Tor Hidden Services Discovery and Analysis: A Literature Survey

Tor Hidden Services Discovery and Analysis：A Literature Survey

Jingjing Bian[1,3], Chunjie Cao[2,3,*],Longjuan Wang[2,3], Jun Ye[2,3], Yan Zhao[4], Chaosheng Tang[2,3]

[1]School of Information and Communication Engineering,

[2]School of Computer Science and Cyberspace Security,

[3]Hainan University, No.58, Renmin Road, Haikou 570228, China

[4]Hainan Provincial Public Security Department, No.09, Binya Road, Haikou 570228, China

*18085208210001@hainanu.edu.cn

Abstract: Hidden services are a feature of Tor(The Onion Router)[1]. It provides anonymity for the service requester while maintaining the anonymity of the service provider. Since it is quite difficult to trace back and locate both parties in the communication, the criminals use hidden services mechanisms to construct various illegal activities in the darknet, which has brought adverse effects to society. In order to prevent the abuse of Tor hidden services, the discovery and analysis of hidden services are particularly important. The aim of this survey paper is to review and compare the literature of the past five years, provide the readers with methods for discovering tor hidden services, along with the various content analysis methods developed and proposed from time to time. we explain their key ideas and show their interrelations.

Keywords: The Onion Router(Tor), anonymous communication, hidden services, darknet

1. Introduction

Tor's core technology "onion routing" was first proposed by the US Naval Research Laboratory, which released its official open source version in 2004, and since then it has supported hidden services. Its strong anonymity and confidentiality make it an opportunity for criminals while not exposing user

privacy, making it a hotbed for some cybercriminals. Compared with cybercrime on the surface network, cybercrime on the darknet is more concealed and more difficult to detect, which poses a huge challenge to the supervision of law enforcement agencies. The aim of this survey paper is to review the literature on the discovery and analysis of hidden services in the past five years.

The rest of this paper is organized as follows. In Section 2, we introduce the basic principles of the Tor hidden services. For the Tor hidden services: In Section 3, we present the methods of collecting domain names, along with their comparison. In Section 4, we present the types of analysis content and core algorithms. The paper is concluded in Section 5.

2. Tor Hidden services

Since 2004, Tor supports hidden services. The server can generate a domain address in the form of "ABC.onion" based on its own public key, and the client can use this domain to access the Tor hidden services:

①Server randomly selects some nodes (three by default) as Introduction Point(IP), and informs them of their public key information when establishing circuits with them.

②Server will generate a hidden service descriptor, which contains the public key of the hidden service, IP information, etc. The descriptor will be uploaded to the Hidden Service Directory Server.

③Before accessing the hidden service, the client must first obtain the domain address of the server, and then randomly select a Hidden Service Directory Server to download the hidden service descriptor corresponding to the address. The client will obtain the public key and IP information of the hidden service from the descriptor.

④Then the client will select a node as Rendezvous Ponit(RP) and establish circuits to this node.

⑤Use the public key of the server to encrypt the selected RP information and a one-time secret cookie and send it to one of the IP nodes selected by the server. After the IP2 passes the verification, the encrypted information is forwarded to the server.

⑥The server decrypts the message forwarded by IP2, extracts the cookie and RP information, and establishes circuits to the RP

The RP uses a cookie to verify the identity of the server. After the verification is passed, the data is forwarded to the client. At this time, the client and the server complete the handshake through the RP.

3. Address Collection

The Tor domain name address is different from the surface network. It has no regularity and is almost impossible to remember. The form of the Tor domain address is "ABC.onion"， when the protocol version is V2, ABC is composed of 64 letters, which we call short links; when the protocol version is V3, ABC consists of 128 letters, which we call long links. Before accessing a specific hidden service, we must first obtain the corresponding domain address.

In view of previous studies, we present the methods of collecting domain names, along with their comparison, which mainly focuses on the surface network, dark network, and server deployment, including their combination, as shown in Tab.1.

