FANCI : Feature-based Automated NXDomain Classification and Intelligence全文翻译

摘要

FANCI是一个新的系统，它通过监视DNS流量中不存在的域（NXD）响应来检测基于域生成算法（DGA）的恶意软件感染。它依赖于基于机器学习的NXD分类（即，包含在失败DNS响应中的域名）为DGA相关的良性NXD。分类特征仅从要分类的单个NXD中提取。我们对59个DGAs从dgaarchive中生成的恶意数据、一个大型大学校园网中记录的数据以及一个大公司内部网络上记录的数据进行了评估。结果表明，该系统在较低的误报率下具有很高的分类精度，具有很好的泛化能力，能够识别未知的dga。

1简介

现代僵尸网络依赖域生成算法（DGAs）来建立与指挥控制（C2）服务器的连接，而不是使用固定的域名或固定的IP地址[14,2]。根据DGArchive1，到目前为止，已知有72种不同的dga，并且随着dga显著提高僵尸网络抵御攻击的能力，这个数字有望进一步增加[14]。DGA生成一组恶意的算法生成域（magd），作为与C2服务器的潜在共同域。bot随后会查询域名系统（DNS）以获取这些域的IP地址。依靠DGA，每天生成的域数量在1到10000之间变化[14]。botmaster注册了一些这样的mAGD。如果这些被机器人查询，bot将获得其C2服务器的有效IP地址。bot的许多其他查询都将导致不存在的域（NXD）响应。

过去，监视DNS流量（成功解析和/或不解析）被用作检测网络中恶意活动的主要或附加信息源（例如[2、16、18、9、4]）。这些方法中有些集中在识别C2服务器上，有些则集中在识别受感染的设备或检测一般的恶意域上。然而，这些先前的方法都需要从DNS查询和/或响应组中提取的信息的相关性，因此通常需要广泛的跟踪。此外，许多先前的方法都是基于聚类的，这涉及到手动标记已识别的聚类。虽然这些先前的工作显示了有希望的检测能力，但很少有关于检测过程在时间和内存需求方面的效率的信息。

这项工作提出了FANCI：基于特征的自动NXDomain分类和智能，一个通过监测NXD响应来检测基于DGA的恶意软件感染的新系统。FANCI的分类模块使用机器学习（ML）分类器（随机森林（RFs）或支持向量机（SVM））将NXD分为良性不存在域（bNXDs）和MAGD。该分类器使用少量独立于语言的特征，这些特征可以单独从NXD响应中包含的域名中有效地提取出来。不需要从包含域名的完整NXD响应、其他相关DNS响应或任何其他源提取的其他上下文信息。

我们对FANCI的分类模块进行了广泛的评估，这些数据来自DGArchive[14]以及RWTH亚琛大学校园网和西门子AG3内部网络中记录的数据。结果表明，FANCI能够以很低的假阳性率检测出未知的dga，检测准确率达到99%以上。与以往的工作不同，我们还表明FANCI具有很好的泛化能力，即即使应用于不同于训练网络的数据，FANCI也能保持其检测质量。通过使用FANCI，我们能够识别出撰写本报告时DGArchive中未包含的10个DGA。我们估计其中至少有四种是完全未知的，而其他的很可能是由未知的种子或已知的dga的变体造成的。最后，我们的系统在训练（92102个样本上5.66分钟）和预测（每个样本0.0025秒）方面都是非常有效的，因此它甚至可以在没有采样的情况下在大型网络中进行离线检测。

FANCI的轻量级功能设计及其通用性允许多种应用场景，包括将其分类为服务，并将其用于大规模网络和家庭级硬件。

2准备工作

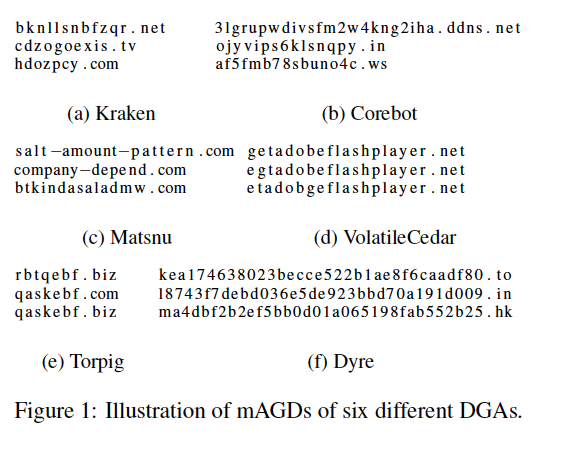
在本节中，我们将简要概述mAGDs的类型不同的dga生成并分类NXD响应中由于良性原因而出现的不同类型的域名。接下来是我们在本文中使用的监督学习分类器的概述。注意，在整个工作中，我们总是使用NXD响应来引用包含DNS响应的整个UDP4包。相比之下，我们将NXD称为此类响应中包含的裸域名。

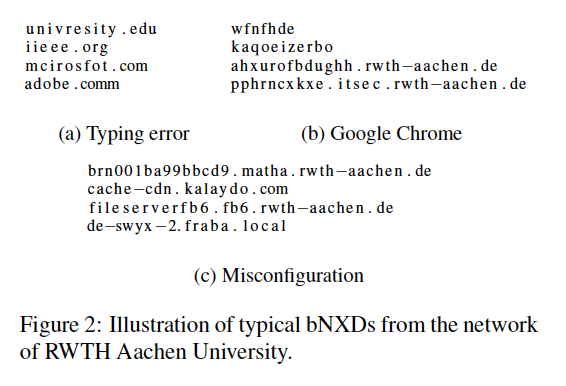
2.1 NXD响应中的域名

为了突出不同dga使用的生产方案的多样性，图1展示了六个不同dga的示例magd。在由Kraken、Corebot和Torpig生成的mAGDs看起来完全是随机的，Matsnu的mAGDs是真正的英语单词的串联。mAGDs of Volatile Cedar都是相同基础域名的置换，Dyre生成长度相等的magd，由3个字符的前缀后跟类似哈希的字符串组成。

除了由DGAs生成的NXDs（即mAGDs），主要有三组良性的不存在域（bNXDs），分别源于键入错误、错误配置和误用，其中错误配置和误用属于良性算法生成的域组（bAGDs）。bagd和mAGDs一样，都是通过算法生成的，但源于良性软件，只具有良性的目的。输入错误bNXDs是由人类引起的

拼写错误的现有域名。错误配置badds是由设备或软件试图解析由于配置错误或错误而不存在的域名引起的。误用bAGDs通常是由软件出于非预期目的而使用DNS引起的。例如，执行签名检查的防病毒软件[17]或谷歌Chrome，它使用随机域名来探测其DNS环境并检测DNS劫持企图[19]。图2显示了三个类别中每个类别的示例bNXDs。





2.2监督学习分类器

在我们的工作中，我们重点研究监督学习分类器，更具体地说是随机森林（RFs）和支持向量机（svm）使用两个标签良性和恶意。这些标签用于培训目的。RF是多个决策树（DTs）的集合，它克服了单个DT的局限性。使用RF预测未知样本的标签是由森林中所有DTs的多数票来执行的。RFs最初是在[10]中引入的，后来经过改进，例如在[5，6]中

