doi:10.3772/j.issn.1002-0470.2016.10 – 11.004

基于相对密度的 DNS 请求数据流源 IP 异常检测算法①

王靖云② 史建焘③ 张兆心④ 沈英洪

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 研究了域名系统(DNS)的异常检测。通过对基于相对密度的离群点检测算法的研究,提出了一种基于相对密度的 DNS 请求数据流源 IP 异常检测算法。该算法计算每个源 IP 的相对密度,并将该密度的倒数作为其异常值评分;在计算相对密度时,从查询次数、源端口熵值、所请求非法域名占比等 9 个维度来表示一个源 IP。试验结果表明,这种基于相对密度的源 IP 异常检测方法,能正确地根据各个源 IP 不同的异常程度,给出其相应的异常值评分。

关键词 域名系统(DNS),相对密度,离群点,异常检测

0 引 言

域名系统(domain name system, DNS)是互联网的一个核心服务系统,它通过对域名和 IP 地址相互映射,完成两者之间的转换,从而方便用户进行互联网的访问。DNS 协议在设计之初,没有在安全方面进行考量,未能提供加密和认证机制,很难保证信息的保密性、真实性和完整性。同时 DNS 作为互联网中关键的基础设施之一,在互联网的正常运行中起着重要的作用,防火墙对于 DNS 协议一般不会进行拦截。因此,DNS 很容易被攻击者利用,来进行一些非法的网络行为。在日常的网络应用中,比较常见的 DNS 攻击行为有 DNS 缓存污染、DNS 通信劫持、DNS 的 DDoS 攻击等。使用 DNS 协议来构建隐蔽的信道,也被广泛应用在僵尸网络、木马与 C&C 服务器之间进行通信。用户也可以通过 DNS 隧道来绕过官方的认证系统,从而实现免认证联网的目标。除此之外,关于 DNS 的重大安全事件也时有发生,从而危及国家的网络安全。

关于 DNS 异常检测,Pomorova 等通过在 DNS 数据流中寻找相似 DNS 请求模式的群体来发现僵尸网络[1];Manasrah 等基于僵尸网络周期性成群出现的特点来检测异常 DNS 数据流来发现僵尸网络[2];严芬等使用信息熵来发现利用 DNS 进行的 DoS 攻击[3];Bilge 等人使用 DNS 的时间信息、应答信息以及 TTL 值等作为特征,利用 J48 决策树进行分类,实现了一套恶意域名检测系统[4];章思宇等使用数据分组、域名、DNS 消息等特征,同样利用 J48 决策树进行分类,提出了一种关于 DNS 隐蔽信道的检测方法[5];林成虎等提出了一种基于 W-Kmeans 的算法用于检测 DNS 查询流量的异常[6];李杰等利用信息熵模型进行 DNS 欺骗和缓存中毒的检测[7]。目前这些研究主要集中在针对某类特定的恶意行为,在 DNS 流量中检测出这类行为。而关于在 DNS 流量中,对具有明显异常的源 IP 检测,目前还没有对应地解决方案。针对这种情况,本文提出了一种基于相对密度的离群点检测方法,用于

① 国家科技支撑计划(2012BAH45B01),国家自然科学基金(61100189,61370215,61370211)和国家信息安全计划(2014A085,2015A072)资助项目。

② 男,1993 年生,硕士;研究方向:域名体系安全;E-mail: cloud.aha@gmail.com

③ 男,1980 年,博士;研究方向:计算机网络,云计算,信息安全;E-mail: shijiantao@hit.edu.cn

④ 通讯作者,E-mail: heart@hit.edu.cn

(收稿日期:2016-07-21)

发现在 DNS 流量中存在异常的源 IP 地址。

1 基于相对密度的离群点检测算法

基于相对密度的离群点检测是一种离群点发现的算法,在异常检测中经常得到应用。该算法在进行检测时,可以定量得给出每个点的离群点得分,并且当数据具有不同的密度区域时,算法也能很好地处理。其主要思想是首先求得数据集合中每个点各自的相对密度,然后将每个点的相对密度的倒数作为每个点各自的离群点得分,得分越高的点,就被认为越异常。通过选取得分排名靠前的点,即可获得数据集合中的异常点。

