计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220846
61(4): 1038−1048, 2024

新通用顶级域名解析行为分析与恶意域名检测方法

杨东辉[1,2] 曾 彬[3] 李振宇[1,2]
[1](中国科学院计算技术研究所 北京 100190)
[2](中国科学院大学 北京 100049)
[3](长沙学院 长沙 410022)
(yangdonghui@ict.ac.cn)

New gTLD Resolution Behavior Analysis and Malicious Domain Detection Method

Yang Donghui[1,2], Zeng Bin[3], and Li Zhenyu[1,2]
[1](*Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049)
[3](*Changsha University*, *Changsha* 410022)

Abstract Since ICANN initiated the delegation of new generic top-level domains (new gTLDs) in 2013, more than a thousand of new gTLDs have been added to the domain name system (DNS). Previous work has shown that while new gTLD domains bring flexibility to registrants, they are also commonly used for malicious behavior because of their low registration costs, and it is important to identify malicious new gTLD domains. However, because of the unique characteristics (e.g., domain length) of new gTLD domains, the accuracy is low when applying existing malicious domain identification methods to malicious new gTLD domain identification. To address this issue, we first characterize the resolution behavior of new gTLD domains based on massive domain name resolution data from five aspects including the number of associated SLDs per new gTLD, query volume, query failure rate, content replication and hosting infrastructure sharing. Then we analyze the resolution behavior of malicious new gTLD domains and find their unique behavioral characteristics in terms of content hosting infrastructure concentration, the number of FQDNs per SLD, the number of queries, the distribution of end users' network footprints, and the distribution of the length of SLDs. Finally, according to these features, we design a malicious new gTLD domain identification method based on random forest. The results of the experiment show that the proposed method achieves 94% accuracy, which is better than the existing malicious domain identification methods.

Key words domain name system (DNS); new gTLD; Internet measurement; behavior analysis; malicious domain detection

摘 要 自2013年ICANN发起新通用顶级域名(new gTLD)的授权以来,域名系统(domain name system, DNS)中已增加了上千个new gTLD.已有工作表明new gTLD在为域名注册者带来了灵活性的同时,由于注册成本低等原因也经常被用于恶意行为,识别恶意new gTLD域名具有重要的意义.然而,由于new gTLD域名在域名长度等方面的独有特征,已有恶意域名识别方法应用于new gTLD恶意域名的识别时准确率低.针对这一问题,首先基于海量域名解析数据,从顶级域名对应二级域名(SLD)数量、查询量、查询失败率、内容复制和承载基础设施共享5个方面刻画了new gTLD域名解析行为.然后分析恶意域名的

收稿日期: 2022-10-09; **修回日期**: 2023-06-26
基金项目: 国家自然科学基金项目 (U20A20180, 62072437)
This work was supported by the National Natural Science Foundation of China (U20A20180, 62072437).
通信作者: 曾彬(13974880055@139.com)

解析行为并发现其在内容承载基础设施集中性、SLD 对应的完全限定域名（FQDN）数目、域名查询次数、请求用户网络空间分布、SLD 长度分布等方面的特征. 最后根据这些特征设计了一种基于随机森林的 new gTLD 恶意域名检测方法. 实验结果表明，所提方法达到了 94% 的准确率，优于已有恶意域名检测方法.

关键词 域名系统；新通用顶级域名；互联网测量；行为分析；恶意域名检测

中图法分类号 TP393

域名系统（domain name system, DNS）是互联网最重要的基础服务之一，它提供人类可读的域名和其相关的 DNS 记录之间的映射. 域名由“.”分隔的字符构成，子域名以父域名结尾，完整的域名被称为完全限定域名（fully qualified domain name, FQDN），其中忽略掉最后的点号时最右边的一组字符被称为顶级域名（top-level domain, TLD），例如 com, net, org, cn 等，顶级域向下数一级的子域称为二级域（second-level domain, SLD）. 顶级域名中不代表一个国家或地区的顶级域名被称为通用顶级域名（generic top-level domain, gTLD）. 为了扩展域名系统、增加域名注册者的选择，互联网名称与数字地址分配机构（Internet Corporation for Assigned Names and Numbers, ICANN）推出了新通用顶级域名（new generic top-level domain, new gTLD）计划，自 2013 年 10 月的首批授权以来，new gTLD 计划已经使上千个新顶级域名，包括使用 ASCII 字符和不同字母系统的域名加入了互联网的根区（root zone）.

new gTLD 为内容发布者（域名注册者）带来了灵活性，使内容发布者能够为其网站创建易于记忆的定制名称. new gTLD 还为 DNS 生态中的利益相关者，如注册机构和注册商提供了新的机会. 近年来，注册 new gTLD 域名逐渐成了一种潮流，这些 new gTLD 域名[①]日渐增加的使用给互联网带来了新的变化. 已有研究指出，尽管与传统顶级域名相比，new gTLD 域名解析量依然存在较大差距，但 new gTLD 呈现逐年增长的活跃趋势，且大部分注册的域名被投入了实际使用[1]. 但同时域名滥用[2]随之出现，new gTLD 域名用于开展恶意行为的情况也越来越多，new gTLD 域名成为各种恶意行为的“热土”. 因此，研究 new gTLD 域名的解析行为将提供有关其使用情况的态势，并进一步促进 new gTLD 的推广、相关基础设施的加强. 然而，已有研究都集中在注册人的行为[3]或其他安全问题，例如中间人攻击[1,4-6]，缺少对于 new gTLD 域名尤其是恶意域名的解析行为分析；此外由于 new gTLD 域名的特殊性，已有恶意域名检测方法可能不适用于 new gTLD 恶意域名的检测. 因此，需要分析 new gTLD 域名的解析行为，寻找 new gTLD 恶意域名的重要行为特征，并据此设计 new gTLD 恶意域名检测方法.

本文使用包含对 815 个 new gTLD 的 930 万次查询的被动 DNS 日志，对 new gTLD 的解析行为进行了深入分析. 同时，在相同测量期内收集了包含所有类型域名的 30 亿次 DNS 查询，以便进行比较. 本文首先通过考察查询量、SLD 数量和查询失败率来分析 new gTLD 的活跃状况. 然后从内容复制和基础设施共享等角度探讨了 new gTLD 域名的承载基础设施. 最后分析了恶意 new gTLD 域名的行为，发现按承载域名量排序的前 5 个 IP/24 网段所承载的域名之中有 73.6% 是恶意的. 根据上述分析设计了 new gTLD 恶意域名检测方法.