支持向量机在训练过程中计算一个超平面，根据训练数据的标签来分离训练数据。然后，通过确定观测样本相对于该超平面的位置，可以预测未知数据。SVM由Vapnik引入[7]。

3．特征

在本节中，我们将介绍FANCI用于将NXD分为bNXDs和mAGDs的21个特性。我们将这些特征分为三类：结构特征、语言特征和统计特征。我们关注的是计算量轻的w.r.t.特征提取，不需要预先计算，也不需要先验知识，并且独立于特定的自然语言。

我们的特征设计自然受到相关工作中使用的特征的启发[14,2,16]。然而，我们关注的是可以从单个域名中提取的特征。特别是，我们去掉了以前工作中使用的所有需要附加上下文信息的特性，而不会损失（事实上相当于增加）准确性（见第6节）

3.1定义和符号

在本文的其余部分中，我们将使用下面详述的符号。

域名d是字母表中的一系列字符。它由一系列由点分隔的子域组成：d=sn….s2:s1，其中si表示d的第i个子域。请注意，合法域名中允许的字母表S取决于当地的注册机构。理论上，几乎所有的Unicode字符都是允许的[13]。

有效的顶级域（TLD）是一个TLD，它是由Internet分配号码管理局（IANA）维护的TLD的官方列表的一部分，例如，org、com、eu和edu[3]。目前，1547个有效的TLD列在根区域[11]。

公共后缀是域可公开注册的后缀。这包括有效的TLD和后缀，例如dyndns.org或者co.uk.。Mozilla基金会维护了一个超过11000个有效公共后缀的列表5[8]。

特征定义为样本d的函数，其中F（d）表示提取的特征。F（d）可以是单个标量，也可以是标量的向量。将所有提取的特征串联起来，得到d的特征向量。在下面的部分中，我们的一些特征（用\*标记）忽略分隔点，而一些（用†标记）忽略有效的公共后缀。忽略这两者的特性在一个称为无点公共后缀自由域（dot-free public-suffix-free domain）并由d-dsf表示的字符串上操作。例如考虑域名d =itsec.rwth-aachen.de产生了dds f = itsecrwth-aachen.。

请注意，我们忽略了一些特征中的分隔点，因为子域数量特性已经反映了域名的子域数量，而这些点本身并没有提供任何附加信息。我们忽略了一些特性中的公共后缀，因为它们不是通过算法生成的。虽然DGA可能会改变其magd的公共后缀，但它只能从官方可用的公共后缀库中进行选择，否则产生的域名将无法在公共互联网上解析。由于良性域名必须从完全相同的官方可用公共后缀池中选择公共后缀，公共后缀没有提供有价值的附加信息来区分mAGDs和bNXDs。

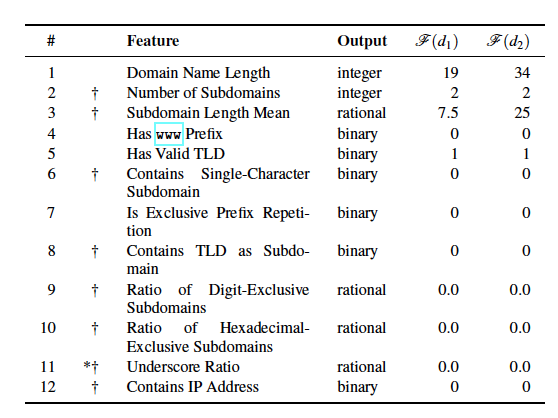


表1：在示例域d1和d2上评估的12个结构特征说明，其中d1 = bnxd.rwth-aachen.de d2 = dekh1her76avy0qnelivijwd1.ddns.net. 有些特征（用\*标记）忽略分隔点，有些（用†标记）忽略有效的公共后缀。

3.2结构特点

第一个特性类别集中于域名的结构属性。表1概述了我们的结构特征，包括对域名的一个示例评估d1 =bnxd.rwth-aachen.de d2 =dekh1her76avy0qnelivijwd1.ddns.net, 其中d1为良性，d2为已知mAGD。下面，我们将更详细地讨论非自解释的结构特征#7、#9、#10和#12。

（#7）是独占前缀重复。这是一个二进制特征，当且仅当域由至少重复两次的单个字符序列w组成时，它为1。例如，对于域名rwth-aachen.derwth-aachen.de此特征的计算结果为1，但对于域名rwthrwth-aachen.de计算结果为0。

（#9）纯数字子域的比率。这一特征是由数字组成的子域的数目与子域总数的比值来计算的。它忽略公共后缀。例如考虑域名123.itsec.rwth-aachen.de结果是1=3，因为它有3个子域（不包括公共后缀de），其中一个子域是由数字组成的。

（#10）纯十六进制子域的比率。此特征的定义类似于数字排他子域的特征（#9）比率。

（#12）包含IP地址。这是一个二进制特性，当且仅当域包含IP地址时为1，其中IP地址指IPv4和IPv6地址（包括点）的常用符号。

3.3语言特点

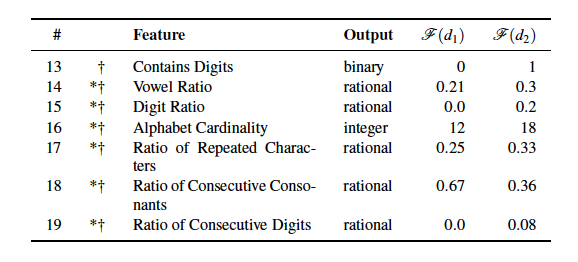


表2：应用于示例域d1和d2的7个语言特征的概述。

为了扩展我们的特征集，我们着重研究了域名的语言特征。这些特征被用来捕捉与域名常见语言模式的偏差。表2概述了所有7种语言特征。下面，我们将详细讨论非自解释的语言特征#17、#18和#19。

（#17）重复字符的比率。重复字符比率在d-dsf上计算，定义为d-dsf中出现多次的字符数除以字母基数（#16）。考虑到示例域名d =bnxd.rwth-aachen.de 特征的计算结果为3/12，其中d-dsf中的重复字符为n、h和a。

（#18）连续辅音的比率。该特征将辅音大等于2的不相交序列的长度相加并除以d-dsf的长度。例如，考虑域名d = bnxd.rwth-aachen.de结果（8+2）/15=0:67，其中d-dsf=bnxdrwth-aachen，连续不相交的辅音序列为：bnxdrwth和ch。

