TRACTS -
TRaffic ACtivity Targeting uSers

Thomas Krauss, Vanessa Arndorfer, Aidan Barton, Alon Bandelac, Anthony Huynh,
Sophie Khan, Ha Nuel Lee, Christina Lin, Sara Malek, and Brian Stoneham

Ted and Kayrn Hume Center for National Security and Technology
Virginia Tech
Blacksburg, VA, 24061 USA
Email: {tkrauss,arndorvf,tonaid,alonb,tonybh97,sophfk1,hanuelee,christie,smalek,brians2}@vt.edu

Abstract—**The purpose of this research project was to investigate network traffic, with the intent to distinguish between user activities (web browsing, audio streaming and video streaming) based on Wi-Fi packet inter-arrival times. Data collection was performed on a semi-isolated Wi-Fi network within the Hume Center laboratory with Wireshark. The resulting data was then used to compute the optimal representative probability density functions for each activity. The Pareto distribution was found to be the best fit regardless of activity. It was found that the audio streaming and video streaming are distinct as shown by the Kullback-Leibler divergences; web browsing, however possesses qualities of both audio and video.**

I. Introduction

Over the past several decades network traffic has been growing rapidly. Of particular interest is the 802.11-based traffic. This wireless Wi-Fi has become the de-facto standard used in virtually every home, business, campus, and cell phone and represents an ever-more difficult challenge for network engineers to deliver the data rates demanded by users. To that end, significant research has been performed to accurately model the statistics of Wi-Fi frames as they are broadcast in order to optimize the network traffic and protocols to improve throughput.

The increased network traffic is reflected directly in both the usage level of consumers consumers are using their networks for more demanding tasks as well as in the raw number of network aware devices. It is estimated that roughly 50% of the world population has an Internet connection today up from 1% two decades ago[1]. The increased usage for sensitive activities including banking, communications, shopping, and political advocacy has driven the adoption of higher security connections including encrypted web (HTTPS) and virtual private networks (VPN). It has been recognized that VPN penetration rates have been increasing, especially in politically restrictive regimes like Saudi Arabia, China, and Russia[2]. A VPN is created by establishing a virtual point-to-point connection through the use of dedicated connections, virtual tunneling protocols, and traffic encryption. In most instances the source address and destination address of network traffic is obfuscated by the VPN. The source or destination of packets appears to be the VPN host with the true destination encapsulated in the encrypted traffic. This reduces, by design, the ability to track user activity by monitoring network traffic directly.

The objective of this study is to investigate the Wi-Fi inter-frame arrival time statistics to identify user network activity. The identification will include the broad class of activity — web browsing, video streaming, etc. — rather than specific destinations such as specific web sites being visited, but the assessment will be based only on inter-frame time statistics and not on source, destination, protocol, etc.

II. Related Work

Many researchers have taken on the challenge of modeling network traffic, analyzing Wi-Fi traffic, and classifying Wi-Fi traffic. The goal of most of this research was to optimize network throughput and generate graphs and statistics that help model the traffic. This work coincides with the first part of our research which captures statistics based on Internet activity: video streaming, audio streaming, or web browsing.

Modeling Wi-Fi Traffic: Research studying how to model Wi-Fi Traffic used statistical analysis to fit data from Wi-Fi traffic as accurately as possible. The majority of research noted that Wi-Fi traffic tends toward a Poisson distribution[3][4]. However in a 1995 paper, studying TELNET traffic, it was found that the Poisson model greatly deviated for many connection arrivals because the Poisson distribution does not take burstiness of network traffic into account[5]. This paper concluded that the Pareto distribution is a better description of Wi-Fi traffic and was backed by research at the University Sv. Kiril i Metodij[6]. Other papers also advocated modeling Wi-Fi traffic with an exponential distribution in accordance with calculations for the Hurst parameter which quantifies the correlation of time series.

Wi-Fi Traffic Analysis: Zhang looks at frame size and traffic direction in addition to inter-frame time to classify web activities to 90 percent accuracy[7]. This study focused on machine learning algorithms including Support Vector Machine and Neural Networks. In contrast, our study classifies different web activities using only statistical models of inter-frame times.

