

Network Flow Data Re-collecting Approach Using 5G Testbed for Labeled Dataset

Gyeoul Lee, Jonghoon Lee, Youngsoo Kim, Jong-Geun Park

Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

{merrywinter, mine, blitzkrieg, queue}@etri.re.kr

Abstract— With the emergence of fifth generation technology (5G) environments, intelligent network security against constantly evolving attacks related to massive Internet of Things devices, user equipment, and various edge services has become more important. Moreover, to employ state-of-the-art learning-based detection methodologies, a labeled dataset is essential. However, it is not easy to obtain such a dataset by collecting the real communication dataset for a 5G network. Hence, in this study, we build a purpose-built 5G testbed that can observe 5G network features by replaying the collected data. Additionally, we implement a specialized network collector system that can 5G edge network. Subsequently, the network traffic data collected in the configured 5G testbed are analyzed. It is discovered that a re-collecting methodology using the proposed 5G testbed and network collector can be sufficiently utilized to construct a 5G-based labeled dataset for supervised learning methods.

Keywords— 5G, GTP, Network Collector, Testbed, AI Dataset

I. INTRODUCTION

Fifth generation technology (5G) research and development aim to achieve various advanced characteristics, such as a capacity higher than that of the current 4G technology, a higher density of mobile broadband users, and support for device-to-device communications and massive machine-type communications [1]. In addition, 5G services are an essential communication infrastructure technology for the future of the fourth industrial revolution and are expected to revolutionize various services and areas compared with the previous 4G LTE [2,3].

5G research and development aim to achieve various advanced characteristics, such as a capacity higher than that of the current 4G technology, a higher density of mobile broadband users, and support for device-to-device communications and massive machine-type communications [1]. In addition, 5G services are an essential communication infrastructure technology for the future of the fourth industrial revolution and are expected to revolutionize various services and areas compared with the previous 4G LTE [2,3].

Hence, in this study, to provide a learning dataset by collecting network data in real environments for AI methods, we built a 5G testbed that can observe 5G network features by replaying the collected data.

The remainder of this paper is organized as follows. In Section II, we present the background information and

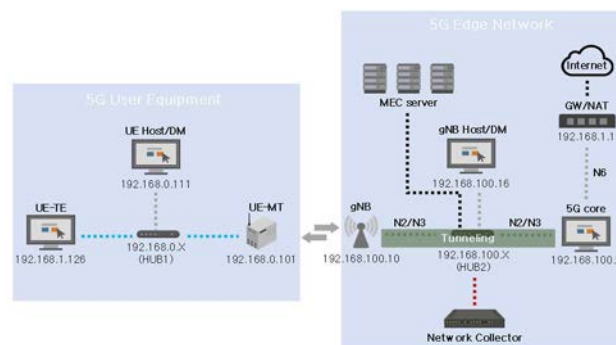


Figure 1. Architecture of 5G testbed for network security

architecture of the 5G testbed for network security. Section III provides the 5G network data used for the testbed. In section IV, we present the data collected in a model factory and re-collecting environments. In Section V, we describe the labeling methodology using the implemented testbed. Finally, the conclusions and future work are discussed in Section VI.

II. 5G TESTBED ARCHITECTURE AND ENVIRONMENT

In this study, to observe 5G features for network security, we implemented a 5G testbed that can collect the network flow data by replaying various network data in 5G environments. In this section, we present the background and main systems for the 5G environment and the configured 5G testbed, as shown in Fig. 1.

To build the test environments for 5G network features, we designed and constructed a testbed that is primarily composed of three systems: a 5G Core network (5GC), gNodeB (gNB), and user equipment (UE) as follows:

1) 5GC: The 5G core network manages subscriber information such as the subscriber's number and current location; it serves as a server for connecting to a wired telephone network service and providing other additional services. In addition, it manages the entire network and access networks by region. The primary network functions in the 5G core network are as follows:

- Access and mobility management function (AMF): The function is applied to manage access control and mobility, and it is implemented in MME for legacy cellular networks.



