# 恶意流量的快速提取

## 摘要

网络中充斥着大量的恶意请求，其中绝大部分是是放大攻击、随机子域名攻击和僵尸网络产生的流量，在利用DNS流量进行一些恶意行为的分析的时候，往往需要对每一个域名单独进行检测，而数据量是很大的，简单的过滤无法迅速降低需要检测的域名数量，这样一来在计算资源有限的前提下，需要花费大量的时间。本文介绍了一种针对DNS流量的提取方案，我们针对流量中占比最大的三种攻击流量，分别设计了简单快捷的提取手段。我们使用了统计和分类的方法对所有流量进行处理，实现了一个系统原型，并在真实的流量环境下进行了实验，在召回率为100%的前提下，我们将待检测二级域名数量降低到原本数量的1%，待检测DNS记录降低到原本数量的1%。

## 1 简介

IP地址是由IP协议提供的数字型统一地址标识，作为一种逻辑地址来定义一台设备在网络之中的位置，网络设备逐渐增多IP地址的记忆困难显现出来，保罗·莫卡派乔斯 (Paul Mockapetris) 在1983年的第882和在南加州大学里资讯科学研究院所提出的883号因特网标准草案中提出DNS的架构，提议将其改进为分布式和动态的数据库域名系统，也就是我们今天所用的域名系统的雏形。域名作为一个工具，给我们带来方便的同时，也给网络上的恶意行为提供了便利。

针对这些恶意行为，有DGArchive、Notos、DSBL等黑名单与相应的信誉系统，也有Pleiades、EXPOSURE、FluxBuster等检测系统，更有利用LSTM、word2vec等深度学习模型解决问题的系统，在我对山西省和广东省电信的被动DNS数据处理的过程中，由于数据量过大，完整处理全部数据所需时间较长。如果直接使用白名单进行过滤，导致忽略了那些利用良性域名进行攻击或者针对良性域名的攻击行为。我们希望可以通过简化或者借鉴这些检测方法来构建出一个高效的恶意流量提取模型。

在本文中，我们提出了一个流量定位模型，利用一些统计特征以及预先训练好的模型，将恶意流量快速定位出来，该模型部署在本地递归服务器(RDNS)或网络边缘节点下，监控网络下的主机的DNS请求及应答，定时进行分析，快速将包含恶意域名的流量定位出来，用于进一步的分析与检测。根据报告（APRICOT 2015， Nominum），2015年9月2日的统计，恶意的DNS查询中，随机子域名占据80%，放大攻击占15%，C&C流量占5%，我们的模型也将针对这些方面进行过滤。

该模型从海量数据中定位到的恶意流量具有极高的召回率，同时可以大大缩减之后检测恶意域名的时间花销，由于将诸多干扰去除，也对检测的效果有了极大地提升，我们利用山西省和广东省不同时间的流量进行实验，证明该模型具有良好的适应性。

本文做出了一下贡献：

1. 我们提出了一个轻量级的恶意流量定位模型，它可以在大型ISP网络下有效的将恶意流量定位出来，快速缩减进一步流量监测的量级；
2. 我们给出该模型的原型实现，并在两个省不同时期的网络环境中实验，具有极好召回率，流量规模缩减也十分明显，也具有很高的适用性；

## 2 背景和相关工作

**基于DNS放大攻击背景**

DNS放大攻击，就是通过DNS服务器作为跳板放大攻击流量。正常的DNS查询过程为源IP地址发送DNS查询到达DNS服务器，查询结果再返回源IP地址，而攻击者在攻击时伪造源IP地址（即攻击目标），查询结果将返回到伪造的IP地址，通常一个DNS查询包大小为60byte左右，如果发起一个请求类型为ANY的DNS查询，则表示对该域名的所有记录（A记录、MX记录、CNAME记录、PTR记录等等）的请求，那么返回的数据包通常达到几百byte到上千byte。Akamai研究人员在2014年发现了利用TXT记录来进行的DNS放大攻击，攻击者使用名为DNS Flooder的工具，从guessinfosys.com获得TXT记录，攻击峰值高达4.3Gbps。

Tama等人利用异常检测的手段，根据报文头属性对网络数据流进行建模，采用朴素贝叶斯算法给每个到达的数据流评分，评价报文的合理性。Karnwal等人通过维度转换，将一维时序转换为多维的AR模型参数时序，采用支持向量机算法对数据流进行学习和分类。Wang等人利用异常检测手段，利用隐马尔科夫模型来描述数据流中报文头的变化情况。