Classifying WiFi Traffic: The original motivation for conducting this research was inspired by research conducted at Hanyang University in Korea which fit inter-frame arrival time to various distributions including Pareto, Log-normal, and Weibull[8]. Their research traced WiFi across a campus whereas our research focuses on an isolated network. Another paper conducted at the University of Cambridge took a machine learning approach toward traffic classification[9]. Their work used machine learning and traffic statistics to identify Wi-Fi traffic across a network.

III. Data Collection

Data collection was performed on an semi-isolated Wi-Fi network within the Hume Center laboratory. This network is depicted in 1. The network consisted of a set of dedicated computers connected via Wi-Fi to a Linksys access point. No other connections or computers were using the Wi-Fi access point, however the access point was hard-wired to the Hume Center and Virginia Tech networks and, thereby, the Internet. The distance between computers and access point was under 3 m ensuring a very high signal-to-noise ratio and, therefore, use of the highest 802.11 data rate.

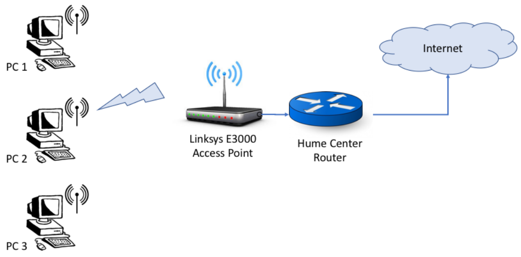

Fig. 1: Wireless Network Configuration

For this first stage, no VPN was established . Over a six week period, students were asked to perform specific network activities and record the network traffic using the Wireshark[10] network protocol analyzer. Specific activities included web browsing, video streaming, and audio streaming. No constraints were placed on the web locations nor on the content streamed, rather, relatively typical user behavior was captured. Sixteen datasets spread roughly equally across the three activities were collected. In addition, several collections made from other, uncontrolled, locations such as the universitys public library, dorm rooms, personal apartments, etc.

IV. Statistical Analysis

Before fitting any specific probability density functions, it was speculated that the data would follow a heavy-tailed distribution. Some known properties of Internet traffic such as the tendency to have a burst of frames at the beginning, or scale invariance fits closely into characteristics of heavy-tailed distributions.

Collected Wi-Fi frame data was processed via a combination of Python and Matlab scripts. The data was first processed via Python to extract only frames to or from a specified IP address and to compute the inter-frame times. Multiple probability density functions (PDF) were then fit using Matlabs “fitdist” function to the resulting inter-frame times to determine the most appropriate model. Figure 2 shows an example of one such fitting for a single collection of web browsing activity. As seen in the legend of Figure 2, the Pareto distribution is the best fit (in the mean-square-error sense) for this particular collection. A direct comparison of the best PDF and the distribution parameters for multiple activity types was undertaken. This comparison involved the computation of the best PDF for the particular activities. An example PDF comparison is shown in Figure 3.

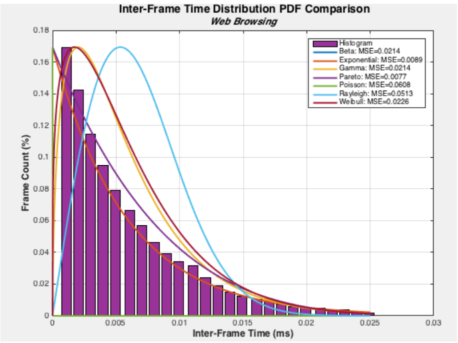

Fig. 2: Example PDF Fitting

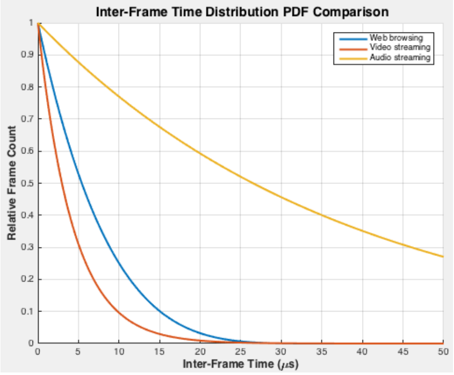

Fig. 3: Activity PDF Comparison

Multiple data sets of each activity were collected, specifically video streaming, audio streaming, and general web browsing. Comparison of the “best” fit, in the mean square error sense, PDFs for each activity are shown in Figures 4, 5, and 6.