本文的主要贡献包括 4 个方面：

1）发现每个 new gTLD 的查询量和对应 SLD 数量符合重尾分布. 例如，按 DNS 查询量排名前 3 的 new gTLD 合计贡献了 48.7% 的查询量.

2）大多数 new gTLD 域名只将它们的内容复制到 1 或 2 个 IP/24 网段，这些 IP/24 网段被域名共享，因为有 89.1% 的域名与其他域名使用相同的 IP/24 网段，并且 new gTLD 域名倾向使用云作为内容承载基础设施，导致内容承载基础设施的集中性，即少量的 IP/24 网段专门承载了大部分域名.

3）与正常域名相比，恶意域名有独特的行为. 例如请求用户网络空间分布独特、内容承载基础设施更集中（存在某些 AS 或 IP/24 承载大量恶意域名）、SLD 长度分布偏斜（更短）、对应的 FQDN 数量更多、每个 FQDN 的平均查询次数少以及倾向使用默认生存时间（TTL）设置（例如 10 min）.

4）基于上述发现设计并实现了一种 new gTLD 恶意域名检测方法，以 new gTLD 域名解析行为特征为输入，使用随机森林作为分类器，实现恶意域名的快速识别. 实验结果表明，该方法的准确率可达 94%，优于传统恶意域名检测方法.

① 下文使用 new gTLD 域名指代顶级域名为新通用顶级域名的 SLD 或 FQDN.

1 相关工作

在 new gTLD 的行为方面，已有研究多集中在注册人的行为或安全问题，如域名滥用行为等. Halvorson 等人[3]提出了一种自动识别注册目的的方法，并分析了 new gTLD 域名的注册类型. Chen 等人[4-5]刻画了在 new gTLD 域名时代通过域名碰撞（domain name collision）实现的中间人（man in the middle, MitM）攻击，并讨论了缓解策略. Korczyński 等人[1]对比了 new gTLD 与传统通用顶级域名的域名滥用行为，发现 new gTLD 已经成为了恶意行为者的“吸铁石”（a magnet for malicious actors）.Pouryousef 等人[6]发现在引入 new gTLD 之后，误植域名（typosquatted domains）的数量增加了几个数量级.

在恶意域名检测方面已经有许多研究. Hao 等人[7-8]利用域名注册信息尝试在发生大量查询之前识别恶意域名. Manadhata 等人[9]利用域名-IP 映射建立二部图，在已知少量域名标签前提下，使用置信度传播推测其他域名是否为恶意域名. Schüppen 等人[10]基于域名文本特征检测 DGA 产生的恶意域名. Yu 等人[11]将域名字符串输入到 CNN 进行恶意域名检测. Lei 等人[12]用二部图建模域名解析行为，用图嵌入学习域名的向量表示，最后用 SVM 分类. 然而，现有研究仍然缺乏针对 new gTLD 恶意域名检测的方法.

2 数据集与预处理

本文所使用的被动 DNS 日志数据集覆盖我国 3 个主要互联网服务提供商（Internet service provider, ISP）. 每条 DNS 应答对应 1 条日志记录，每条日志包含用户的 BGP（Border Gateway Protocol）前缀、自治系统号（autonomous system number, ASN）、域名、DNS 查询类型、所有应答资源记录、TTL 和时间戳. 本文使用 nTLDStats[13]的列表来识别 new gTLD，将 ICANN 发起的 new gTLD 计划之后引入的通用顶级域名称为 new gTLD，在该计划之前引入的通用顶级域名，例如 com, net, info 则被标记为传统通用顶级域名.

为了进行对比分析，同时收集了包括传统通用顶级域名、国家代码顶级域名（country code top-level domain, ccTLD）和 new gTLD 的所有类型域名共 30 亿条 DNS 日志.

在进行数据预处理时，使用 public suffix 库[14]把 FQDN 映射到对应的 SLD. 由于数据集不包括应答码（response code），在判断一次查询是否成功时无法直接从 NOERROR 或 NXDOMAIN 这样的应答码中进行判断. 因此为了计算查询失败率，采取一种启发式的方法过滤掉大概率是 NXDOMAIN 的或者无法准确标记的应答. 这里不检查返回的应答 IP 地址是否正确，而只考虑一次查询是否包含有效的 DNS 应答. 如果一次查询返回了有效的 DNS 应答，就把这次查询标记为成功. 例如对某一次 A 类型查询，如果应答中存在所查询域名的 A 记录，则认为该次查询成功. 在进行完上述标记之后，对于每个查询类型 $QTYPE$，检查数据集中每一个被查询的域名 $QNAME$ 是否至少成功过 1 次，将从未成功过的二元组 $\langle QTYPE, QNAME\rangle$ 对应的日志过滤掉.

经过上述预处理之后，获得了 9 295 368 次对 815 个 new gTLD 的 611 769 个 FQDN 的查询；作为对比使用的总体数据集则包含约 28 亿条日志.

需要说明的是，数据集中所有可能关联到用户的信息均已删除或者匿名化. 特别地，数据集不包含最终用户的 IP 地址，用户所在网络的 BGP 前缀也经过匿名化处理. 此外，所有分析以聚合的方式进行统计分析，而不进行任何针对单个用户的分析.

3 new gTLD 域名的解析行为分析

本节从查询量和域名数量的总览分析 new gTLD 域名的解析行为，并与非 new gTLD 域名进行对比. 为了分析 new gTLD 计划首批授权对顶级域名带来的扩展，首先使用 whois 来获取数据集中每个顶级域名的创建日期，并收集了 547 个顶级域名的成功回复. 图 1 展示了每年创建的顶级域名的数量，可以看出 new gTLD 快速增长.

3.1 域名与查询量

本节分析各个 new gTLD 的 DNS 查询量和 SLD 的数量，并将它们的分布绘制在对数图中，如图 2 所示，其中 x 轴是分别按查询量和 SLD 排序的顶级域名排名，y 轴是相应的查询量或 SLD 数量.

从图 2 中可以看到一个重尾分布，表明查询量和注册的域名数量集中在少数的顶级域名上. 事实上，如表 1 所示，前 3 个顶级域名吸引了大约一半的与 new gTLD 相关的查询. 表 1 也列出了查询失败率，可以看出，不同的 new gTLD 失败率为 1.6%~45.8%，这意味着顶级域名的可靠性存在较大差异，其中 help 的失败率最低，表明其最为可靠；而 win 的失效率最高，意味着对其下域名的查询更容易失败，可靠性最差.