设数据集合 $S = \{a_1, a_2, \cdots, a_n\}$,其中的每个点 a 为 m 维向量,即 $(x_1^{(a)}, x_2^{(a)}, \cdots, x_m^{(a)})$。定义每个点 a 的 k 近邻为集合 $N(a,k) = \{a_{i_1}, a_{i_2}, \cdots, a_{i_k}\}$,其中 a_i 为与点 a 相距最近的 k 个点,且 $a_i \neq a$。点 a 与 b 之间的距离定义为欧式距离,即

$$dis(a,b) = \sqrt{\sum_{i=1}^{m} (x_i^{(a)} - x_i^{(b)})^2} \quad (1)$$

根据点 a 的 k 近邻,定义点 a 的周围密度为,点 a 与其 k 近邻的中各个点的平均距离的倒数,即

$$density(a,k) = \left(\frac{\sum_{b \in \mathbf{N}(a,k)} dis(a,b)}{|N(a,k)|} \right)^{-1} \quad (2)$$

其中,$|N(a,k)|$ 表示该集合的大小。当一个点与其周围的点距离较大时,其密度较小;反之,当一个点与其周围点的距离较小时,其密度较大。

当数据点的分布在不同区域具有不同密度时,仅仅使用周围密度可能会造成在密度稀疏的区域中其中的点周围密度普遍较低,更容易被判定为离群点;而在密度稠密的区域中,其中的点周围密度普遍较高,不太容易被判定为离群点,即使在该高密度区域中,一个点已经离该区域其他点较远。

在此引出相对密度的概念。基于上述点 a 的周围密度,定义点 a 的相对密度为点 a 的周围密度与其 k 近邻中所有点的平均周围密度之比,即

$$relative\_density(a,k) = \frac{density(a,k)}{\sum_{b \in \mathbf{N}(a,k)} density(a,k) / |N(x,k)|} \quad (3)$$

点 a 的离群点得分即为该点相对密度的倒数。根据每个点的离群点得分,选取分大于阈值的点,即可得到数据集合中的异常点。

2 源 IP 地址异常检测特征提取

在使用相对密度进行域名系统(DNS)请求源 IP 异常检测时,需要对源 IP 的特征进行描述。因此,本文以校园中真实的 DNS 请求数据为例,通过对 2015 年 11 月 22 日 21 时这一个小时内的 DNS 请求流量进行分析,提取了 9 个特征来描述源 IP。数据共包含 200 万次 DNS 请求,其中涉及到 3148 个源 IP,每个源 IP 均有自己的序号。所提取的 9 个特征分别为:单个源 IP 的查询次数、单位时间内的查询峰值、源端口信息熵、DNS 报文头部 ID 的信息熵、目的 IP 个数、查询不同域名的个数、单位时间内查询域名种类的峰值、畸形数据包的比例、非法域名的比例等。

(1)查询次数。正常 IP 在进行网络活动时,其 DNS 请求的次数不会太高,如果 DNS 请求次数过多,则该源 IP 很有可能具有异常行为,如进行 DoS 攻击。因此,每个源 IP 在一个小时内的 DNS 请求的次数,可以作为判断源 IP 是否异常的因素之一。从图 1 中可以看出,源 IP 序号为 2571 的请求次数远远高于其他 IP 地址,并在一个小时内,发出了 15 万次以上的 DNS 请求,这个源 IP 地址很明显是异常的。