Year	Proposal	CR			IT
		SN	DN	SD	
2016	Li et al.[2]	√	×	×	Tor2web
2016	Owen et al.[3]	×	×	√	DHT
2017	Mohaisen et al.[4]	√	×	×	DNS resolution
2017	Yang et al.[5]	√	√	×	the search engine
2018	Marques et al.[6]	×	×	√	Descriptor segmentation

Tab.1 Comparison of Tor hidden service domain name collection methods

Note：CR(Collection Resources), SN(Surface Network), DN(Dark Network)SD(Server Deploy ment), IT(Implementation Technology) "√" means the corresponding CR is applied；"×" means the opposite of "√"

Based on comprehensive analysis, we believe that the following problems still exist in the collection of hidden service domain names: For hidden services that are not configured with Tor2web, they cannot be searched in the surface network's search engine, and the hidden services are obtained from the DNS has large randomness. At present, the collection of the tor hidden service domain name address is mainly concentrated on the short links of the V2 protocol. After the algorithm improvement of the V3 protocol, the acquisition is improved. The difficulty of the long links of the V3 protocol.

4. Content Analysis

Having collected some of the hidden service addresses from the previous section, we can use them as data sources for content classification and analysis steps.

Based on the previous studies, we present the types of analysis content and core algorithms. We find that more than half of the literature were analyzed for text content, and there were also structural analysis and image analysis for the Tor darknet. In terms of core algorithms, some researchers have proposed new algorithms, and some have promoted and applied the old algorithms, they provide a good foundation for tor hidden services content analysis. Details are shown in Tab.2.

Year	Proposal	FO			CA
		CC	IC	SA	
2016	Nunes et al.[7]	√	×	×	Label propagation (LP)[21]，Co-training (CT)[22]
2016	Kaati et al.[8]	√	×	×	random indexing[23]
2017	Sanchez-Rol a et al.[9]	×	×	√	Graph Analysis, web tracking analyzer[24]
2017	Ghosh et al.[10]	√	×	×	Term Frequency – Inverse Corpus Frequency（TF-ICF[25], Automated Tool for Onion Labeling

					（ATOL） Classifier
2017	Al Nabki et al.[11]	√	×	×	Bag-of-Words（BOW）[26], Term Frequency Inverse Document Frequency model（TF-IDF[27], Support Vector Machine (SVM)[28],Logistic Regression(LR)[29], Naive Bayes (NB)[30]
2019	Al Nabki et al.[12]	√	×	×	ToRank
2019	Yoon et al.[13]	√	×	×	content grouping algorithm
2019	Takaaki et al.[14]	√	×	×	Naive Bayes (NB)[30]
2019	Trivedi et al.[15]	×	×	√	in-degree and out-degree Sublink-based Analysis (SLBA) Sublink with Keyword-based Analysis (SLKBA)
2019	Bernaschi et al.[16]	×	×	√	Graph Analysis
2020	Jeziorowski et al.[17]	×	√	×	image hashing, Structural Similarity Index Metric (SSIM)
2020	Faizan et al.[18]	√	×	×	Named Entity Recognition（NER）
2020	Biswas et al.[19]	×	√	×	Frequency-Dominant Neighborhood Structure (F-DNS)
2020	Al-Nabki et al.[20]	√	×	×	Local Distance Neighbo（LDN）

Tab.2 Types of analysis content and comparison of core algorithms.

Note：FO(Focus), CC(Contextual Content), IC(Image Content), SA(Structure Analysis), CA(Core Algorithm) "√" means the corresponding FO is applied； "×" means the opposite of "√"

Based on comprehensive analysis, we consider that due to the timeliness of Tor hidden services links and the repeatability of the content, the quality of obtaining Tor hidden services content is closely related to the link sources. In addition, it is still very difficult to obtain information about the most inaccessible and unpublished sites on the darknet.

5. Conclusion

Throughout the recent 5 years of Tor hidden services discovery and analysis related research, it is mainly focused on the analysis and classification of text content, including the classification of illegal text content and picture content, aiming to automatically identify illegal content in the Tor darknet and obtain timely warning. However, the supervision of Tor hidden services still faces great difficulties

and challenges at present, such as location positioning of cloud server users, crawling of site contents requiring login, dynamic monitoring of Tor hidden services status, and other aspects, which still need in-depth and extensive research. The domain addresses of Tor hidden services are acquired in a way that is quite different from those of the surface network and generally not regular. Therefore, the acquisition of data link sources in the first step is particularly important in the whole analysis framework.

