Anon17: Network Traffic Dataset of Anonymity Services

Khalid Shahbar A. Nur Zincir-Heywood
Faculty of Computer Science
Dalhousie University
Halifax, Canada
{Shahbar, Zincir}@ cs.dal.ca

***Abstract*—One of the difficulties that face researchers on the anonymity networks field is the lack of anonymity dataset. Researchers need to collect their data to conduct a research or to use a simulated environment to collect the data they need. The lack of such a dataset is due to the nature of the anonymity networks. These anonymity networks aim to provide certain level of privacy for the users. In this paper, we present Anon17, a traffic flow dataset of different anonymity services. Anon17 specifically contains data collected from three anonymity networks; Tor, JonDonym, and I2P.**

Keywords— Dataset; Tor; I2P; JonDonym; Traffic Flow; Anonymity

I. INTRODUCTION

Anonymity tools aim to provide the users with some level of privacy. These tools enable the users on the Internet to freely access websites or run applications on the Internet without revealing their identity to any site that observes the network. In addition, anonymity tools hide the users' identity from the final destination (webserver). Some of the most used and known anonymity tools are Tor [1], JonDonym [2], and I2P [3].

There are many studies conducted on these anonymity tools. The researches include a wide range of aspects related to the anonymity field such as improving the design, performing attacks on the anonymity tool, analyzing the users behavior on the anonymity network, studying the performance and delay, revealing the users identity, and many others.

In some of the anonymity researches, the used data are collected in a simulated environment. Others used real data collected by the researchers themselves. The most common issue that faces researchers in the anonymity field is that these anonymity tools provide anonymity to the users; thus collecting the data and making it publicly available might affect the privacy of users of the anonymity tools. Therefore, the researches on the anonymity field count on data collected from a simulated environment or collected by the researchers themselves.

The traffic on the anonymity networks relies on passing the users' data through multiple stations on the network (nodes for example). These stations pass traffic for multiple users; collecting the data from these stations will include traffic for other users. Usually that means the researches need to run a station (node) and modify the way they collect the data to include only their traffic.

In this paper, we present a dataset for three anonymity tools: Tor, JonDonym, and I2P. We provide these data and make it publicly available without affecting the privacy of the users. To this end, the IP addresses of the users have been removed. The payload information only used for statistical measurement and then are removed. This is because we aim to provide a publicly available anonymity dataset that could be used to study the aforementioned anonymity tools. The dataset includes several applications used on these anonymity tools and also includes several obfuscation techniques that are used on some of these tools. Therefore, the dataset could be used for multiple types of researches.

The rest of this paper is organized as follows. Related work is reviewed in Section II. The Anon17 dataset description is introduced in Section III. Section IV presents the features and the format of the dataset. Finally, conclusions are drawn and the future work is discussed in Section V.

II. RELATED WORK

To the best of our knowledge, there is not any publicly available dataset that exclusively focuses on anonymity networks. Therefore, researches in this filed used simulation environments to experiment or collected their own data.

Barker et al. [8] collected Tor data using a simulated environment. The goal of their research is to study the possibility to differentiate between encrypted traffic and Tor traffic. The data included HTTP and HTTPS traffic over the simulated Tor network and HTTPS traffic.

Bauer et al. [7] presented a Tor network emulation tool, namely ExperimenTor. The tool provides a test environment for the Tor researchers by modeling Tor routers, bandwidth, users, and applications. The ExperimenTor tool is available as standalone and a VMware image that has the tool installed and configured.

There are many other researches on Tor, JonDonym, and I2P anonymity networks [16] [17] [18] [19] where researchers used their own data to conduct their research. In this work, we also collect our own data on three anonymity tools, namely Tor, JonDonym, and I2P. Furthermore, we make this dataset publicly available to be used by different researches.

III. ANONYMITY NETWORKS

Anonymity networks (services) have much in common in more than one aspect. They provide services to the users on the Internet while keeping their identity hidden. In addition, most of the time, anonymity services provide anonymity to the users by forwarding the users' traffic through multiple stations until the users' data reach to their destination. During this journey, the data are encrypted multiple times. This way, the users' data stay anonymous where each station during the journey knows only part of the information. Therefore, it becomes difficult to trace the users' data in these networks.

At the same time, these networks have differences between them. They differ in the design; this includes how the data is forwarded between the stations, the type of the encryption used, the way to manage the networks, the protocol to select the stations, the operators of the stations, and other details regarding the design. On the other hand, these networks have different goals under the design, and different supported applications within the networks. For example, the anonymity networks could be designed specifically to browse the Internet websites anonymously. Others could be designed to be a private network that users trade information internally. Such a network is not optimized to browse Internet websites. The supported applications on these networks are different, such as browsing, file sharing, IRC, web hosting, etc. One type of difference is the operators of the networks; some anonymity networks count on volunteers to provide the service to the users. Others use companies to provide the service. In addition, the threat models of these networks are not the same based on their design. What could be a threat to one of these networks might not form any threat to the others.