（#19）连续数字的比率。这个特征的定义类似于连续辅音的特征（#18）比率。

3.4统计特征

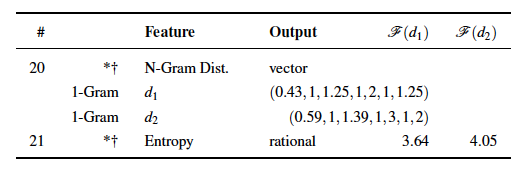


表3：在示例域d1和d2上评估的2个统计特征的概述。

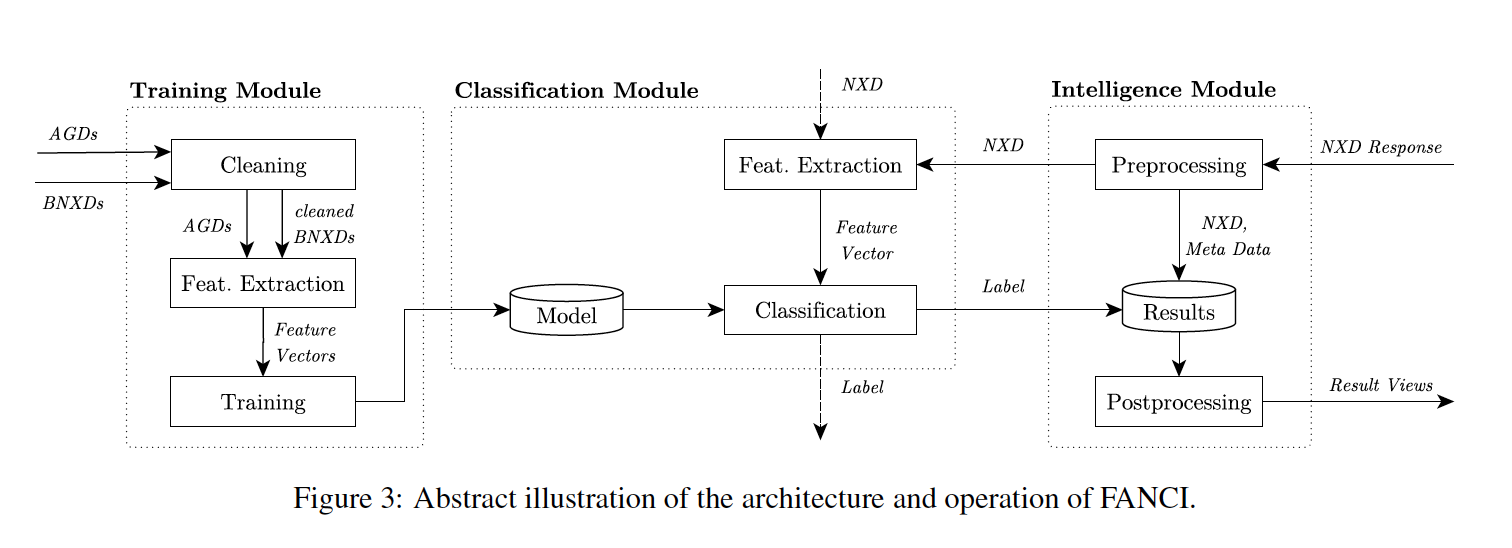
FANCI使用的两个统计特征如表3所示。下面将对两者进行详细说明。

（#20）N-Gram频率分布[2]。一个域名d的n-gram 是一个多组字符序列e ，e ddsf , |e|=n 。表示相应n-gram的频率分布。N-gram频率分特征定义为,表示fn的算术平均数，是对应的标准差min（f n）是最小值，max（f n）是最大值，是中位数， 下四分位数，是上四分位数。表3例示了在域d1和d2上对1-gram这一特性的评估。FANCI使用g1；g2；g3作为特征#20，其结果是21个输出值总和的数组。。

（#21）熵[14，2]。熵（根据香农）的定义是考虑到d的1-gram频率分布:其中pc是根据f 1的字符c的相对频率。表3显示了域d1和d2的示例评估。

4 FANCI

在本节中，我们将介绍基于特征的自动NXDomain Classification and Intelligence（FANCI）。FANCI是一个轻量级的系统，它只根据域名将任意nxd分类为良性的和DGA相关的。它包括三个模块：训练、分类和情报。图3概述了FANCI的体系结构、所需的输入、输出以及FANCI内部处理数据的方式。下面将更详细地描述这三个模块和潜在的应用场景。



4.1训练模型

由于FANCI是基于监督学习分类器的，因此需要使用标记数据进行训练。训练模块实现分类器的训练，并需要输入标记的mAGDs和bNXDs（参见图3的左上角）。我们从DGArchive获得用于培训目的的贴有标签的Magd。假设FANCI在校园网或商业网中运行，例如可以从网络的DNS解析程序获得bNXDs。为了获得一组尽可能干净的bnxd用于训练，我们在一个清洁步骤中根据DGArchive[14]中的所有已知magd对它们进行过滤。在清洁步骤之后，对每个输入执行特征提取，如第3节所述。

训练模块的输出是一个经过训练的模型，可以用于分类模块中未知NXDs的分类。

4.2分类模块

培训模块（参见图3的中间部分）。分类模块在NXD上运行，也就是说，在单个域名上作为输入，由智能模块（参见第4.3节）或任何其他来源提交分类，如图3中虚线箭头所示。分类模块的输出是提交的NXD的一个标签，它可以采用良性或恶意两个值中的任何一个。

在第一部分中，对特征向量进行训练，然后对特征向量进行提取。分类模块可以单独使用，也可以与智能模块结合使用

4.3智能模块

智能模块的任务是根据分类结果提供情报，特别是发现受感染的设备和识别新的dga或未知种子。与仅将NXD本身作为输入的分类模块不同，智能模块还将每个NXD响应的源和目标IP地址以及时间戳作为输入，以便能够将作为分类结果的恶意标签映射回发起查询的设备。

在第一个预处理步骤中，该模块从NXD响应中提取域名和上述元数据。它使用分类模块来确定相应NXD的标签，并将结果（包括元数据）存储在数据库中。为了处理和改进结果，进行了后处理，可分为过滤和转换。

过滤被执行以进一步减少误报（FPs），并且通过根据两个白名单过滤所有良性域名来执行。如果NXD以一个白名单中存在的域名结尾，则从列表中删除。

第一个白名单是全球性的，总是适用的。它由顶级的X Alexa域6组成，在这个步骤中使用的确切数量X是可配置的。将Alexa顶级域名列入白名单是基于这样一个普遍的假设，即罪犯无法在最流行的域下托管指挥与控制（C2）服务器[4，1]。为了避免白名单域名，例如dyndns.org网站根据Mozilla的8个域名列表，我们将其排除在公共域名列表之外。

第二个白名单是地方性的。它考虑了FANCI运行的网络中出现频率较高的域。这个列表是完全可配置的，我们在本文的评估部分提供了两个网络的示例（见第5.2.4节）。

过滤后，对结果应用转换，以生成关于此数据的不同视图，并便于分析结果。这些转换主要包括按TLD或二级域对所有阳性的分组、按请求设备的IP地址分组nxd以及按时间戳分组。此外，还可以对nxd执行基于字符串的搜索和过滤。现在，数据已经准备好了，可以进行人工审查和结论性解释。

4.4使用场景

FANCI是一个多功能、灵活的系统，适用于各种不同的场景。我们主要区分两个主要用例。第一种情况考虑在一个操作现场使用FANCI及其所有三个模块，而第二种情况利用了FANCI的模块化设计并考虑了FANCI的分布式使用。

本地的。这种部署场景是典型的企业级或校园级网络，在这种情况下，FANCI可以作为一个功能齐全的系统在本地使用。这种规模的网络通常有一个集中的DNS基础设施，这有助于FANCI的部署，特别是获取bNXDs来训练分类器，以及随后使用NXD响应进行实时检测。在这种部署中，以前训练过的模型用于标记nxd，并向网络管理员和事件处理程序提供有关受感染设备的见解。