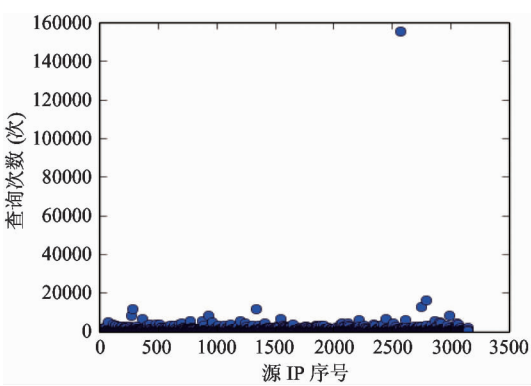

图 1 查询次数

(2)单位时间内的查询峰值。尽管有的源 IP 在一个小时内查询次数不多,但是在很短的一个时

间范围内,该源 IP 发出了大量 DNS 请求,这种现象也表明该源 IP 可能在具有异常行为,如进行 DoS 攻击。因此,单位时间内的查询峰值也可以作为判断源 IP 是否异常的因素之一。在一个时间范围内,计算每个源 IP 在单位时间内的 DNS 查询次数,取最大值作为该源 IP 的查询峰值,本文选取单位时间为 1s。图 2 显示的是在一个小时内,所有 3148 个源 IP 的 DNS 查询峰值。结合图 1 可以看出,有些源 IP 虽然总查询次数不多,但是在某个时间点内,却发出了大量的 DNS 查询请求。

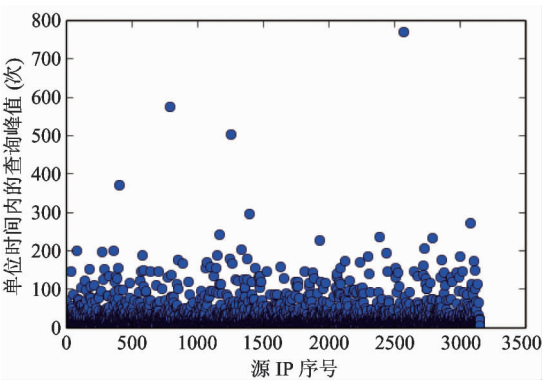

图 2　单位时间内查询峰值

(3)源端口的信息熵。源 IP 在发出 DNS 请求时,为了防止被劫持,源端口的使用需要有足够好的随机性。而信息熵则可以很好地衡量源端口的分布的随机程度,随机性越好,信息熵越大,而随机性越差,信息熵越小。信息熵的计算方法为

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log \frac{m}{m_i} \tag{4}$$

其中,n 为端口数,m_i 为端口 i 的查询次数,m 为所有端口总的查询数。图 3 显示了不同源 IP 之间源端口信息熵的差异,具有较低源端口信息熵的源 IP 很有可能具有异常行为。

(4)DNS 报文头部 ID 的信息熵。为了防止 DNS 劫持,在发送请求报文时,DNS 报文头部的 ID 号也是随机的。同时,在利用 DNS 协议构建的隐蔽信道进行通信时,所构造的 DNS 报文通常会有一个固定的报文 ID,以便于使用者进行捕获,这样会造成 DNS 报文头部 ID 的信息熵低于正常值。所以,

