摘要许多僵尸主机使用域生成算法（DGA）生成大量恶意算法生成的域（MAGD），然后选择多个MAGD进行实际的指挥控制（C2）通信。僵尸主机使用不同的种子（例如时间戳）来生成不同的magd，这使得通信机制能够适应黑名单。因此僵尸大师可以很好地隐藏C2通道。如果我们能从DNS流量中快速检测到这些magd，我们将有效地阻止通信。在本文中，我们提出一个新的系统，称为pontus，从DNS流量检测磁碟。Pontus只从单个域名中提取特征。我们将pontus系统与最先进的系统进行了比较，发现庞图系统的精度至少提高了4.7%。

A、 背景和动机

在僵尸网络中，攻击者（botmaster）通过命令和控制（C2）通信通道远程控制一组恶意软件受损机器（bot）。要联系僵尸主机，机器人通常会访问C2服务器的硬编码域名或IP地址，以从僵尸主机接收命令。这种方式有一个明显的缺点，即网络安全工作人员在找到固定域名或IP地址时，可以摧毁整个僵尸网络。为了有效地隐藏C2服务器，许多先进的僵尸网络，如Mirai[2]，都采用了域生成算法（DGAs）作为一种新的通信机制。在这种机制中，bot和botmaster周期性地使用相同的种子（例如，时间戳）执行相同的域生成算法，以生成候选恶意域名的列表。僵尸主机将选择几个域名进行注册。机器人程序查询所有这些恶意算法生成的域（magd），直到其中一个域名可以解析为C2服务器的有效IP地址，然后才能与僵尸主机建立连接。很快，攻击者就会放弃域名和IP地址，使用新种子生成的magd。因此，要及时发现用于C2通信的magd并阻断僵尸网络是非常困难的。这就是为什么很多类型的僵尸网络仍然存在于互联网上。本文设计了一个基于域名字符串的差分gps检测系统，该系统能够有效地从域名系统（DNS）流量中识别mAGDs。

B、 现有技术的限制

深入理解DGA检测是网络安全中的一个重要问题。这一领域的研究兴趣已经持续了十年[3]-[6]。最近和相关的工作是FANCI[3]。FANCI利用有监督的学习分类器来检测magd，它产生了非常高的分类精度。然而，我们发现FANCI有两个

基本缺陷。

首先，FANCI在二级域（2LDs）和三级域（3LDs）中具有高假阳性。域名由一系列由点分隔的子域组成。最右边的子域称为TLD，2LD表示由点分隔的两个最右边的子域，依此类推（例如。，baidu.com是2LD，api.baidu.com网站是3ld）。FANCI采用DGArchive中的全限定域名（FQDN）[12]和2ld或3ld的magd作为训练数据。FANCI可以根据域名的长度、子域的数目等结构上的差异很容易区分它们。这种方法在企业级或校园级网络中是有效的，因为2ld和3ld的数量很少。但是在ISP网络中有太多的2ld和3ld，FANCI将许多良性的2ld和3ld判断为magd。

其次，FANCI对基于wordlist的magd的检测性能较差。我们大致将所有的magd分为两类：随机浏览的magd和基于单词列表的magd。随机查找的magd的字符分布看起来是随机的，因为magd是通过从字母表中随机选择字符、散列或排列初始域名生成的。基于wordlist的magd由wordlist中的几个单词连接起来。基于wordlist的magd遵循常规的语言模式，因为许多良性域名也由多个单词连接起来。因此，基于wordlist的magd很难被检测到。FANCI的特性不足以反映良性域名和基于wordlist的mAGDs之间的区别，因此FANCI无法准确区分基于wordlist的mAGDs和良性域名。

C、 提议的方法

本文提出了一种基于强大的语言学特征的DGA检测系统Pontus。Pontus的特征完全是从单个域名中提取出来的，仍然具有良好的分类性能。

我们的系统基于这样一个关键的观点：良性域名和magd在语言方面有很大的不同。良性域名通常代表一些特定的含义，如商标名、人名。这些域名通常遵循规则的语言模式，以便流畅地阅读或容易记住。然而，这些随机出现的杂志违背了正常的语言模式。虽然基于单词列表的magd遵循规则的语言模式，但它们可以完全分成2个或3个单词。有时，两个或三个单词之间用连字符隔开。

Pontus的输入是域名，输出是每个域名的标签，良性的或磁性的。如图1所示，Pontus有两个主要阶段：1）训练阶段和2）分类阶段。训练阶段使用标记的域名训练有监督的分类器。分类阶段从DNS流量中检测magd。