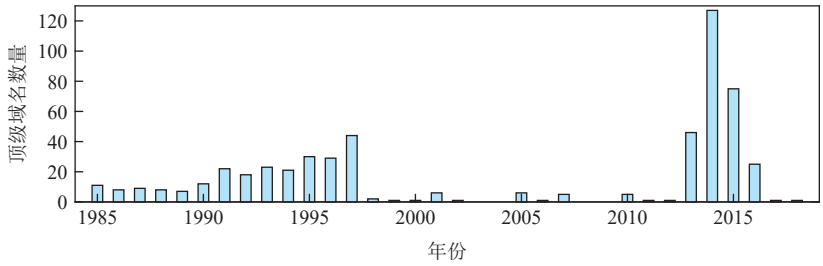

Fig. 1 The number of TLDs created per year

图 1 每年创建的顶级域名个数

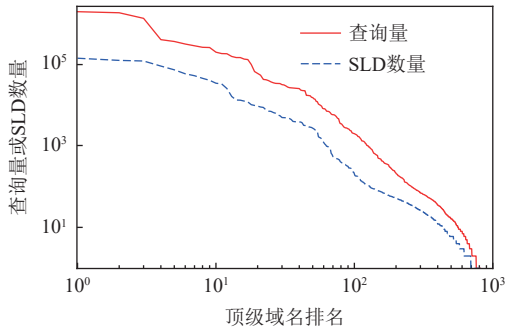

Fig. 2 Distribution of the number of queries and SLDs across TLDs

图 2 顶级域名对应的查询量和 SLD 数量分布

Table 1 Top 10 New gTLDs Ranked by the Number of Queries

表 1 按查询量排序的前 10 个新通用顶级域名 %

顶级域名	查询量占比	查询失败率
top	31.1	16.0
xyz	11.1	19.0
help	6.5	1.6
win	6.2	45.8
link	6.1	2.3
club	5.1	20.9
vip	4.2	10.8
space	2.8	17.4
online	2.1	9.6
loan	1.9	3.7

接下来分析域名查询失败量，表 2 呈现了 2 种主要的 DNS 查询类型：A 类型和 AAAA 类型的查询量占比和查询成功率. 为了进行比较，还计算了所有类型的顶级域名（包括传统通用顶级域名、国家代码顶级域名和新通用顶级域名）的成功率. 可以看出，AAAA 类型查询的成功率明显较低，且只占了 10.4% 的查询量，与 2012 年观察到的情况相似[15]. 此外，new gTLD 域名的查询中 AAAA 类型的查询量占比略大.

Table 2 Percentages of the Number of Queries and Success Rates for A and AAAA Queries

表 2 A 类型和 AAAA 类型查询量占比与查询成功率 %

指标	所有域名		new gTLD 域名	
	A	AAAA	A	AAAA
查询量占比	86.2	10.4	84.3	14.8
查询成功率	93.1	35.8	88.6	25.9

图 3 进一步分析查询失败量是否集中在少数顶级域名上. 可以看到曲线是一个近似对角线的趋势，这表明 new gTLD 的失败查询量也集中在少数顶级域名上.

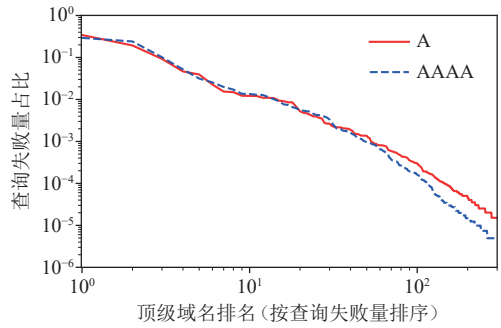

Fig. 3 Distribution of the number of failures for TLDs

图 3 顶级域名的查询失败量分布

综上，在查询量、每个顶级域名的相关 SLD 数量和查询失败率方面，new gTLD 域名的行为呈现重尾分布，并且 new gTLD 域名的成功率比所有域名低.

3.2 内容承载基础设施

本节分析 new gTLD 域名的内容承载基础设施. 由于 AAAA 类型查询在 DNS 查询中占比较小，而且经常失败，在接下来的分析中使用应答含有 IPv4 地址的 A 类型查询.

通过测量承载每个 SLD 使用的 IP/24 网段的数量，考察 new gTLD 域名的内容在互联网上的复制情况，其分布情况如图 4 所示，其中按照 DNS 查询次数对域名进行排序. 图 4 分析了 3 组结果，即前 1 000 域

名、前 10 000 域名和全部 new gTLD 域名. 可以看出, 流行域名被复制到更多的 IP/24 网段. 然而, 在排名前 1 000 的域名中, 只使用一个 IP/24 网段来承载其内容的域名多达 40%; 当考虑所有 new gTLD 域名时, 这一占比上升到 93.7%. 相比之下, 即使是排名第 100 000 的非 new gTLD 域名(该域名是 mattel.com)也在 10 个 IP/24 网段上进行复制. 这些结果表明, new gTLD 域名的内容复制非常有限.

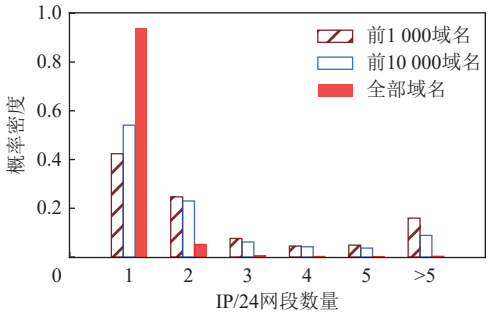

Fig. 4 Distribution of the number of IP/24 segments used by each SLD

图 4 每个 SLD 使用 IP/24 网段数量的分布

进一步分析了这些 IP/24 网段的内容承载特征. 为此, 参考文献 [16] 定义了 2 个指标: 内容分发潜力(content delivery potential, CDP)和内容垄断指数(content monopoly index, CMI). CDP 定义了 IP/24 网段能服务的域名(即域名解析后得到的目标地址在该 IP/24 网段中)占比, 以此衡量潜在地可以从一个 IP/24 网段提供的内容数量. 具体地, 给定一个 SLD 集合(记为 R), IP/24 网段 i 的 CDP 计算公式为

$$CDP_i = \frac{|S_i|}{|R|}, \tag{1}$$

其中 $S_i \subseteq R$ 是该 IP/24 网段承载的 SLD 集合.