在一些网络中（例如，在典型的大学网络中），可以以使查询设备的IP地址可见的方式监视设备的DNS通信量。在这种情况下，FANCI的智能模块能够

将NXD响应中检测到的magd映射到查询它们的受感染设备。检测到受感染的设备可以触发对来自这些设备的成功解析的DNS通信量的监视。使用FANCI在成功解析域上训练的分类模块（见第5.5节），然后可以检测成功解析的MAGD并识别C2服务器，从而将相应的IP地址列入黑名单。请注意，从只监视NXD响应开始，其优点是在这一步中需要处理的流量比我们监视整个DNS流量要少得多。由于DGA通常会产生比解析magd更多的magd，因此监控NXD响应是发现受感染设备的最有希望的方法。受感染的设备在查询非解析mAGD之前能够联系C2服务器的机会似乎非常渺茫。

在许可较少的网络中（例如，在大型企业网络中），DNS流量可能不允许直接映射到设备，例如，因为层次DNS基础设施，其中中央DNS服务器只与从属域控制器通信。在这种情况下，被感染设备的识别就不那么直接了，但在某种程度上可以借助于FNACI检测到的sinkholing mAGD做到。FNACI也可以集成到现有的监控软件中，通过提供可直接利用的威胁情报，显著增加其检测能力的价值。被FANCI归类为mAGDs的域可以被认为是高可信度的威胁指标（IOC）。因此，FANCI可以触发并支持各种后续措施。这可能包括代理日志和DNS日志分析，例如，回顾性地检测进一步的感染，以及对已识别的C2域的sinkhole或黑名单。此外，在基于主机的代理或网络边缘设备（如路由器或防火墙）上使用检测到的magd有可能发现更多受感染的设备并同时中断C2通信

外包。FANCI很好地适用于未知环境，这意味着某些部件可以外包。特别是，可以使用从某个校园级网络获得的数据进行训练，并使用由此产生的模型在其他网络中执行检测。这使得FANCI能够在难以进行培训的网络中使用。例如，这可以是小型网络（例如，小型企业的网络），在这种网络中，获取必要数量的培训数据需要花费太长的时间，或者这可能是一个网络，在这种网络中，获取一组干净的BNXD用于监督学习是一项非常重要的任务（例如，ISP网络）。

此外，FANCI的分类模块可作为服务使用，例如，通过API或安全软件或安全研究人员使用的web服务访问。注意，在这种情况下，只有有问题的域名必须提交到服务器。还可以利用各种机制进一步分享全部贴有标签的mAGD，例如作为威胁情报来源，这可以再次纳入大中型公司现有的保护工作中。

5评价

在本节中，我们将对FANCI的分类模块进行全面评估。我们比较了SVM和RFs，以找到用于检测mAGD的最佳分类器设置，并表明在本用例中，RFs的性能略优于SVM。我们证明了FANCI的分类模块可以很好地推广到未知的网络环境中，并给出了一个真实的应用测试，由此我们能够报告新的DGA。最后，我们评估了FANCI的分类模块在整个DNS流量中检测解析Magd的能力。在详细介绍我们的结果之前，我们首先描述我们的评估过程，包括我们评估所基于的数据集的描述。

5.1数据集

由于FANCI的分类模块依赖于监督学习分类器，因此我们需要标记的数据集进行训练和评估。此外，由于分类只对域名执行，因此我们只需要一组标记的唯一域名来评估分类性能。我们使用的三个数据源是RWTH Aachen大学校园网、西门子股份公司的内部网络和DGA Archive [14]

亚琛大学。亚琛RWTH亚琛大学的中心DNS解析程序是bNXD响应的第一个来源，包括各种学术机构、eduroam7、几个行政网络、学生公寓和RWTH亚琛大学医院。校园网还与亚琛应用科学大学和尤利希研究中心互联[15]。由于强制执行，绝大多数设备使用网络的中央DNS解析程序。我们的bNXD数据集是在中心DNS解析程序上记录的NXD响应的连续一个月的记录。我们总共记录了31天，更准确地说是从2017年5月22日到2017年6月21日。在这一个月内，我们记录了NXD响应的pcap文件，大小为98:9GB，包含大约7亿个NXD响应，即平均每天记录3:2GB或2260万个NXD响应。总的来说，这个数据集包含3580万个独特的nxd。。

西门子。作为bNXDs的第二个来源，我们从西门子的DNS基础设施获取数据。请注意，我们只获得了NXD，而不是完整的NXD响应，因为这对于FANCI的分类模块来说已经足够了。这些数据来自西门子公司的几个中央DNS服务器，覆盖三个地区：欧洲、亚洲和美国。这种广泛的国际覆盖保证了来自不同实体和设备的不同数据。我们获得了2017年9月和10月两个月的数据（即61天），包括3120万个独特的NXD。

两个良性数据集的长记录周期保证了一个具有代表性的数据集，包括一天中的不同时间、一周中的不同日期以及工作和非工作日。为了尽可能清理我们的良性数据集，我们对照DGArchive[14]检查了良性数据，并删除了所有已知的magd。

DGArchive。为了获得一组已知的mADGD，我们使用了DGArchive[14]。DGArchive中的magd是通过使用逆向工程dga的重新实现和使用相应的已知种子来计算的。因此，DGArchive是恶意数据集的一个非常可靠的来源。我们的数据集包括撰写本文时DGArchive提供的所有数据。从2018年1月30日到2014年2月12日，我们获得了1340天的数据。总共，这个集合包含72个不同的DGA。由于我们选择的ML算法至少需要几百个nxd的集合才能获得良好的性能，因此我们决定通过消除所有具有少于250个唯一magd的dga来减少集合。这将导致剩余59个DGA。对于我们的恶意数据集，我们只考虑这些dga的唯一mAGDs。这包括49738973个独特的mADGD。在这些dga中，唯一的magd的数量在251到13488000之间。

5.2分类精度

在本节中，我们首先确定用于检测magd的最佳性能分类器或分类器集合。接下来，我们进行了几个实验，每个实验都证明了FANCI分类模块的某种能力。这包括检测未知种子和未知dga的能力，以及显示FANCI的分类模块很好地通用化。

5.2.1实验设置

由于我们的数据集相当大，我们进行随机抽样，以生成我们的评估集。每个数据集由与mAGDs一样多的bnxd组成，并通过对我们的良性数据集中的每个单独集执行新的均匀随机采样来创建。根据相应的实验，恶意数据可以从所有DGA的唯一mAGDs中随机抽取，也可以从单个DGA的唯一mAGDs中随机抽取。对于考虑所有dga的集合，我们力求尽可能统一地表示所有dga。这里集合的大小表示样本总数，即bNXDs和mAGDs的总和。

根据实验的不同，我们进行5折交叉验证（CV）或只剩一组（LOGO）CV。在5折CV中，数据集被分成5个大小相等的子集，其中4个用于训练，1个用于预测。每一个子集都被精确地用于预测一次。结果统计指标在所有5次运行中取平均值。LOGO CV在其基本过程中类似于k-fold CV，但不是构建k个随机子集，而是根据预定义的分组（例如，通过种子或dga）来定义子集。

