A Forensic Insight into Duolingo

Karan Nihalani
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, United Kingdom
17023122@stu.mmu.ac.uk

***Abstract*— As human beings, we seem to have a necessary and emotional connection to communication due to its representation in many forms, where examples include, music, writing and speaking. According to Greek philosophy, communication is viewed as, "a process bringing together humans to consider a shared reality through the world" [1] and conversing in different languages as a method of communication between individuals seems to be represented as a more personal and private form of connecting with others, where normally, people are unintentionally shut out from experiencing this with foreign languages because of their lack of knowledge. This is where Duolingo is proven to be of great use. This multi-platform application is adapted into many businesses and into many people's lives, as a free accessible tool for free language-learning. Professional workplaces employing this tool are bound to find themselves in a beneficial position, due to Duolingo's easy accessibility and online courses, where employees and organizations allow themselves to expand to a global scale via the ability of learning a new language only a couple of clicks away. This paper presents a network and forensic point of view on the famous application, where further subjects, such as, data privacy and confidential information, along with a technical network analysis will be explored further.**

Keywords—language learning, Duolingo, TCP, HTTP, IP address, Wireshark, data, examination, network, analysis, network forensics

I. INTRODUCTION

The will or ability to learn one or more foreign language in the professional and modern world is deemed of great importance, as a great amount of effort and focus is emphasized on business relationships with the purpose of connecting through a native language to create long-lasting relationships, which can end up being beneficial to the learner and from a more personal perspective, understanding and undergoing the process of learning a new language, allows for a deep dive into the discovery of new cultures. On the other hand, having learnt a foreign language is proposed to be a newer, competitive skill, where "we must foster an environment from a young age that promotes multilingual learning" [2], with the end goal of setting up children for "success, security and ultimately, prosperity [2]."

Duolingo is a mobile and desktop application that has the purpose of teaching its users a very large amount of languages, over 30+, through the aid of online courses and activities. Modern day forensics requires the understanding of many languages due to the Internet being available to the entire world, regardless of where an attacker or analyst is on the globe. Apart from this, analysts and investigators themselves are required to learn or have the capability of comprehending more than one language to the variety of global threats being carried out constantly. Organizations like GCHQ are current examples of recruiters looking for multilingual employees looking to work with global terrorist-influenced issues and furthermore, it can be said that Duolingo can be employed similarly from the "enemies" point of view, suggesting that languages are easily available to be learnt. Within this report, Duolingo will be analyzed from a forensic point of view, where most of its network traffic will be examined to potentially filter, capture and inspect packets transmitted when Duolingo is running. Duolingo's terms of service will also be examined, as well as, it's data protection methods, when conserving confidential user data. The results from the examinations will be demonstrated alongside the concluding ideas.

THE DUOLINGO APPLICATION

A. Duolingo's storage of information

Duolingo is considered to be one of the leading language learning applications used in the modern world by citizens, consisting of a wide range demographic. The application is able to target people of all ages and cultures, due to its flexibility with language difficulty, ranging from the more commonly spoken English to more rare languages, like Esperanto or Scottish Gaelic.

From a more technical perspective, Duolingo is mentioned to be run on Amazon Web Services, utilizing MySQL when storing data in their database. According to Brendan Meeder, an employee at Duolingo, the program DynamoDB is also used, for "storing user vocabularies and a combination of Redis and Memcache for caches [3]." It is further mentioned that their "backend stack is written in Python [3]", it being one of the most common used languages lately. When collecting information, the user will be asked for their name, e-mail address and date of birth or age. Along with this, if the user chooses to pay for further Duolingo services, more confidential information is handed, such as, payment information.

