FINDING PATTERNS OF DNS PACKETS FROM THE CLUSTERED INTERNET TRAFFIC DATA

Rohini Badgujar*

Shrikant Lade**

Abstract:

The Domain name Service (DNS) offers a discriminating capacity in pointing Internet targeted visitors. Shielding DNS hosting space from move speed approaches is supported from the capability for you to viably my personal DNS log information intended for measurable designs. Processing as well as handling DNS log information can be classified because data-intensive issue, and the idea represents issues special to this particular class involving issue. On the point as soon as issues come about in finding log info, or if the DNS server encounters a blackout (booked or maybe unscheduled), the standard pattern involving traffic for your server obtains blurred. Basic direct insertion from the openings in the information isn't going to save gimmicks, by way of example, crests with traffic (which can happen throughout the ambush, doing them especially note). In this particular paper a novel regular trend mining method is proposed, or a periodic tendency pattern based traffic conjecture method. Clustering is adopted for you to partition numerous websites into different groups from the characteristics in their query targeted visitors time series.

Keywords: Clustering, Data mining, Pattern extraction, Pattern matching, Rule mining etc.

* M.Tech Scholar, Department of Computer Science and Engineering Engineering, RKDF Institute of science andTechnology. (R.G.P.V), Bhopal.

** Asst.Professor, Department of Computer Science and Engineering Engineering, RKDF Institute of science andTechnology. (R.G.P.V), Bhopal.

I. INTRODUCTION

The always expanding intricacy of current communication systems has tossed huge difficulties in adequately dealing with the

Divergent sub-components and administrations advertised. The pressing need felt by the organizations to address these difficulties has impelled powerful movement in this field. Traffic activity modelling has an essential influence in network management. A few methodologies have been proposed, each with its one of a kind qualities and shortcomings.

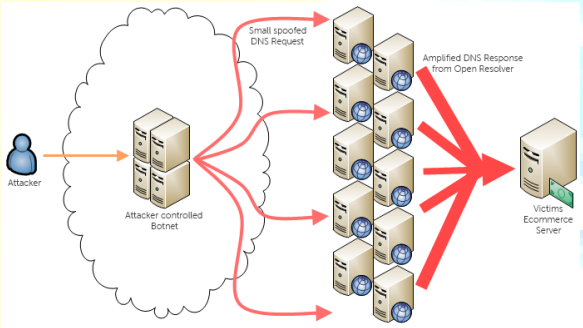

Figure 1: DNS amplification

The Domain Name Service (DNS) is a fine example of a framework that is profoundly vulnerable to denial of service (DOS) sort assaults. Various illustrations have been recorded, including [4]. At the point when examining DNS activity, catching log information and utilizing data mining methods to uncover patterns and other learning can prompt the improvement of apparatuses and systems for anticipating such attacks [2] [3]. Throughout the data mining methodology, care must be taken to check the nature of the source information. On account of DNS log information, one inadequacy that emerges is the vicinity of gaps in the information where for one reason or an alternate DNS server neglected to log the action throughout a period of time. Because of the trouble of acquiring log information, it is frequently impractical to basically acquire an alternate dataset, and information must be cleaned before it is utilized to make inferences.

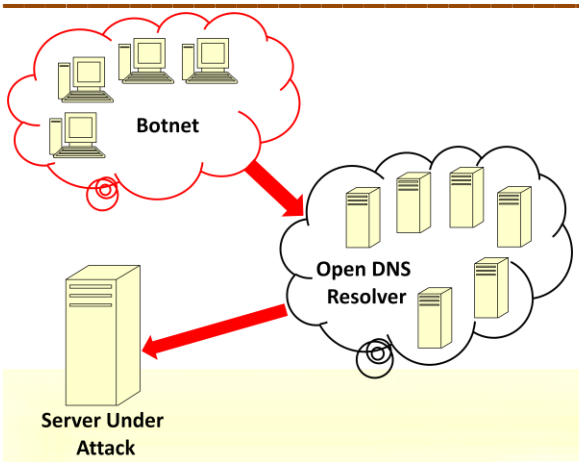

Figure 2: DNS DoS attacks

DNS amplification assaults are a manifestation of distributed denial of service (DDOS) attacks, where name servers are mishandled to send extensive replies to DNS questions to the exploited person. This is proficient by sending a small inquiry that will bring about an extensive reaction, utilizing a satirize source IP location to a name server. The proportion between the reaction and inquiry size is the amplification component. The name server then reacts with this huge reply to the satirize IP address, conceivably flooding the connection towards that.

II. INTERNET TRAFFIC DATA

Even though DNS is often a repository, above all it is a sent out repository. Every single DNS server is made up of simply a little part of the actual number title in order to IP target mappings (relative in order to the volume of information for the complete Internet).Each DNS server is actually configured using a exclusive report in which conveys to the actual DNS server exactly where (the IP handle involving another DNS server) it's going to perform a seek intended for information it won't get throughout the percentage of the actual DNS data source. For this reason agreement, every DNS server retains just a modest percentage of the overall DNS host to IP handle mappings.