D、 主要贡献

我们在两个ISP的DNS递归服务器上收集良性域名的数据集上对Pontus进行评估，并通过DGArchive中的61个DGA生成MAGD[11]。Pontus的有效性是通过精确性、精确性和召回率来衡量的。总之，本文有三个主要贡献：

1） 我们提出了一种ISP规模的DGA检测系统。我们的系统只依赖于从单个域名中提取的语言学特征，不需要额外的信息，比如Whois。我们的系统仍然能够从庞大复杂的ISP规模的DNS流量中准确地检测mAGDs。

2） Pontus的特性准确地反映了良性域名和基于单词表的mAGDs之间的区别。基于Wordlist的mAGDs可以分成几个单词。因此，我们对一个基于单词表的AGD进行分割，然后从分割后的单词中提取一些特征。因此，Pontus具有很好的检测基于单词表的DGAs的能力。

3） Pontus对mAGDs的检测能力很强。在相同的数据集上，FANCI的准确率为92.4%，FANCI的准确率为92.8%，FANCI的准确率为96.7%，召回率为96.5%。Pontus的表现明显优于FANCI。

论文的其余部分安排如下。在第二节中，我们将介绍Pontus的结构。在第三节中，我们详细介绍了Pontus的特点。在第四节中，我们进行了实验来评估Pontus的性能，并将Pontus的实验结果与FANCI的实验结果进行了比较。最后，我们在第五节中对本文进行了总结。

二。Pontus结构

Pontus由训练阶段和分类阶段组成。图1显示了Pontus的建筑。

A、 培训阶段

培训阶段是本系统的基础。该阶段包括四个模块，即域过滤、良性域名抽取、特征抽取和模型训练器。

1） 域筛选器。域过滤的目的是过滤掉典型的良性域名和无效域名。在这个模块中，我们过滤掉域名：a）没有无效tld的域名；b）域

白名单上的名字。白名单由Alexa top 100000域名后缀[10]的域名、一些CDN域名和反向DNS域名组成。c） 现有magd。d） nLDs（n>3）。因为现有文献表明，大多数magd是2ld，少数dga生成3ld，通常是动态域名（例如。，ugrpuoo.dyndns.org网站)[七]

2） 良性域名提取程序。良性域名抽取器获取良性数据用于训练分类器。首先，我们获得最常访问的2ld和3ld。然后我们在一个可靠的安全分析网站上查看这些域名。VirusTotal收集恶意软件生成的许多恶意域名。我们删除被VirusTotal标记为恶意的域名，以确保我们的良性数据集尽可能不包含magd。

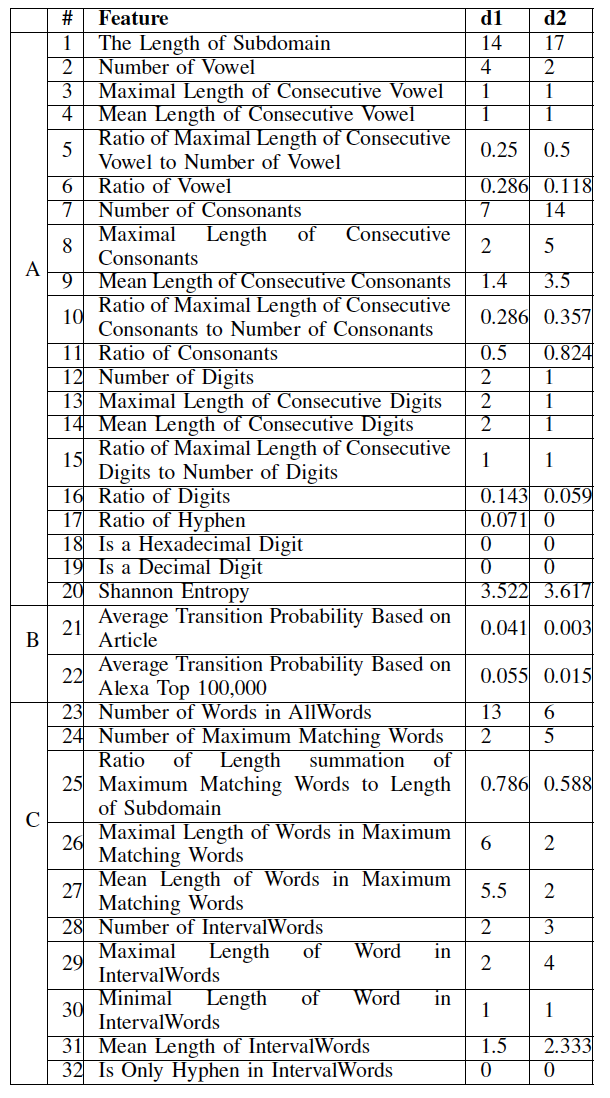
3） 特征提取器。该模块的输入为良性域名抽取器抽取的良性2ld和3ld以及61个DGAs从DGArchive生成的magd。输出为输入域名的特征向量和对应标签。首先，Pontus从没有公共后缀的域名子域中提取特征[9]。如果域名有两个子域名，我们将两个子域名的特征向量连接起来。如果域名有一个子域名，我们将使用零向量替换空白空间。例如api.baidu.com网站是api的特征向量和百度的特征向量的组合，但是雅虎网由零向量和yahoo的特征向量组成。