**随机子域名相关背景**

随机子域名在域名的表现上是在二级或三级域名下有大量不同随机子域名，而该二级域名或者三级域名通常是合法的，随机子域名攻击是一种针对域名服务器的DDoS攻击。攻击者使用受感染的网络设备，构造合法域名下的随机子域名进行DNS查询，这些查询首先到达ISP递归DNS服务器，服务器中没有相应记录，然后递归服务器向上级DNS服务器转发，一直到达顶级域名服务器或者权威DNS服务器。这一些过程消耗的是查询资源而不是带宽，但是会明显拖慢甚至阻止域名的正常查询，造成对服务器的过载。

现在针对随机子域名攻击还没有成熟的实时检测手段，例如secure64给出的解决方案为加大内存，加大递归并发数，自动封禁有太多失败请求的IP。现在大多数机构只能采取这样的被动防御，难以建立起细粒度的有效过滤系统。

**DGA域名相关背景**

DGA（Domain generation algorithms）是域名生成算法，被大量的恶意软件家族所使用，用以建立受感染主机与C&C服务器（command and control servers）的联系。传统僵尸网络使用固定IP或者域名与C&C服务器建立连接，隐蔽性很差，极易被发现。后来出现的例如Nugache，Storm，Waledac，Zeus等基于P2P的僵尸网络虽然具有较好的鲁棒性与稳定性，但实现难度和维护成本较高。如今大部分活跃的僵尸网络都采用了DGA，依赖于集中的C&C服务器，相比于前两者具有简单易行，兼顾稳定性与隐蔽性的优点。

针对DGA算法的检测主要有黑名单、机器学习方法、逆向工程。L.Bilge[7]从DNS数据中分别基于时间、基于DNS响应、基于生存时间值（TTL）、基于域名提取了这4类共计15种特征，使用J48决策树训练分类器。使用十分完善的特征选取，弥补了无法检测一个IP地址只被恶意使用一次的恶意域名，对恶意域名的检测也更加广泛。并据此建立EXPOSURE系统。B.Rahbarinia[9]提出了一种基于行为的系统Segugio，Segugio在大的ISP网络中通过追踪被恶意软件感染主机的DNS请求行为来高效的发现新增的malware-control域名。相比于Notos和EXPOSURE，前两者建立的都是domain-IP映射关系模型（使用域名字符串的特征、域名承载的恶意内容等信息）而没有利用本地DNS服务器下游的主机请求行为，Segugio通过监测ISP网络用户的DNS请求行为，重在精确的追踪新增的“malware-only”域名。J. Woodbridge[13]等使用LSTM预测DGA产生的域名，可以实时运行并且不需要人为创造特征。D.Plohmann，F.Fkie等人针对DGA做了大量细致的工作[27]，他们对43种DGA恶意软件家族和变种进行一个综合性的研究，针对DGA提出一个分类学方法，并用它对所研究的DGA进行分类与比较。这些利用机器学习的检测方法普遍使用Alexa top域名作为白名单进行初步过滤，过滤域名数量有时很有限。

**相关工作**

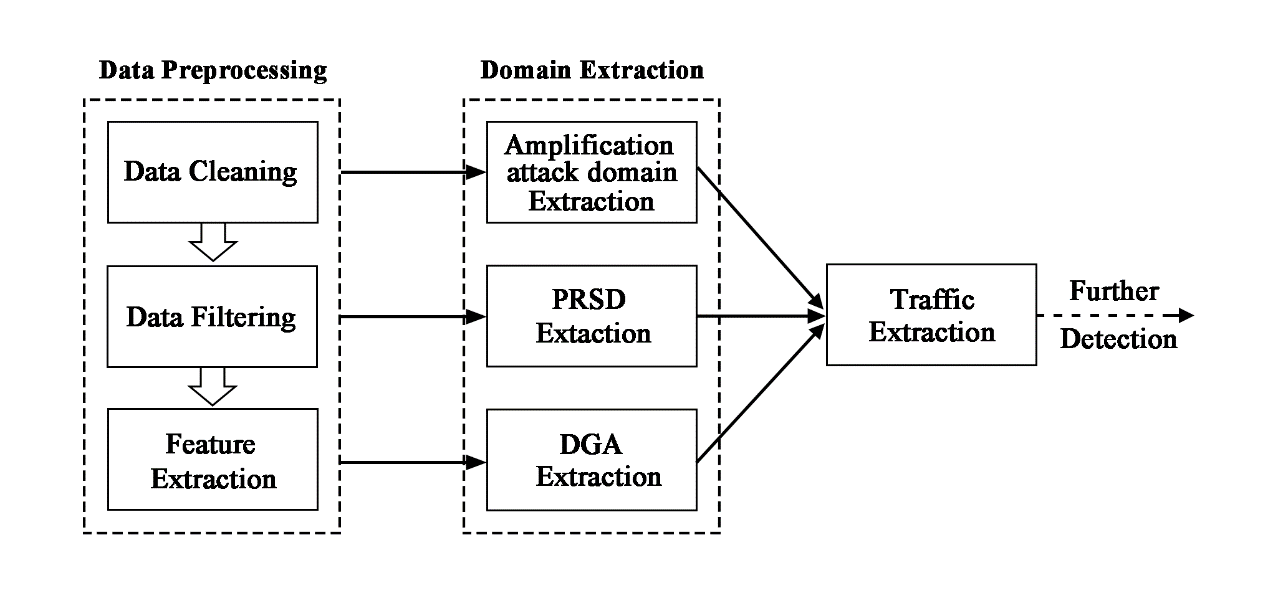
我们的工作和针对流量的检测模型的最大一个不同点是我们不关注是否出现误报，我们希望得到尽可能高的召回率、提取率（保留下来的恶意流量占完整流量的比例）与提取效率。针对放大攻击的工作中，最为显著的一点在于我们只需使用pdns而不是完整报文，