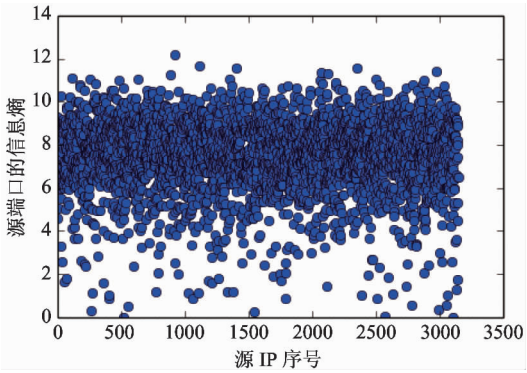

图 3　源端口的信息熵

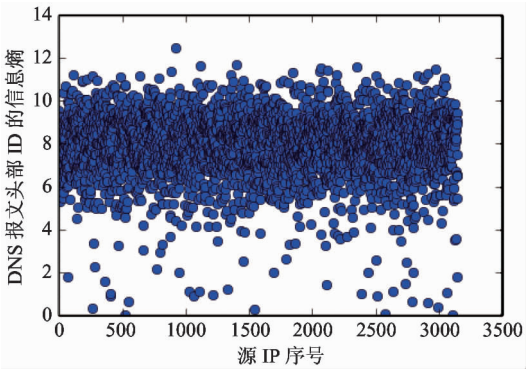

图 4　DNS 报文头部 ID 信息熵

该熵值也可以作为源 IP 异常检测的因素之一。图 4 显示了不同源 IP 之间 DNS 报文头部 ID 信息熵的差异。

(5)目的 IP 个数。正常的源 IP 在一段时间内,访问的 DNS 服务器个数并不会太多。如果一个源 IP 在请求中涉及到过多的目的 IP 地址,则该源 IP 很有可能具有异常行为,如利用 DNS 协议向不同的 IP 传输非法数据。图 5 显示了不同源 IP 在一个小时内所涉及到的目的 IP 个数。从中可以看出,有四个源 IP 地址,在一个小时内访问的目的 IP 个数超过了 200。

(6)请求域名种类。这个特征类似于查询次数,从另一个角度反映了一个源 IP 的 DNS 查询的频繁程度。这里的域名种类是指该域名所属的权威域的种类,即 ssl. ptlogin2. qq. com 和 www. qq. com 均属于 qq. com,只记录一次。图 6 显示了不同源 IP 一个小时内请求的域名数量。

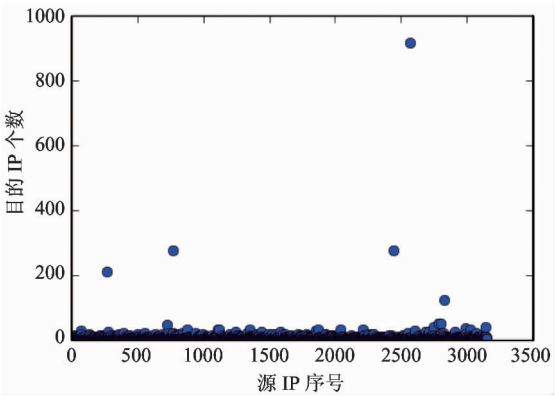

图 5　目的 IP 个数

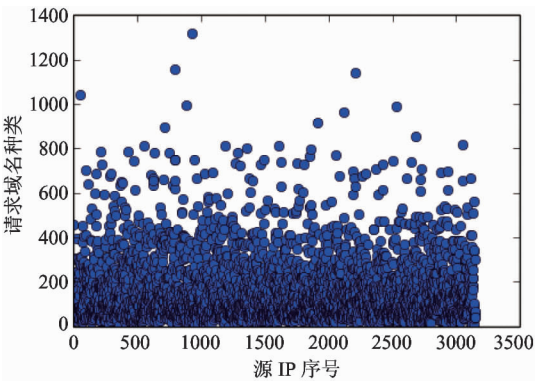

图 6　请求域名种类数

(7)单位时间内域名请求种类的峰值。正常的网络行为,在较小的一个时间段中不会请求特别多的域名种类。如果短时间内源 IP 请求了很多的域名,则其很有可能是异常的,比如进行网络代理或信息探测。图 7 展示了不同源 IP 在一小时内域名请求的峰值。结合图 6 可以看出,一些源 IP 在一个小时内请求域名数量不是很多,但是在单位时间内,会发出大量的域名请求。