4） 模特教练。模型训练器使用特征提取器中的特征和标签为分类阶段训练模型。

B、 分类阶段

分类阶段是Pontus的第二阶段，主要用于对未标记的域名进行分类。分类阶段包括三个模块：域过滤器、特征提取器和域分类器。在域过滤器中，我们过滤掉已被识别为良性的域名和已出现的mdga。然后我们从其他2ld和3ld中提取特征。最后，我们使用训练阶段提到的模型来检测领域分类器模块中的magd。

三、 特征



这些特征是从子域中提取的，而不是从FQDN中提取的。很简单，因为百度，功能是从百度中提取的，而不是百度. 我们将这些特征分为三类：分布特征、转换概率特征和词汇特征。分布特征着重于元音、辅音和数字的字符分布。过渡

概率特征表示相邻字符之间转换概率的平均值。词特征是指子域进行分词时从词和区间中提取的特征。在表一中，我们通过对域名d1=“grand-casino50.com”和d2=“trlekmynqihxxhy6k.com”的示例评估，对所有功能进行了概述，其中d1是一个良性域名，d2是一个mAGD。

A、 分布特征

由于许多DGA通过从字母表中随机选择字符来生成域名，因此良性域名和MAGD之间的字符分布会有很大的差异。因此，我们可以通过分析字符分布来区分良性域名和magd。所有分布特征见表I-A。我们只解释那些不易理解的特征。（#18）是十六进制数字。如果子域的所有字符都在0-F中，我们认为子域是十六进制数字。特别是在（#12、#13、#14、#15、#16和#19）中，数字是指0-9。（21）香农熵。S是一个域名的1g集，c是S中的一个字符，P（c）是c的概率，Shannon熵–∑c∈sp（c）×log2（P（c））。

B、 转移概率特征

当域名遵循规则的语言模式时，域名的某些字符对通常出现在其他单词或域名中，因为字符对可能是音节或有意义的缩写（例如，Google、Yahoo、good具有相同的字符对-oo因为oo是音节）。根据已有的文献[8]，我们设计了转移概率特征来说明域名是否遵循规则的语言模式。首先，我们使用一个语料库来获得一个字符出现时的转换概率。然后计算域名所有相邻字符的平均转移概率。算法1给出了转换概率特征提取的细节。我们执行名为getFeature的函数来获取值。一个字符对频繁出现在语料库中，意味着从前一个字符到后一个字符的转换概率很高。因此，域名的字符对在语料库中出现的频率越高，转换概率特征值越高。所有转移概率特征如表I-B所示。

（21）基于文章的平均转移概率。我们使用文献[13]作为语料库来计算域名的转移概率。

（22）基于Alexa Top 100000的平均转移概率。我们以AlexaTop100000为语料库计算域名的转移概率。

C、 文字特征

随机查找的magd通常不包含单词，基于单词表的magd包含2或3个单词，这激励我们通过分割和最大匹配来提取特征。

首先，我们在子域上进行分割。我们称子域中包含的所有单词为all words。然后执行最大匹配操作。我们将子域按从左到右的顺序尽可能长地划分成单词。我们称这些词为最大匹配词。同时，我们最大限度地调用单词中的所有字符串