在提取DDoS的流量中，我们首先尝试了使用逻辑回归，希望可以利用少量属性，进行二分类，并尝试人工标注一些可疑域名视为恶意的以提高召回率，但是即使这么做，正例数量与反例数量差别还是很大，导致我们始终无法得到稳定可用的各特征对应的权值。因此我们考虑将更为实用的统计方法与观察经验相结合进行分类。对于放大攻击，我们关注TXT查询与ANY查询的占比。txt记录用来保存域名的附加文本信息，txt记录的内容按照一定的格式编写，最常用的是spf格式，spf用于登记某个域名拥有的用来外发邮件的所有ip地址，攻击者利用txt记录进行放大攻击时，会利用自行注册的域名，并将该域名的txt记录内容设置的尽可能长，以提高放大倍数。Any查询将对该域名所有记录进行查询，攻击者选择使用any查询，可以获得尽可能大的放大倍数。只关注TXT查询和ANY查询的比例，是为了尽可能减少提取的时间，在做一遍遍历的过程就可以记录下每个域名TXT查询与ANY查询的数量，以及所有查询的数量。对于随机子域名攻击流量的提取与上文类似，我们主要关注一个域名的子域名数量与查询成功率，在提取的过程中要注意清洗掉反向域名解析记录和一些配置错误的域名，前者后缀为.in-addr.arpa，后者后缀为通常为‘localhost’或者‘local’，两者都会产生大量的子域名和无法解析的记录。

对于DGA域名的提取，只使用黑名单过滤效果很差，简单的使用统计数据进行拟合会由于DGA种类的多样性和复杂性而导致提取比率很低，而构建复杂的模型和检测方法耗费时间又太多，因此我们选择使用黑白名单训练模型，同时只选取域名的字符特征。

我们的结构很好，可以达到怎样的效果，时间、比例。。。

## 3 系统介绍

该系统的目标是提取放大攻击、随机子域名攻击与DGA产生的DNS流量，给定一定连续时间的流量，甚至只有一个小时的流量数据，我们就可以初步判断其中恶意的部分。直观上看，我们可以将一批DNS数据分成两个部分，一部分数据只包含合法的访问，另一个部分数据包含了所有的恶意流量。我们的数据来源是被动DNS数据，其中包含了递归域名服务器响应的历史信息，我们收集了分别来自于山西和广东的递归DNS服务器数据，山西省平均每小时的数据量达到8千万条数据，而广东省每小时平均达到一亿七千万条数据。在提取DGA域名的流量中，我们需要黑名单与白名单，黑名单来自于360安全实验室公布的，白名单采用alexatop100万。

### 3.1 系统概述

在本节中，我们对我们的系统进行一个高层次的概括。如图一所示，我们将系统分为两个部分：pdns预处理模块和流量提取模块。我们将对每个模块功能进行表述，并讨论如何共同达到目的并且尽可能提高召回率与效率。

预处理的过程主要达到三个目标，一是剔除包含错误的域名的DNS记录，即进行数据清洗，例如请求的域名中出现了‘？’，‘！’等符号；二是剔除无关流量，一方面将良性的流量和恶意的流量中都不会出现的流量剔除，例如逆向解析流量。另一方面我们在针对某一类型流量提取过程中，也会尝试剔除不会出现在恶意流量中的域名，例如DGA流量的提取可以剔除白名单中的域名。三是为流量提取模块计算特征值，例如每个域名的ANY查询的比例等。