Figure 2. Photo of configured 5G testbed

- Session management function (SMF): Based on network policy, this function can establish and manage sessions. For a single AMF, multiple SMFs can be assigned to manage different sessions of a single user.

- Policy control function (PCF): This function provides roaming and mobility management, quality of service, and network slicing. The AMF and SMF are controlled by the PCF. Furthermore, the PCF can provide differentiated security.

2) **gNB**: The access network manages subscriber number processing, service connection, and information transmission and reception, and it is directly connected to the subscriber. The main base of this access network is gNB, which is installed by a mobile communication service provider, and the cell is a range that allows the gNB to provide service. Adjacent gNBs can provide a handover technology that provides continuous service even when a customer's location changes, i.e., when a user moves from one cell to another.

3) **UE**: Each UE has a unique number, and in cases without a call, it sends a continuous signal to gNB to recognize its own number. The transmitted terminal information is registered in the 5G core network such that the user's current location can be identified.

We collected network flows from a model factory that uses 5G networks. Subsequently, the collected data were replayed on the 5G testbed described above.

Based on the 5G network structure, the implemented 5G testbed comprised a UE-MT for wireless communication, a UE as a user terminal, a UE-DM for verifying the UE's status, a gNB as a base station, and a gNB-DM for verifying the gNB's status. Fig. 2 shows the components for our implemented testbed.

The UE-MT and gNB are radio sections that are connected to the AMF through the N2 interface, and a data transmission device is connected to the UPF through the N3 interface. In the 5G testbed, SCTP and UDP protocols are primarily used in the gNB and 5G C sections, and data passing through the section is tunneled to the UDP with a 4-byte header.

The backhaul network comprises network elements and communication channels between the gNB as the base station and 5GC as the core network. In particular, similar to 4G LTE, these channels are encapsulated by a tunneling protocol, which is typically known as GTP. Because of the tunneling mechanism, when the network collector captures the network raw data, the original end-to-end network information is contained to the payload data of the captured data instead of to the IP. Therefore, to be aware of the original network

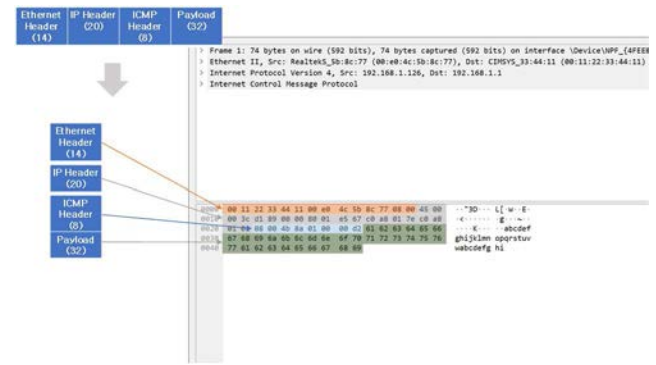


Figure 3. Normal ping packet

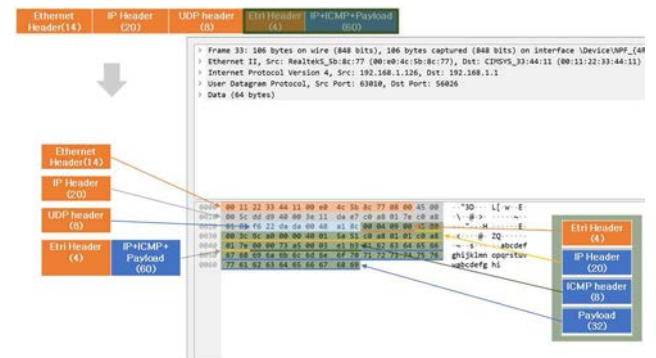


Figure 4. Tunneled ping packet

information, the tunneled header must be eliminated, except in some cases. In the case of control plane data similar to SCTP, because they are not designed to be encapsulated, the tunneling header must be maintained.