匹配词作为间隔词。所有文字特征见表I-C。

图2清楚地显示了应用于d1=“grand-casino50.com”的分段和最大匹配过程。对我们来说，理解定义的含义是非常直观的。

（23）所有单词中的单词数。隐藏在子域中的字数。对于d1，数字是13。

（25）最大匹配词长度和域名长度之比。最大匹配字中字的总长度与子域长度的比率。对于d1，比率=11/14=0.786。

（#26）最大匹配单词中单词的最大长度。最大匹配字中最长字的长度。对于d1，值为6（“casino”的长度）。

（#29）间隔词的最大字长。该功能类似于（#26）。对于d1，长度为2（长度为“50”）。

（#32）只是间字中的连字符。间隔词中的所有字符串都是连字符。对于d1，该值为0，因为“50”不是连字符。

A、 数据集

在我们的数据集中，我们谨慎地将访问频率最高的2ld和3ld视为良性域名，并在VirusTotal中将其标记为良性域名，找到可靠和权威的AGD源DGArchive[11]。简而言之，我们的良性域名来自以下来源：

•Alexa Top List：该列表包含全球最受欢迎的100万个域名。我们需要前10万个作为一个语料库来提取特征，所以我们只考虑排名从100001到1000000的域名。

•中国电信原始榜单：我们统计中国电信（一家互联网服务提供商）递归DNS服务器上捕获的DNS流量中的域名频率。过滤掉Alexa top之后

100000和一些CDN域名，我们选择了前300002lds和3LDs作为良性域名。

•中国移动原始榜单：相似之处在于，该榜单包含了中国DNS服务器在DNS流量中查询最多的前30000个2LD和3LD中国移动移动是另一个互联网服务提供商。

我们的恶意磁碟来自DGArchive：

•DGArchive：到目前为止，DGArchive从现在起收集89个DGA，并提供这些DGA生成的磁盘。因为有些DGA只产生极少量的磁碟，所以我们只选择61个DGA。这些DGA要么有500多个磁碟，要么产生3LD。我们在一周内（从2018年4月27日到2018年5月3日）从ISP DNS服务器收集DNS流量，并获得每天的最高列表。换句话说，我们每个ISP有7个数据集。当我们建立7天的数据集时，我们随机地

从每个mAGD中选择500个mAGD，并采样相同总数的良性域名。如果某些DGA不能生成500个磁碟，我们将选择所有磁碟。我们尝试为每个数据集制作不同的magd。由于不同的用户习惯，只有40%的域名在不同的ISP数据集中重叠。

B、 分类器选择

我们将三个分类器作为候选分类器：梯度提升决策树（GBDT）、支持向量机（SVM）和随机森林（RF）。分类准确度（ACC）和训练时间（time）是我们最关注的两个指标。精确性是指Pontus能否找到Magd来实现我们的目标，而培训时间则反映了实际环境中的可用性。在本节中，我们将并行检查这些分类器在两个数据集上的性能。第一个数据集是中国移动的一天数据集。另一组数据来自中国电信。表二给出了三个分类器的分类精度和训练时间。无论在分类能力还是训练时间上，梯度提升决策树（GBDT）都是最适合的分类器。因此，在接下来的所有实验中，我们都使用GBDT作为分类器。

C、 概括

接下来，我们测试Pontus的泛化。一方面，我们利用在一个数据集中训练的模型来测试属于同一ISP的其他数据集。这样，我们就可以知道当DNS流量随时间变化时，Pontus是否能够很好地推广。另一方面，我们在一个ISP数据集中训练一个模型，并预测另一个ISP的数据集，以验证部署环境变化时系统的通用性。同一ISP上的训练与预测。我们在一天的数据集上训练一个模型，然后使用该模型预测其他天的数据集。我们从一天的数据集中提取80%的域名。在过滤掉训练域名并平衡其他数据集中的良性域名和magd之后，我们预测其余数据以评估Pontus的泛化能力。表三显示了这个实验的结果。在本实验中，庞图仍然具有很高的分类精度。中国移动和中国电信的平均准确率分别为96.6%和95.5%。此外，由于标准差在两个网络中都很小，所以Pontus是稳健的。

不同ISP的训练和预测。在本实验中，我们藉由训练在不同ISP资料集上的模型来预测ISP资料集。我们关注的是庞图是否能很好地推广到不同的网络。与上面的实验类似，我们从一天的数据集中提取80%的域名，然后预测另一个ISP的其他数据集。根据表四的结果，我们发现无论是使用中国移动训练的模型预测中国电信的数据集，还是使用中国电信训练的模型预测中国移动的数据集，均具有较高的平均精度和较小的标准差。因此，Pontus可以很好地推广到不同的网络。