### 3.2 统计特征

这一部分我们会定义关键的统计特征和计算方法。

我们针对每一个域名计算9个特征，分别是Bigram、Trigram、Fourgram，熵值、长度、any比例、txt比例、子域名数量、解析失败率

定义与符号

域名由两组或两组以上的ASCII或各国语言字符构成，各组字符间由‘。’分隔开，最右边的字符组称为顶级域名，二级域名是指顶级域名之下的域名，在国际顶级域名下，它是指域名注册人的网上名称。三级域名同理在二级域名之下，可以称其为子域名。

n-gram 特征

将白名单中的域名只取出二级域名，每一个二级域名在开头加上‘^’，结尾加上‘$’，例如buaa.edu.cn变成了‘^buaa&’，以此作为训练语料库测量N-gram频率分布，其中n=2,3,4。然后每给定一个域名，可以利用其二级域名分别获得相应的Bigram、Trigram、Fourgram特征

熵的特征

给定一个域名，我们提取出二级域名，按照公式1来计算其熵值，其中p（c）为二级域名中每一个字符的出现概率。熵值可以表现出一个域名构成的随机程度。

比例相关的特征（qn,qan,qtn,子域名数量、nxdr）

通过DNS记录可以计算一些比例相关的特征，通过比例而不是绝对数量来表现特征，例如ANY类型的查询占总查询数量的比例等，

### 3.3 PDNS预处理

预处理模块对采集来的原始数据进行三层处理。分别是数据清晰、数据过滤和特征提取。

首先对所有数据进行清洗，一个合法的域名只包含26个英文字母（包括大写和小写）、数字、中划线和用来分割成每一段的点。通过构建正则表达式很容易可以将这些数据清理干净。还有一部分是无顶级域名的域名，产生这种情况的原因比较多，有配置或者人为的各种原因。

其次，进行数据的过滤操作。第一部分先过滤反向解析域名，域名反向解析是指从IP地址到域名的映射，其主要应用于邮件服务器阻止垃圾邮件。为了实现逆向域名解析，因此有一个专门DNS服务器负责反向解析，返回数据包较小，不适合用作放大攻击，因此恶意流量中没有出现利用反向解析记录进行攻击的行为。第二部分要过滤配置错误产生的域名，这类域名极为常见，这其中以‘local’,’localhost’为后缀的域名出现最多。第三部分过滤国际化域名，国际化域名是指非英语国家推广本国语言域名系统的一个总称，使用punycode编码。

该模块另一个主要的功能就是将一条条DNS记录转化成对应的特征。按照一小时将一天的DNS记录划分，构成一个DNS序列，其中每一条DNS记录都包含有域名、源IP、请求类型、返回类型、rcode、时间戳等信息，在顶级域名下的每一个二级域名都是一个独立的分支，由不同的域名注册人管理，因此我们的统计特征也都是针对二级域名。每个二级域名在一定时间内请求总量qc，Any类型请求数量qac，TXT类型请求数量qtc，我们设定any查询比例qar=qac/qc,txt记录查询比例qtr=qtc/qc，同样的根据rcode来判断该域名是否能够成功解析，并计算其中不存在域名出现的比例nxdr，统计每一个二级域名下在当前时间段中出现的子域名数量sdc，这里我们时间间隔选择为1小时，将在4.2节讨论。然后根据得到的二级域名分别计算其熵值、Bigram、Trigram、Fourgram，长度，并标注是否在alexatop100万白名单中。

最终我们得到每一条记录为二级域名、qar、qtr、nxdr、sdc、熵值、Bigram、Trigram、Fourgram、长度、inwhite。

### 3.4 域名的提取

这一模块的工作是要利用二级域名及其特征将其中涉及到恶意行为的域名提取出来，提取过程针对三类攻击的模式，分别是放大攻击、随机子域名攻击、DGA域名。

3.4.1 放大攻击域名

这一部分主要目的是要将放大攻击中作为跳板的那部分域名找出来。攻击者想要利用这些域名，必然是利用其TXT记录或者ANY查询返回该域名所有资源记录。我们获取ANY类型请求比例qar和TXT类型请求比例qtr，带入公式2中，β为我们设定的一个参数，当qar+qtr<=β时，结果为0，当qar+qtr>β时qar+qtr与s1成正相关，同时设定阈值α，其中s1>α，我们认定为疑似放大攻击的流量。这里α取值为，β取值为，阈值的取值将在下一章仔细讨论。

3.4.2 随机子域名攻击域名

这一部分目的是将随机子域名攻击所使用的那部分域名找出来，攻击者会在二级域名下伪随机的生成大量子域名，这些域名都是不存在的。因此我们用sdc与nxdr相乘来表示被恶意使用的可能性大小，而这个值范围较大，我们使用公式3将结果变到0到1之间。这里我们的θ取值为0.3.