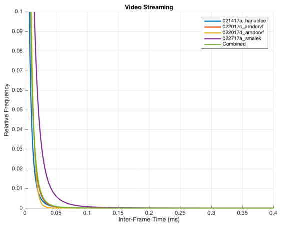

Fig. 4: Video Streaming Pareto Distribution

Comparison of multiple activity datasets shows the Generalized Pareto distribution which is specified by the three

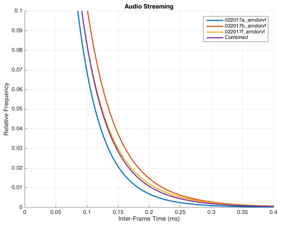

Fig. 5: Audio Streaming Pareto Distribution

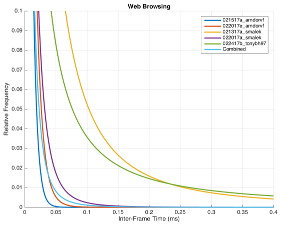

Fig. 6: Web Browsing Pareto Distribution

parameters location μ, scale σ, and shape ξ and has a PDF of the form

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\left(1+\frac{1}{\xi}\right)} \tag{1}$$

The composite PDF estimate was made by combining the inter-frame times from all activity collections. The composite Pareto distribution parameters for the activities (see Table I):

TABLE I: Activity Pareto Parameters

	ξ	σ	μ
Audio Streaming	0.09535	3.9398e-05	0.0
Video Streaming	0.24128	3.9935e-06	0.0
Web Browsing	0.38399	6.8662e-06	0.0

To identify and quantify the differences between the activity distributions, the Kullback-Leibler divergence[11] was computed. This provides a measure of the (non-symmetric) difference between two probability distributions P and Q. For distributions P and Q of a continuous random variable, the Kullback-Leibler divergence is defined to be the integral (Equation 2)

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx \tag{2}$$

The question then arises, for the data collected here, are the Kullback-Leibler divergences within an activity more similar to each other than the divergences between activities.

$$\text{Video Streaming:}\begin{bmatrix} 0 & 0.0127 & 0.0552 & 0.1685 \\ & 0 & 0.0264 & 0.1185 \\ & & 0 & 0.2995 \\ & & & 0 \end{bmatrix}$$

$$\text{Audio Streaming:}\begin{bmatrix} 0 & 0.0154 & 0.0078 \\ & 0 & 0.0023 \\ & & 0 \end{bmatrix}$$

$$\text{Web Browsing:}\begin{bmatrix} 0 & 0.0693 & 3.1020 & 0.3098 & 3.7103 \\ & 0 & 1.2611 & 0.0677 & 1.4383 \\ & & 0 & 0.4871 & 0.0209 \\ & & & 0 & 0.5228 \\ & & & & 0 \end{bmatrix}$$

Compare the inter-activity divergences with those between activities

$$\text{Video-Audio:}\begin{bmatrix} 1.5954 & 1.8207 & 1.7061 \\ 1.5977 & 1.8621 & 1.7351 \\ 3.3895 & 4.2447 & 3.9441 \\ 0.7730 & 0.9451 & 0.8589 \end{bmatrix}$$

$$\text{Video-Web:}\begin{bmatrix} 0.1179 & 0.2537 & 1.5655 & 0.3812 & 1.5075 \\ 0.0533 & 0.1725 & 1.6453 & 0.3174 & 1.6430 \\ 0.0665 & 0.2758 & 4.7463 & 0.6805 & 5.5676 \\ 0.1074 & 0.0273 & 0.7879 & 0.0444 & 0.7956 \end{bmatrix}$$

$$\text{Audio-Web:}\begin{bmatrix} 2.1300 & 0.9873 & 0.0808 & 0.4539 & 0.1849 \\ 2.7357 & 1.2706 & 0.0543 & 0.5922 & 0.1391 \\ 2.5216 & 1.1528 & 0.0468 & 0.5233 & 0.1294 \end{bmatrix}$$

In general, the inter-activity divergences are relatively small indicating good agreement between the PDF representations. The cross-activity divergences are, in general, significantly larger indicating identifiable differences between the distributions. Of particular note, however, are the web browsing distributions. It is not entirely clear that a distinction could be made between audio or video and web browsing. In fact, the web browsing activity appears to behave similarly to video streaming at times, and audio streaming at others. More investigation into the cause of this is needed. It is possible that the type of web page may play a role in this. For instance, large, graphic-heavy or advertisement-heavy pages may appear more like video streams in their inter-frame distributions. Some other types of pages may include messaging sites, or even collaboration sites that brings multiple users together into one virtual space.