CMI 通过对 IP/24 网段能服务的所有域名求权重平均值(其中每个域名的权重为能承载该域名的 IP/24 网段数量的倒数, 这样设置权重是因为如果能承载该域名的 IP/24 网段越多, 则该域名越无法体现该 IP/24 网段的内容垄断性, 因此权重越低)衡量一个 IP/24 网段承载其他 IP/24 网段所没有的新通用顶级域名的程度. 一个 IP/24 网段的 CMI 为

$$CMI_i = \frac{1}{|S_i|}\sum_{j \in S_i}\frac{1}{m_j}, \tag{2}$$

其中 m_j 是承载 SLD $j \in S_i$ 的 IP/24 网段数量. 一个 IP/24 网段的 CMI 大, 意味着一些 new gTLD 域名是由该 IP/24 网段独家提供服务的.

图 5 分析了按 CDP 排序的前 15 个 IP/24 网段. 可以看出, 除了前 2 个 IP/24 网段之外, 其他 IP/24 网段的 CDP 都相当小. 大约 30.3% 的 SLD 可以由第 1 个 IP/24 网段提供服务, 该网段属于中国电信天翼云(CTCloud). 第 2 个 IP/24 网段属于阿里巴巴云, 它可以为 26.5% 的 new gTLD 的 SLD 提供服务. 由此可见, new gTLD 域名的所有者倾向于将网站外包给云服务进行内容承载, 这些域名广泛使用了共享的内容承载基础设施. 另一个观察结果是, 这些 IP/24 网段的 CMI 值都相当高(接近 1), 表明 new gTLD 域名几乎唯一地承载在对应 IP/24 网段. 这也是符合预期的, 因为大多数 SLD 只使用 1 个 IP/24 网段进行内容承载, 如图 4 所示.

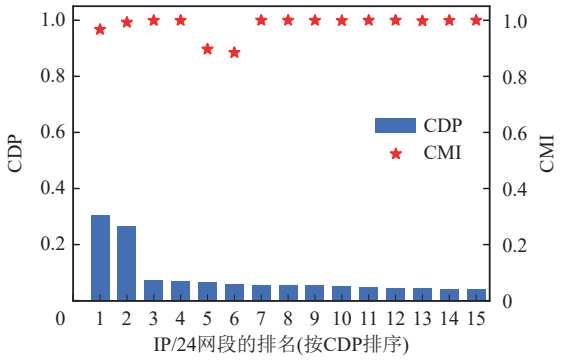

Fig. 5 CDP and CMI of the top 15 IP/24 segments ranked by CDP

图 5 按照 CDP 排序前 15 个 IP/24 网段的 CDP 和 CMI

进一步, 图 6 展示了 new gTLD 域名解析到的 IP 地址范围分布. 对于一个由多个 IP 地址承载的域名, 选择使用次数最多的 IP 地址用于作图, 以确保 y 轴取值不超过 1. 观察到 3 条线中每条都有几个突起, 再次证实了 new gTLD 域名的承载服务器集中在某些 IP 地址范围内. 此外, new gTLD 域名和 IP 地址有不同的集中范围. 这表明, 一些内容承载服务商使用

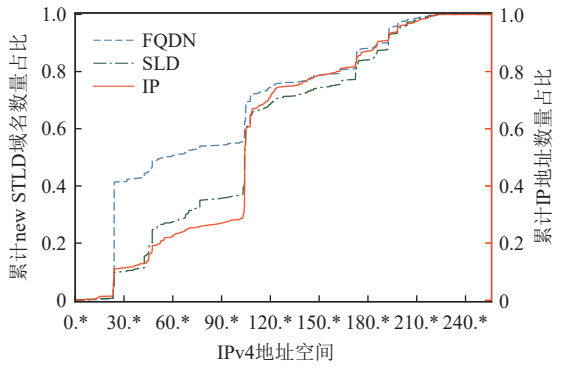

Fig. 6 new gTLD domain and response IP address distribution across IPv4 address space

图 6 new gTLD 域名和应答 IP 地址在 IPv4 地址空间上的分布

少数 IP 地址承载大部分的域名，而其他服务商则使用一个 IP 地址池进行内容承载.

3.3 new gTLD 恶意域名

为研究 new gTLD 恶意域名，本文将 IP/24 网段按其承载的 new gTLD 域名数量进行排序，并提取前 5 个网段所承载的 new gTLD 域名，其中这 5 个网段对应的自治系统（autonomous system, AS）分别属于 Enzu, Leaseweb, Psychz Net, Alibaba 等数据中心、云服务提供商，所承载的 new gTLD 域名对应了约 24.3 万个 FQDN 和 25.5 万次查询，然后使用 VirusTotal 和 360 这 2 个黑名单对域名进行检查，如果这 2 个黑名单中的任何 1 个将域名分类为恶意域名，就将该域名标记为恶意域名. 由于域名数量较大，只检查了 SLD 而没有检查完整的 FQDN，最终有 1 171 个 SLD 被标记为恶意域名，对应 17.9 万个 FQDN（73.6%）和 18.8 万次查询（73.7%）.

如此高的恶意域名数量占比和查询量占比，再结合 3.2 节中的发现，说明存在承载大量 new gTLD 恶意域名的基础设施. 此外，可以观察到 new gTLD 恶意域名的查询量小、与域名个数非常接近，说明这些域名存在时间短、变化快.

接下来从域名个数、源（用户）IP、域名长度和 TTL 等角度分别考察识别到的 new gTLD 恶意域名的 DNS 行为特征，这些特征对恶意域名的早期检测至关重要[7]. 为了叙述方便，由所有 new gTLD 域名组成的域名集合用S_{all}表示，而仅包含恶意 new gTLD 域名的集合用$S_{\text{malicious}}$表示.

图 7 给出了 new gTLD 恶意域名的 SLD 对应 FQDN 数量分布，可以看出恶意的 SLD 对应更多的 FQDN. 具体地，对于$S_{\text{malicious}}$来说，88.5% 的 SLD 有超过 10 个 FQDN. 与此相比，当考虑S_{all}时，87.2% 的 SLD 只对应 1 个 FQDN. 其原因有可能是 DGA 被用来生成恶意域名的 FQDN.