Tor [1] [15], JonDoNym [2], and I2P [3] are the three well-known anonymity networks that we included on the proposed dataset. Tor provides the users with anonymity by passing the traffic through three stations (nodes). These nodes run by volunteers shared their bandwidth with the anonymity network. The selection of these nodes is done by a protocol called path selection protocol. The users get the list of the currently available nodes on the networks from directory authorities servers. The path that the users use is not fixed, it continuously keeps changing. The user has the option to define the first or the last node on the path instead of counting on the path selection protocol to select the nodes.

JonDonym provides anonymity service to the users by passing the traffic through multiple stations (Mixes). The path (cascade) on the JonDonym network is fixed. The user can select the cascade that will relay the traffic but cannot change the mixes on the path. This is due to the difference in design of the anonymity networks. The mix on the JonDonym network multiplexes traffic from different users and sends it to the next mix on the cascade.

The Invisible Internet Project (I2P) is a packet switched network that uses multilayer encryption to provide anonymity to its users. The path (Tunnels) on the I2P network is unidirectional. The user builds an encrypted tunnel to the final destination to send messages using the created tunnel (outbound tunnel). The messages travel in one direction only; therefore, messages from destination to the sender must use another tunnel (inbound tunnel). Tunnels on the I2P networks are used for communication and management. By default, the users on the I2P network share bandwidth by participating on building I2P Tunnels. The user has the option to reduce or increase the amount of participation on building I2P Tunnels.

The connection between the user and any of these anonymity networks is not hidden. These networks aim to provide anonymity to the user but they do not hide that the user is connecting to such anonymity network. Therefore, blocking these anonymity networks is possible. Thus, Tor employed different obfuscation techniques (Pluggable Transports) to hide the connection to the Tor network. The implemented pluggable transports are Scramblesuit [9], Flashproxy [10], Obfs3 [11], Meek [12], Format-transforming encryption (FTE) [13]. Furthermore, JonDonym offers two options to resist the network blockage: TCP/IP forward and Skype tunnel. Both could be used to connect the user to the JonDonym network if the network is blocked. On the I2P network, the obfuscation options are considered but not implemented yet.

IV. ANON17

Anon17 [14] is collected at the NIMS lab [6] between 2014-2017 in a real network environment. The dataset is labeled based on the information available on the anonymity services themselves. For example, in the Tor network, the IP addresses of the Tor nodes are available. Therefore, whenever we collect data related to a node on the Tor network, we use the IP address to label this traffic as "Tor". The same applies for all the labeling in our data; we did not use any application classification tools to label our data. The Anon17 dataset contains the following:

A. Tor

The Tor dataset contains Tor traffic. The traffic includes the circuit establishment and the user activities on the Tor network such as browsing the Internet websites.

B. TorApp

The TorApp dataset contains flows for three machines (computers) running three applications on the Tor network (Browsing, Video streaming, and File sharing). Therefore, there are three classes on the TorApp dataset (Browsing, streaming, and BitTorrent). The Browsing class contains connection flows between a user and an entry node on the Tor network when the user is using Tor to browse different Internet websites. The Streaming class is the connection flows between the user and the entry node when the user is watching videos on Tor. The last class, the BitTorrent class contains flows between the user and the first node when the user is using Torrent files on the Tor network.

TABLE I. THE NUMBER OF TRAFFIC FLOWS IN EACH DATASET

Tool	Type of Traffic	Dataset Name	Classes	Number of Instances	Total Number of Instances
Tor	Normal Tor Traffic	Tor	Tor	5283	5283
	Applications on Tor	TorApp	Browsing	84	252
			Video	84	
			BitTorrent	84	
	Tor Pluggable Transports	TorPT	Obfs3	14718	353391
			Meek	43152	
			FTE	106237	
			Scramble suit	16953	
			Flash proxy	172331	
I2P	I2P Applications Tunnels with other Tunnels – 80% Bandwidth	I2PApp80BW	Eepsites	149992	449987
			jIRCii	149998	
			I2PSnark	149992	
	I2P Applications Tunnels with other Tunnels – 80% Bandwidth	I2PApp0BW	Eepsites	127349	195081
			jIRCii	29357	
			I2PSnark	38375	
	I2P Users Traffic	I2PUsers	Pc1	150000	449998
			Pc2	150000	
			Pc3	149998	
	I2P Applications	I2PApp	Eepsites	145	640
			jIRCii	221	
			I2PSnark	62	
			Exploratory Tunnels	86	
			Participating Tunnels	126	
JonDonym	JonDonym Traffic	JonDoNym	JonDonym	5440	5440

C. *TorPT*

The TorPT dataset contains flows for Tor pluggable transports. The TorPT has five classes: Obfs3, Meek, Flashproxy, FTE, and Scramblesuit. The TorPT is collected by connecting to the pluggable transports form the NIMS lab and capture the traffic. The Obfs3 is collected from two different Obfs3 bridges. The Scramblesuit traffic is collected by connecting to 22 different Scramblesuit servers. This is to ensure that we include the effect of changing the flow behavior that Scramblesuit pluggable transport aims to achieve.