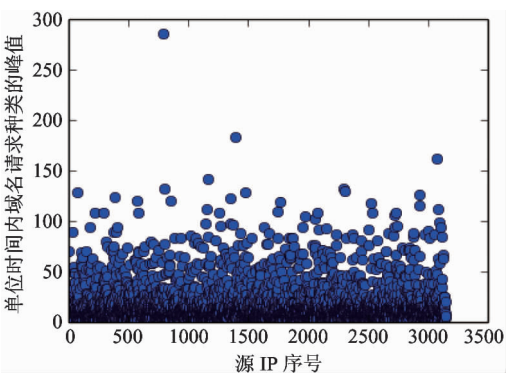

图 7　单位时间内域名请求种类峰值

(8)畸形包的比例。请求报文,如果不能严格的按照 DNS 协议规定的格式进行解析,则会被判定为畸形包。畸形包中很小的一部分来自于网络设备传输过程中的故障,而更多的来源于源 IP 机器上客户端错误地构造了数据包或者是一些恶意软件故意构造出畸形包来进行攻击行为。所以,畸形包的比例也可以作为源 IP 异常检测的因素之一。图 8 展示了不同源 IP 所发出的畸形包比例。

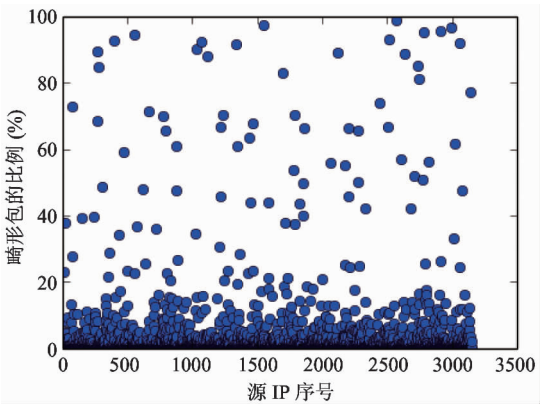

图 8　畸形包比例

(9)非法域名的比例。正常的域名是由字母(A ~ Z,a ~ z 不区分大小写)、数字(0 ~ 9)和连接符(–)组成,各级域名之间用实点(.)连接,并且每一个级不超过 63 个字符。如果一个域名不满足这些规则或者域名字段不能正常解析时,那么就它就是非法域名。过多的非法域名意味着这个源 IP 在利用 DNS 协议中域名字段在传输信息。因此,非法域名的比例也可以作为源 IP 异常检测的因素之一。图 9 展示了不同源 IP 所发出的异常域名所占比例。

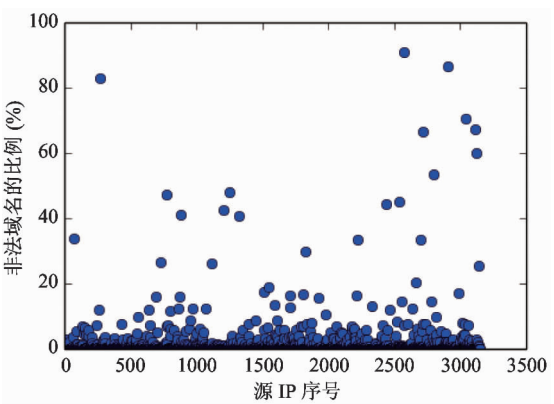

图 9　非法域名比例

3 DNS 请求数据源 IP 异常检测

在使用相对密度来进行 DNS 请求数据源 IP 的异常检测时,首先需要进行数据预处理,即从 DNS 请求数据流中提取出部分信息来进行分析。接着从这些信息中提取本文第 2 节中所提到的有关源 IP 的 9 个特征,并对这些特征进行规范化处理。进而在 k 不同取值时,对比在小范围数据内求得的异常值得分,来选取合适的 k 值。最后通过求得不同源 IP 的相对密度,来计算其异常值得分。

3.1 数据预处理

对于 DNS 请求报文中的信息,提取出其中的部分字段来进行存储,DNS 请求报文存储记录的格式为(src _ ip, src _ port, dst _ ip, dst _ port, domain, dns _ id, timestamp)。表 1 为记录的各项数据含义描述。

