Connection Type Identification and Uplink Speed Estimation of Malware Infected Hosts

XUPING HUANG¹,²,a) SHUNSUKE MOCHIZUKI³,b) KATSUNARI YOSHIKA¹,c)

Received: March 8, 2022, Accepted: September 2, 2022

Abstract: IoT malware Mirai and its variants continue to evolve and their activities consume network resources, particularly radio resources. This paper proposes a method to identify connection types and estimate the wireless uplink speed of malware-infected hosts observed by IoT honeypot by using the Connection Type Database of Maxmind’s GeoIP2, a well-known industrial resource for IP address related information, and Network Diagnosis Tool (NDT) database, a measurement data set of the uplink speed of various networks. The proposed Mobile Network Identification method divides IP addresses into IP ranges assigned to each Autonomous System (AS), and then employs the NDT database based on the IP ranges. We analyzed the infected hosts observed by IoT honeypot to assess and validate the precision of the proposed technique. Our method estimates the maximum average uplink speed of the infected cellular host to be 40.6 Mbps, which is between two reference measurement results of cellular networks, indicating the adequacy of the proposed method.

Keywords: IoT malware, connection type identification, uplink speed estimation

1. Introduction

1.1 Background

IoT Botnets and their enormous distributed denial-of-service (DDoS) attacks have become a concern of the rapid growth of wireless devices and communication technologies. They have evolved to become platforms for harmful activities for cyber-attacks. Symantec has reported more than 6 million bots detected, and 85% of spam is from botnet[1]. As these stats are not for IoT botnet but general windows-based botnet, we should omit this reference. More than 29 million IoT malware samples have been captured by honeypots in project [2] from 2016 to March 2022 in total. Ubiquitous IoT devices enhance daily life and improve the efficiency of businesses based on mobile communication services, however, many of them are also vulnerable to cyber-attacks[3], [4]. Malware-infected devices can be used as a bastion host to generate a large amount of unauthorized wireless communication.

Most current cellular IoT services operate on 4G cellular networks, and with the expected deployment of billions of IoT devices, the traffic characteristics of IoT devices differ significantly from those of smartphones, thus increasing the danger of resource consumption.

The threat of IoT-based malware keeps increasing, thus we focus on attacks from these ubiquitous IoT devices.

In this research, we aim to develop a mechanism to identify connection types and estimate the uplink speed of devices infected by IoT malware. In our future work, we plan to utilize the mobile uplink speed information obtained by the proposed method to estimate the total volume of mobile resources that could be wasted by these attacks.

1.2 Related Works

Many industrial security schemes have been proposed to monitor the consequences of malicious attacks, such as the research results in NICT’s work in Japan[5], which supplies real-time monitoring of malicious attacks from variable infected devices. Using these monitoring logs, ATR Wave Engineering Laboratories developed a network-type estimation method using RTT response time and port scanning[6], [7]. Additionally, the uplink speed of the network is widely used to estimate the pattern of data transmission[10], and to calculate the scale of malicious access to abuse the Internet resources[11], etc. As a reference, uplink speed is analyzed in alternative conventional works[8], [9], [12], [13], [14], [15], [16], [17], [18], [19]. In detail, the work[8] showed the value of combining information from several sources to spatially distinguish significant sites and roughly locate the user utilizing network artifacts, like IP addresses, which are gathered from mobile applications. Work[9] based on uplink sensing, in which training signals for channel estimation are widely exploited. Work[12] proposed a method to predict the delay of LTE communication, using the uplink speed estimation. The focus is different from the proposed method. Work[13] relied on geo-referenced satellite imaging to extract the environmental characteristics for improved wireless network deployments. Uplink speed is used as a parameter, rather than the core technique. Work[14] analyzed datasets to reveal pre-
dicted network performance by using uplink speed, rather than uplink estimation. Work [15] suggested that more connected devices in a network might make the devices become more reliable, and the global average speed of broadband is still growing and will double from 45.9 Mbps to 110.4 Mbps from 2018 to 2023. Work [16] worked on multiple linear regression methods to predict hourly downlink speed. Work [17] investigated the uplink bandwidth prediction in the cellular networks using the machine learning method, which focused on quality of service based on a signal-noise ratio. Work [18] focused on the cellular uplink speed prediction between IoT devices in vehicle connection. Work [19] explored the cellular speed and capacity in a limited region. Works [20], [21], [22] discussed uplink speed of unreliability in ubiquitous IoT devices, particularly in the connected cars, focusing on the quality of service. Though these works focused on cellular speed, most of the works proposed the prediction method using the machine learning method rather than using multiple big-scale databases, covering domestic and worldwide data to calculate the uplink speed, which is one of the main contributions of our work.