D. *I2PApp80BW*

These traffic flows are collected while running three applications on the I2P network. The applications (classes) are Eepsites (the websites browsing on the I2P network), jIRCii (Internet Relay Chat (IRC) plugin on the I2P network), and I2Psnark (the file sharing plugin on the I2P network). The bandwidth sharing on the I2P client is set to default which is 80% sharing rate of the user bandwidth. In this dataset, each class contains the application flows in addition to the management traffic flows. For example, the Eepsites flows contain flows for the Eepsites Tunnels in addition to the Tunnels used for the management of the I2P network and ttunnels used to share bandwidth such as the Exploratory and the Participating Tunnels.

E. *I2PApp0BW*

This dataset is similar to the I2Papp80BW; the difference is that the amount of shared bandwidth is set to 0%. This will reduce the amount of management traffic flows on each class.

F. *I2PUsers*

This dataset contains the traffic flows for three users on the I2P network. The classes are named PC1,PC2, and PC3. The dataset I2PUsers is the same dataset used in I2PApp80BW. The difference is that the data is classified differently. Instead of labeling the dataset based on the application; the dataset is labeled based on the user traffic. Therefore, any class on this dataset will contain the three applicaiotns used on the I2Papp80BW dataset. For example, the PC1 flows contain

TABLE II. ANON17 DATASET FEATURES

Tranalyzer Features	Description
dir time_first, time_last, duration	Flow direction, time, and duration of the flow
numPktsSnt numPktsRcvd numBytesSnt numBytesRcvd minPktSz maxPktSz avePktSize pktps bytps pktAsm bytAsm	Counting of Packets and Bytes
ip_mindIPID ip_maxdIPID ip_minTTL ip_maxTTL ip_TTL_Chg ip_TOS ip_flags ip_Opt ip_OptCnt	The IP Header related features such as TOS, TTL etc.
tcp_PSeqCnt tcp_SeqSntBytes tcp_SeqFaultCnt tcp_PAckCnt tcp_FlwLssAckRcvdBytes tcp_AckFaultCnt tcp_InitWinSz tcp_AveWinSz tcp_MinWinSz tcp_MaxWinSz tcp_WinSzDwnCnt tcp_WinSzUpCnt tcp_WinSzChgDirCnt tcp_AggrFlags tcp_AggrAnomaly tcp_AggrOptions tcp_MSS tcp_WS tcp_OptCnt tcp_S-SA/SA-A_Trip tcp_S-SA-A/A-A_RTT tcp_RTTAckTripMin tcp_RTTAckTripMax tcp_RTTAckTripAve tcpStates	The TCP Header related features such as Window size, sequence number etc.
connSrc connDst connSrc<->Dst	Counting of number of connections between source and destination/ source to different destinations.
min_pl max_pl mean_pl low_quartile_pl median_pl upp_quartile_pl iqd_pl mode_pl range_pl std_pl stdrob_pl skew_pl exc_pl	Packet length statistics
min_iat max_iat mean_iat low_quartile_iat median_iat upp_quartile_iat iqd_iat mode_iat range_iat std_iat stdrob_iat skew_iat exc_iat nfp_pl_iat ps_iat_histo	Inter arrival time statistics
TrafficType	The classes

flows for the machine of the first users when the users used Eepsites, jIRCii, and I2Psnark on the I2P network.

G. I2PApp

This dataset contains traffic flows for the same three applications used in I2PApp80BW. The difference is that this dataset contains separate classes for the management tunnels. The total number of classes on this dataset is five: Eepsites, jIRCii, I2Psnark, Exploratory Tunnels, and Participating Tunnels. Therefore, the application tunnels do not contain any management tunnels flows.

H. JonDonym

The JonDonym dataset contains traffic flows for the JonDonym network. The dataset contains flows for the whole free mixes on the JonDonym network.

Table I shows the Anon17 dataset and the number of instances on each part of the dataset.