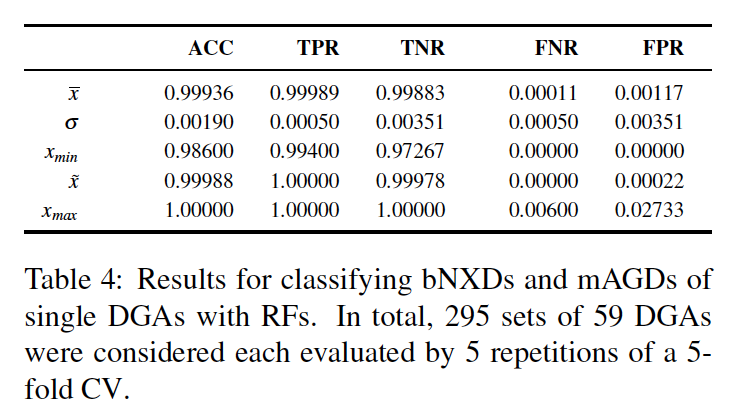
我们通过对独立于评估数据集的广泛网格搜索，确定了两种不同场景下ML算法的最佳参数设置。第一种情况考虑单个DGA检测（即，针对一个特定DGA的一个分类器），其中第二个场景针对多个DGA检测（即，训练一个分类器来检测所有DGA）。我们修复了结果参数，并将其用于所有后续的评估场景，包括在真实场景完成的场景。有关网格搜索结果的摘录，请参见附录B

所有计算都在RWTH Compute Cluster8上进行。

在所有的实验中，我们把准确度（ACC）作为描述分类器性能的主要指标，定义为ACC=|TP|+|TN|/|populationj，其中|TP|是真阳性的数量，|TN|是真阴性的数量。这意味着ACC表示正确预测样本的分数。然而，对于每个实验，我们还提供了以下四个指标的统计数据：真阳性率（TPR）、真阴性率（TNR）、假阴性率（FNR）和假阳性率（FPR）。对于每个指标，我们考虑算术平均值x、标准偏差s、最小xmin、中值∮x和最大xmax。

5.2.2分类器选择

在这一部分中，所介绍的实验反映了为一个实际应用程序选择性能最好的分类器的过程。在下面的实验中，我们只考虑来自RWTH-Aachen的良性数据。我们对支持向量机和RFs进行了每个实验。由于我们的目标是找到性能最好的分类器，并且在大多数情况下，RFs的性能略好于支持向量机，因此我们在下面详细介绍RFs的结果。支持向量机的结果见附录A



单一DGA。第一个实验包括使用专用分类器检测特定的单个DGA。我们考虑了所有59个DGA，并按照第5.2.1节中的程序为每个DGA创建了5个不同的集合，最大集合大小为100000。这意味着每个数据集始终包含相同数量的magd和bnxd。根据DGA的不同，可能会有不到50000个独特的mAGD可用。在这些情况下，相应地调整设定尺寸。综上所述，这将产生295套最大规模为100000套。对于每一组，我们进行5折CVs，我们用新的随机折叠重复5次。

表4给出了RF探测单个DGA能力的统计描述。平均ACC为0:99936，标准偏差为0:00190。在Bobax的检测中，最小的ACC达到了0:98600，这是唯一的异常值。RFs检测59个DGA中的6个（Bamital、Blackhole、Dyre、Sisron、Tofsee和UD2），ACC为100%。

未知的种子。在这个实验中，我们重点评估了一个DGA生成的mAGDs用一个新的种子进行检测，其中模型是用同一个DGA生成的mAGDs使用已知的种子进行训练的。

为了评估这个场景，我们执行一个LOGO CV，也就是说，我们使用一个特定DGA的一个种子以外的所有种子的magd进行训练，对跳过的一个进行预测，并对每个种子和DGA重复此过程。同样，我们使用最大大小为100000的数据集，每个DGA使用5个不同的数据集。我们考虑所有至少有两个已知种子的DGA，总共产生30个DGA和550个种子。总的来说，对于所有可用的种子和dga，这将导致5 \*550=2750次迭代。

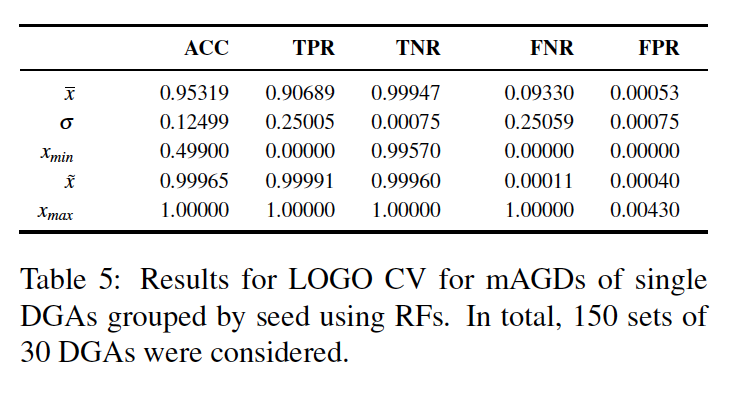
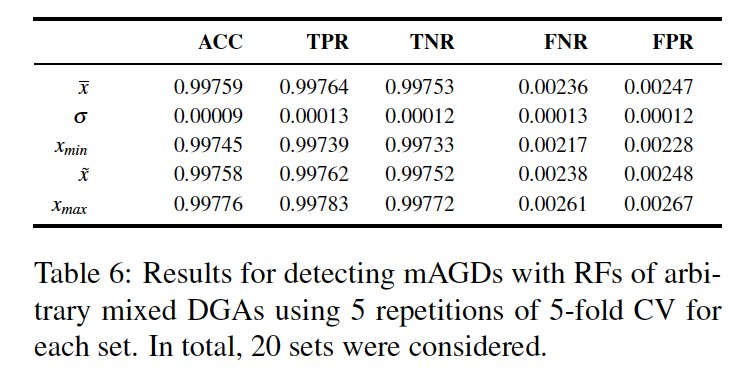


表5给出了本试验RFs评估结果的统计汇总。ACC的平均值为0:95319，显著的标准偏差为0:12499。ACC值介于0:49900和1:0之间，其中75%的测量值显示ACC高于0:98193。由于只有6个dga与低于98%的ACC相关，ACC的大范围可以用异常值来解释。

这个实验是唯一的一个实验，其中支持向量机的性能略优于RFs。SVM的平均ACC为0:98315，标准偏差为0:06166，但范围相似，范围从0:49850到1:0。支持向量机实验的详细结果见表14。SVM也会受到与RFs相同的异常值（即相同的DGA导致问题）的影响。与RFs相比，支持向量机并不总是漏掉这些特定dga的所有新种子，因此平均ACC略高。

