Research on Feature Construction for Polymorphous Clone of Distributed Reflection Denial of Service Attack Traffic

Chanjuan Zhang[1,2,a], Xiaorui Gong[1,2,b,] and Zhenyu Song[1,c,*]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

[a]zhangchanjuan@iie.ac.cn, [b]gongxiaorui@iie.ac.cn, [c]songzhenyu@iie.ac.cn

*Corresponding author

Keywords: DRDoS, Clone, Feature selection, Apriori, Filter, Wrapper.

Abstract. Polymorphous clone of Distributed Reflective Denial of Service Attack traffic has great significance for DRDoS's demonstration, verification, defense, evaluation and scenarios that require DRDoS attack traffic. This paper studies a construction method of features in polymorphous clone of DRDoS attack traffic. Based on knowledge of information entropy and mutual information, traffic features were established from two perspectives of the content and statistical properties for clone of DRDoS attack traffic. It used a machine learning algorithm called Apriori, the filter feature selection and the wrapper mode of Random Generation plus Sequential Selection. By analysis of the NTP Distributed Reflective Denial of Service attack traffic which were collected and as the sample traffic, a feature set including 27 features are used as the attack traffic's features.

DRDoS攻击流量多态克隆过程中流量特征构建方法研究

张婵娟[1,2,a]，龚晓锐[1,2,b]，宋振宇[1,c,*]

[1]中国科学院信息工程研究所，北京，中国

[2]中国科学院大学网络空间安全学院，北京，中国

[a]zhangchanjuan@iie.ac.cn, [b]gongxiaorui@iie.ac.cn, [c]songzhenyu@iie.ac.cn

[*]通讯作者

关键词: 分布式反射拒绝服务攻击; 多态克隆; 流量特征; Apriori; 过滤式; 封装式

中文摘要. 分布式反射拒绝服务攻击流量的多态克隆对分布式反射拒绝攻击的演示验证、防御测评等需要DRDoS攻击流量的场景具有重要意义。本文研究了DRDoS攻击流量形态各异克隆过程中流量特征的构建方法。根据信息熵、互信息等知识，利用Apriori机器学习算法、过滤式特征选择算法、封装式的随机序列选择算法，从反射式拒绝服务攻击流量的内容特征和整体特征两个方面构建攻击流量特征。以采集的NTP协议DRDoS攻击样本流量为例，获得了包含27个特征的流量特征集。

1. 引言

2012年到至今，越来越多的攻击者[1,2,3]利用服务器、路由器等设备反弹响应数据包从而淹没被攻击机器，逐步演变为分布式拒绝服务攻击(DDoS)中的分布式反射拒绝服务(DRDoS)攻击 (Distributed Reflection Denial of Service Attack)。分布式反射拒绝服务攻击如图1

所示，一般采用伪造源地址技术，利用NTP、DNS、SSDP等具有放大效果的协议，向NTP、DNS等服务器发送字节数较小的请求数据包，通过这些服务器反弹大量的响应数据包到靶机，从而发动威力大、隐蔽性强的DRDoS攻击。

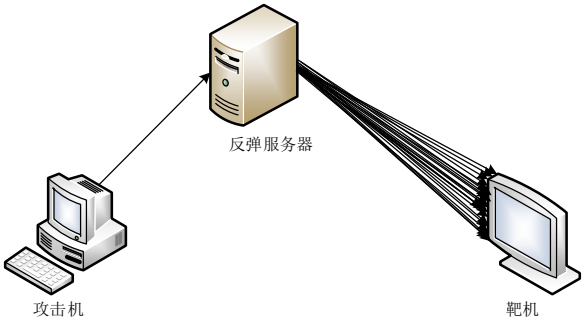

图1 DRDoS攻击示意图

频繁发生的DRDoS攻击事件和信息安全公司Arbor network、CDN服务商Akimai等发布的报告表明[4,5,6,7]，分布式反射拒绝服务攻击简单有效，发生频繁，带宽总量大，破坏性强。因此，分布式反射拒绝服务攻击对互联网安全有长期和深远的影响，越来越多研究者深入研究分布式反射拒绝服务攻击及其防御[8,9]。因直接在现实网络中进行DRDoS攻击演示、验证、测试或评估具有破坏性，所以对分布式反射拒绝服务攻击流量进行多种形态、逼真地模拟具有现实意义。一方面，模拟克隆出的逼真DRDoS攻击流量可以支持网络测试床中反射式拒绝服务攻击场景的快速构建。另一方面，也适用于入侵检测系统、安全审计系统、系统上线前的测试等，从而提供真实性高的研究环境和测试条件。

