An exploration of user recognition on domestic networks using NetFlow records

Anthony Brown
School of Computer Science,
University of Nottingham, UK
psxab@nottingham.ac.uk

Tom Rodden
School of Computer Science,
University of Nottingham, UK
Tom.Rodden@nottingham.ac.uk

Richard Mortier
School of Computer Science,
University of Nottingham, UK
Richard.Mortier@nottingham.ac.uk

UbiComp '14, September 13 - 17 2014, Seattle, WA, USA Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-3047-3/14/09$15.00.
http://dx.doi.org/10.1145/2638728.2641560

Abstract

In this paper, we describe HomeNetViewer, a system for collecting, visualising and annotating domestic network NetFlow records from a domestic network gateway. HomeNetViewer is designed to collect ground truth data which, enables the linking of users to low level network traffic. We present our first annotated dataset from a real household in the UK and the results of our preliminary work to build a user identification system. Our initial classifier achieves a true-positive rate of 64% with false-positive rate of 28% when compared to the ground truth annotations. This work attempts to address the lack of transparency and accountability within the domestic network infrastructure by identifying the user behind the device.

Author Keywords

Domestic networks, user recognition, NetFlow

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction

Home networks are now common place and have become an unremarkable feature of everyday life [3]. These networks are used to access a vast array of content and services. However, despite their widespread deployment, the management, configuration and control of this resource imposes a significant technical burden. Domestic networks have been criticised for their lack of transparency and local accountability [4]. This lack of transparency obstructs the process of policing usage and troubleshooting problems on the network.

Policing and defining acceptable behaviour on domestic networks is normally accomplished via locally negotiated rules and norms. The lack of transparency and accountability hinders the oversight of this process [9]. In this work, we address the lack of transparency by developing a system that identifies users interacting with networked devices from the traffic they generate. Figure 1 provides an overview of the actors in the system for interacting with remote applications over a domestic network and the sources of data that we collect and analyse.

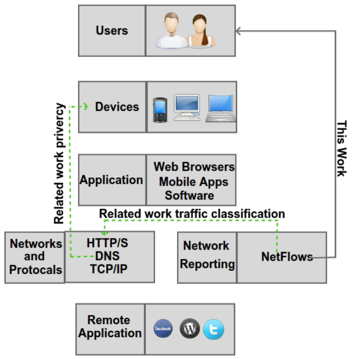


Figure 1: An overview of the actors in the system for interacting with remote applications over a domestic network (related work shown with a dotted line).

Interventions by the HCI community have shown that improving transparency enhances the user's ability to manage the domestic network. For example, Chetty et al explore the issues around household bandwidth management and the consequences of improving its visibility within the household [1]. Their participants indicated "frustration, annoyance and general unhappiness about the quality of service" when tasks they were performing were affected. For example, media being jittery or call quality using voice over IP being degraded". They explored participant's theories and reasoning around these reported problems before and after the introduction of their technical probe (Home Watcher). They found that exposing bandwidth usage, made the concept of bandwidth more accessible to the household. Allowing bandwidth to be discussed and reasoned about just like any other household resource. This enabled users to consider the fair sharing and usage of the network within the normal politics of daily life.

Chetty et al, considers the reporting of bandwidth usage at the device level, but householders develop rules around the users of the household as their main point of reference. For example, a rule generated might be: Alice should not access Facebook before she has done her homework. A simplistic solution to this problem is to assume a one-to-one relationship between users and devices. However, this is unlikely to hold as most devices are seen as a shared resource with any family member allowed to have access within the framework of locally negotiated rules [7].

Our goal is to explore the possibilities for detecting the unique behaviours and traffic characteristics of users within their NetFlow traces. This would enable the automatic assignment of traffic to specific users, enabling reporting and control of the network using the "user behind the device" as the focus rather than the device itself. Our work is related to the field of traffic classification in core and backbone networks [6], but has two main differences. First, our target classification is different. Core and backbone network traffic classification focuses on assigning traffic to classes of use to trained network professionals, often based on network protocols, which are of limited use to domestic users. Second, our viewpoint is different; we capture flows for a single household rather than the flows of numerous unconnected users.