1.3 Our Contributions

We develop the mobile network identification method in this paper to identify the connection types of infected devices via IP addresses, and then calculate the uplink speed of the traffic of attacks by analyzing the IoT honeypot log.

The main contributions of this paper are: (1) we propose a novel mobile Network Identification to classify the connection types of the malware-infected devices using IP addresses in honeypot log data; and (2) we estimate the uplink speed of the access by analyzing multiple big-scale datasets.

The details of the proposed model will be discussed in Section 2. The results of connection type identification and uplink speed of communication are discussed in Section 3, and Section 4 concludes.

2. Mobile Network Identification and Uplink Speed Estimation

The proposed method aims to identify network type and estimate the uplink speeds of hosts infected by IoT malware. Uplink speed estimates, in particular, may be used as the upper bound of possible DoS attacks that can be conducted by the infected hosts.

2.1 Mechanism for Mobile Network Identification

We use Maxmind’s GeoIP2 [23] to decide which IP address ranges provide mobile communication services. More specifically, we use Connection Type Database that provides network connection types such as “Cable/DSL”, “Corporate”, “Cellular” and “Unknown” for each IP address range. We also use GeoLite2 ASN Database to map IP address ranges to ASN. Finally, each address range is assigned ASN and connection type.

Furthermore, we evaluated the coverage ratio of GeoIP database for IP addresses, which is being used in domestic Japan, by referring the IP addresses released in APNIC [24] as the preliminary step. The APNIC dataset is available as plotted in Fig. 1, which includes 190 million IP addresses, and GeoIP2 connection type database which covers 85% of IP addresses in domestic in Japan.

2.2 Uplink Speed Estimation

We use the Network Diagnostic Tool (NDT) dataset in MLab project data, supplied in Google Cloud Platform (GCP) [25], where the particular table “measurements-lab.ndt.uploaded_uploads”, is used. Figure A-2 plotted an example of NDT dataset, with columns related to uplink speed estimation, including “MeanThroughputMbps” (connection speed), “MinRTT”, client IP and Server IP, etc. The dataset is collected during 2009/02/18–2020/08/18. We collected the IP addresses and upload speed from the dataset with a size of 947.4 GB.

To evaluate the coverage of NDT dataset, we checked how many IP address ranges assigned to Japan are covered by the dataset. We downloaded and analyzed the provided data and found that it covers 54.8% of ASNs in Japan.

2.3 Identifying Connection Types and Estimating Uplink Speed

The overview of the proposed method is illustrated in Fig. 2. The input data is the IP address of the infected devices and the output data includes the connection types and the uplink speed. The target IP address is extracted from honeypot log as the input data. On the one hand, the IP addresses are linked to GeoLite2 ASN database to get AS number, which is mapped to NDT dataset to get the maximum uplink speed. On the other hand, the same IP addresses are linked to GeoIP Connection Type database to get the connection type: “Cellular”, “Cable/DSL”, “Corporate”, and “Unknown”.

ASN information is used for mapping, since not all the infected IP addresses are included in NDT dataset. Using ASN means including more infected hosts for mapping to get uplink speed. By doing this, more uplink speeds of infected IP addresses can be estimated.

An identification method using multiple databases to get uplink speed and connection type has not been proposed in previous works. Estimating the uplink speed using ASN rather than IP addresses is another enhanced part of the method. Furthermore, crosschecking promises a lower false identification of connection type, since uplink speed value is a reference for connection type that false identification may be aware if uplink speed is outra-
3. Analysis Results of Honeypot Log by Cross-checking Mechanism

3.1 Results of Connection Type Identification

We extract IP addresses of the infected hosts in honeypot log data [2], identify their connection types provided from GeoIP2, and estimate uplink measurement data from MLab project. As a preliminary, we processed the tags from the log file to extract the timestamp and source IP address of the observed attacks. We extracted infected IP addresses observed during 2020/11/18–2020/12/1 with 995 unique IP addresses. The classification of ASN and connection types by GeoIP2 is shown in Fig. 2. Connection types of 759 IP addresses (76.3%) are classified as “Cable/DSL”, 224 (22.5%) as “Corporate”, and 8 (0.8%) as “Cellular”. The results are listed in Table 1.