接下来分析请求这些恶意域名的网络地址区域

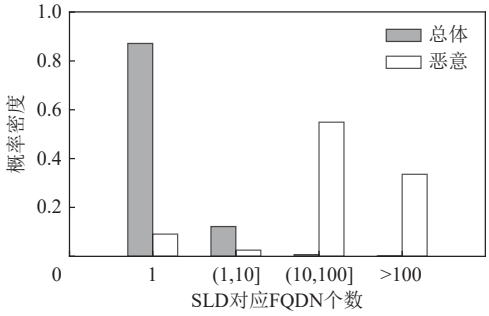

Fig. 7 Distribution of the number of FQDNs across SLDs

图 7 SLD 对应 FQDN 数量的分布

是否与整体分布不同. 为此，计算请求 2 类域名的用户 BGP 前缀的排名，即分别根据S_{all}和$S_{\text{malicious}}$中域名的发送查询次数对用户的 BGP 前缀进行排序，并计算 2 种排名的 Kendall 距离指标[17].Kendall 距离是从 Kendall's tau（衡量 2 个排名列表的相似程度）推广出来的指标，该指标放宽了 Kendall's tau 对于“被比较的 2 个排名列表必须有相同的元素”这一要求，为此，Kendall 距离引入了一个惩罚参数p，并引入一个参数k表示对 2 个排名列表的前k个元素进行比较. 具体地，带有惩罚参数p的 Kendall 距离定义为：

$$K^{(p)}(\tau_1,\tau_2)=\sum_{i,j\in\tau_1\cup\tau_2}\overline{K}_{i,j}^{(p)}(\tau_1,\tau_2), \tag{3}$$

其中τ_1和τ_2表示被比较的 2 个前k排名列表. 本文使用“乐观法”（optimistic approach），具体地，将 p 设置为 0，并且判断 3 个条件是否成立：1）i 和 j 都出现在 2 个列表中，但它们在 2 个列表中的顺序是相反的；2）i 和 j 都出现在某一个列表中，其中 i 的排名高于 j，而 j 出现在另一个列表中；3）只有 i 出现在一个列表中，而 j 出现在另一个列表中. 当这 3 个条件之一成立时，令$\overline{K}_{i,j}^{(p)}(\tau_1,\tau_2)=1$，否则，令$\overline{K}_{i,j}^{(p)}(\tau_1,\tau_2)=0$. 最后，计算归一化的 K（normalized K）[18]：

$$K=1-\frac{K^{(0)}(\tau_1,\tau_2)}{k^2}, \tag{4}$$

所得到的 K 取值范围在 0~1，如果列表τ_1和τ_2不包含相同元素，则 K=0；如果 2 个列表τ_1和τ_2完全相同，则 K=1.

表 3 给出了归一化 K 随考虑的域名个数 k 的变化情况，其中 k=20, 40, 60, 80, 100. 从表 3 中可以看出 2 个列表的距离小于 0.1 时表明 2 个排名差距大，也就是说S_{all}和$S_{\text{malicious}}$中域名的用户前缀差距大.

Table 3 Comparison of Users' BGP Prefixes

表 3 用户 BGP 前缀比较

k	K
20	0.08
40	0.06
60	0.07
80	0.06
100	0.08

进一步分析 new gTLD 域名的 SLD 长度，结果如表 4 所示，可以发现超过 90% 的 new gTLD 恶意域名长度为 4，而考虑所有域名时，域名更长且域名长度分布更均匀. 在传统域名下，考虑到大量的短域名已经被注册，为保证恶意域名可以被攻击者注册，恶意

域名长度一般较长,已有 DGA 域名检测方法往往忽略较短的域名,因此这些方法无法应用于 new gTLD 恶意域名的检测.

Table 4 Fraction of new gTLD Domains of Different SLD Lengths

表 4 不同 SLD 长度的 new gTLD 域名占比

长度(字符个数)	在 new gTLD 域名中的占比/%	在恶意 new gTLD 域名中的占比/%
≤3	8.5	0.1
4	25.2	93.0
5	17.2	6.1
≥6	49.1	0.8

域名解析行为的另一个重要特征是域名解析应答的 TTL 值. 图 8 给出了 A 类型查询的 TTL 值分布,其中包括全体 new gTLD 域名、恶意 new gTLD 域名、非恶意(合法)new gTLD 域名的 TTL 分布,以及作为对比的数据集中全部域名的 A 类型查询 TTL 值分布. 考虑到由于可以从递归解析器的缓存中返回应答,日志中看到的 TTL 值可能小于权威域名服务器设置的原始值. 因此,本文参考文献 [19−20] 中的做法,使用每个域名观察到的最大 TTL 值作为权威域名服务器设置的原始 TTL 值的估计.

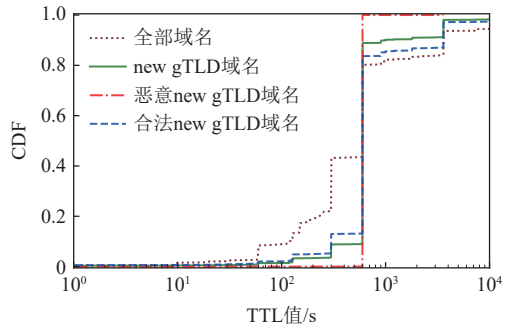

Fig. 8 TTL value distribution of A type queries

图 8 A 类型查询的 TTL 值分布

从图 8 中可以看出,大约 45% 的全体域名 TTL 值小于 120 s,而 new gTLD 域名倾向于使用 600 s 的 TTL 值,并且几乎所有的 new gTLD 恶意域名都把 TTL 值设置为 600 s(这也是许多内容承载提供商的默认设置). 因此,TTL 值可能不适用于 new gTLD 恶意域名的检测.

3.4 对 new gTLD 恶意域名检测的启发

对恶意 new gTLD 域名行为的分析发现,按承载域名量排序的前 5 个 IP/24 网段所承载的域名之中有 73.6% 是恶意的,且与正常域名相比,恶意域名有独特的行为,例如请求用户网络空间分布独特、内容承载基础设施更集中(存在某些 AS 或 IP/24 承载大量恶意域名)、SLD 长度分布偏斜(更短)、对应 FQDN 数量更多、每个 FQDN 的平均查询次数少以及倾向使用默认 TTL 值设置(例如 TTL 值为 600 s). 上述发现从多角度对恶意域名检测的输入特征选择有所启发. 例如在查询量角度,发现恶意域名 FQDN 查询量少,启发使用 SLD 对应的 FQDN 查询量的统计量作为恶意域名检测的输入特征;在域名个数角度,发现恶意 SLD 对应的 FQDN 更多,启发使用 SLD 对应的 FQDN 个数作为输入特征;在请求用户网络空间和应答地址角度,发现恶意域名有独特的网络足迹,启发使用域名与请求用户地址(应答地址)的映射关系,从查询量和域名个数入手,用统计量展开维度作为输入特征;在域名长度角度,发现恶意域名短,启发避免使用域名的文本特征作为输入;在 TTL 角度,发现集中在 600 s,启发不使用 TTL 值作为输入.