Our work is also related to that of Herrmann et al. who study the re-identification of users from the point of view of a passive attacker, with the ability to monitor Domain Name System (DNS) requests [5]. They use data containing DNS records for more than 3600 users from the student living accommodation at their university, collected over a two month time-frame. Herrmann et al. assume that a single IP address in the dataset is linked to one user, which is valid for their data as each user is assigned a static IP for their room. However, they note that some noise is present in their data due to students visiting friends using devices they do not own. They build profiles for each user over a 24-hour period and test a number of classifiers and filtering techniques (sub-linear transform, normalisation, inverse document frequency). Their best reported classifier can re-identify 85.4% of DNS requests to users.

The first challenge we address is how to collect a valid ground truth dataset linking NetFlow data to users within the domestic environment. To this end, we designed HomeNetViewer to collect, visualise and annotate network data within the home. We describe HomeNetViewer in the next section. Then, we report some results from our preliminary investigations conducted with NetFlow traces collected in a previous study, annotated retrospectively using HomeNetViewer by one occupant of the household.

System Design

Our tool HomeNetViewer has been designed to achieve three goals. First, we wish to allow domestic users an improved view into to their household's network usage, beyond bandwidth as explored in [1]. Secondly, we wished to collect information on the types and ownership of devices on the network to start to unpack device sharing within the home. Thirdly, we wish to collect the ground truth from users regarding who was using a device at a particular time and what activities caused the traffic that they can see. This approach is similar to that recently taken by Costanza in their exploration of domestic energy usage [2]. Figure 2 shows an overview of the system described below.

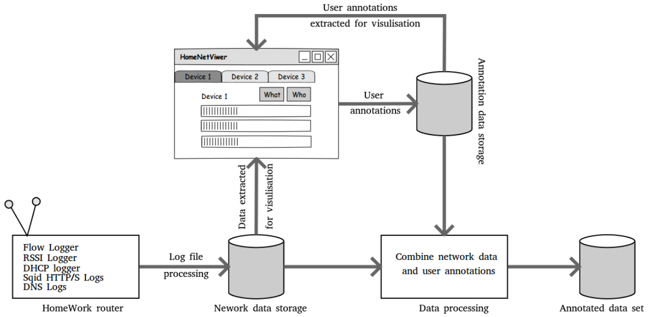


Figure 2: An system overview of HomeNetViewer data collection and annotation platform.

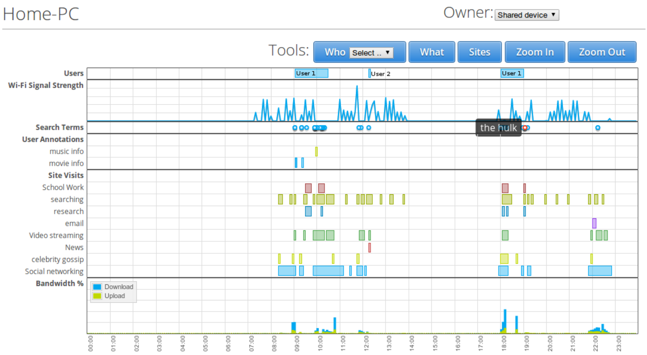
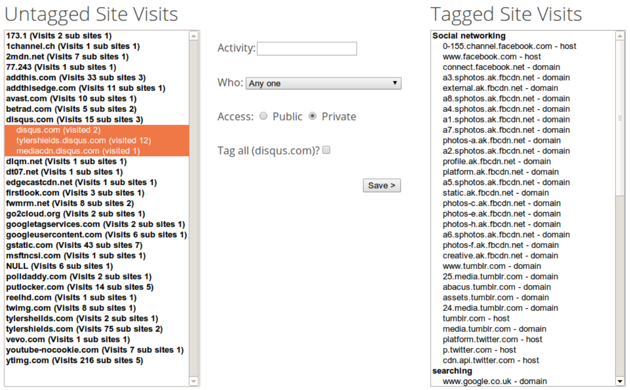


Figure 3: Final annotation interface for a single device.