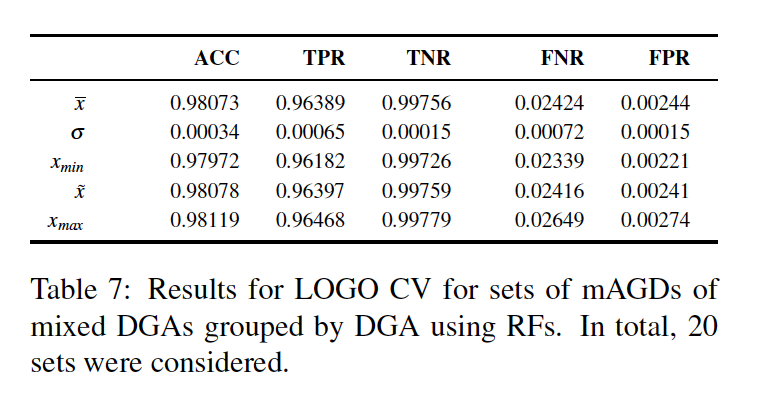


混合DGA。下一步，我们将研究一个训练在已知dga的mAGDs上的单个分类器如何能够检测由这些dga中的一个生成的其他magd。

我们创建了20组目标大小为100000的数据集，其中包含59个DGA中相同数量的Magd。对于数量太少（即少于50000/59～ = 847）的DGA，我们包括了此类DGA的所有可用Magd，这导致有效集大小为92102。在这20组中，我们对每一组进行了5次重复的5折CV。

在其趋势中，在包含多个dga的magd的集合中检测magd的结果类似于对每个dga使用专用分类器的检测。表6显示了RFs的测量结果。ACC的平均值为0:99759，标准差很小，为0:00009。最小和最大ACC值分别为0:99745和0:99776。

综上所述，用多个dga的mAGDs训练一个分类器，在检测任意mAGDs时可以达到非常高和稳定的ACC。



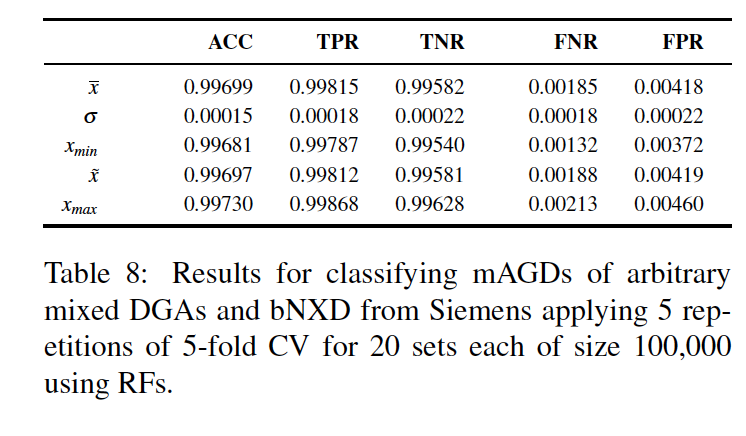
未知的DGA。本实验验证了该方法在检测未知mDGA方面的能力。为了验证我们的分类器是否能够泛化为未知DGA的mAGDs，我们对一个DGA分组进行LOGO-CV，即使用除一个DGA外的所有DGA的mAGDs进行训练，并预测出遗漏DGA的mAGDs。本实验所考虑的集合与上一个实验中的集合是等价的，也就是说，我们考虑20个集合，每个DGA具有相同数量的mAGDs。这意味着，对于20个集合中的每一个，我们执行了59次训练和预测迭代，一次只剩下一个DGA。

表7描述了RFs检测未知mDGA的结果统计汇总。ACC在0:97972和0:98119之间，平均值为0:98073，标准差很小，为0:00034。RFs检测到59个漏失的dga中的55个，ACC与先前提出的实验相当。我们的结论是，我们能够检测到未知dga的mAGDs.

分类器选择。在实际应用中，我们的目标是可靠地检测已知的dga以及未知的种子和dga。此外，我们希望达到最大的分类精度。因此，我们必须选择性能最好的分类器或分类器组合来实现这些目标。为此，我们还评估了几个专用于单个dga的分类器的逻辑组合。特别是，我们测试了几个or和and组合，不同阈值的阈值投票，多数投票，甚至RFs和SVM的组合。然而，使用所有已知的dga训练的单个RF分类器的性能优于上述任何一个组合。这就是为什么FANCI使用一个单一的RF分类器，该分类器由所有已知DGA的MAGD训练而成。

5.2.3概括

到目前为止，我们用rwt亚琛大学的含bNXDs的测试集进行了所有的实验。在本节中，首先，我们将展示FANCI在不同网络中训练和部署时的性能相同。第二，我们证明甚至可以使用一个网络中记录的数据进行训练，并在另一个网络中使用得到的分类模型。这意味着FANCI很好地适用于新环境。



混合dga；训练和预测西门子。为了说明FANCI的检测能力独立于某个网络，我们重复了第5.2.2节中的混合DGA实验，但使用的是西门子数据集的bNXDs生成的数据集。该实验得出的ACC值与在相同设置下获得的RWTH数据相当。平均ACC为0:99699，标准偏差为0:00015，最小值为0:99681，最大值为0:99730。表8显示了使用西门子网络数据时的详细检测性能。

接下来，我们进行了两个实验，证明我们训练的分类器能很好地推广到未知网络，也就是说，我们研究了使用来自某个网络的数据训练分类器，但在其他地方使用该分类器的情况。为了评估我们在使用在国外网络中训练的分类器时在ACC中的损失，我们将ACC与我们使用来自同一网络的bNXDs进行训练和预测的场景进行比较。

混合DGAs，训练RWTH，预测西门子第一个实验考虑使用rwt亚琛的bNXD进行训练，并对由西门子bNXDs组成的集合进行预测。第二个实验反过来进行。这些实验是基于这样一个事实：mAGDs在不同的网络中没有不同，但只有bNXDs可能不同。对于这两个良性数据源，我们考虑20个数据集，每个数据集都是在前面的实验中生成的。每个数据集用于一次训练，其中对另一个bNXD源的20个集中的每一个进行预测。这两个实验的结果是20 \* 20=400次通过。

表9列出了考虑使用RWTH亚琛公司的BNXD进行培训的数据集和包含西门子公司的BNXD数据集进行预测的结果。平均ACC为0:99534，标准偏差为0:00018。与对包含西门子BNXD的数据集进行培训和预测（见表8）相比，平均ACC仅略小，即0:00165个百分点。这可以解释为FPs的增加。然而，假阴性（FNs）甚至减少。

混合DGAs、训练西门子、预测RWTH表10显示了考虑包含西门子bNXD用于培训和RWTH亚琛bNXD数据进行预测的结果。在这个实验中，平均ACC为0:99785，与仅RWTH（见表6）的实验相比甚至稍大，即0:00026个百分点。虽然FPs略有增加，但FNs减少。这证实了先前实验的趋势。

同样，我们用支持向量机进行了所有的实验，RFs和RFs的性能始终优于支持向量机。SVMs的结果见附录A。综上所述，先前的实验表明，FANCI一般独立于某个网络，能够很好地推广到未知环境，甚至可以包括实际的分类。

5.2.4附加假阳性减少

如第4.3节所强调的，FANCI在智能模块中执行过滤，以减少FPs。为了评估我们的过滤方法的效率，我们考虑在前面的章节中所介绍的实验中出现的所有唯一的FP-bnxd集合。当我们在第二个过滤步骤中使用特定于本地的白名单时，我们考虑两个数据集，一个用于RWTH Aachen FP bNXDs（6522），另一个用于Siemens FP bNXDs（11431）。我们使用Alexa top 100、top 10000或top 1000000评估了全局过滤步骤。针对每个网络使用适当的白名单执行特定于本地的过滤。例如，对于RWTH-Aachen大学网络，该列表包括域，例如rwth-aachen.de, sophosxl.net，和fh-aachen.de。对于西门子网络，该列表包括：siemens.net, trendmicro.com,mcafee.com, 和 bayer.com.，. 这些特定于本地的白名单假定校园网中不存在C2服务器。此外，我们假设某些公司，如Sophos、McAfee和TrendMicro不托管C2服务器。