Hence, we developed a specialized network collector system that can extract the original end-to-end network features to collect network traffic through tunneling in a 5G network. The network collector was positioned between the gNB and 5GC. In the testbed, both normal and malicious traffic between the UE and the application server were collected by the network collector in the backhaul scope.

For network data re-collecting environments, we used Tcpreplay through the testbed.

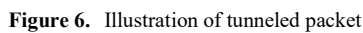
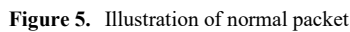
III. 5G NETWORK DATA IN TESTBED

A. Common IP packet in 5G testbed

We performed an ICMP ping test to the gateway from one host that is positioned in the core network scope in the testbed. We observed that the ICMP packet comprised eight bytes for the ICMP header, 20 bytes for the IP header, 24 bytes for the Ethernet header, and 32 bytes for the payload, as shown in Fig. 3.

B. Tunneled packet in 5G testbed

Figure 4 presents the structure of the tunneling packet. The structure of the tunneling packet captured between the base station (gNB) and the 5G core network comprised 20 bytes for the IP header, eight bytes for the UDP header, and four bytes for the tunneling header, similar to the GTP. The original data



C. Comparison between normal and tunneled packets

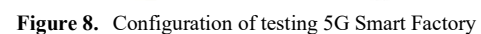
As shown in Figs. 5 and 6, a typical ping packet comprises a payload, which is transmission data, 14 bytes for an Ethernet header, 20 bytes for an IP header, and eight bytes for an ICMP header. In the 5G network, the ping packets collected from the gNB and the 5GC section are encapsulated with the tunnel header. Additionally, it contains the 5GC and gNB's IPs as the source and destination IPs, respectively.

To replay the 5G testbed built by collecting actual data from the model factory, an additional task of removing the tunnel header is required. This task was performed by the network collector.

Owing to the characteristics of mobile communication, only registered users (telecommunication customers) can use the network.

Even in the case of a packet whose tunnel header has been removed, if information (IP and MAC addresses) that is not registered in the 5GC is discovered, it must be converted into information (telecom customer IP address and MAC address) registered in the corresponding 5GC.

For real data collection, we positioned the network collector in one model factory, as shown in Fig. 8. The network collector collects traffic between the base station gNB and 5GC (IP: 192.168.100.10 - IP: 192.168.100.2). The source and destination IPs of all packets are the gNB and the 5GC’s IPs, respectively. The 5G network encapsulates the original source



IP and destination IP with a 4-byte tunnel header in the UDP protocol. Fig. 9 shows the structure of packets collected in the model factory.

Building a successful machine learning model that can recognize cyber threats requires large amounts of training data. Building labeled datasets for 5G network security is significant and essential. Therefore, we collected network data from the 5G-based real smart-factory to build the datasets. Subsequently, we replayed the traffic on the 5G testbed that we built and re-collected the generated traffic.

The network collector collects traffic between the model factory's gNB and the 5GC. It is capable of storing the traffic data after removing the tunnel header of the packet.

By eliminating the GTP tunnel header of the collecting packet, the original source IP and the destination IP can be extracted for analyzing network flow.

Even if the tunnel header has been removed, the network cannot be used if it is not the information (IP and MAC addresses) of the carrier user owing to the characteristics of

mobile communication. Hence, we modified the source IP and source MAC to the information (IP and MAC addresses) registered in the 5GC. In addition, because the UE performs wireless communication with the gNB through the UE-MT, we changed the destination information to the IP and MAC addresses of the UE-MT.

We implemented modified packets on the 5G testbed which comprised a user terminal composed of N virtual machines, a 5GC, and a gNB that replayed normal and abnormal data collected based on the 5G network. Subsequently, the data generated in the testbed were re-collected by the collector. By repeating collection and replay, a dataset for AI-based intrusion detection was created.