Our system was built around the Homework router platform [8] which, provided reliable routing and data logging functionality. Early attempts at visualizing the data quickly became cluttered and overloaded with technical data like port number, IP address and MAC address. It was clear that we needed to reduce the complexity of the data to build a usable visualization. We noted that over 53% of traffic was on port 80 with another 6% on port 443. These correspond to Hypertext Transfer Protocol (HTTP) and HTTP Secure (HTTPS) respectively. The other 41% of traffic was distributed widely over 57 thousand other ports. For example, DNS (port 53) accounted for just 0.7% of flows. For this reason it was decided to focus the interface on exploring HTTP traffic rather than traffic on all ports. HTTP also has the advantage that users are somewhat familiar with its main identifying feature, the Uniform Resource Locator (URL) and the host/domain names contained within them. HTTPS was not included as the encryption prevented the extraction of the required data from the packets as they passed through the router. It was also observed that non-HTTP activity on the network, for example, Skype and Bittorrent often left clues in the HTTP logs that they were in use. These clues take the form of either visits to related websites (e.g. torrent search engines) or accessing interface components and/or software updates over HTTP (e.g. Skypes interface is an embedded web page).



Untagged Site Visits

173.1 (Visits 2 sub sites 1)
1channel.ch (Visits 1 sub sites 1)
2mdn.net (Visits 7 sub sites 1)
77.243 (Visits 1 sub sites 1)
addthis.com (Visits 33 sub sites 3)
addthisedge.com (Visits 11 sub sites 1)
avast.com (Visits 10 sub sites 1)
betrad.com (Visits 5 sub sites 2)
disqus.com (Visits 15 sub sites 3)
disqus.com (visited 2)
tylershields.disqus.com (visited 12)
mediacdn.disqus.com (visited 1)
dlqm.net (Visits 1 sub sites 1)
dt07.net (Visits 1 sub sites 1)
edgecastcdn.net (Visits 1 sub sites 1)
firstlook.com (Visits 3 sub sites 1)
fwmrm.net (Visits 8 sub sites 2)
go2cloud.org (Visits 2 sub sites 1)
googletagservices.com (Visits 2 sub sites 1)
googleusercontent.com (Visits 6 sub sites 1)
gstatic.com (Visits 43 sub sites 7)
msftncsi.com (Visits 1 sub sites 1)
NULL (Visits 6 sub sites 1)
polldaddy.com (Visits 2 sub sites 1)
putlocker.com (Visits 14 sub sites 5)
reelhd.com (Visits 1 sub sites 1)
twimg.com (Visits 8 sub sites 1)
tylersheilds.com (Visits 2 sub sites 1)
tylershields.com (Visits 75 sub sites 2)
vevo.com (Visits 1 sub sites 1)
youtube-nocookie.com (Visits 7 sub sites 1)
ytimg.com (Visits 216 sub sites 5)

Activity:

Who: Any one ▾

Access: ○ Public ● Private

Tag all (disqus.com)? □

Save >

Tagged Site Visits

Social networking
0-155.channel.facebook.com - host
www.facebook.com - host
connect.facebook.net - domain
a3.sphotos.ak.fbcdn.net - domain
external.ak.fbcdn.net - domain
a8.sphotos.ak.fbcdn.net - domain
a4.sphotos.ak.fbcdn.net - domain
a1.sphotos.ak.fbcdn.net - domain
a7.sphotos.ak.fbcdn.net - domain
photos-a.ak.fbcdn.net - domain
a2.sphotos.ak.fbcdn.net - domain
profile.ak.fbcdn.net - domain
platform.ak.fbcdn.net - domain
a5.sphotos.ak.fbcdn.net - domain
static.ak.fbcdn.net - domain
photos-c.ak.fbcdn.net - domain
photos-e.ak.fbcdn.net - domain
photos-h.ak.fbcdn.net - domain
a6.sphotos.ak.fbcdn.net - domain
photos-f.ak.fbcdn.net - domain
creative.ak.fbcdn.net - domain
www.tumblr.com - domain
25.media.tumblr.com - domain
abacus.tumblr.com - domain
assets.tumblr.com - domain
24.media.tumblr.com - domain
tumblr.com - host
media.tumblr.com - domain
platform.twitter.com - host
p.twitter.com - host
cdn.api.twitter.com - host
searching
www.google.co.uk - domain