3.4.3 DGA域名

AGDs意思是由DGA算法生成的域名，

域名生成算法产生的域名也是二级域名的部分，所以这部分的检测关注针对流量中的二级域名，

我们模型使用黑白名单训练，数据来源在4.1节中介绍。将名单中每个域名的二级域名提取出来，分别计算其长度、熵值、n-gram，其中n=2,3,4。分类器我们选择随机森林，训练每棵树时，从全部训练样本（样本数为N）中选取一个可能有重复的大小同样为N的数据集进行训练（即bootstrap取样）。随机森林的训练速度快，并且可以平衡误差。

我们从上一个模块中获得的域名，对于每一个二级域名，我们先剔除白名单中的域名，然后对于剩余域名依次放入分类其中进行分类，获得疑似的ADGs。

### 3.5 流量提取

在做进一步的检测时，往往需要完整的流量进行处理，因此在通过3.4节获得可以的域名之后，这一模块将这一组二级域名还原成对应的流量，即原始的DNS资源记录。操作过程非常简单，这批二级域名作为一个名单，与数据清洗之后的DNS记录相比对，保留二级域名一致的即可。

### 3.4 随机子域名分析模块

在做这部分提取的时候，我们先要选择对数据进行过滤，在保证召回率的同时，也要尽可能提高提取率，而有些类型的域名会在记录中频繁出现造成干扰。首先是反向域名解析域名，在分析DNS记录中的域名项时，‘.in-addr.arpa’将被认为是一个完整的二级域名，而需要解析的ip地址作为其子域名前缀，因此误认为产生大量随机子域名；其次是配置错误域名，出现最多的三类后缀为‘.Localhost’、’.local’、’.localname’。

对于这一部分流量我们主要关注一个域名的子域名数量和解析成功率，和对放大攻击的的处理类似，我们先选取分别为10、30、60、120的时间间隔（分钟），针对每一个二级域名我们统计在每一个时间间隔之下的不同子域名个数sdc和该域名的解析成功率nxdr，随机子域名攻击所利用的域名会出现大量不同的并且无法访问的子域名，因此我们使用g=sdc\*nxdr来表示随机子域名攻击的可能性，并构建了一个类似于sigmoid函数的函数用于将g(x)的值映射到0到1之间，。同样我们也设定阈值Θ，当s2>θ时，我们认为存在随机子域名攻击的行为，反之认为不存在相关行为。

### 3.6 流量提取模块

通过

## 4 评估

在这一章节中，我们报告一下整个系统的评估结果。首先我们介绍一下我们所使用的数据，然后在4.2节中讨论一下参数的确定。最后把流量提取的结果展示出来。

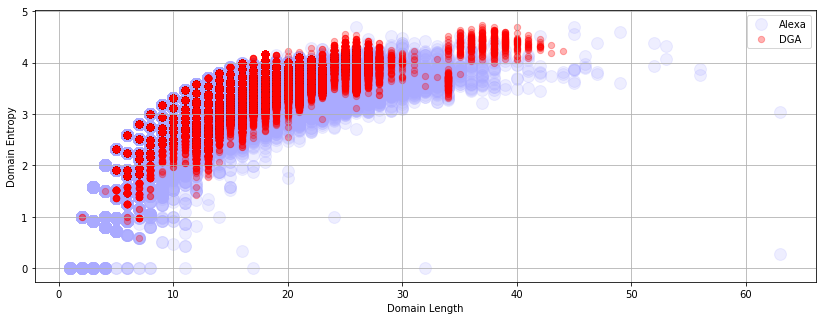
### 4.1 数据收集

我们使用了来自CNCERT/CC提供的数据，包括了山西省的中国电信pdns数据和广东电信的PDNS数据。

其中山西省数据为2015.10.15 23个小时的数据，如图所示，DNS记录的总量接近20亿条，每小时不重复的二级域名数量在10万到20万之间，共标注DDOS相关恶意域名101个，DGA相关域名322个。

广东省的数据为2017.4.14 9点到16点共计9个小时的数据，如图所示，DNS记录总量达11亿余条，每小时不重复的二级域名数量达到60万左右，共标注DDOS相关恶意域名163个，DGA相关域名265个。

在进行DGA流量的提取中，我们使用alexatop100万的域名列表作为白名单，黑名单我们将从360安全实验室下载的DGA黑名单作为黑名单，包含1037304条二级域名。图展示了黑名单与白名单中的二级域名长度和熵值之间的关系，在长度一致的时候，dga的域名往往与有更大的熵值。