Next, to evaluate the identification of connection types by GeoIP2, we analyzed 39,867 IP addresses in the mobile network of Softbank. As shown in Table 2, 99.5% of the given mobile addresses are identified as “Cellular” correctly, and only 0.5% is identified to be “Cable/DSL” incorrectly.

3.2 Results of Uplink Speed Estimation

Now, we look at the results of uplink speed estimation. Figure 3 illustrates the uplink speed of network 126.204.192.0/21, one of the Softbank mobile network. Among the IP address range, 8,608 addresses had their uplink speed measurement data in NDT dataset. Among them, 7,657 addresses are correctly identified as “Cellular” and 951 are incorrectly identified. Both groups have very similar distribution of uplink speed with a maximum value of around 30 Mbps. According to Softbank report [26] of mobile phone (cellular connection) in 2019, the average uplink speed is approximately 32 Mbps.

Figure 4 plots the uplink speed of different connection types with honeypot log data, including Cable/DSL, Corporate, Cellular, and Unknown. The average uplink speed of cellular connection is 40.6 Mbps, which is relatively close to the uplink speed of 45.9 Mbps shown in the previous work [15].

Since the literal reference datasets about communication speed of a certain IP address in cable/wireless connection are not released, we refer to theoretical value to verify the correctness of the results of the proposed method. Work [27] indicated the 4G wireless connection had a theoretical speed of around 200 Mbps with LTE/WiMAX connection, and work [28] revealed that a cable connection worked with a theoretical speed under 1 Gbps, and with 920 Mbps in the local investigation of the network. Work [29] suggested that wire communication is probably faster than wireless connection. The results obtained from the proposed method shown in Fig. 4 indicate a maximum uplink speed of 210.69 Mbps in cellular networks, and a maximum uplink speed of 920.48 Mbps in Cable/DSL networks, which aligns the theoretical value of cable and wireless connections.

Figure 5 plots the uplink speed for each network for each IP range with Softbank IP addresses data. The mainstream trend of each IP range is similar as one can confirm with the histograms of uplink speed of each network plotted in Fig. 6. Since the distribution feature of each network is similar, we extract the maximum uplink speed of each network with more than 1,000 IP addresses to estimate the maximum uplink speed with Softbank IP addresses data, which is plotted in Fig. 7. According to the results, a maximum uplink speed is 32 Mbps, which matches the open result of 32 Mbps from Softbank report [26].
3.3 Discussion

The uplink speed of different networks by nations/regions and connection types is available from IP addresses by the proposed method. Figure 8 plots the results of uplink speed for variable connection types and nations/regions partially from the malware activity log. The detected uplink speed values depend on the period and data monitored during the period. This mechanism can be applied to any malware activity log with IP address of infected devices and timestamps. Specific cyber security incident occurred from particular nations/regions can be observed, and furthermore, the potential scale of attack can be estimated by the proposed network identification and uplink speed, which could be a reference to make a measurement against a cyber pandemic.

Furthermore, all of these results above are estimated and analyzed by AS numbers in the same IP range. Besides, in order to verify whether the proposed method can identify and divide the cable and cellular traffic precisely, we also experimentally analyze the uplink speed by each IP independently. The dataset we used is plotted in Table 3 from GeoIP2 connection type database, including 190,128,384 IP addresses. The results of the uplink speed of each IP address by different network types are plotted in Fig. 9. According to the results, the uplink speed of cable, cellular, and corporate are totally different, which matches the theoretical values approximately and makes cellular traffic, including IoT access distinguishable from all of the network communications using the proposed method, with only the IP address as the input data. It suggests our proposed method could be applied to almost all honeypot log and other malware logs to monitor infected devices in Japan.

However, the challenge is that there is no ground truth dataset to confirm the connection type of a particular IP. We investigated multiple datasets, including GlobalComms database, which can-
not identify the connection type of a particular IP address. We also explored SURFPOINT data, which covers most domestic IP addresses, however, it cannot distinguish cable/wireless connection in the same AS.

4. Conclusion

In this paper, we proposed a novel mobile network identification method to identify the connection types (Cable/DSL, Corporate, Cellular, or Unknown) using IP addresses from the collected honeypot log data, by crosschecking multiple databases for IoT Botnet attack monitoring. As a result, 99.5% are successfully identified by verifying the results referring to smartphone access IPs from Softbank, Ltd. Furthermore, we estimated the uplink speed of cellular communication by crosschecking mobile access datasets as a preliminary research. We also verified the precision of our method, by comparing the uplink speed of specified IP addresses in IP ranges specified in Softbank report, partially. According to the statistical analysis results, the average of the maximum uplink speed is 40.6 Mbps, which align with the result of 32 Mbps released in Softbank report. Thus, the effectiveness of the proposed cellular identify model has been verified.