B. The use of Artificial Intelligence

Ever since its development in 2011, Duolingo's creator Luis von Ahn, who is also the developer of the reCAPTCHA authentication method used by Google on most websites, "verifying your humanity" [4] and he has taken the next step in manufacturing the language application via the use of artificial intelligence, to bring forward more effective language learning. Duolingo reportedly make use of their AI to "verify the person's identity, generate the test items, score them and administer the test adaptively [5]" and "to get as close as possible to having a human-to-human experience" [5], where all of this data is stored with the purpose of enhancing content development, personalization and cognitive modelling. As a reader, a question of whether Duolingo's developers make use of some form of reCAPTCHA technology as their AI when assessing their users' skills is raised.

C. Data Privacy

Within Duolingo, the user's data is promised to be kept safe, although how much data is kept for safekeeping is not mentioned, when accessing the Desktop application. When loading the application, the user is taken to the login screen, where only a form page with the options to insert a username and password is shown, as well as, the option to create an account, prompting the user to select a language course and difficulty.

Although the option to select an account has been chosen, the new user is able to start the selected language course right away, but all progress would be discarded once the application is shut down, resulting in forcing the end user to create the account and submit their name, e-mail address and password.

At first, it seems that the only information taken from the user is the data mentioned previously, but according to Duolingo's updated Privacy Policy, this isn't entirely true, for example, when logging into the application, Duolingo collects information on the device type, the device's IP address and the device ID. Furthermore, it is also stated that "Duolingo may use or share anonymous data collected through the Service, including Activity Data without limitation [7]", to provide the user with the "best service possible [7]".

II. APPLICATION NETWORK FORENSIC ANALYSIS

A. Set Up for Application Analysis

The aim of this examination is to carry out a forensic analysis on the network traffic generated by the Duolingo desktop application when running. The analysis was undertaken using the device mentioned below:

DESKTOP – I91OP71

12GB RAM

1TB Storage

Windows 10 (ver. 1909)

The device was connected to a public Wi-Fi network that was used so that the application could be fully functional, as the application would fail to load without a Wi-Fi network connection. Network traffic was also retrieved by making use of Wireshark (version 3.2.2) [8], an industry network packet analysis tool. Only the Duolingo desktop application and Wireshark were left open, to reduce the unnecessary amount of traffic that would be generated, such as, DHCP traffic.

The examination of the network traffic was undertaken as a live forensic examination, although the packets recorded by Wireshark were used for further examination at another time. The ACPO guidelines [9] were not taken into consideration, as well as other law enforcement procedures, as this was seen as more of a research activity, rather than a criminal investigation. Nevertheless, the research carried out and recorded in this instance is to be used in the future to aid law enforcement in the use of this product. For the prevention of any leaked packets, packet recording commenced right after the application was launched from the device's taskbar. This is influenced by Anderson's rule where the examiner would have to focus on the performance of the product versus the security. Also, if any information regarding data packets were to be leaked, signifying a weakness, it would allow for more digital forensic opportunities.

The Duolingo application was downloaded from the Microsoft Store [10] and was installed as instructed. Once installed, an account was created, and a language course was selected to be ready for examination. Other settings such as, silencing listening exercises, notifications or sound effects were left untouched.

B. Examination Process:

The first step undertaken was that the appropriate atmosphere was set to carry out the examination, where only Duolingo and Wireshark (ver. 3.2.2) were left open. The recording for data packets was then initiated within Wireshark, led by the login action of the Duolingo application right after. The log in details consisted of an e-mail, 'karan1999.kn@gmail.com', followed by the password. In the application, several language learning activities present were carried out, as an attempt to demonstrate a transmission of data packets across the

network. These language activities were completed consecutively for a total of 16 minutes. Immediately after this, the data packet recording was stopped by the examiner and the state of that recording was saved in the device's 'Documents' directory. The data packet analysis Wireshark file was saved as a Wireshark capture file with a '.pcapng' file extension, containing "a 'dump' of data packets captured over a network" [11].

III. RESULTS OF THE EXAMINATION

The examination was undertaken between the 10th and 11th of March 2020.