表 1 请求报文存储记录项

字段名	描述
src _ ip	源 IP 即发起请求的 IP
src _ port	源端口
dst _ ip	目的 IP 即服务器 IP
dst _ port	目的端口
Domain	请求解析的域名
dns _ id	DNS 数据包头部的 dns _ id
Timestamp	捕获数据包的时间戳

3.2 特征提取

按照本文第 2 节所述方法,从数据中提取出与源 IP 相关的 9 个特征。

由于所选取的 9 个特征的数据在取值范围上有较大的差距,因此对在特征提取完成之后,需要进行规范化处理。因为特征中具有请求次数、查询域名个数等,这些特征的取值范围不能确定,所以使用 Z 分数规范化来对数据进行处理。对于每个特征 f_i,其中的每个数据 $v_{i,j}$ 进行规约化得到 $v'_{i,j}$,具体公式为

$$v'_{i,j} = \frac{v_{i,j} - mean_i}{\mathrm{var}_i} \tag{5}$$

其中 $mean_i$ 为特征 f_i 对应数据的均值,而 var_i 为特征 f_i 对应数据的标准差。

3.3 选取 k 值

算法中 k 值如果太小,则各个点的相对密度会十分敏感,比较容易造成误判。因此,本文提出一种基于结果平稳性的 k 值选择方法。

首先令 k 取最小值 1,然后依次增加 k 的取值。记本次 k 取值的情况下,异常值得分最高的 10 个点为集合 $S = \{a_1, a_2, \cdots, a_n\}$,下一次 k 取值时,异常值得分最高的 n 个点为集合 $S' = \{a_1', a_2', \cdots, a_n'\}$。通过比较集合 S 和 S' 的相似性,来进一步看当前 k 值下,结果是否平稳。若两个集合相似性越大,表示结果已经趋于平稳,若两个集合相似性越小,表示结果还处于变化之中。集合 S 与 S' 的相似性由 Jaccard 相似度求解,即

$$Jaccard(S, S') = \frac{|S \cap S'|}{|S \cup S'|} \tag{6}$$

其中 $|S \cap S'|$ 表示 S 和 S' 交集元素的个数,$|S \cup S'|$ 表示 S 和 S' 并集元素的个数。当 $Jaccard(S, S')$ 足够大时,选取当前 k 值为最终取值。

3.4 算法描述

基于相对密度的 DNS 请求源 IP 异常检测首先会求出每个数据点的 k 近邻,然后根据式(2),求得每个点的周围密度,再根据式(3),求出其周围密度,并将该密度的倒数作为这个点的异常值得分。算法的伪代码描述如下:

```
Input
    源 IP 特征数据集 D 、近邻个数 k
Output
    每个源 IP 的异常值得分
BEGIN
  for a in D do
     a. kSet = 计算离 a 最近的 k 个点
     sum = 0
     for b in a. kSet do
          distance = a 与 b 之间的距离
          sum + = distance
     end for
     a. density = k / sum
```

```
    end for
    for a in D do
      sum = 0
      for b in a.kSet do
        sum += b.density
      end for
      a.rele_density = a.density / sum
      a 的异常值得分 = 1 / a.rele_density
    end for
END
```

4　试验结果及分析

试验使用的流量样本来自于哈尔滨工业大学校园网的 DNS 流量。其中数据为 2015 年 11 月 22 日 20 时 0 分至 2016 年 1 月 2 日 23 时 59 分的 DNS 请求数据报文,其中共包含 8.3 亿次请求数据。

4.1　k 值的选取

通过对 2015 年 11 月 22 日 21 时这一个小时 DNS 的数据,按本文提出的方法进行处理。将 k 值从 1 递增计算得到了在不同 k 值下,异常检测结果的变化趋势如图 10 所示。其中在进行结果相似性计算时,所选取的集合大小为 30。