表11给出了随后对这两组唯一的FP-bnxd应用两个过滤步骤的结果。它以百分比表示FPs的减少以及剩余FPs的数量。对于来自RWTH亚琛的数据，我们可以将FPs降低75.53%至89.88%，这将分别产生1596或660帧的剩余FPs。考虑到西门子网络，我们将FPs至少减少了47.85%，产生了5961个域，在最好的情况下，我们将FPs减少了77.74%，产生了2544个域。

结果清楚地显示了我们随后的FP滤波的效率。尽管FANCI的分类准确率已经非常高，但即使只考虑Alexa top 100作为白名单，我们至少可以将FPs的数量减半。在最好的情况下，我们甚至可以将FPs减少到初始数量的十分之一。

现在，我们已经看到了FANCI在检测Magd方面的能力，并证明了我们的假阳性减少的有效性，我们将在下一节介绍FANCI的实际应用。

5.3现实世界

在这一部分中，我们将介绍FANCI在亚琛RWTH大学网络中的应用。设置。对于我们的FANCI的实际应用测试，我们考虑从中央进行一个月的新记录

RWTH亚琛大学DNS解析程序包括31天，更准确地说是从2017年10月13日到2017年11月12日，其中数据量与第5.1节中的记录类似。这意味着FANCI总共要处理大约7亿个NXD响应，其中包含3500万个唯一的NXD。FANCI与一个单一的RF分类器一起使用，该分类器由一组92102规格的mAGD组成，该mAGD来自RWTH-Aachen网络，来自第5.1节所述的数据集。该装置包含bNXDs和MAGD，且每个DGA的Magd数量相等。我们应用了FANCI，首先对来自新记录的所有NXD响应使用分类模块，然后使用智能模块的过滤功能，使用Alexa的top 1000000进行FP减少。

结果。应用这两个步骤，我们得到22755个唯一的阳性NXD（ॸ0:065%），共有45510个NXD反应（ 0:0065%）。在对这些剩余阳性进行半自动检查后，我们能够报告405个未知mAGDs，这些mAGDs分别对应于10个不同的组，它们分别表示未知的DGA（UD）或未知的种子（US）。为了找到未知mAGD，我们使用了FANCI智能模块提供的不同视图，如第4.3节所示。请注意，这里的“未知”意味着，在撰写本文时，发现的mAGD既没有列在DGArchive中，也无法通过其他常见来源找到。我们将把所有调查结果提交给DGArchive。图4显示了10个组中的每个组的代表，其中包括一个标签，表明我们是否认为该组是UD，还是与我们一样，或者两者都可能。我们在DGArchive、领域知识和手工研究的帮助下对组进行了标记。

言下之意，在我们一个月的实际测试中，我们看到最多22345个独特的FPs，导致最坏情况下的FPR约为0:00064。由于在实际应用中很难确定正确的地面真实性，所以这种FPR的意义非常有限。对于有关FANCI分类能力质量的说明，更详细地分析潜在的FPs更有希望。潜在FPs集合的特点是nxd之间的高度多样性。图5显示了在我们的实际评估中看到的12个潜在fp。它们可以分为两类：人生成的和机器生成的。在人类生成的nxd通常显示自然语言模式或与现有领域非常相似的地方，机器生成的nxd往往具有随机性或技术来源。如果没有额外的信息，将NXD分配给这些类之一并不总是可能的，例如考虑潜在的FP NXD c.ssl-cd.com，它可能属于每个类

由于在潜在的FPs集合中没有显著的相似nxd，这使得我们可以得出结论，FANCI没有出现系统的分类错误，这突出了FANCI非凡的分类性能。

由于RWTH亚琛的网络由商业安全软件和设备保护，使用已知mAGD的黑名单，因此在我们的实际测试中几乎找不到已知的mAGD也就不足为奇了。准确地说，使用DGArchive，我们只能识别出31种已知的独特的mAGD。

FANCI在亚琛rwt大学网络上的应用，在一个月的时间里，充分说明了它在现实世界中的检测能力。此外，这项测试强调了FANCI探测未知mAGD和已知mAGD的能力。为了进一步支持FANCI在实际的大规模网络中的适用性，我们在下面对FANCI的分类速度进行了考虑。

5.4训练和分类速度

本节简要概述了训练和分类速度，以证明FANCI在现实世界中的适用性。所有测量都是在带有Intel i7的Dell OptiPlex 980上单线程执行的870@2.93GHz运行UbuntuLinux16.04的CPU和16GB内存。我们对第5.2节中用于评估的92102大小的混合集进行了10次培训和分类。时间测量包括特征提取。

平均而言，RF训练时间为339:71秒（5；66分钟）。

RF能够在234:76秒内对92102个未知样本进行分类。这意味着对单个未知样本执行分类平均需要0:0025秒（包括特征提取）。

基于上述测量结果，FANCI能够使用一个线程在一台通用计算机上执行每秒400个数据包的分类。如第5.1节所述，在亚琛RWTH大学的网络中，平均每秒有164个NXD响应，最大峰值为每秒900个NXD响应，我们可以说FANCI是现实世界中适用的，甚至可以在大型网络中执行实时检测而无需采样。

5.5成功解析域名

如果FANCI检测到设备感染了机器人，它最终将成功查询其C2服务器的IP地址。如果可以检测到这样一个成功的查询（例如，在被感染的设备被识别后，使用FANCI对成功的查询），这将显示所述僵尸网络的C2服务器的IP地址。

因此，我们对FANCI如何将MAGD与成功解决查询区分开来进行了初步评估。特别是，我们使用随机森林和类似于第5.2.2节中提出的混合DGA情况的设置来进行测试测量。我们不使用bNXDs，而是从西门子网络和已知的任意dga的mAGDs中合成成功解析域的数据集。如第5.2.2节所述，我们对20组进行了5次重复的5倍CVs。如果没有进一步的优化或适应成功解析域的新特性，我们实现了平均ACC为0:94962，小标准差为0:00071，最小为0:94809，最大为0:95060。表12给出了使用RFs进行概念验证实验的详细结果。SVMs的结果见附录A。

考虑到我们只需要处理单个设备或小组设备的成功解析域的事实，前面提出的方法对于C2服务器的识别是非常有前景的。

6相关工作

过去，监视DNS流量（成功解析和/或不解析）被用作检测网络中恶意活动的主要或附加信息源（例如[2、16、18、9、4]）。其中一些方法集中于识别C2服务器（例如，[18，16]），其他方法专注于检测magd（例如，[2]），识别受感染的设备（例如[9]），或检测一般的恶意URL（例如[4]）。

这些先前的方法与FANCI最显著的区别在于，它们都或多或少地需要对DNS流量进行广泛的跟踪，也就是说，它们需要从DNS组中提取的信息之间的相关性

查询和/或响应（例如，特征提取）。相比之下，FANCI的分类模块在预测特定NXD时使用的特征仅从该NXD中提取，这样FANCI就不需要任何跟踪。此外，许多先前的方法都是基于聚类的，这使得人工标记已识别的聚类。与此相反，FANCI（如[4]）使用ML分类器。