V. Conclusions and Future Work

Initial comparison of the inter-frame time probability density functions shows an apparent difference between the activities investigated. The data suggests that all of the activities studied were best modeled with a Pareto distribution, further there is parametric difference between the distributions of the activities. The Kullback-Leibler divergence of distributions within each activity suggest they are consistent within an activity while the divergence suggests statistically significant differences in distributions between activities. These differences imply, at least with the limited data set, that the inter-frame arrival times distributions can be used to distinguish between these particular activities with the exception of web browsing.

The divergence within and between activities for web browsing do not show a strong identifiable distribution. It is postulated that many web pages encountered include both audio and video either through normal web content or advertisements, or perhaps the graphical content or the websites appear similar to audio or video. For example, web browsing will often load videos and audio, especially when not using an ad blocker. Thus, web browsing could be broken down into multiple categories such as graphic-heavy webpages, audio-heavy web pages, text-heavy webpages, etc. Further investigation into this is warranted.

In order to increase the reliability of this analysis, more data is required. This would include more examples of each activity as well as longer time spans. Additional types of activities could also be captured including web chatting, voice-over-IP, and the use of applications such as video games that send large amounts of network data.

Once characterization of the activity distributions is complete, investigation of the activities when performed over a VPN should be investigated. It is postulated that the presence of a VPN does *not* alter the distributions, however that remains to be shown. Including data collections of non-controlled, more active "hotspots" (e.g., Starbucks) will give an indication of the impact of many, simultaneous users as well as the impact of VPN and other encryption technologies.

References

[1] Apr. 2017. URL: http://www.internetlivestats.com/internet-users/.

[2] Apr. 2017. URL: https://www.statista.com/statistics/301204/top-markets-vpn-proxy-usage/.

[3] Jin Cao, William S. Cleveland, Dong Lin, et al. "Internet Traffic Tends Toward Poisson and Independent as the Load Increases". In: *Nonlinear Estimation and Classification*. Ed. by David D. Denison, Mark H. Hansen, Christopher C. Holmes, et al. New York, NY: Springer New York, 2003, pp. 83–109. ISBN: 978-0-387-21579-2. DOI: 10.1007/978-0-387-21579-2_6. URL: http://dx.doi.org/10.1007/978-0-387-21579-2_6.

[4] T. Karagiannis, M. Molle, M. Faloutsos, et al. "A nonstationary Poisson view of Internet traffic". In: *IEEE INFOCOM 2004*. Vol. 3. Mar. 2004, 1558–1569 vol.3. DOI: 10.1109/INFCOM.2004.1354569.

[5] V. Paxson and S. Floyd. "Wide area traffic: the failure of Poisson modeling". In: *IEEE/ACM Transactions on Networking* 3.3 (June 1995), pp. 226–244. ISSN: 1063-6692. DOI: 10.1109/90.392383.

[6] A. Tudjarov, D. Temkov, T. Janevski, et al. "Empirical modeling of Internet traffic at middle-level burstiness". In: *Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference (IEEE Cat. No.04CH37521)*. Vol. 2. May 2004, 535–538 Vol.2. DOI: 10.1109/MELCON.2004.1346984.

[7] Fan Zhang, Wenbo He, Xue Liu, et al. "Inferring Users' Online Activities Through Traffic Analysis". In: *Proceedings of the Fourth ACM Conference on Wireless Network Security*. WiSec '11. Hamburg, Germany: ACM, 2011, pp. 59–70. ISBN: 978-1-4503-0692-8. DOI: 10.1145/1998412.1998425. URL: http://doi.acm.org/10.1145/1998412.1998425.

[8] Dashdorj Yamkhin and Youjip Won. "Modeling and Analysis of Wireless LAN Traffic". In: *Journal of Information Science and Engineering* 25 (2009), pp. 1783–1801.

[9] W. Li and A. W. Moore. "A Machine Learning Approach for Efficient Traffic Classification". In: *2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*. Oct. 2007, pp. 310–317. DOI: 10.1109/MASCOTS.2007.2.

[10] Apr. 2017. URL: https://www.wireshark.org.

[11] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. ISSN: 00034851. URL: http://www.jstor.org/stable/2236703.