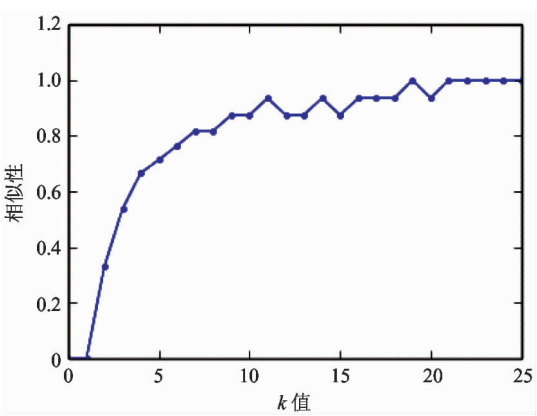

图 10　不同 k 值下的结果相似性

从图中可以看出,在 k 的取值大于 21 之后,异常值得分的前 30 名便不再变化,说明当 k 值为 21 时,可以很好地求出各个源 IP 的异常值得分。

4.2　异常源 IP 的检测

设置 k 值为 21 时,对 2015 年 11 月 22 日 21 时这一个小时 200 万次 DNS 查询数据进行源 IP 异常检测,程序运行时长为 31.07s。最后,异常值得分最高的 100 个数据分布如图 11 所示。

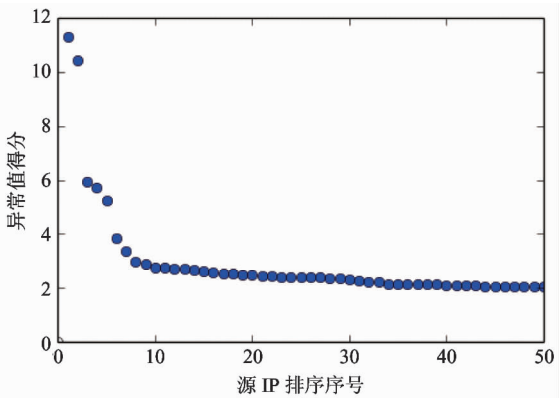

图 11　异常值得分 Top50 数据分布

从图中可以看出,异常值得分在小于 3 之后,各个源 IP 之间的差距已经很小,大于 3 的 7 个点,在异常值得分上,与其他值之间的差距较大。异常值得分前 7 名的具体得分及部分特征如表 2 所示。

表 2　异常值得分最高的 7 个源 IP 及其部分特征

源 IP	异常值得分	请求次数	端口信息熵	畸形包比例
172.30.226.7	11.3129	155574	0.01880	98.88%
172.30.174.77	10.4346	16390	0.5335	95.45%
172.30.195.66	5.9139	8381	0.2840	82.93%
172.30.202.34	5.7378	2336	2.0659	0
172.30.153.88	5.2503	1886	10.7589	70.47%
172.30.205.92	3.8306	2462	9.0464	1.67%
172.30.203.98	3.3462	2973	10.8216	0.07%

从上表中可以看出,排名前三的源 IP,具有很明显的特征就是 DNS 请求次数异常的高,然而在这么高的请求次数同时,端口信息熵异常的低,也就意味着这些大量 DNS 请求的源端口随机性很差,大多请求集中在某一个或某几个端口上。同时,这些域名请求中畸形包的占比达到 80% 以上。由此可以

看出,异常值排名前三的源 IP,各个方面都体现出了这些源 IP 可能存在恶意行为。

排名第四的源 IP 为 172.30.202.34,请求次数和畸形包比例均正常,端口信息熵低于正常值。未在表 2 中列出的单位时间内请求峰值和域名种类峰值,其值为 575 和 286,均远远高于正常值。从图 7 中可以看出,286 次的域名种类峰值,已经是该时间段内,所有源 IP 峰值的最大值。

排名第 5 到第 7 的源 IP,在请求次数方面,处于正常范围,且端口信息熵很高,具有很好的端口随机性。172.30.153.88 的请求中具有较高的畸形包比例;172.30.205.92 的情况类似于 172.30.202.34,具有 504 和 454 次单位时间内请求峰值和域名种类峰值。172.30.203.98 在源端口熵值高达 10.8216 的情况下,只具有 2.005 的 DNS 报文头部 ID 熵值,远远低于平均熵值。图 12 显示了端口信息熵位于 10 到 11 之间的源 IP,它们的 DNS 报文头部 ID 熵值分布情况,从中可以明显的看出,172.30.203.98 在 DNS 报文头部的构造上具有异常。