B. Replaying in Testbed

In response to the increasing 5G security threat, a dataset for AI learning must be built. To build a dataset for AI learning, data were collected from the model factory. Subsequently, the data collected were replayed and collected on the 5G testbed.

In Fig. 9, the IPs of the source and destination of packets collected in the gNB and the 5GC section were the IPs of the gNB and 5GC, respectively. Because they are tunneled, they cannot be executed in other environments even when the flow is collected.

Figure 7 shows the flow with the tunnel header removed by the collector. Even if the tunnel header is removed from the collector, the data collected from the existing network collector cannot be reused because it has an IP address and a MAC address that are not registered in the 5GC. In this study, the tunnel header was removed using a network collector in the tunneling section. Subsequently, by modifying the packet based on the mobile communication features, the data used in the actual 5G network were applicable to the 5G testbed.

In the experiment, the data generated in the model factory were applied to the testbed. We observed that the generated data can be replayed and re-collected on the 5G testbed to build a dataset.

V. USER TOOL FOR DATA LABELING

We implemented a dataset conversion tool that can assist in labeling the data records, as shown in Fig. 10. The operating procedure of the dataset conversion tool for labeling data records is as follows:

First, in one model factory comprising a gNB, 5GC, and UE, we collected the network data by performing simulated attacks using an actual UE. At this time, the network collector instantly stored raw network data using data collected through simulated attacks and abnormal behaviors in the same manner as in actual collections.

Second, in the 5G testbed comprising a UE composed of N virtual machines, a gNB, and a 5GC, we replayed normal and abnormal data collected in the 5G network. The network collector stored the data collected at this time. In addition, using a network packet replay program such as tcpreplay, it must change the packet's source IP and source MAC, destination IP, and destination MAC. Subsequently, they were transmitted to the 5G tunneling section, which is known as backhaul.

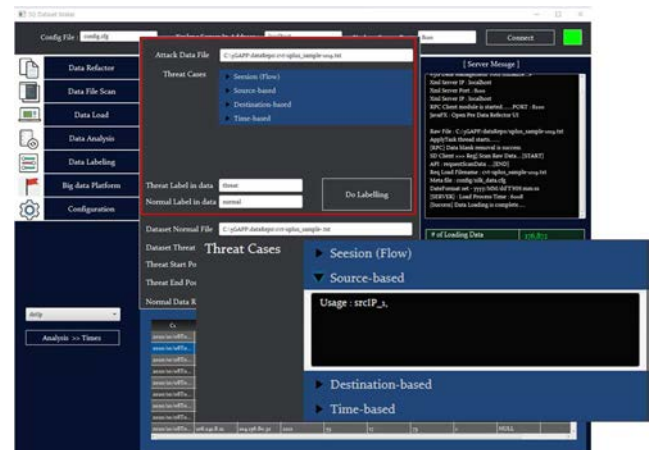


Figure 10. Dataset conversion tool GUI for data labeling

Third, when the replayed and collected data exhibit similarity exceeding a threshold value during a simulated attack, the data can be labeled as attack data. In addition to this similarity threshold, our implemented data tool can perform labeling based on network session information, the source IP group, the destination IP group, and attack timestamp information.

VI. CONCLUSIONS

Herein, we proposed a testbed specialized for 5G network data collection that can replay attack data and ensure network security. Additionally, a network collector was designed and implemented to address the limitation due to the tunneling issue when extracting network flows. The purpose of the testbed was to collect real 5G data and build a dataset for AI learning using the collected data. The collected network data can be converted to a labeled dataset using the implemented dataset conversion tool.

We collected network data by performing simulated attacks using an actual 5G user equipment in one smart factory, and by replaying the collected data, we iteratively collected the network data. Using the data from the proposed methodology, we constructed a labeled dataset that can be utilized as learning data for supervised learning methods. In the future, to address the evolving problem of 5G edge cyberattacks, we will develop a 5G security solution for enhancing intelligent network security using the proposed re-collecting approach and various AI technologies.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00952, Development of 5G Edge Security Technology for Ensuring 5G+ Service Stability and Availability).