Comparing the results to multiple real database with IP addresses and connection types is listed as a future work.

Acknowledgments This research was conducted under a contract of “MITIGATE” among “Research and Development for Expansion of Radio Wave Resources (JP000254)”, which was supported by the Ministry of Internal Affairs and Communications, Japan.

References

[1] Symantec Security Response team: Internet Security Threat Report, Vol.24 (Feb. 2019), available from (https://docs.broadcom.com/doc/istr-24-executive-summary-en)

[2] IoTPTO project of Yoshioha Lab, available from (https://sec.ynu. codes/iot/)

[3] Stock, B., Pellegrino, G., Li, F., Backes, M. and Rossow, C.: Didnt You Hear Me Towards More Successful Web Vulnerability Notifications, The Network and Distributed System Security Symposium (NDSS) (2018)

[4] Shen, Y., Vervier, P.A., and Stringhini, G.: Understanding Worldwide Private Information Collection on Android, The Network and Distributed System Security Symposium (NDSS 2019), pp.1–16 (2019).

[5] Cetin, O., Ganan, C., Altuna, L., Inoue, D., Kasama, T., Tamiya, K., Tie, Y., Yoshioha, K., et al.: Cleaning Up the Internet of Evil Things: Real-World Evidence on ISP and Consumer Efforts to Remove Mi- rai, The Network and Distributed System Security Symposium (NDSS) (2019).

[6] Teng, R. et al.: Identification of IoT Network Type Based on The Response Delay Property, Proc. ICACT2020, pp.319–324 (Feb. 2020).

[7] Teng, R. et al.: Statistical study on response delay characteristics of network scan to IoT devices – the base for estimation of IoT device attributes, IEICE Technical Report, NS2019-142, pp.45–49 (Dec. 2019).

[8] Mirza, M.M. and Karabiyyi, U.: Enhancing IP Address Geocod- ing, Geolocating and Visualization for Digital Forensics, 2021 Inter- national Symposium on Networks, Computers and Communications (SNCCC), pp.1–7, IEEE (2021).

[9] Ni, Z., Zhang, J.A., Huang, X., Yang, K. and Yuan, J.: Uplink sensing in perceptive mobile networks with asynchronous transceivers, IEEE Trans. Signal Processing, Vol.69, pp.1287–1300 (2021).

[10] Eyceyurt, E. and Isko, Z.: Uplink Throughput Prediction in Cellular Mobile Networks, International Journal of Electronics and Communi- cation Engineering, Vol.16, No.6 (2020).

[11] Tan, Z.W., Zhao, J.H., et al.: Device-Based LTE Latency Reduction at the Application Layer, 11th USENIX Symposium on Networked Sys- tems Design and Implementation, pp.471–486 (2020).

[12] Bui, N., Michelinaikis, F. and Widmer, J.: Fine-grained LTE radio link estimation for mobile phones, Pervasive and Mobile, Computing, Vol.49, pp.76–91 (2018).

[13] Otero, Y. and Egi, C.: Machine-Learning and 3D Point-Cloud Based Signal Power Path Loss Model for the Deployment of Wireless Communication Systems, IEEE Access, Vol.7, pp.42507–42517 (2019).

[14] Yue, C., Jin, R., Suh, K., Qin, Y., Wang, B. and Wei, W.: LinkForecast: Cellular Link Bandwidth Prediction in LTE Networks, IEEE Trans. Mobile, Computing, Vol.17, pp.1582–1594 (2018).

[15] Khedkar, S.P. et al.: Prediction of Traffic Generated by IoT Devices Using Statistical Learning Time Series Algorithms, Wireless Communica- tions and Mobile Computing, Vol.21, Article ID 5366222, 12 pages, DOI: 10.11055/2021/5366222 (2021).

[16] Lee, D., Lee, D., Choi, M. and Lee, J.: Prediction of Network Throughput using ARIMA, International Conference on Artificial Intelli- gence in Information and Communication (IAIC1) (2020).

[17] Oussakel, L., Owezarski, P. and Berthou, P.: Cellular Uplink Band- width Prediction Based on Radio Measurements, MobiWac (2019).