需要说明的是,本文在标记数据集中的域名时使用 VirusTotal 和 360 黑名单等恶意域名标记工具,其工作原理是聚合多家杀毒引擎、网站扫描器、URL 分析工具和域名屏蔽列表判断所提交域名是否为恶意域名. 这些工具的恶意域名识别准确度高,但是在域名覆盖度上以及检测时效性上存在不足. 因此,它们被广泛应用在基准数据集的构造上,以评估所设计方法的准确性,而不能直接作为恶意域名检测的工具.

本文旨在设计一种基于解析特征的简单有效的 new gTLD 恶意域名识别方法,实现恶意域名的快速准确检测,降低恶意域名造成的危害.

4 new gTLD 恶意域名检测方法

本节基于第 3 节分析发现的 new gTLD 恶意域名行为特征,设计 new gTLD 恶意域名检测方法. 恶意域名检测本质上是一个二分类问题,主要包括输入特征、分类器、输出 0/1 标签(0 为正常域名(本问题中作为负类);1 为恶意域名(本问题中作为正类)).

new gTLD 恶意域名检测的主要挑战在于,new gTLD 恶意域名相比传统恶意域名有特殊性,需要选取更合适的特征. 现有恶意域名检测方法中缺少针对 new gTLD 域名的检测方法,而传统恶意域名检测方法设计中往往利用域名的文本特征(如可发音单词数、数字符号字符占比等)进行检测. 但是,第 3 节的分析发现了若干 new gTLD 恶意域名的独特行为特征,使得传统恶意域名检测方法并不适用. 如 new

gTLD 域名长度分布偏斜(恶意域名相比正常域名更短)，因此域名的文本特征区分性不强，进而导致传统方法对于 new gTLD 恶意域名检测效果不佳. 因此，本文基于第 3 节的分析结果，选择具有区分度的特征设计针对 new gTLD 的恶意域名检测方法，这些特征包括请求用户网络空间分布独特、内容承载基础设施更集中、对应 FQDN 数量更多、每个 FQDN 的平均查询次数少等.

4.1 输入特征

结合 3.2 节与 3.3 节的发现，本节避免使用域名的文本特征，也不使用 TTL 值作为输入，而是分别从域名的查询量、域名个数、发起查询的用户(源 IP)、应答 IP 的角度提取了特征. 表 5 总结了使用的特征.

Table 5 Feature Sets

表 5 特征集合

特征集合	特征名称	维度
域名的查询量	SLD 对应 FQDN 查询量的统计量	8
域名个数	SLD 对应 FQDN 的个数（总数+每天）	15
客户端（源 IP）	SLD 对应源 AS 和 BGP prefix 的个数	2
	SLD 对应源 AS 和 BGP prefix 查询量的统计量	16
	SLD 对应源 AS 和 BGP prefix 在查询方面的特征	14
应答 IP（AS）	SLD 对应应答 AS 的个数	1
	SLD 映射到 AS 次数的统计量	8
	SLD 映射到 AS 在内容承载方面的特征	7

在域名查询量角度，使用每个 SLD 对应的 FQDN 查询量的统计量作为输入特征. 例如，某个 SLD 对应了 n 个 FQDN，这 n 个 FQDN 的查询量分别是 $x_1, x_2, \cdots, x_n$，则对 $(x_1, x_2, \cdots, x_n)$ 计算统计量，这里使用的统计量包括了最大值、最小值、均值、标准差、下四分位点、中位数、上四分位点和熵，共 8 维. 这样做是因为，3.3 节中观察到这些 new gTLD 恶意域名的查询量与域名个数非常接近，说明这些域名查询量少、存在时间短、变化快.

在域名个数的角度，使用了每个 SLD 对应的 FQDN 数量作为输入特征. 除了总量之外，也统计了每天对应的数量，以反映域名数量在时间上的变化. 这样做的原因是，在图 7 中发现恶意的 SLD 比 new gTLD 域名整体有更多的 FQDN.

在发起查询的用户(源 IP)的角度，使用查询每个 SLD 的源 AS 和 BGP prefix 数量、查询量的统计量以及在查询方面的特征作为输入. 考虑这些特征是因为发起恶意域名查询次数多的用户 BGP 前缀与发起非恶意域名查询次数多的用户 BGP 前缀有很大不同，如表 3 所示. 这启发了使用 SLD 与源 IP 的映射关系，从查询量和域名数量入手，用统计量展开特征维度作为 new gTLD 恶意域名早期检测的一部分输入特征.

除了 BGP prefix 之外，本文也考虑把 AS 级别的统计量加入. 查询量的统计量与域名查询量使用的统计量相同，都是 8 维，具体计算的是查询某 SLD 的每个源 AS(或 BGP prefix)查询该 SLD 次数的统计量. 例如，有 n 个 AS(或 BGP prefix)查询过某 SLD，查询该 SLD 的次数分别是 $x_1, x_2, \cdots, x_n$，则计算 $(x_1, x_2, \cdots, x_n)$ 的统计量，即最大值、最小值、均值、标准差、下四分位点、中位数、上四分位点和熵. 源 AS 和 BGP prefix 在查询方面的特征则是这样考虑的，对查询某个 SLD 的每个源 AS(或 BGP prefix)，计算它们查询 SLD 数量的统计量. 例如，有 n 个 AS(或 BGP prefix)查询过某 SLD，它们查询过 SLD 的数量分别是 $x_1, x_2, \cdots, x_n$，则对 $(x_1, x_2, \cdots, x_n)$ 计算统计量，即计算最大值、最小值、均值、标准差、下四分位点、中位数、上四分位[①].

在内容承载基础设施角度，从应答 IP 映射到 AS，使用了每个 SLD 映射到的 AS 个数、映射到 AS 次数的统计量以及所映射到的 AS 在内容承载方面的特征作为输入. 考虑这些特征是因为发现了存在承载大量 new gTLD 恶意域名的基础设施(见 3.3 节)，并且 new gTLD 域名使用有限的内容复制和共享的基础设施(如图 5 和图 6 所示). 同样，使用 SLD 与应答 IP 的映射关系，从查询量和域名数量入手，用统计量展开特征维度. 在计算 SLD 映射到 AS 次数的统计量时，假设某 SLD 能被 n 个 AS 所承载，且在应答中映射到各 AS 的次数分别是 $x_1, x_2, \cdots, x_n$，则对 $(x_1, x_2, \cdots, x_n)$ 计算统计量. 而在考虑 SLD 所映射到的 AS 在内容承载方面的特征时，对该 SLD 映射到的每个应答 AS 所承载 SLD 的个数计算统计量. 例如，有 n 个 AS 能服务某 SLD，它们所能承载的 SLD 的个数分别是 $x_1, x_2, \cdots, x_n$，则对 $(x_1, x_2, \cdots, x_n)$ 计算统计量.

