Determining The C&C Protocol Utilized By The Malicious Botnet Network

Posted on October 17, 2018

Abstract

Botnets are a network of malicious software headed by a botmaster, from which they get all the instructions to execute. In this paper, a novel technique is proposed to infer malware network specifications given a sample of malware binary, which is a file consisting of instructions. (bin file is with .bin execution). Botmasters use Command-and-Control protocols to control malware-infected hosts for executing malicious activities. Every family of malware has its own set of instructions called fingerprints, which are followed to execute malicious activities. In the paper, they tried to extract the fingerprint, which acts as a unique identity for the malware family. They have employed reverse engineering in which they explore the messages and infer all the fields in them. But before applying the inference algorithm to the messages, they need to be decrypted since they are encrypted. For this purpose, they have proposed a system to extract C&C encryption keys by applying the dynamic analysis of the malware binary (MITRE, 2026; Shin et al., 2021). So basically, they're trying to extract the protocol specification from binary malware communication samples and decrypt them before applying the inference algorithm. So basically, the main two contributions of their work are:

1. A technique for rich malware protocol inference that requires only decrypted network traffic (you'll obviously need to decrypt the encrypted traffic before applying the inference algorithm).
2. Proposing an encryption algorithm that is used by the malware networks. This obviously is needed for the first step.

Critical Review Of Past Work

The main concept employed to achieve the above-defined steps is Reverse Engineering. In this field, quite a lot of work has already been done, but since it is a very vast field, I would divide it into three sub-categories and highlight the respective work done in that area. The first category generates the specifications of the protocol from the network. In 2005, C. Leita, along with other members, introduced ScriptGen, which observes the traffic over the network and then makes a state machine based on that which produces approximate responses to the protocol requests. W. Cui, in 2007, proposed a technique for reverse engineering by inferring the message formats using pre-defined sets of field semantics. However, it has been quite rare for binary protocols to be explored. Also, in the previous work, the encryption has not yet been dealt with, while the proposed approach in this paper does. Also, the protocol specifications that are inferred from this paper are quite detailed and

include rich field types that are not pre-defined and generic (Li et al., 2023).

The second group of reverse engineering deals with the technique used to analyze the traces of execution. Those traces are run as a program, and their behavior is analyzed, including how they are used to communicate over a network. A lot of work has been done in this area. G. Wondracek, in 2008, proposed a technique through which one can generate the network specifications by running an algorithm during the process of the message. This approach had its focus mostly on the server side which is a bit difficult to access in the world of malware. J. Caballero in 2009 attempts to work on the malware binary, He employed an automatic reverse engineering algorithm for C&C protocols through the dynamic analysis on the executable binary files containing malware. He called this Dispatcher. He analyzed the data coming from the network and derived a detailed description of the semantics in the message field by first identifying the function prototype of the data. Z. Wang in 2009 considers the encrypted protocols too, where they employ encryption buffers fed with the plain-text information extracted from the incoming data. The proposed method in this paper is a little different from Dispatcher since the technique can learn and make generic encryption details through which all kinds of samples of the messages can be analyzed. The paper also deals with signature generation and the decryption of messages that are delivered online. Instead of ever-long algorithms and files, the proposed technique only addresses passive binary malware files, which can also be avoided if the same malware binary is encountered twice (Shin et al., 2021; Yu et al., 2022).

The third type of work is based on hybrid approaches. Many people have worked in this area as well, highlighting the benefits of amalgamating several approaches. P. M. Comparetti, in 2009, proposed Prospex, which can infer the malware protocol for the analysis of traffic being communicated over the internet and the traces of execution. However, the biggest distinction between the proposed approach in this paper and Prospex is that the latter cannot handle the encrypted data. Most of the data communication is done over the internet and is encrypted. It is the foremost step to first decrypt the message and then infer the protocols from them (Shin et al., 2021).

A lot of approaches have also been employed to generate the signatures or unique patterns that can be used to identify different families of malware. One of the closest works to the one proposed in this paper is ProVex. C. Rossow, the author of ProVex, proposed the technique to detect botnet malicious networks through encrypted C&C channels. Encryption and decryption are also catered to. It decrypts the packets through pre-defined algorithms for encryption. After that, statistics about how the bytes are distributed in the payload are extracted. The difference with the proposed technique lies within the encryption keys. In ProVex, prior knowledge of encryption keys is required. In addition to this, the signatures generated are very probabilistic, which makes them more prone to getting false positive results. This issue is also catered to in the proposed paper (MITRE, 2026; Li et al., 2023).

Conclusion

The overall proposition stated in the paper presents a good solution to infer the protocol of C&C being

used in the malicious network of botnets. It also helps to alleviate the task of manually understanding malicious communications by providing detailed specifications of the protocol (Yu et al., 2022).

References

Li, J., Cheng, G., & Yang, G. (2023). Private protocol reverse engineering based on network traffic: A survey. *Journal of Computer Research and Development, 60*(1), 167–190. https://crad.ict.ac.cn/en/article/cstr/32373.14.issn1000-1239.202110722

MITRE. (2026). *Malware content, data component DC0011*. MITRE ATT&CK. https://attack.mitre.org/datacomponents/DC0011/

Shin, K., Jung, D., Choi, M., & Cho, H. (2021). A study protocol reverse engineering research trend and construction of optimal environment for malware analysis. *Journal of the Korea Institute of Information Security & Cryptology, 31*(2), 175–186. https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002711181

Yu, T., Xin, Y., Tao, Y., Hou, B., & Zhu, H. (2022). Network communication protocol reverse engineering based on auto-encoder. *Security and Communication Networks, 2022*, 2924479. https://onlinelibrary.wiley.com/doi/10.1155/2022/2924479