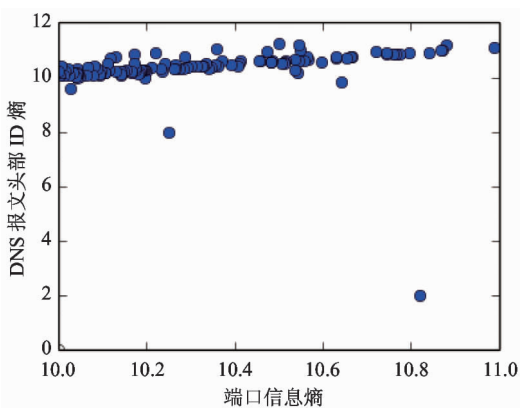

图 12　端口信息熵对应的 DNS 报文头部 ID 熵

综上所述,通过相对密度求得的异常值得分,可以很好地反映源 IP 的异常程度。

5　结　论

通过对基于相对密度的离群点检测算法的研究,提出了一种使用相对密度对 DNS 请求数据流中源 IP 进行异常检测的方法。试验结果表明,数据中具有明显异常的源 IP,均被赋予了较高的异常值得分。因此,通过使用该异常检测算法来计算每个源 IP 的异常值得分,可以很好地反映出各个源 IP 的异常程度。接下来的研究,主要围绕在如何进一步提升算法在大数据规模下的执行效率。

参考文献

[1] Pomorova O, Savenko O, Lysenko S, et al. A Technique for the Botnet Detection Based on DNS-Traffic Analysis. In: Proceedings of the 22nd International Computer Networks, Brunów, Poland, 2015. 127-138

[2] Manasrah A M, Hasan A, Abouabdalla O A, et al. Detecting botnet activities based on abnormal DNS traffic. *International Journal of Computer Science & Information Security*, 2009, 6(1):97-104

[3] 严芬, 丁超, 殷新春. 基于信息熵的 DNS 拒绝服务攻击的检测研究. 计算机科学, 2015,42(3):140-143

[4] Bilge L, Kirda E, Kruegel C, et al. Exposure: finding malicious domains using passive DNS analysis. In: Proceedings of the 18th Annual Network and Distributed System Security Symposium, San Diego, USA, 2011. 1-17

[5] 章思宇, 邹福泰, 王鲁华, 陈铭. 基于 DNS 的隐蔽通道流量检测. 通信学报, 2013,34(5):143-51

[6] 林成虎, 李晓东, 金键, 尉迟学彪, 吴军. 基于 wkmeans 算法的 dns 流量异常检测. 计算机工程与设计, 2013,34(6): 2104-2108

[7] 李杰. DNS 欺骗和缓存中毒攻击的检测:[硕士论文]. 成都:电子科技大学计算机科学与工程学院, 2015. 41-55

An algorithm for detection of source IP anomalies in DNS query based on relative density

Wang Jingyun, Shi Jiantao, Zhang Zhaoxin, Shen Yinghong

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract

The study focused on the anomaly detection for domain name systems (DNS). Through the investigation of the outlier detection algorithm based on relative density, an algorithm for detection of source IP anomalies in DNS query data streams based on relative density was proposed. The algorithm calculates the relative density of each source IP, and uses the inverse of the density as an abnormal value. When calculating the relative density, it uses the nine dimensions of number of query, entropy of source port, proportion of queried illegal domain name and so on to represent a source IP. The experimental results show that the proposed source IP anomaly detection algorithm based on relative density can put forward the corresponding abnormal value accurately according to the abnormality of each source IP.

Key words: domain name systems (DNS), relative density, outlier, anomaly detection