4.2 分类器选择

我们观察发现 new gTLD 恶意域名的活跃天数很短：97.4% 的 FQDN 只活跃 1 天. 因此，需要选择简单、快速、高效的分类器. 本文考虑随机森林(random forest, RF)、支持向量机(support vector machine, SVM)

① 对查询量计算统计量时使用熵反映分散程度，而在计算域名个数的统计量时去掉了熵，只计算其他 7 维统计量.

和 Boosting 类 3 种典型方法.

本节用来研究 new gTLD 恶意域名检测方法的 SLD 数据是由 3.3 节识别到的共 1 171 个恶意域名和 4 770 非恶意域名组成. 本节使用的评价指标包括准确率(*accuracy*)、精确率(*precision*)、召回率(*recall*)和 *F*1 分数(*F1-score*),其计算公式分别为

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{5}$$

$$precision = \frac{TP}{TP+FP}, \tag{6}$$

$$recall = \frac{TP}{TP+FN}, \tag{7}$$

$$F1-score = \frac{2\times precision\times recall}{precision+recall}, \tag{8}$$

其中TP, FP, TN, FN分别代表真阳性、假阳性、真阴性和假阴性. 由于本文的分类任务更关心对恶意域名的分类性能,因此将恶意域名作为正样本,将正常域名作为负样本.

对 3 类分类器进行了性能评估,经过 5 折交叉验证并计算了各指标的平均值,结果如表 6 所示. 其中 SVM 使用径向基核函数(RBF 核),可以看出 SVM 和 RF 已经取得了较好的实验效果,SVM 在召回率上比 RF 略高,而 RF 在精确率和 *F*1 分数指标表现更好. 而实验结果表明,集成学习中的 3 种 Boosting 类算法(AdaBoost, GBDT, XGBoost)并没能带来明显的性能提升,但是增加了额外的开销. 综上所述,基于实现简单、运行高效且综合性能表现更好的事实,本文选择了 RF 作为分类器.

Table 6 Average Value of Each Metric for Different Classifiers Under 5-Fold Cross-Validation

表 6 不同分类器经过 5 折交叉验证得到各指标的均值 %

指标	RF	SVM	AdaBoost	GBDT	XGBoost
准确率	94	94	93	94	93
精确率	88	83	87	88	87
召回率	84	86	77	80	73
*F*1 分数	86	84	81	83	78

4.3 对比实验

本节分别从基于域名文本特征和基于解析行为特征的两大类方法中,选择 3 种有代表性的方法进行对比实验. 在基于域名文本特征的方法中,选择了文献 [10−11] 提出的方法;而在基于行为特征的方法中,选择文献 [12] 提出的方法. 下面首先简单介绍这 3 种方法的实现.

文献 [10] 提取域名的结构特征(如下划线数量占比)、语言特征(如元音占比)和统计特征(如信息熵),然后将手动提取出的特征输入到作为分类器的 SVM 和 RF 中进行分类,输出对每个域名的预测结果. 文献 [11] 将域名字符串作为输入,利用深度学习自动提取域名特征并完成分类. 文献 [12] 利用 DNS 解析数据提供的映射关系,通过二部图建模域名之间在源 IP、应答 IP、查询时间 3 个维度的相似性,用图嵌入得到域名的向量表示,再输入到 SVM 分类器进行分类. 文献 [12] 提到这样做相比从领域知识、网络流量或者文本词汇提取特征更加鲁棒、稳定,因为恶意域名的基本行为特征倾向于高度一致性. 已有研究表明,领域知识在不同网络中有所区别,网络特征,例如恶意域名设置的 TTL 值会随时间改变(为了避免被检测到),恶意域名也会模仿正常域名的文本特征,比如用相似的字符个数或者可发音的单词.

本文复现了文献 [10−12] 的 3 种方法,并与本文方法对比,结果如表 7 所示. 可以看出:1)在基于域名文本特征的方法中,对于文献 [10] 提出的方法,其输出将所有域名标记为正常域名,说明在 new gTLD 恶意域名检测问题上,使用这些人为提取的域名文本特征难以区分恶意与正常域名;而文献 [11] 的方法优于使用人为提取的域名文本特征的方法[10],但各项指标明显弱于基于解析行为特征的方法.2)在基于 DNS 解析行为特征的方法中,文献 [12] 提出的方法优于基于域名文本特征的方法,但由于 new gTLD 域名活跃天数短,减弱了查询时间维度相似性的效果,且没有考虑域名个数、查询量方面的特征,综合指标不如本文方法;此外,文献 [12] 提出的方法中的召回率最高但是精确率偏低,说明将更多域名标记为恶意域名,漏检最少但是误检较多.3)综合各个指标,本文方法取得了最好的效果,更适用于 new gTLD 恶意域名检测. 其原因在于本文方法在输入特征的选择上具有更强的针对性与区分性.

Table 7 Comparison with Other Malicious Domain Detection Methods

表 7 与其他恶意域名检测方法的对比 %

指标	基于域名文本特征		基于 DNS 解析行为特征	
	文献 [10]	文献 [11]	文献 [12]	RF
准确率	80	41	79	94
精确率	0	65	53	88
召回率	0	16	88	84
*F*1 分数	0	25	66	86

进一步地，调整恶意域名和正常域名的比例，原始数据集中恶意域名与正常域名的比例约为 1∶4，通过随机采样正常域名的方法分别将此比例调整为 1∶3，1∶2，1∶1，并对比数据集正负样本数量的比例对本文方法的影响；此外还通过随机采样恶意域名的方法将比例调整为 1∶10，以考察样本较为不均衡的情况，实验结果如表 8 所示. 可以看出，本文方法的性能较为稳定，而样本中正负例数量的比值确实会影响检测结果，样本数量越均衡，分类性能指标越好. 当比例调整为 1∶10 时，精确率、召回率和 $F1$ 分数都有一定的下降，而准确率反而比原始比例（1∶4）更高了，这是因为此时由于负样本太多，分类器倾向于输出分类结果为“正常域名”，反而提高了准确率.