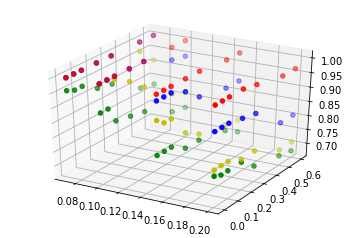


### 4.2 参数的讨论

我们先3.4.1节中的参数进行讨论，这里的参数涉及到了三个，分别是时间间隔、判断域名是否可疑的阈值α、公式1中的参数β。其中时间间隔取值为10、30、60、120，α取值为0、0.05、0.1、0.2、0.3、0.5、0.6，β的取值为0.07、0.1、0.15、0.2，如图所示，z轴代表召回率，当时间间隔为10时，召回率始终在0.98以上，当时间间隔为120时，召回率最高为0.94，都无法满意的对α与β取值。表1展示了在不同参数的取值与可疑域名数量之间的关系。可以看到时间间隔选为30和时间间隔为60所能达到的效果差别不大，而前者的执行次数为后者的两倍，因此我们将时间间隔定为60分钟，为了获得尽可能小的域名数量，我们将α设定为0.1，β设定为0.05。

另一个需要讨论的参数是3.4.2节中公式3中的θ，θ的值越小，这个函数的曲线越平滑。如图3所示，描述了θ的取值和提取出来的域名的数量之间的关系。其中当θ为0.1时，召回率为50%。当θ取值大于0.2时，召回率达到100%。从图中可以看出θ的取值域名数量保持正相关，为避免过拟合，我们将θ取值为0.3。

|  |  |  |  |
| --- | --- | --- | --- |
| 0.07 | 0.05 | 784 | 764 |
| 0.07 | 0.1 | 779 | 758 |
| 0.07 | 0.2 | 768 | 749 |
| 0.07 | 0.3 | 762 | 738 |
| 0.1 | 0 | 751 | 731 |
| 0.1 | 0.05 | 750 | 730 |



[0.1, 37, 0]

[0.2, 50, 1]

[0.3, 66, 1]

[0.4, 77, 1]

[0.5, 90, 1]

[0.6, 98, 1]

[0.7, 112, 1]

[0.8, 121, 1]

[0.9, 130, 1]

[1.0, 142, 1]

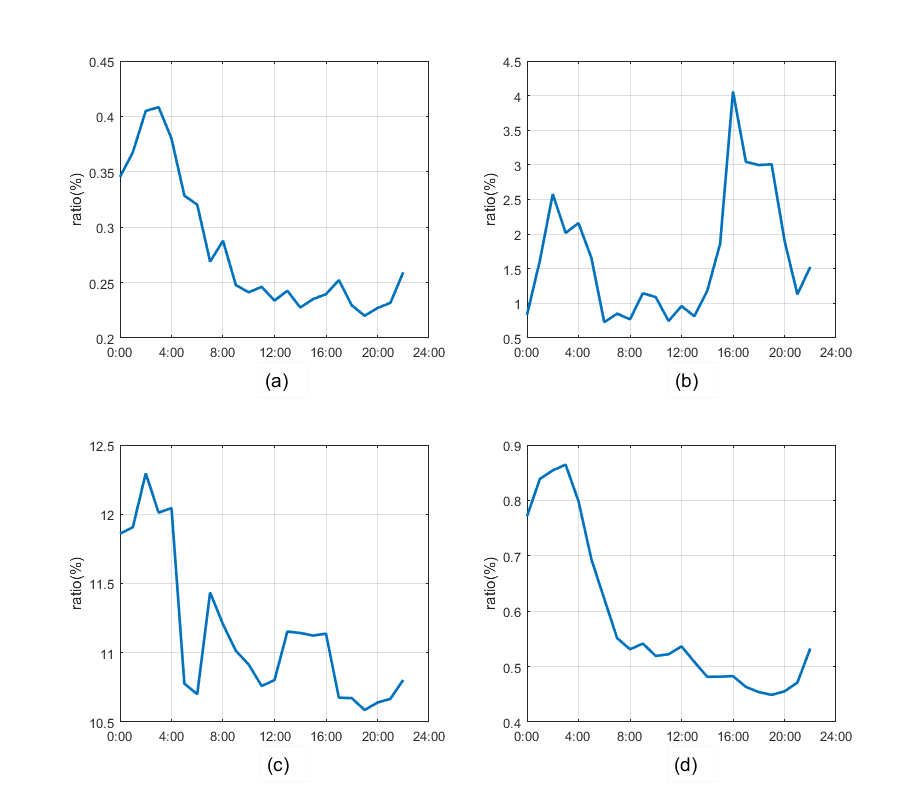
最后讨论一下我们在DGA域名的处理过程中的特征选择，一方面我们不能选择过多的特征，以保证模型的训练时间和分类时所花费的时间不能过长，同时也不能选择过少的特征，不然难以达到满意的召回率。最后我们选取了二级域名的长度、熵值以及bigram、trigram、fourgram，这五个特征。使用随机森林作为分类器。如图7所示，我们通过混淆矩阵来观察结果，召回率达到96以上，并且进行了十折交叉验证，召回率都在85%以上，最高可以达到98%以上。