2．分布式反射拒绝服务攻击多态克隆

分布式反射拒绝服务攻击流量多态克隆是基于分布式反射拒绝服务攻击流量样本，克隆出形态各异但流量特征相似的大规模DRDoS攻击流量，多态克隆是指多样化克隆。类似与电影特效中根据少量的人物动作，利用数字合成、人工智能等技术合成千军万马的逼真场景[10]，反射式拒绝服务攻击流量的多种形态克隆是基于采集的DRDoS攻击流量样本，然后建立精简的流量特征集。下一步，从流量的字段内容、特征属性两方面，储存或者根据一定的规则自动化产生变化字段、整体流量属性可替换成的合理数值范围，再组合数据包的可变部分，从而实现分布式反射拒绝服务攻击流量地多样化、真实性克隆，详细逻辑结构图请见图2。

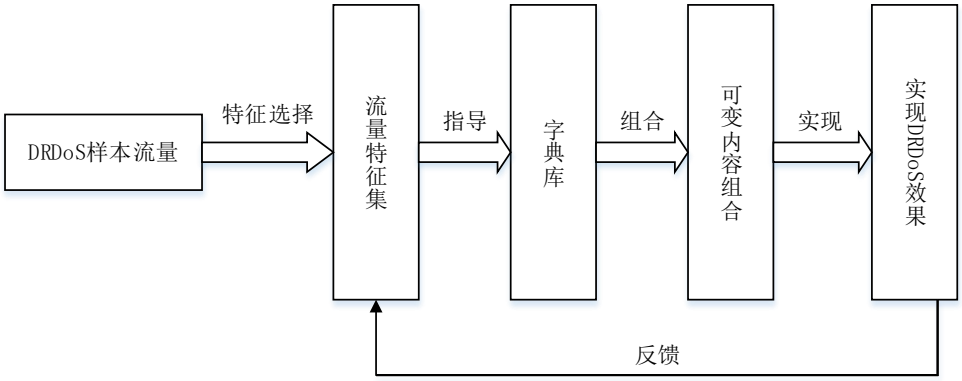

图2 DRDoS攻击流量多态克隆逻辑图

根据采集的NTP协议的DRDoS攻击样本流量，主要从协议内容字符串和流量的整体特征属性值等方面，利用相关算法提取出全面而精简的流量特征，形成流量模型。流量特征提取和选择的目的是去除无关、冗余的特征，最终获得完备精简的分布式反射拒绝服务攻击流量特征子集，从而降低机器学习算法的复杂度，提高字典库中数据的准确率以及高效性。形成

的流量特征集过滤出冗余信息，其规模小，是反射式拒绝服务攻击多种形态克隆过程中流量特征的典型代表特征[11]，从而简化信息、提高效率。

3．分布式反射拒绝服务攻击流量特征构建方法

分布式反射拒绝服务攻击流量地构建主要是从协议的内容特征和流量的整体统计特征两个方面进行建立。

3.1 分布式反射拒绝服务攻击流量的内容特征

对于流量的内容进行流量特征提取并选择，首先，对于已采集的NTP协议分布式拒绝服务攻击流量样本进行分析，并借助tshark[12]，编写相关代码，实现NTP反射式拒绝攻击流量样本从以太网协议、网络层IP协议、传输层UDP协议到应用层等NTP DRDoS攻击数据包中各个字段对应的字段值。详细字段请见表1。

表1 NTP DRDoS 攻击数据包2-4层的字段

协议	字段名
Ethernet II	源MAC地址
	目的MAC地址
	类型
IPV4	版本
	IP首部长度
	服务类型
	总长度
	标识符
	标志
	分段偏移
	生存时间
	协议
	IP首部校验和
	源IP
	目的IP
UDP	源端口
	目的端口
	包长度
	校验和

同时，也获得了应用层NTP协议的各个字段以及各个子字段，请见表2。

表2 NTP DRDoS 攻击应用层字段以及子字段

协议	字段名	子字段
NTP	Flags	
	Auth,scquence	
	Implemetation	
	Request code	
	Err	
	Number of data items	
	Reserved	
	Size of data items	
	Monlist item	Avgint
		Lsint
		Restr
		Count
		remote address

续表2

		local address
		flags
		port
		mode
		Version
		ipv6

然后，利用Apriori算法[13]求得每个字段值的频繁项集和各个字段值之间的关联规则。Apriori算法从单元素开始，迭代组合满足最小支持度要求的项集形成更大的集合。算法的原理是如果一个项集是非频繁的，那么它的超集也是非频繁的。最小支持度是数据集中包含该集合的记录所占的比例，是用来度量一个集合在原始数据中出现的概率。Apriori算法求频繁项集的过程是：