Table 8 The Effect of the Ratio of the Number of Malicious and Legitimate SLDs on the Metrics

表 8 恶意 SLD 和正常 SLD 数量的比例对指标的影响 %

指标	1∶10	1∶4	1∶3	1∶2	1∶1
准确率	96	94	95	96	96
精确率	83	88	91	93	95
召回率	79	84	91	94	97
$F1$ 分数	80	86	91	94	96

5 总 结

本文使用被动 DNS 日志对 new gTLD 的解析行为进行了分析. 主要发现包括：1）new gTLD 的查询量和对应 SLD 数量符合重尾分布；2）大多数 new gTLD 域名只将它们的内容复制到 1 或 2 个 IP/24 网段，且使用共享的承载基础设施；3）与正常域名相比，恶意域名在内容承载基础设施集中性、SLD 对应的 FQDN 数目、域名查询次数、请求用户网络空间分布、SLD 长度分布等方面具有独特的特征. 基于这些发现，设计了一种 new gTLD 恶意域名的检测方法，充分考虑了 new gTLD 恶意域名在解析行为方面的独特特征，并使用 RF 作为分类器，以实现快速、高效的恶意域名检测. 实验结果表明，与已有方法相比，本文方法具有较高的准确率.

作者贡献声明：杨东辉、曾彬、李振宇提出了研究思路和实验方案；杨东辉负责完成实验并撰写论文初稿；曾彬获取原始数据与完成部分数据分析；李振宇负责修改论文.

参 考 文 献

[1] Korczyński M, Wullink M, Tajalizadehkhoob S, et al. Cybercrime after the sunrise: A statistical analysis of DNS abuse in new gTLDs[C]//Proc of the 13th Asia Conf on Computer and Communications Security. New York: ACM, 2018: 609−623

[2] Fan Zhaoshan, Wang Qing, Liu Junrong, et al. Survey on domain name abuse detection technology[J]. Journal of Computer Research and Development, 2022, 59(11): 2581−2605 (in Chinese)
（樊昭杉，王青，刘俊荣，等. 域名滥用行为检测技术综述[J]. 计算机研究与发展，2022，59(11)：2581−2605）

[3] Halvorson T, Der M F, Foster I, et al. From . academy to . zone: An analysis of the new TLD land rush[C]//Proc of the 15th Internet Measurement Conf. New York: ACM, 2015: 381−394

[4] Chen Q A, Osterweil E, Thomas M, et al. MitM attack by name collision: Cause analysis and vulnerability assessment in the new gTLD era[C]//Proc of the 37th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2016: 675−690

[5] Chen Q A, Thomas M, Osterweil E, et al. Client-side name collision vulnerability in the new gTLD era: A systematic study[C]//Proc of the 24th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 941−956

[6] Pouryousef S, Dar M D, Ahmad S, et al. Extortion or expansion? An investigation into the costs and consequences of ICANN’s gTLD experiments[G]//LNCS 12048: Proc of the 21st Int Conf on Passive and Active Measurement. Berlin: Springer, 2020: 141−157

[7] Hao Shuang, Feamster N, Pandrangi R. Monitoring the initial DNS behavior of malicious domains[C]//Proc of the 11th Internet Measurement Conf. New York: ACM, 2011: 269−278

[8] Hao Shuang, Kantchelian A, Miller B, et al. Predator: Proactive recognition and elimination of domain abuse at time-of-registration[C]//Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 1568−1579

[9] Manadhata P K, Yadav S, Rao P, et al. Detecting malicious domains via graph inference[C]// Proc of the 7th Workshop on Artificial Intelligent and Security Workshop. New York: ACM, 2014: 59−60

[10] Schüppen S, Teubert D, Herrmann P, et al. FANCI: Feature-based automated NXdomain classification and intelligence[C]// Proc of the 27th USENIX Security Symp. Berkeley, CA: USENIX Association, 2018: 1165−1181

[11] Yu Bin, Gray D L, Pan Jie, et al. Inline DGA detection with deep networks[C]//Proc of the 17th IEEE Int Conf on Data Mining Workshops (ICDMW). Piscataway, NJ: IEEE, 2017: 683−692

[12] Lei Kai, Fu Qiuai, Ni Jiake, et al. Detecting malicious domains with behavioral modeling and graph embedding[C]//Proc of the 39th Int Conf on Distributed Computing Systems (ICDCS). Piscataway, NJ: IEEE, 2019: 601−611

[13] greenSec GmbH. nTLDStats[EB/OL]. [2019-05-23]. https://ntldstats.com

[14] nexB Inc. public suffix2 2.20191221[EB/OL]. [2020-03-01].https://pypi.org/project/publicsuffix2/

[15] Gao Hongyu, Yegneswaran V, Chen Yan, et al. An empirical reexamination of global DNS behavior[C]//Proc of the 27th ACM SIGCOMM Conf. New York: ACM, 2013: 267−278

[16] Ager B, Mühlbauer W, Smaragdakis G, et al. Web content cartography[C]//Proc of the 11th ACM SIGCOMM Internet Measurement Conf. New York: ACM, 2011: 585−600

[17] Fagin R, Kumar R, Sivakumar D. Comparing top *k* lists[J]. SIAM Journal on Discrete Mathematics, 2003, 17(1): 134−160

[18] McCown F, Nelson M L. Agreeing to disagree: Search engines and their public interfaces[C]//Proc of the 7th ACM/IEEE-CS Joint Conf on Digital Libraries. New York: ACM, 2007: 309−318

[19] Callahan T, Allman M, Rabinovich M. On modern DNS behavior and properties[J]. ACM SIGCOMM Computer Communication Review, 2013, 43(3): 7−15

[20] Allman M. Putting DNS in context[C]//Proc of the 20th Internet Measurement Conf. New York: ACM, 2020: 309−316

Yang Donghui, born in 1994. PhD. His main research interests include the domain name system and Internet measurement.

杨东辉, 1994 年生. 博士. 主要研究方向为域名系统、互联网测量.

Zeng Bin, born in 1979. PhD. His main research interests include artificial intelligence for IT Operations, network security, and big data analysis.

曾　彬, 1979 年生. 博士. 主要研究方向为智能运维、网络安全、大数据分析.

Li Zhenyu, born in 1980. PhD, professor. His main research interests include Internet measurement, network protocol and networked systems.

李振宇, 1980 年生. 博士, 研究员. 主要研究方向为网络测量、网络协议与网络系统.