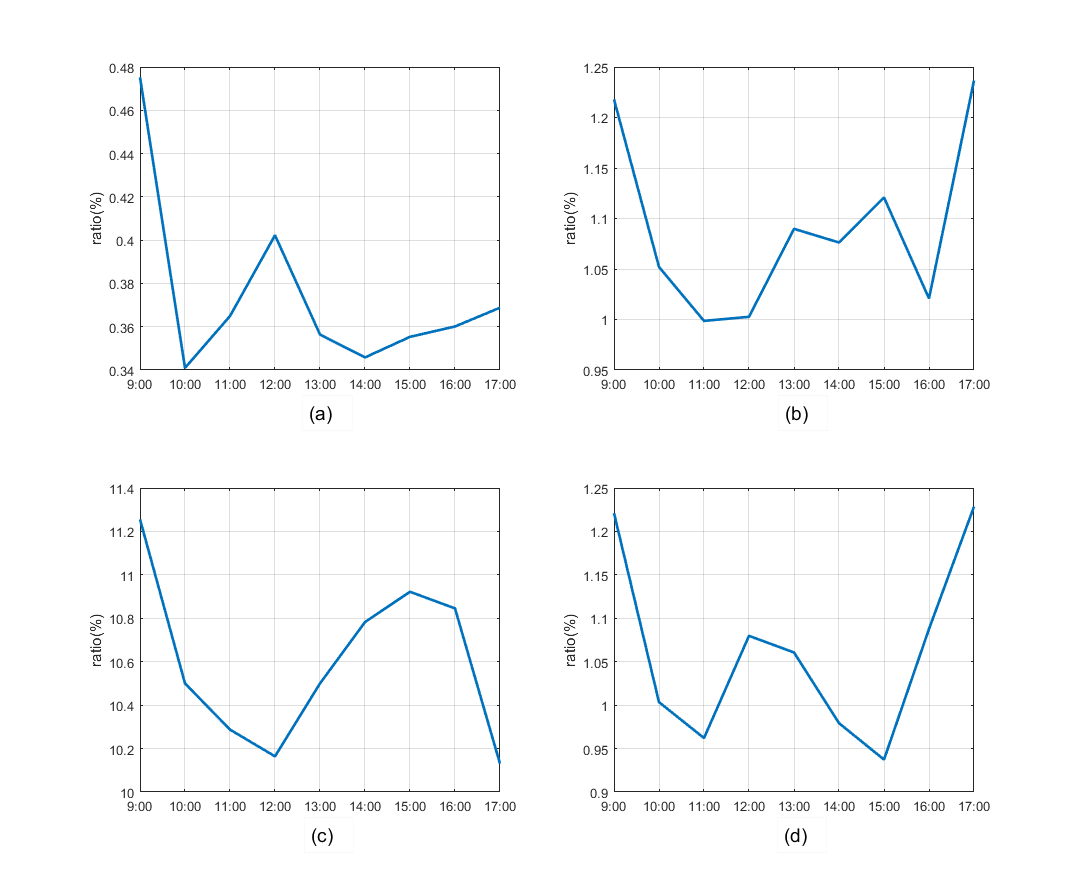
### 4.3 模型的效果

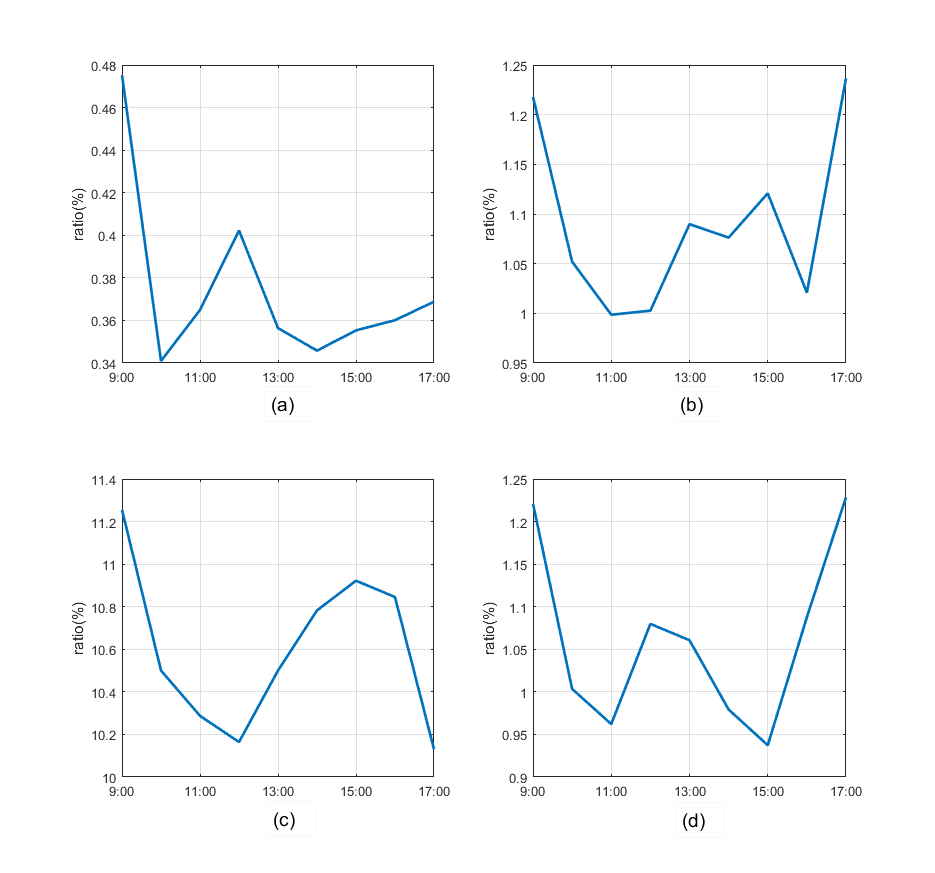
在4.2节中，我们利用山西省的数据进行实验，确定了相关的参数，如图8所示。使用同样的参数我们对广东的数据也进行了实验，如图9所示。图8（a）展示了随即子域名攻击和放大攻击所涉及的域名数量的变化趋势，峰值约为谷值的两倍左右，而观察图7（b），变化趋势与域名数量的变化不尽相同，峰值有谷点的7倍左右，这个更能反映放大攻击和随机子域名攻击的一个攻击情况。图7（c）展示了提取的DGA域名的变化趋势，由于是我们只通过域名的字符特征进行判断，所以其中包含了大量合法的域名，因此其DNS的记录的大小并不能真实反映真实攻击流量的大小，另外，这里我们的DGA域名的召回率为92%，

对广东的数据的实验结果中，放大攻击和随机子域名攻击的召回率同样达到了100%，但是对DGA的召回率之后接近90%。关于如何进一步提高召回率，我们在第五章中提出了一些想法并进行了一些讨论。

变化趋势、ddos域名和dns记录的关系、







## 5 限制与讨论

在这篇论文中，我们构建了一个恶意流量的定位系统，他可以快速的将放大攻击、随机子域名攻击和DGA产生域名的DNS记录从递归服务器的被动DNS数据中提取出来，而不是仅仅通过黑名单过滤，可以极大的降低进一步作为检测的DNS数据量。