V. DATASET FEATURES AND FORMAT

We used Tranalyzer [4] to extract the flows from the PCAP files we captured in the NIMS lab. Tranalyzer has 92 features such as Number of bytes sent, Number of bytes received, Statistics about the Interarrival time, and Number of connection etc. Some of the unrelated features are removed from the dataset such as the ICMP features and VLan features because they do not provide useful information for our purposes. IP addresses and payloads of the packets are also removed from the dataset to protect the privacy of the users. In the dataset, the values of some features might have zeros. For example, the I2P network works on both TCP and UDP. Therfore, if the I2P dataset contains UDP connections then all the TCP related features will have zero values. The data is formatted into arff file format used in the open source data mining software tool, Weka [5]. Table II summarizes the features included in Anon17 dataset.

VI. CONCLUSION AND FUTURE WORK

Anon17 is an anonymity network dataset that contains three anonymity network data: Tor, JonDonym, and I2P. In addition to the traffic flows of these three anonymity networks, the dataset includes applications traffic flows run on Tor and I2P. Furthermore, the dataset contains traffic flows for the obfuscation techniques used on the Tor network; the pluggable transports. Anon17 includes also data on the tunnels used on the I2P network. To the best of our knowledge, Anon17 is the first publically available anonymity network dataset that covers three anonymity networks. For future work, we will expand Anon17 to include additional applications run on these anonymity networks.

REFERENCES

[1] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: the second-generation onion router," in *Proceedings of the 13th conference on USENIX Security Symposium - Volume 13*. USENIX Association, 2004, pp. 21–21.

[2] Project: AN.ON – Anonymity [Online]. Available: http://anon.inf.tu-dresden.de/index_en.html.

[3] The Invisible Internet Project (I2P) [Online]. Available: https://geti2p.net/en/

[4] *TRANALYZER2* [Online]. Available: http://tranalyzer.com/

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten,"The WEKA data mining software: an update," SIGKDD Explorations, vol. 11, no. 1, pp. 10-18, 2009.

[6] NIMS: Network Information Management and Security Group [Online]. Available: https://projects.cs.dal.ca/projectx/

[7] K. Bauer, M. Sherr, D. McCoy, D. Grunwald, "ExperimenTor: a testbed for safe and realistic tor experimentation," in *Proceedings of the 4th conference on Cyber security experimentation and test*, p.7-7, August 08, 2011, San Francisco, CA

[8] J. Barker, P. Hannay, and P. Szewczyk, "Using traffic analysis to identify the second generation onion Router," in *the 9th IFIP International Conference on embedded and ubiquitous computing*, Melbourne, AUS, 2011, pp.72-78.

[9] P. Winter, T. Pulls, and J. Fuss. "ScrambleSuit: A Polymorphic Network Protocol to Circumvent Censorship," In Workshop on Privacy in the Electronic Society, Berlin, Germany, 2013. ACM.

[10] D. Fifield, N. Hardison, J. Ellithrope, E. Stark, R. Dingledine, D. Boneh, and P. Porras, "Evading Censorship with Browser-Based Proxies," In PETS, 2012.

[11] Obfs3. [Online]. Available: https://gitweb.torproject.org/pluggable-transports/obfsproxy.git/tree/doc/obfs3/obfs3-protocol-spec.txt

[12] Meek. [Online]. Available: https://trac.torproject.org/projects/tor/wiki/doc/meek

[13] K. Dyer, S. Coull, T. Ristenpart and T. Shrimpton, "Protocol Misidentication Made Easy with Format-Transforming Encryption," ACM SIGSAC Conference on Computer and Commu- nication Security, CCS'13, pp. 61-72, ACM, 2013.

[14] Anon17: Anonymity Networks Dataset. [Online]. Available: https://web.cs.dal.ca/~shahbar/dataset/anon17

[15] Tor project: Anonymity online. [Online]. Available: https://www.torproject.org/

[16] A. Chaabane, P. Manils and M. A. Kaafar, "Digging into Anonymous Traffic: A Deep Analysis of the Tor Anonymizing Network," *2010 Fourth International Conference on Network and System Security*, Melbourne, VIC, 2010, pp. 167-174.

[17] M. AlSabah, K. Bauer, and I. Goldberg, "Enhancing Tor's performance using real-time traffic classification," in Proceedings of the 2012 ACM conference on Computer and communications security, Raleigh, USA, 2012, pp. 73-84.

[18] B. Westermann, D. Kesdogan, "Malice versus AN.ON: possible risks of missing replay and integrity protection," in Proceedings of the 15th international conference on Financial Cryptography and Data Security, 2011, Gros Islet, St. Lucia.

[19] P. LIU, L. WANG, Q. TAN, Q. LI, ,X. WANG, and J. SHI, "Empirical Measurement and Analysis of I2P Routers". Journal of Networks, North America, 9, sep. 2014.