When reviewing the traffic generated during the recording of data packets, it is seen that a SYN-ACK request was made repeatedly between the host's IP address and Duolingo's server IP address in Iowa, USA. The traffic also shows that the HTTP protocol is used on port 80, every time the user interacted with an activity through advancing levels. More specifically, HTTP/1.1 is used, allowing for "client-server connections to be pipelined", where many requests are sent, "without waiting for a response from the server" [12]. This is shown in Fig. 1 below:

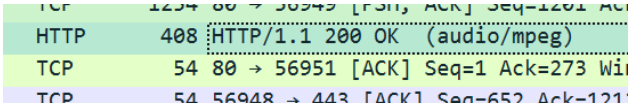

Fig. 1

Furthermore, the results also demonstrated that traffic constantly flowed through port 443, using the TCP protocol. This suggests that the traffic is portrayed to be "secured HTTP traffic" [13], enabling "websites to be available over both HTTP and HTTPS" [13]. The fact that the TCP protocol is made use of to transfer data packets is advantageous to how the Duolingo application handles its data, as it "guarantees the integrity of data sent over the network, regardless of the amount" [14]. This is due to the high-level encryption provided in the TCP/IP layer. Below, Fig. 2 demonstrates the TCP protocols when transmitting data packets in the data packet activity visualizer, in Wireshark.

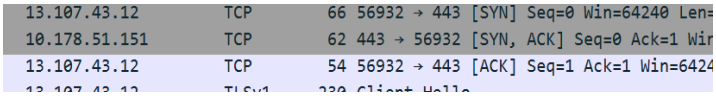

Fig. 2

IV. RECOMMENDATIONS

According to the examination that was undertaken on the application, the following recommendations are presented.

1. To make sure that Duolingo has not been used before and to carry out the network analysis over a shorter period of time, as data can result in a great number of unnecessary logs, therefore, the analysis process should be planned beforehand.
2. The examiner should be aware that data correlation could be causal or temporal, therefore, for temporal data correlation, "timestamps should be logged as well" [15].
3. The examiner should perform a further in-depth research into the 'pcapng' file saved. This is due to the fact that these files contain a great amount of data, for example, mappings from IPv4 and IPv6 addresses to host names" [16].
4. The examiner should restrict network connections to prevent "putting the computer at risk" if the network is "outside of the administrator's control" [17] due to the possibility of the device's security settings being weak.
5. The examiner should perform the network analysis with only Wireshark (with the latest version updated) and Duolingo (latest version) open in the device. This is to prevent any other traffic from interrupting the investigation.

These recommendations should allow a future digital forensic practitioner in examining this product.

V. CONCLUSIONS

In this paper, the Duolingo application was examined in a forensic manner, in terms of the different ports used and the network traffic generated. The results displayed that Duolingo makes sure to transfer data through the network in a safe manner, utilizing the TCP protocol, as the protocol "is known more for its reliability" [18].

To conclude, Duolingo's data seemed to be transmitted efficiently and although the application retrieves a lot of user information that remains confidential, it was proven by this examination that the data was not declared as vulnerable after all and from a forensic point of view, a data privacy leak is deemed as a low risk.

REFERENCES

[1] P. Cobley, "Communication: Definitions and Concepts", The International Encyclopedia of Communication, First Edition, Website, 2008. https://books.google.co.uk/books?id=UceOCQAAQBAJ&pg=PA1589&lpg=PA1589&dq=Communication:+Definitions+and+Concepts+The+International+Encyclopedia+of+Communication,+First+Edition&source=bl&ots=GcY5ltXul9&sig=ACfU3U0ihaw_XFlpqIjLeKAS-scapuG1Sw&hl=es&sa=X&ved=2ahUKEwjIocHImpHoAhV6QkEAHai7DvoQ6AEwCHoECAoQAQ#v=onepage&q=Communication%3A%20Definitions%20and%20Concepts%20The%20International%20Encyclopedia%20of%20Communication%2C%20First%20Edition&f=false [Accessed on 4th March 2020].