我们的系统也有一些限制。例如由于缺乏连续的长时间的数据来源，缺乏充分的历史数据支持，我们认为利用历史数据中的一些统计特征构建数据仓库，可以更为清晰的判断域名。利用一个域名在过去30天内是否出现过或者是否每天都出现来对DGA流量过滤，很明显DGA域名的生存时间通常远远小于30天，这一特征可以更大限度的对域名进行甄别。相同原因也造成我们无法对该模型的时效性进行探讨，无法确定参数的有效性，但是我们在文章中提供了对参数进行探索的方法。

对于DGA流量的提取中，召回率并没有达到100%。对于如何提升召回率，我们也存在一些想法。上一段中提到的增加特征是一方面，另一方面是对DGA更好的分类，分清DGA是利用字典生成的还是hash生成的，亦或是根据时间伪随机生成的。但同时也要考虑平衡时间与召回率的关系，越接近一个完整的检测模型，所花费的时间越多，这也是我们需要重点考虑的问题。

再针对DDOS流量的提取中，我们的方法只能选取出网站的二级域名。一旦涉及到一些正常流量比较高的网站，转化到实际的流量中的时候，无法将同一个网站下的合法流量和恶意的流量区分开来，这可能会降低域名的提取效率

## 6 总结

在这篇文章中，我们提出了一个恶意流量的定位模型系统，该系统参考相关的检测系统，选取特征，初步对一些攻击行为涉及到的域名和流量进行一个筛选。相比于简单的预处理过程，该系统可以将恶意流量选取到一个较小的范围中，在保证召回率的同时，也能很快捷的达到目的。为进一步检测缩小范围。我们的评估使用真实的数据，来自省级电信的不同时间的被动DNS数据。放大攻击和随机子域名攻击涉及到的域名召回率均达到100%，DGA域名召回率达到90%以上。