Figure 4: Host/domain annotation interface

We constructed an interface to visualise the data and collect annotations, including; device ownership, device usage and user activity ground truth annotations. It consisted of a set of time-lines one for each device connected to the network. Each set of time-lines was labelled with the device hostname where available, or the device MAC address if no hostname was captured. The devices could be renamed by the users to a more human friendly name if required. It was also possible to assign an owner to each of the devices in the interface. The timelines represented Wi-Fi RSSI bandwidth (inbound and outbound) and annotated HTTP visits along with annotations added by the users (see Figure 3). Any section of the time line is selectable and the users can apply a number of annotation types to the selected time range. These are:

Host/Domain activity tags: a nested list of all the hosts that were contacted in the selected time frame grouped by domain name (see Figure 4). Where users were instructed to examine the list to find domain/hostnames that they recognised. They were then asked to add "tags", relating to the activity most likely to have caused them to visit that domain/host. Tags are entered as free text but a list of previously entered tags is displayed to improve consistency and reduce duplication. **Host/Domain user tags:** it was noted that some hosts/domains have a strong link to an individual in the household so this data was also collected. **Search activity tags:** View and assign search terms to activities. **Search user tags:** View and assign search terms to users. **Free user:** Annotate the time-lines with who was using a device at a given time **Free activity:** Annotate the time-lines with the activity being performed at a given time.

Experimental set-up

Our initial experiment was developed using five months of traffic captured within a single-family household in mid 2012 using the Homework router platform [8]. The household consisted of two adults of working age with two children at secondary school. These traces were collected from 24 devices. A breakdown of these can be seen in table 1. The traces contained:

- 32.6 million IP flow records, 15.6 million in and 17.0 million out the imbalance is due to failed connections.
- 17.1 million HTTP requests.
- 32 million RSSI readings.
- 23 thousand DHCP lease requests.

Using the interface described above we instructed the participant to annotate three weeks of network traces from their household. This generated 547 annotations. Table 2 shows which type of annotations were collected to be used as the ground truth for our classifiers.

Device Type	Num devices	Num flows (Millions)
Mobile phone	4	0.4
Desktop PC	2	11.3
Laptops	4	9.5
Tablets	6	7.1
Music players	2	0.1
Games consoles	1	0.8
Unknown	5	2.7

Table 1: breakdown of the device in our dataset

Annotation Type	Number of annotations
host/domain activity	107
host/domain user	28
Search activity	195
Search user	157
free user	59
free activity	0

Table 2: Number of each type of annotation collected

Training the classifiers

Using the generated annotations, it has been possible to perform a preliminary analysis to ascertain the feasibility of using these annotations to assign traffic to users. The annotations allowed the assignment of 99 thousand of the 2.9 million flows (33%) recorded on port 80 and 443 to users, creating our ground truth dataset. We chose to examine inbound and outbound flows separately, using inbound flows to look for variations in the behaviour of end hosts and outbound flows to look for user behaviours. This dataset allowed us to formulate a simple supervised machine learning classification problem. We evaluated several classifiers (Naive Bayes, k-Nearest Neighbors, Random forest) and the tree based random forest classifier provided the best initial results. The dataset was split into to two sections the first 4000 flows per user were designated as the training set and the remaining flows were used to verify the resulting classifier. The classifiers were trained on eight features extracted from the flow records these were; number of packets, flow duration, time since the last flow to same hosts, packet rate, average packet bytes, source address or destination address and destination port or source port. The next section presents the results obtained using these classifiers.

Experimental results

The classifiers were trained as set out in the previous section and then used to classify the verification dataset. Table 3 shows the accuracies achieved for each user against the ground truth verification dataset at the NetFlow level. Combining the in flow and out flow classifiers using a simple mean improved the overall accuracy for all but one user who showed a slight decrease in classification accuracy. After the classifiers were trained the combined classifier was used on the original unfiltered dataset of 8.7 million flows to generate a set of automatic user annotations. To generate the annotations, we set a class membership cut-off of 25%. This cut-of, discarded all results with a class membership probability less than this threshold. The classifier assigned 41 thousand (35%) flows to users.