[2] L. De Valoes, Adjunct Faculty, Trinity Washington University, "Importance of Language – Why Learning a Second Language is Important", Website, 2014. https://discover.trinitydc.edu/continuing-education/2014/02/26/importance-of-language-why-learning-a-second-language-is-important/ [Accessed on 5th March 2020].

[3] B. Meeder, "What is the technology stack behind Duolingo?," Quora.com, Website, 2013. https://www.quora.com/What-is-the-technology-stack-behind-Duolingo [Accessed on 10th March 2020].

[4] A. Griswold, "How Luis Von Ahn Turned Countless Hours Of Mindless Activity Into Something Valuable", Business Insider, Website, 2014. https://www.businessinsider.com/luis-von-ahn-creator-of-duolingo-recaptcha-2014-3?r=US&IR=T [Accessed on 10th March 2020].

[5] N. Gagliordi, "How Duolingo uses AI to disrupt the language learning market," *Between the Lines*, ZDNet, Website, 2018. https://www.zdnet.com/article/how-duolingo-uses-ai-to-disrupt-the-language-learning-market/ [Accessed on 10th March 2020].

[6] P. Sawers, "How Duolingo is using AI to humanize virtual language lessons", Website, 2019. https://venturebeat.com/2019/07/05/how-duolingo-is-using-ai-to-humanize-virtual-language-lessons/ [Accessed on 10th March 2020].

[7] Duolingo, "5. Use of information obtained by Duolingo," Privacy Policy, Website, 2018. https://www.duolingo.com/privacy [Accessed on 10th March 2020]

[8] Wireshark, Download Wireshark, *Windows Intsaller (64-bit)*, Website, 2020. https://www.wireshark.org/download.html [Accessed on 10th March 2020].

[9] ACPO (Association of Chief Police Officers), ACPO Good Practice Guide for Digital Evidence, Website 2012. https://www.digital-detective.net/digital-forensics-documents/ACPO_Good_Practice_Guide_for_Digital_Evidence_v5.pdf [Accessed on 10th March 2020].

[10] Microsoft Store, Education, Language, *Duolingo – Learn Languages for Free*, Application, 2020. [Accessed on 11th March 2020].

[11] FileInfo, *.PCAPNG File Extension*, "What is a PCAPNG file?", Website, 2020. https://fileinfo.com/extension/pcapng [Accessed on 11th March 2020].

[12] Wireshark, *Hyper_Text_Transfer_Protocol*, "Hyper Text Transfer Protocol (HTTP)," Website, 2011. https://wiki.wireshark.org/Hyper_Text_Transfer_Protocol [Accessed on 11th March 2020].

[13] N. Congleton, Lifewire, "What is Port 443?", Website, 2019. https://www.lifewire.com/what-is-port-443-4690657 [Accessed on 11th March 2020].

[14] M. Rouse, Techtarget, "(TCP) Transmission Control Protocol", Website, 2019. https://searchnetworking.techtarget.com/definition/TCP [Accessed on 11th March 2020].

[15] InfoSec Institute, "Computer Forensics: Network Forensics Analysis and Examination Steps [Updated 2019]," Website, 2019. https://resources.infosecinstitute.com/category/computerforensics/introduction/areas-of-study/digital-forensics/network-forensics-analysis-and-examination-steps/#gref [Accessed on 11th March 2020].

[16] NetreSec, "Exctracting Metadata from PcapNG files", Website, 2013. https://www.netresec.com/?page=Blog&month=2013-02&post=Extracting-Metadata-from-PcapNG-files [Accessed on 11th March 2020].

[17] KrypSys, "Ten Basic Network Security Recommendations", Website, 2020. https://www.krypsys.com/network-security/ten-basic-network-security-recommendations/ [Accessed on 11th March 2020].

[18] Extrahop, *Transmission Control Protocol (TCP)*, "What is (TCP) Transmission Control Protocol?", Website, 2020. https://www.extrahop.com/resources/protocols/tcp/ [Accessed on 11th March 11, 2020].