Acknowledgments
This work was supported by Natural Science Foundation of Hainan Province under grant no.618MS025 and National Natural Science Foundation of China under grant no. U19B2044.

References

[1] Roger D, Nick M and Paul S 2004 *Proc. Symp. on USENIX Security* (San Diego:Naval Research Lab Washington DC)
[2] Kang L, Peipeng L, Qingfeng T, Jinqiao S, Yue G and Xuebin W 2016 *Proc. Symp. on* Applied Computing (Italy:ACM) pp2057-2062
[3] Gareth O and Nick S 2016 *J. Iet.Inf. Secur* **10** p113
[4] Aziz M and Kui R 2017 *J. IEEE.Acm.T.Network* **25** p3059
[5] Yi Y, Han G, Yijun W and Zhi X 2017 *J. Communications Technology* **50** p2304
[6] João M, Leandro V and Rik v D 2018 *Universiteit van Amsterdam*
[7] Eric N, Ahmad D, Andrew G, Ericsson M, Vineet M, Vivin P, John R, Jana S, Amanda T and Paulo S 2016 *Conf. on Intelligence and Security Informatics* (Tempe:IEEE) pp7-12
[8] Lisa K, Fredrik J and Elinor F 2016 *Int. Conf. on Cybercafe and Computer Forensic* (Stockholm:IEEE) pp1-7
[9] Iskander S, Davide B and Igor S 2017 *Proc. of the 26th Int. Conf. on World Wide Web* pp 1251-1260
[10] ddShalini G, Phillip P, Vinod Y,Ken N and Ariyam D 2017 *Conf. on Artificial Intelligence*
[11] Mhd Wesam A, Eduardo F, Enrique A, and Paz I 2017 *Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics* pp 35-43
[12] Mhd Wesam A, Eduardo F, Enrique A, and Laura F 2019 *J. Expert Systems with Applications* **123** p212
[13] Changhoon Y, Kwanwoo K, Yongdae K, Seungwon S and Sooel S. 2019 *Conf. on The World Wide Web* pp2225-2235
[14] Sugiu T and Inomata A 2019 *Proc. the ACM International Workshop on Security and Privacy Analytics* pp53-59
[15] Tarun T, Vinod P, Manas K and B M M 2019 *M.in Data Mining and Information Security* (Singapore:Springer) pp567-578
[16] Massimo B, Alessandro C, Stefano G, Flavio L and Enrico M 2019 *conf. On The World Wide Web* pp105-115
[17] Susan S, Muhammad I and Ambareen S 2020 *Proc. of the Sixth International Workshop on Security and Privacy Analytics* pp15-22
[18] Mohd F, Raees Ahmad K and Alka A 2020 *J. Applied Computing and Informatics*

[19] Rubel B, Víctor G, Eduardo F and Enrique A 2020 *Neurocomputing Elsevier* **383** p24
[20] Mhd Wesam A, Eduardo F, Enrique A and Laura F 2020 *Neurocomputing* **382** p1
[21] Xiaojin Z, John L and Zoubin G 2003 *Conf. on he continuum from labeled to unlabeled data in machine learning and data mining*
[22] Blum A and Mitchell T 1998 *Proc.Conf.on Computational learning theory* pp92-100
[23] Magnus S 2005 *Conf.on terminology and knowledge engineering*
[24] Iskander S and Igor S 2016 *University of Deusto Technical report*
[25] Joel W R, Yu J, Thomas E P, Brian A K, Mark T E and Ali R H 2006 *Conf. on Machine Learning and Applications* pp258-263
[26] Yin Z, Rong J and Zhihua Z 2010 *J.Int.J.Mach.Learn.Cyb* **1** p43
[27] Akiko A 2003 *J.Inform.Process.Manag* **39** p45
[28] J A K Suykens and J Vandewalle 1999 *J.Neural.Process.Lett* **9** p293
[29] Steven L G 1994 *J.Contemporary sociology* **23** p159
[30] Andrew M and Kamal N 1998 *Conf. on learning for text categorization* vol **752** pp41-48