After the flows were assigned to each user by the classifier, the flows were grouped by time and new annotations generated for each user per device. We then applied a filter to remove overlapping annotations between the users, keeping the annotation with the highest class membership probability in each overlapping set. After filtering 259 annotations remained These were then compared back to the ground truth. Compared to the ground truth annotations, we achieved a true-positive rate of 64% and a false-positive rate of 28% across all users and devices. Figure 5 shows the generated annotations and the ground truth for two devices over a 14-day period. The top section of each graph gives an indication of network activity in flows per minute; the bottom section shows the ground truth annotations in blue, with the generated annotation red. The topmost figure has areas of true-positive (green boxes) and false-positive (red boxes) highlighted for illustration.

	In flow (%)	Out flow (%)	Combined (%)
user1	46.6	49.8	56.8
user2	37.7	37.9	37.3
user3	51.0	52.0	60.0
user4	78.9	77.8	88.0

Table 3: Achieved classification accuracy on verification dataset at NetFlow level

Discussion

In some cases, the flow classifiers are able to attain a reasonable level of accuracy, for example, they achieve an 88% accuracy for user4. However, for some users, the classifier achieves an accuracy only slightly above that of random guessing. The classifiers have particular difficulty distinguishing between user2 and user3. User2 and user3 were teenage girls of similar age. We believe this confusion is due to them having related interests and hence visiting a large number of similar websites. While we have shown for some users that this technique can provide reasonable results the accuracy of the classifiers needs to be significantly improved before it would be acceptable for deployment in a real system.

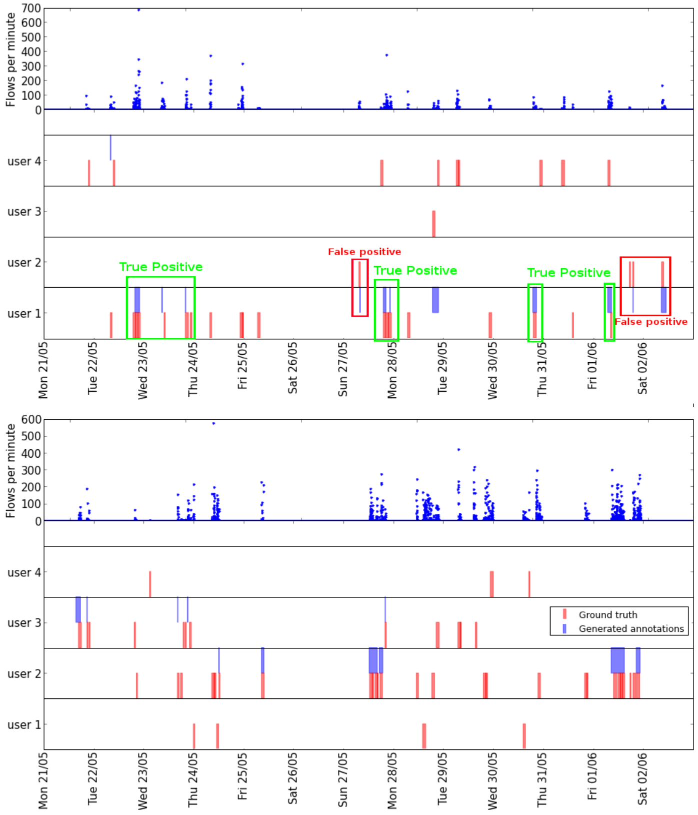


Figure 5: Generated annotations compared to ground truth for two devices over 14 days

The human annotations collected for this analysis occurred a significant time after the acquisition of the data. This passage of time affected the quality and density of our ground truth because it was not always apparent to our participant who caused the observed traffic. Hence, some sections of the traffic are un-annotated. These un-annotated sections can be seen in Figure 5 as periods of high traffic without any annotations. Within the un-annotated sections of traffic it is imposable to calculate the accuracy of the generated annotations as we do not have the required ground truth for comparison.