[18] Liu, L., Zhang, S. and Zhang, R.: Exploiting NOMA for Multi-Beam UAV Communication in Cellular Uplink, ICC 2019 – 2019 IEEE International Conference on Communications (ICC), pp.1–6, DOI: 10.1109/ICC.2019.8761670 (2019).

[19] Jha, S.K., Rokaya, R., Bhagat, A., Khan, A.R. and Aryal, L.: LTE Network: Coverage and Capacity Planning — 4G Cellular Net- work Planning around Banepa, 2017 International Conference on Networking and Network Applications (NeNA), pp.180–185, DOI: 10.1109/NeNA.2017.23 (2017).

[20] Obidou, E., Abubakar, A., Ramam, A., Sastry, N. and Mangiante, S.: IoT share or not to share: Reliability assurance via redundant cellular connectivity in Connected Cars, 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS), pp.1–6, IEEE (2021).

[21] Bo, H. et al.: An uplink throughput optimization scheme for uav- enabled urban emergency communications, IEEE Internet of Things Journal (2021).

[22] Nguyen, T.V. et al.: A comparison of cellular vehicle-to-everything and dedicated short range communication, IEEE Vehicular Net- working Conference (VNC), pp.101–108 (2017).

[23] available from (https://www.maxmind.com/en/geop2-connection- type-database) (accessed 2022-06-05).

[24] available from (https://www.apnic.net/) (accessed 2022-07-05).

[25] Google, Inc.: Network Diagnostic Tool, available from (https://www. measurementlab.net/tests/ndt/)

[26] SoftBank Group Corp.: Speed survey, available from (https://www. softbank.jp/mobile/network/service/speed-survey/)

[27] Vaigandla, K.K. et al.: A Survey On Wireless Communications: 6g And 7g, IEEE Communications Magazine, Vol.59, No.6, pp.20–25 (2021).

[28] Lai, J. and Widmar, N.O.: Revisiting the Digital Divide in the COVID- 19 Era, Applied Economic Perspectives and Policy, Vol.43, No.1, pp.458–464 (2021).

[29] Baliga, J. et al.: Energy Consumption in Wired and Wireless Access Networks, IEEE Communications Magazine, pp.70–77 (2011).

[30] Huang, X., Mochizuki, S., Fujita, A. and Yoshioka, K.: Simulation of the nations of attack sources in order to monitor assaults during a target application, the proposed technique can concentrate on different connection types. Figure A-1 plots a part of remain- ing analysis results of uplink speed by different connection types of more nations with malicious access for further reference. As a target application, the proposed technique can concentrate on the nations of attack sources in order to monitor assaults during certain events, such as hybrid cyber wars, Olympic games and bitcoin trading, etc. Figure A-2 plotted an example of the NDT dataset, with columns related to uplink speed estimation, including “Mean-ThroughputMbps”. We focus on measurement-lab:ndt.unified,
since this table can be used to crosscheck the infected hosts to get the connection speed, which could be a criteria to evaluate effectiveness to send security notification to infected users. Uplink speed values, before and after taking countermeasures to infected IoT devices, are used to calculate the internet resource mitigation ratio with simulation in our work [30].

Xuping Huang received B.S. and B.A. degrees from the Department of Software Science, Dalian JiaoTong University, China in 2007 and an M.S. degree from the Department of Information Science, Graduate School of Iwate Prefectural University, Japan in 2009. Since 2009, she has been a Ph.D. candidate at the Graduate University for Advanced Studies (SOKENDAI) and a research assistant at the National Institute of Informatics, Japan. She is currently an assistant professor at the Advanced Institute of Industrial Technology and a visiting lecturer in Yokohama National University. Her research interests include information hiding and audio signal processing. She is a member of IPSJ.

Shunsuke Mochizuki received his B.S. and M.S. degrees in Chemistry from Tokyo Metropolitan University in 2004 and 2006, respectively. Since 2006, he has been a researcher at NTT DATA Mathematical Systems Inc. in Japan. His research interest is computer simulation in various fields of science and engineering.

Katsunari Yoshioka received his B.E., M.E., and Ph.D. degrees in Computer Engineering from Yokohama National University in 2020, 2002, and 2005, respectively. From 2005 to 2007, he was a researcher at the National Institute of Information and Communications Technology, Japan. Currently, he is an Associate Professor for Division of Social Environment and Informatics, Graduate School of Environment and Information Sciences, Yokohama National University. His research interests cover a wide range of information security, including malware analysis, network monitoring, intrusion detection, and information hiding.