a) 输入上文从NTP协议的分布式反射拒绝服务攻击流量样本获得各个字段值数据集和最小支持度。

b) Apriori算法首先生成所有单个数据包字段值的项集列表，然后扫描查看哪些字段的数值集合满足最小支持度要求，去掉不满足最小支持度的字段值集合。

c) 下一步，对剩余协议字段进行组合生成包含两个字段内容的列表。再重新扫描集合，去掉不满足最小支持度的数据包字段项集。

d) 重复迭代此过程，直到所有协议字段都被扫描和处理。

下一步，再根据关联规则原理即如果某条规则不满足最小可信度要求，该规则的所有子集也不会满足最小可信度要求。获得反射式拒绝服务攻击流量各个字段的规则列表的步骤是：基于已获得的分布反射拒绝服务攻击流量的频繁字段项集，创建一个规则列表，其中规则右部只包含一个字段元素。然后，对这些规则进行对比分析并去掉不满足最小关联规则的数据包内容字段集合。下一步，合并剩余数据包内容字段规则并创建一个新的字段规则列表，其中规则右部包含2个数据包字段因子。以此类推，直到遍历所有可能的流量字段规则列表子集。

3.2 分布式反射拒绝服务攻击流量的整体特征

对于流量的整体属性特征方面，主要是利用了过滤式特征选择算法(Filter)和封装式特征选择算法(Wrapper)进行特征提取和选择。过滤式特征选择算法具有较强的通用性，算法复杂度低，可用于快速地过滤大量不相关的流量特征[14]，可作为预筛选器。因此，分布式拒绝服务攻击流量的特征识别和选择先利用过滤式策略对流量特征进行预筛选，过滤一些冗余特征、干扰信息，再利用封装式策略随机序列搜索算法再进行搜索，从而形成恰当、准确完备但又不冗余的DRDoS攻击流量特征集[15]。

首先，根据Moore的经典数据包特征的全部列表[16],分别计算出已采集样本流量的69个特征值，主要从包级别、三元组流两个粒度，包括服务器端口号，客户端端口号，数据包长度，数据包载荷的字节数以及字节数的平均值、最大值、最小值、中位数、方差，流中数据包数，流中相邻数据包的时间间隔，三元组流的持续时间等流量特征值。

然后，利用过滤式特征选择算法，过滤掉冗余、不相关的信息，主要是运用了特征集的信息熵和信息的相关性。根据香农定理[17]知识即信息熵在信息论中代表随机变量不确定性的度量。假设特征 X 中 x 的概率为 $p(x)$，信息熵的公式为

$$H(X) = -\sum_{x \in X} p(x) log\ p(x) \tag{1}$$

同时，基于互信息知识，利用信息的对称不确定性衡量流量特征集中各个特征属性的相关性[11, 18]。良好的特征子集应该所包含的特征彼此相关度较低。计算两两特征之间的对称不确定性，采用计算公式(2)， SU 越大代表相关性越高,

$$SU(X,Y)=\frac{2I(X,Y)}{H(X)+H(Y)} \tag{2}$$

其中 X,Y 分别代表多个样本组成的特征， I 为其两个特征之间的互信息.

基于过滤式特征选择算法筛选出来的流量特征集，再利用随机序列选择算法进一步筛选。随机序列选择算法[19]基于过滤式选择算法得到的分布式反射拒绝服务攻击流量特征集作为基础特征集合，然后在此特征集中随机选择DRDoS攻击流量特征作为初始特征子集，然后每添加一个特征并与原始特征子集进行评价和筛选，直到子集最优。

具体的计算过程为，第一步，先计算特征集的信息熵，按照熵的大小，对原始特征集中特征的重要性进行排序，信息熵越小，特征的重要性越大，从而选取前 61个特征组成的初次筛选的新特征集。第二步，对初次筛选后的新特征集计算彼此相关度，即削弱特征之间的互信息，从而得到第二次筛选的包含40个特征的流量特征集。下一步，利用封装式特征选择算法的随机序列选择算法再进一步筛选特征，从而提高流量特征的精确性和完备性。最后，得到最终流量特征集。

3.3 分布式反射拒绝服务攻击流量特征集

基于已采集的NTP协议分布式反射拒绝服务攻击流量样本，利用DRDoS攻击流量的内容特征提取方法和流量整体特征属性选择方法，再根据分布式反射拒绝服务攻击时效果反馈至流量特征集进行完善和参数的优化，最后构建的反射式拒绝服务攻击流量特征集合请见表2。

表3 流量特征集

源端口	帧长度	TTL	流中载荷字节数的标准差
目的端口	IP层的数据长度	TTL的平均值	流的持续时间
载荷字节数	相邻数据包时间间隔平均值	流中数据包的个数	流中相邻时间数据包的时间间隔的最大值
载荷字节数平均值	相邻数据包时间间隔的中位数	流中载荷字节数	流中相邻数据包的时间间隔的最小值
载荷字节数方差	相邻数据包时间间隔的方差	流中载荷字节数的最大值	流中相邻数据包的时间间隔的方差
载荷字节数最大值	相邻数据包时间间隔的最大值	流中载荷节数的最小值	流中相邻数据包时间间隔的平均值
载荷字节数最小值	相邻数据包时间间隔的最小值	流中载荷字节数的平均值	