Several sources of noise have been identified within the data which, may have a impact on the accuracy of the classification. One source of noise originates from the underlying structure of web based applications. A single-page visit will download many resources (html, images, js and css), generating many flows per page view. Even if the page is customized to a user a significant number of the resources related to the operation of the site will be identical between users. Therefore, only a small number of flows for a particular site visit will provide useful information to the classifiers. We also observed a large number of the generated annotations were at times when only automatic traffic was observed (e.g. windows update and ICloud backup). We attribute this to automatic traffic being present as noise in the classifier training set. Filtering out these common flows and removing automated traffic before classification may lead to a better result. How to achieve this and whether it will have any impact on accuracy is still an open question.

Our classifiers also worked across all devices on the network. This may have a bearing on accuracy as different devices generate distinctive traffic patterns for visits to the same web application. For example, a mobile optimised website may deliver different content to mobile and desktop clients. Hence, training one classifier per device may prove beneficial. Another approach to improve accuracy would be to change the way the classifiers operate. Currently, our classifier considers a single flow at a time. A better strategy may be to look at several flows at once, for example, grouping flows by time and using multiple flows as a single observation.

Conclusion and future work

We have presented our preliminary work exploring how to identify the user actively using a device on a domestic network from NetFlow records captured by the household gateway. We have described our system for collecting, visualizing and annotating NetFlow data in the home to generate a ground truth dataset. We also present results of our initial classification work on this data. Although the accuracy of our classifiers is disappointing for some users, it shows promise for others achieving an 88% accuracy for one of our test subjects. Compared to the ground truth annotations our system attained a true-positive rate of 64% and a false-positive rate of 28% showing potential but with room for improvement.

We wish to continue this work by exploring the possibilities for improving the accuracy outlined in the discussion and collecting further ground truth data from a number of different households to validate the method. We would also like to perform a longer study to examine the stability of the classifiers over time. If this method can be perfected it will enable new possibilities for communicating information about the domestic network infrastructure to users, improving transparency and accountability facilitating the local negotiation of rules and acceptable behaviours within the home.

Acknowledgements

This work was supported by the Doctoral Training Centre in Ubiquitous Computing, RCUK grant EP/G037574/1 and by Horizon Digital Economy Research, RCUK grant EP/G065802/1.

References

[1] Chetty, M., Banks, R., Harper, R., Regan, T., Sellen, A., Gkantsidis, C., Karagiannis, T., and Key, P. Who's hogging the bandwidth: the consequences of revealing the invisible in the home. In *SIGCHI 2010* (2010).

[2] Costanza, E., Ramchurn, S. D., and Jennings, N. R. Understanding domestic energy consumption through interactive visualisation: a field study.

[3] Crabtree, A., Mortier, R., Rodden, T., and Tolmie, P. Unremarkable networking: the home network as a part of everyday life. DIS '12, ACM (New York, NY, USA, 2012).

[4] Grinter, R. E., Edwards, W. K., Newman, M. W., and Ducheneaut, N. The work to make a home network work. In *ECSCW 2005* (2005), 469488.

[5] Herrmann, D., Banse, C., and Federrath, H. Behavior-based tracking: Exploiting characteristic patterns in DNS traffic. *Computers & Security 39, Part A* (Nov. 2013), 17–33.

[6] Kim, H., Claffy, K., Fomenkov, M., Barman, D., Faloutsos, M., and Lee, K. Internet traffic classification demystified: myths, caveats, and the best practices. CoNEXT '08, ACM (New York, NY, USA, 2008), 11:111:12.

[7] Salmon, B., Hady, F., and Melican, J. Learning to share: a study of sharing among home storage devices. Tech. rep., Technical Report CMU-PDL-07-107, Carnegie Mellon University Parallel Data Lab, 2007.

[8] Sventek, J., Koliousis, A., Sharma, O., Dulay, N., Pediaditakis, D., Sloman, M., Rodden, T., Lodge, T., Bedwell, B., Glover, K., and Mortier, R. An information plane architecture supporting home network management. In *2011 IFIP/IEEE International Symposium on Integrated Network Management (IM)* (2011), 1–8.

[9] Tolmie, P., Crabtree, A., Rodden, T., Greenhalgh, C., and Benford, S. Making the home network at home: Digital housekeeping ECSCW, Springer/Kluwer. In *ECCSW* (2007).