在成功解析DNS通信量中检测magd可以识别C2服务器（参见第5.5节了解FANCI在这方面的初步评估）。然而，只监视NXD响应的优点是可以在较短的延迟内检测到bot的感染，同时处理的流量也大大减少，因为绝大多数dga发出的NXD比注册的名称多得多。

虽然先前的研究显示了对特定数据集的有希望的检测能力，但是关于它们的泛化性和检测过程在时间和内存需求方面的效率方面的信息很少。FANCI在预测（0.0025s/样本）和训练（92102个样本5.66min）方面都是高效的，即使在不同的网络上训练，FANCI在非常大规模的现实场景中显示出高精度和低FPR。

很难在探测精度和效率方面对FANCI和以前的方法进行公平比较，因为它们瞄准的目标略有不同，即使它们确实瞄准同一目标，也使用不同的数据集。这些数据集和系统的实现不是公开的。下面，我们将更详细地讨论与FANCI最密切相关的方法。

Exposure.。Bilge等人[4] 引入一个名为Exposure的系统，该系统旨在检测DNS流量中的恶意域名，也就是说，它们不关注mAGDs，而是旨在检测在网络钓鱼或恶意代码托管环境中使用的域名。与FANCI相比，Exposure监控的是完整的DNS流量，而不仅仅是NXD响应。此外，Exposure.总是需要访问比FANCI更敏感的信息（例如访问模式）。像FANCI一样，Exposure.是基于ML分类，并使用一组精心挑选的特征。然而，这些特征不仅是从单个域名中提取的，而且还包括从关联多个DNS查询或响应中提取的特征。暴露的准确度与FANCI的ACC相似（但通常是针对检测恶意域名），并在实际数据中进行评估。由于需要敏感的和上下文相关的信息，公开并不像FANCI那样通用，尤其是在软件方面，比如在服务部署方面。

Winning with DNS Failures.。Yadav和Reddy[18]是第一个考虑利用DNS响应（成功解析域名）和NXD响应来检测僵尸网络的。他们引入了一个主要针对基于DGA的僵尸网络的C2服务器的IP地址识别系统。该系统基于通过过滤缩小一组潜在的恶意IP地址。此筛选要求访问总体成功解析的DNS流量（以便统计解析到给定IP地址的域数）、成功查询附近的NXD响应以及失败和成功的DNS查询的熵。过滤的输出是一组潜在的C2服务器IP地址。

Pleidas. Antonakakis等人。[2] 介绍了一个称为Pleidas的DGA检测与发现系统。该系统能够通过聚类的方法发现新的dga，并通过使用多类交替决策树的有监督学习来检测已知的dga。在15个月的时间里，他们在一个大型ISP环境中应用了他们的系统，他们发现了12个新的DGA，其中6个是全新的，6个是以前已知的DGA的变体。

Pleidas使用一组统计和结构特征，其中所有特征都是从来自单个主机的NXD响应组中提取的。9统计特征包括熵度量和域名组上的n-gram。结构特征包括结构域长度、TLD的唯一性和频率分布，以及存在的子域能级数。

利用标记数据评价了Pleidas的分类精度。Alexa的前10000个域名是良性的。恶意数据集由四个DGA（即Bobax、Conficker、Sinowal和Murofet）生成的60000个NXD响应组成。对于每个主机5nxd响应的组大小，TPR在95%到99%之间，FPR在0.1%到1.4%之间。每组有10个NXD反应，准确度略有提高。在这种情况下，TPR在99%和100%之间，其中FPR在0和0.2%之间。

由于Pleidas需要跟踪DNS响应以进行特征提取，因此我们期望它比FANCI效率低得多。报告的检测质量与FANCI相似，但FANCI是在更广泛的数据集上进行评估的，该数据集使用了更多的DGA和真实世界的良性流量，而不是Alexa的前10000个域。Pleidas的可推广性没有得到评估。

Phoenix.。Schiavoni等人。[16] 提出了一个基于DGA的僵尸网络跟踪和智能系统称为凤凰。与之前介绍的Pleidas不同，Phoenix专注于情报行动，而不是DGA探测。这特别包括跟踪僵尸网络的C2基础设施的IP地址范围。然而，Phoenix也可以将DNS流量标记为与DGA相关的或良性的。

他们评估了Phoenix在1153516个域上的分类性能，包括从被动DNS获得的三个不同dga和bNXDs的magd。评估得出的TPRs在81.4%和94.8%之间，因此在mAGDs检测方面明显低于FANCI。由于所使用的功能重量较轻，需要跟踪，我们预计Phoenix在速度方面比FANCI效率低。

NetFlow。Grill等人。[9] 提出了一种不同的基于DGA的恶意软件检测方法，其特定目标是以隐私保护的方式适用于大规模网络。他们的系统完全基于NetFlow数据，也就是说，基于在源IP和端口以及目的IP地址和端口的组合之间交换的网络包元数据的聚合。导出的元数据取决于NetFlow的特定实现，但通常包括：IP地址、时间戳、端口号、字节计数器和数据包计数器。Grill等人。使用标准化的IPFIX NetFlow格式[12]。它们执行异常检测是基于这样一个假设：主机的正常行为是通过DNS为某个域名请求一个IP地址，然后是一个或多个连接到这个新解析的IP地址。他们假设一个被DGA恶意软件感染的设备的特征是定期发出DNS请求，而没有随后连接到新的IP地址。

为了进行评估，他们进行了三个实验，考虑了不同类型的主机、网络大小和一天中的时间。他们考虑了六种不同的DGA。ACC值在88.77%和99.89%之间，这取决于所讨论的设置，因此低于FANCI的精度。由于NetFlow基于广泛的跟踪，因此其效率可能低于FANCI。

DGArchive。Plohmann等人。[14] 对目前的DGAs进行了广泛的研究。他们的论文以基于DGA的恶意软件的收集和逆向工程为基础，提供了关于现代DGA功能的详细技术见解，可分为三个主要贡献：DGAs的分类、DGAs和相应mAGDs的数据库dgaarchive以及对注册mAGDs的分析。而Plohmann等人。不要实现自动检测，DGArchive提供了将已知Magd列入黑名单的方法。我们的工作建立在DGArchive的两个方面：我们在训练前使用它来清理我们的良性流量，我们使用它作为恶意mads的来源。

7结论

在这项工作中，我们提出了FANCI，一个通用的系统，用于检测恶意的DGA相关域名在任意NXD DNS流量中基于监督学习分类器。FANCI的多功能性是其轻量级和独立于语言的功能设计的结果，完全依赖于域名进行分类。。在我们的广泛评估中，我们在不同的实验中验证了FANCI的高精度和高效率的mAGD检测能力，包括其可推广性。在一个大型大学网络的一个月的实际应用中，我们发现了10个与DGA相关的mAGDs，其中至少有4个来自全新的DGA。

凭借其经验性的检测能力和成功的真实世界测试，FANCI可以为打击基于DGA的僵尸网络做出决定性的贡献。FANCI能够为现有的安全解决方案提供有价值的信息，并且能够在各种环境中为更高级别的设备和网络安全做出贡献。