Table 1:DNS Sample Data

No.,Time,Source,Destination,Protocol,Length,Info
"66","112.123110000","192.168.2.17","192.168.2.1","DNS","70","Standard query

0x389f A google.com"

"67","112.218235000","192.168.2.1","192.168.2.17","DNS","382","Standard query response 0x389f A 74.125.236.98 A 74.125.236.99 A 74.125.236.100 A 74.125.236.101 A 74.125.236.102 A 74.125.236.103 A 74.125.236.104 A 74.125.236.105 A 74.125.236.110 A 74.125.236.96 A 74.125.236.97"

The actual variety of host-name-to-IP-address mappings was comprised of with all the DNS database can be also called a namespace. Essentially, any time searching for a name throughout DNS, the actual DNS purchaser initial investigations a top-level DNS server database. That server explains to your customer which usually DNS server hosts another the main DNS name, as well as the purchaser and then inquiries in which server. This particular lookup-and-handoff process continues before the purchaser locates the actual DNS server in which hosts the actual DNS record showcased, knowing that server affords the IP address.

Figure 3: Internet traffic data notepad view

Hence, we will take this raw data for pre-processing and process it by making cluster of protocols based data, and then it will send to the pattern identification module. Where it will be processed and result will be generated.

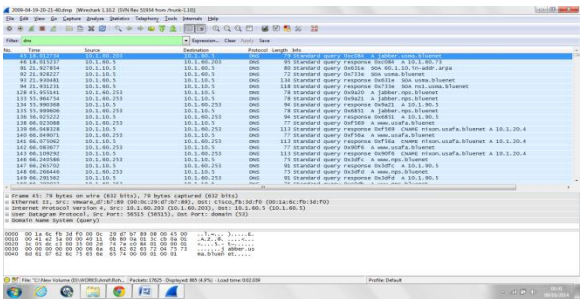

Figure 4: Reading Network Traffic using Wireshark

III. LITERATURE SURVEY

In [1], the authors concentrated on finding helpful patterns from the DNS inquiry activity information to help DNS service suppliers to find out peculiarities and get to take in the conduct patterns of the Web clients of distinctive domain names. Since DNS log information is tremendous in size and developing quick, data mining methodologies which can uncover designs alterably are embraced in this work.

Firstly, the authors proposed a novel issue, periodic pattern mining, and after that a forecast system was proposed to foresee the incoming network traffic volume and distinguish anomaly. Furthermore, the authors partitioned the query movement time arrangement information into divided groups by utilizing SNN grouping calculation, and along these lines further bits of knowledge of each one bunch are considered. The viability of the routines has been showed by the test comes about on a certifiable DNS question activity log dataset. Fascinating patterns are monitored which are valuable for the DNS service suppliers and the National Security Bureau.

Various works have discussed about the subject of filling missing values in information sets being utilized for data mining. In [7] numerous methods are depicted, including various ascription and hot deck attribution. These terms allude to producing a quality to remain in for the missing worth while the information set is constantly handled, and afterward keeping on transforming the information set as though the stand-in was the monitored value. This is like the methodology we are utilizing, aside from that we don't utilize stochastic or factual procedures, yet rather information of the normal patterns of a DNS server and scaling focused around other true perceptions in the time arrangement to create substitution values. This might be viewed as a pattern or model-based attribution system.

Different sources assess the estimation of these methods, for example, [6][8]. We could positively investigate the fluctuation and natures of methodology utilizing comparable systems yet don’t for quickness. Numerical investigation methods, for example, interpolation and linear regression are additionally a significant reference, but many of such methods do not incorporate the knowledge of the expected outcome or in their generality become inefficient or too expensive to employ.

Other testing strategies, for example, list wise deletion[5] are not adequate since almost every specimen has some measure of missing information, and the information we have is judged to be helpful in our issue area even with the gaps. Comparable issues were examined in [9].

Rozekrans and de Koning measured the adequacy of RRL [11] and figured out that more intricate attacks, e.g. strike where the inquiries are spread out over numerous names and zones, are not distinguished nor halted by RRL.

In their 2006 paper "Visualizing DNS Traffic" [10], Ren et al. describe many connected interactive visualizations that were created to realize insight into logs created by resolvers to spot potential security incidents with the users of these resolvers. They perform the visualization of the changes in users behaviour over short periods of time to offer a fast summary to permit the operator to observe the health of their users. In one amongst their case studies, they detected a user that was actively collaborating in SSH brute force attacks by seeing the amount of PTR by that user queries rise quickly.

Since the query traffic flow is a reflection of DNS service, anomaly detection in question traffic has been aid a lot of and a lot of attention. as an example, Carl Gustav Jung et al. [12] planned a unique methodology to discover anomaly in SMTP user by DNS query traffic. Ishibashi et al. [13] planned a way to get direct mail senders by learning ISP DNS. However in some circumstance, DNS itself will be a part of the attack in web like DDOS [14] and DNS cache poisoning [15].

Wang et al. [17] projected a mathematical methodology to discover nation-wide large-scale attacks on the web. A covariance matrix is made to record the variance between the query volumes happened at two completely different provinces at different time stamps. Average variance matrix indicates a traditional scenario. If the present variance matrix deviates from the typical variance considerably, an abnormal event could also be happening. This methodology is appropriate for a nation-wide attack however fails within the detection of attacks towards a selected name.

