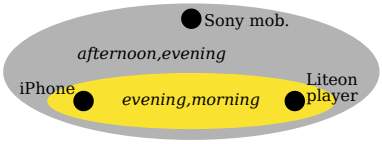


Figure 6. Association rules graph, for our case study.

Figure 6, shows the n-sets of our case study home using hypergraph diagram of rules with support and confidence of at least 0.5. We focus only on the temporal device co-usage, without considering the activity level. The always-connected IPTV and phone devices are excluded from our analysis. Each vertex represents a set of devices that appear either in the head or the body of an extracted rule. A hyper-edge represents all possible temporal periods for which devices in n-sets have been observed co-active. Note that hyper-edges are undirected since the confidences of the extracted rules which have the set of devices represented by a vertex as head or body, are similar. However, there can be exceptions. Specifically, media player and fixed devices show the largest co-activity support at high activity level, in the morning or $s(\{'d16','d12'\}\times\{'hi'\})$. Other exceptions are the portables $d2$, $d9$ which appear co-active at all periods, except for night.

The hypergraph of our case study, includes a media player (d16) and two portable devices (iPhone d2, Sony d9). Rules' supports range from 0.5 to 0.63 and the confidences from 0.5 to 0.78. More precisely, the activity of the media player at a certain time period is correlated (with confidence from 0.5 to 0.78) with the iPhone usage at another period (and vice versa), for evening, morning and afternoon. The rule with the highest confidence (0.78) shows that if the media player is active in the afternoon, then the iPhone will be used in the evening ($d16, afternoon \rightarrow d2, evening$). The largest hyper-edge represents the co-activity of the media player, Sony portable (with confidence from 0.6 to 0.74) and iPhone, Sony portable (with confidence from 0.62 to 0.73), during afternoon and evening.

V. System Architecture

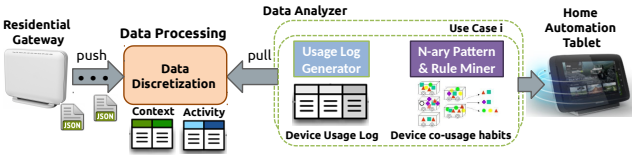


Figure 7. Usage pattern mining system architecture.

Our goal is to perform gateway logs analysis entirely *in-house* using commodity hardware, without the ISP support. Consequently, residents' personal data leakage can be mitigated, since the usage data of their devices can be fully stored and processed inside the residential Intranet. An overview of our system is shown in Figure 7. We observe that our data analyzer processes one day and one week logs in a few seconds (30) and minutes (3), respectively. The vast majority of processing is allocated to the log analyzer module, since PINARD runs in only 3 seconds even when we process 3-month logs. Note that PINARD running time increases with dimensionality; if we increase the problem dimensions from 3 to 4, PINARD running time increases from 3 seconds to approximately a minute. The low processing overhead of our system indicates that our analytics can be produced by home dedicated devices.

VI. Summary

In this paper, we have explored n-ary association rules for mining device usage patterns in residential Intranet of Things. These patterns provide valuable insights to uncover daily practices of residents without employing intrusive home sensors. We plan to extend our system in two key ways. First, we could enhance the dimensions of our analytics with application-level gateway feedback, or reports from other types of devices as home sensors. Second, our proposed use cases mainly apply to devices connected to the gateway. We foresee a unified Intranet of Things architecture where "things" can talk also to each other, thus enabling applications over an even more diverse set of devices.

Acknowledgements

This work was partially funded by the European ICT FP7 User Centric Networking project (grant no. 611001) and by EU IRSES project SemData (ID: 612551).

References

[1] L. Yao, Q. Z. Sheng, A. H. H. Ngu, and X. Li, "Things of interest recommendation by leveraging heterogeneous relations in the internet of things," in *ACM TOIT*, 2015.
[2] H. Ma, H. Cao, Q. Yang, E. Chen, and J. Tian, "A habit mining approach for discovering similar mobile users," in *ACM WWW'12*.
[3] X. Li, H. Cao, E. Chen, H. Xiong, and J. Tian, "Bp-growth: Searching strategies for efficient behavior pattern mining," in *IEEE MDM'12*.
[4] V. Srinivasan, S. Moghaddam, A. Mukherji, K. K. Rachuri, C. Xu, and E. M. Tapia, "Mobileminer: Mining your frequent behavior patterns on your phone," in *ACM UbiComp*, 2014.
[5] S. Rollins and N. Banerjee, "Using rule mining to understand appliance energy consumption patterns," in *IEEE PerCom'14*.
[6] K.-N. Nguyen, L. Cerf, M. Plantevit, and J.-F. Boulicaut, "Multidimensional association rules in boolean tensors." in *SIAM SDM'11*.
[7] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut, "Data-Peeler: Constraint-based Closed Pattern Mining in n-ary Relations," in *SIAM SDM'08*.
[8] G. Poghosyan, I. Pefkianakis, P. Le Guyadec, and V. Christophides, "Mining usage patterns in residential intranet of things," *Procedia Computer Science*, 2016.
[9] G. Poghosyan, "Device analytics in home networks," in *EPFL Master's Thesis*, August 2014.
[10] I. Pefkianakis *et al.*, "Characterizing home wireless performance: The gateway view." in *IEEE INFOCOM'15*.
[11] What speeds do I need for Skype, Netflix, video games, etc.? https://support.speedtest.net/.
[12] H. Cao, T. Bao, Q. Yang, E. Chen, and J. Tian, "An effective approach for mining mobile user habits," in *ACM CIKM '10*.
[13] S. Nath, "Ace: Exploiting correlation for energy-efficient and continuous context sensing," in *ACM MobiSys '12*.
[14] L. Ong, M. Bergés, and H. Y. Noh, "Exploring sequential and association rule mining for pattern-based energy demand characterization," in *ACM BuildSys'13*.