同时，构建的DRDoS攻击流量特征会根据不同的协议、不同的样本流量自动化进行流量特征的选取，随协议、流量样本甚至参数的不同会灵活变化，从而更好地指导字典库的生成和分布式反射拒绝服务攻击流量多种形态克隆时可变内容的组合，使克隆出的流量更加真实可信。同时，也适用于其他DNS、SSDP等其他协议的分布式反射拒绝服务攻击流量的克隆。

4．结束语

分布式拒绝服务攻击流量的多态且真实的克隆研究，为DRDoS攻击进行测评、验证提供了简单有效的逼真环境。从分布式反射式拒绝服务攻击数据包内容方面和流量整体特征属性方面进行特征提取和特征选择，克隆出形态各异的反射型DRDoS攻击流量在数据包的内容字段方面以及流量的整体统计特征具有和DRDoS攻击流量样本相似的特征。但文章只针对IPV4协议分布式反射拒绝服务攻击流量进行研究，并未考虑IPV6情况下NTP协议DRDoS攻击流量，下一步可以进一步研究包含IPV6协议的分布式反射拒绝服务攻击流量的多态克隆。

致谢

本论文获得中国科学院网络测评技术重点实验室和网络安全防护技术北京市重点实验室资助，获得了国家重点研发计划（No. 2016YFB0801004和2016QY04W0905）课题资助。

References

[1] Czyz J, Kallitsis M, Gharaibeh M, et al, Taming the 800 pound gorilla: The rise and decline of NTP DDoS attacks, *Proceedings of the 2014 Conference on Internet Measurement Conference, ACM*, pp. 435-448, 2014.

[2] Rossow C, Amplification Hell: Revisiting Network Protocols for DDoS Abuse, *NDSS*, 2014.

[3] Kührer M, Hupperich T, Rossow C, et al. Exit from Hell? Reducing the Impact of Amplification DDoS Attacks, *USENIX Security Symposium*, pp. 111-125, 2014.

[4] Answers about recent DDoS attack on Spamhaus on

[5] http://www.spamhaus.org/news/article/695/answers-aboutrecent-ddos-attack-on-spamhaus

[6] akamai's [state of the internet] / security Q3 2016 report on

[7] https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/q3-2016-state-of-the-internet-security-report.pdf

[8] akamai's [state of the internet] / security Q1 2016 report on

[9] https://www.akamai.com/es/es/multimedia/documents/state-of-the-internet/akamai-q1-2016-state-of-the-internet-security-report.pdf

[10] Rudman L, Irwin B, Characterization and analysis of NTP amplification based DDoS attacks, *Information Security for South Africa (ISSA), 2015. IEEE*, pp. 1-5, 2015.

[11] WORLDWIDE INFRASTRUCTURE SECURITY REPORT On http://www.asiapacificsecuritymagazine.com/wp-content/uploads/2017/01/2017-01-19-Arbor-WISR-Full-Report.pdf

[12] Behal S, Kumar K, Characterization and Comparison of DDoS Attack Tools and Traffic Generators: A Review, *IJ Network Security*, vol.19, pp. 383-393, 2017.

[13] Aitken M, Butler G, Lemmon D, et al, The Lord of the Rings: the visual effects that brought middle earth to the screen, *ACM SIGGRAPH 2004 Course Notes, ACM*, pp. 11, 2004.

[14] Zhu Wenbo, Application Research of Flow Statistical Features on Network Traffic Classification, Hangzhou Dianzi University, 2015.

[15] Tshark on https://www.wireshark.org/docs/man-pages/tshark.html

[16] Harrington P, Machine Learning in action, edtied by Posts & Telecom Press, Beijing, 2013.

[17] Mao Yong, Zhou Xiaobo, Xia Zheng, et al. A Survey for Study of Feature Selection Algorithms. *Pattern Recognition and Artificial Intelligence*, 2007.

[18] Katal S, Singh A, A Survey of Machine Learning Algorithm in Network Traffic Classification, *Int. J. Comput. Trends Technol.(IJCTT)*, vol.9, 2014.

[19] Moore A, Zuev D, and Crogan M, Discriminators for use in flow-based classification, 2013.

[20] Information theory on https://en.wikipedia.org/wiki/Information_theory

[21] Chandrashekar G, Sahin F, A survey on feature selection methods, *Computers & Electrical Engineering*, vol.40, pp. 16-28, 2014.

[22] Andradóttir S, A review of random search methods, *Handbook of Simulation Optimization, Springer New York*, pp. 277-292, 2015.