REFERENCES

- [1] D. Fang, Y. Qian and R. Q. Hu, "Security for 5G Mobile Wireless Networks," in IEEE Access, vol. 6, pp. 4850-4874, 2018.
- [2] Y. Kim et al, "Security Threat and Response Technology for Multiaccess

- Edge Computing in 5G Environments," Communications of the Korean Institute of Information Scientists and Engineers, Vol.38, No.9, 2020.
- [3] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," in IEEE Journal on Selected Areas in Communications, vol. 35, no. 10, pp. 2181-2195, Oct. 2017.
 - [4] M. Shafi et al., "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," in IEEE Journal on Selected Areas in Communications, vol. 35, no. 6, pp. 1201-1221, June 2017.
 - [5] J. Lee, J. Kim, I. Kim and K. Han, "Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles," IEEE Access, vol. 7, pp. 165607-165626, 2019
 - [6] I. Ahmad, S. Shahabuddin, T. Kumar, J. Okwuibe, A. Gurtov and M. Ylianttila, "Security for 5G and Beyond," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3682-3722, Fourthquarter 2019.
 - [7] J. Park et al, "3GPP 5G Security Architecture Features and Major Improvements," The Korea Institute of Information Security and Cryptology, Vol.29, No.5, pp.21-30, 2019.
 - [8] I. Ahmad, T. Kumar, M. Liyanage, J. Okwuibe, M. Ylianttila and A. Gurtov, "Overview of 5G Security Challenges and Solutions," in IEEE Communications Standards Magazine, vol. 2, no. 1, pp. 36-43, MARCH 2018.
 - [9] R. Chaudhary, N. Kumar and S. Zeadally, "Network Service Chaining in Fog and Cloud Computing for the 5G Environment: Data Management and Security Challenges," in IEEE Communications Magazine, vol. 55, no. 11, pp. 114-122, Nov. 2017.



Gyeoul Lee received the M.S. degrees in Computer and Software from Hanyang University, Korea, in 2019. She has been involved in various research projects for network fields, and she is currently working as a associate researcher trainee of Cyber Security Research Division of the ETRI. Her research interests include mobile network security, cloud security and learning-based network intrusion detection.



Jonghoon Lee received the M.S. and Ph.D degrees in Computer Engineering from Kyungpook National University, Korea, in 2002 and 2020, respectively. He joined the Electronics Telecommunications Research Institute (ETRI) in 2002. Since 2002, he has been involved in various research projects for cyber security and network fields, and he is currently working as a principle researcher of Cyber Security Research Division of the ETRI. His research interests include cyber security, AI-based 5G edge security, 5G network security, learning-based network intrusion detection, and bigdata analytics for 5G security



Youngsoo Kim received his M.E. and Ph.D. degree in Computer Engineering from Sungkyunkwan University, KOREA, in 2000 and 2009, respectively. He is a Principal Researcher of the Electronics and Telecommunications Research Institute (ETRI) since 2000 and an Educational Tutor of Institute of Information Security Education of Kongju National University. He has been an Adjunct Professor in Computer Science and Engineering Department at Chungnam National University for 2012-2015. His research interests include Cryptographic Algorithms and Protocols, Digital Forensics, Cloud Security, Network Security, Cyber Security and 5G Security.



Jong-Gun Park received his B.S. and M.S. in industrial engineering from Sungkyunkwan University, Korea, in 1997 and 1999, respectively. He also received his Ph.D in computer engineering from Chungnam National University, Korea, in 2013. From 1999 to 2001, he was a researcher at Agency for Defense Development (ADD), Daejeon, Korea. Then, since 2001, he has been working at ETRI, Daejeon, Korea, as a principal researcher. Currently, he is interested in mobile network security, SDN/NFV and cloud security.