Xu et al. [18] improved ripper formula to find Botnet, which is commonly used for malicious activities (e.g., DDOS, spam, phishing etc.). It outperforms the standard algorithms, like options matching or applied math strategies, in discovering a lot of less-visited domain names.

IV. PROPOSED WORK

In the proposed work, we'll use the web traffic traces, a form of internet traffic log file that contains abstract information of all the packets. This traffic log file is than clustered and from the suitable cluster, a knowledge mining pattern identification process is applied to extract out the patterns for analysis. As we are employing a real-world DNS log dataset in our experiments. A DNS log record is generated

whenever a DNS request is issued by a client. For example, when a user browses a website name by its URL, a DNS log record are created and recorded within the log. Here within the proposed work, we'll show that the statistical properties of traffic corresponding to differing types of net traffic differ. A traffic type is characterized by variety of parameters, namely, packet length, packet inter time of arrival, connection period, connection packet count and connection byte count.

V. CONCLUSION

We started with a raw dataset of DNS logs capturing volumetricdata regarding frequency of requests of a set of DNS servers.However, this data was not useful in its raw form becauseof irrelevant and useless entries. After pre-processing the data will be used to perform robust, useful analysis.

The method can be applied to other problem domains involving time series data. We believe portions of this technique canalso be adapted to real-time monitoring of streaming data. By detailing areproducible method for analyzing server log data to identifyhabitual patterns for the traffic processed by DNS servers, wewill demonstrate a technique to assist in detecting attacks onDomain Name Service (DNS) servers that rely on statisticalanalysis.

VI. REFERENCES

1. WeizhangRuana, Ying Liub, RenliangZhaob, Pattern Discovery in DNS Query Traffic, "Information Technology and Quantitative Management", Procedia Computer Science-17, 2013, pp. 80-87.
2. SeongSoo Kim and A. L. Narasimha Reddy. Statistical techniques for detecting traf?c anomalies through packet header data. IEEE/ACM Trans. Netw., 16(3):562–575, 2008.
3. Keunsoo Lee, Juhyun Kim, Ki Hoon Kwon, Younggoo Han, and Sehun Kim. Ddos attack detection method using cluster analysis. Expert Syst. Appl., 34(3):1659–1665, 2008.
4. R. Naraine. Massive ddos attack hit dns root servers. www.internetnews.com/dev-news/article.php/1486981, October 2002.
5. Paul D. Allison. Missing data. Sage Publications, Inc., Thousand Oaks, CA, USA, 2002.
6. Marvin L. Brown and John F. Kros. The impact of missing data on data mining. pages 174–198, 2003.
7. Roderick J A Little and Donald B Rubin. Statistical analysis with missing data. John Wiley & Sons, Inc., New York, NY, USA, 1986.

8. IngunnMyrtveit, Erik Stensrud, and Ulf H. Olsson. Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. IEEE Trans. Softw. Eng., 27(11):999–1013, 2001.
9. W. Eric Wong, Jin Zhao, and Victor K. Y. Chan. Applying statistical methodology to optimize and simplify software metric models with missing data. In SAC '06: Proceedings of the 2006 ACM symposium on Applied computing, pages 1728–1733, New York, NY, USA, 2006. ACM.
10. Pin Ren, John Kristoff, and Bruce Gooch. "Visualizing DNS traf?c". In: Proceedings of the 3rd international workshop on Visualization for computer security. ACM. 2006, pp. 23–30.
11. T Rozekrans and J de Koning. Defending against DNS re?ectionampli?cation attacks. 2013. URL: http : / / www . nlnetlabs.nl/downloads/publications/reportrrl-dekoning-rozekrans.pdf.
12. Jung, J., Sit, E. An Empirical Study of Spam Traffic and the Use of DNS Black Lists. Proc. of the 4th ACM SIGCOMM Conference on Internet Measurement, 2004, pp. 370-375.
13. Ishibashi K., Toyono T., Toyama, K., et al. Detecting Mass-mailing Worm Infected Hosts by Mining DNS Traffic Data. Workshop on Mining Network Data at ACM SIGCOMM, USA, 2005, pp. 159-164.
14. Klein, A. BIND 9 DNS cache poisoning. http://www.trusteer.com/docs/BIND_9_DNS_Cache_Poisonding.pdf.
15. US-CERT. The Continuing Denial of Service Threat Posed by DNS Recursion. http://www.us-cert.gov/readingroom/DNS-recursion033006.pdf.
16. Cheng, J., Li, X., Yuan, J., et al. K-means Based Analysis of DNS Query Patterns. Journal of Tsinghua University, 2010.
17. Wang, Z., Li, X., Yan, B. Abnormity Detection of DNS Query Traffic at CN Top Level Domain Server, Technical Report, 2010. http://www.cdns.cn/about/vision-2.html.
18. Xu, H., Li, Z., Zhou, L. Botnet Detection Methods in DNS Traffic. Journal of Xiamen University, 2007.