# 使用 SNARE 检测垃圾邮件发送者：时空网络级自动信誉引擎

*郝爽、纳迪姆·艾哈迈德·赛义德、尼克·费姆斯特、亚历山大·格雷、斯文·克拉瑟*
*佐治亚理工学院计算机学院　　*迈克菲公司*
*{shao, nadeem, feamster, agray}@cc.gatech.edu, sven_krasser@mcafee.com*

## 摘要

用户和网络管理员需要主要根据发件人的信誉筛选电子邮件的方法。不幸的是，传统的发件人声誉机制（尤其是 IP 黑名单）维护起来很麻烦且易于规避。本文研究了仅基于网络级特征推断电子邮件发件人信誉的方法，而不查看邮件的内容。首先，我们研究网络级特征的一阶属性，这些属性可能有助于区分垃圾邮件发送者和合法发件人。我们检查无需查看数据包内容即可确定的特征，例如 IP 空间中与其他电子邮件发件人的距离或发件人和收件人之间的地理距离。我们派生了轻量级的功能，因为它们不需要查看来自单个 IP 地址的大量电子邮件，并且可以在不查看电子邮件内容的情况下收集 – 许多此类功能甚至从单个数据包中也很明显。第二 我们将这些功能整合到分类算法中，并评估分类器自动将电子邮件发件人分类为垃圾邮件发送者或合法发件人的能力。我们基于这些功能构建了一个自动化信誉引擎 SNARE，使用来自已部署的商业垃圾邮件过滤系统的标记数据。我们证明了SNARE可以达到与现有静态IP黑名单相当的准确性：检测率约为70%，误报率低于0.3%。第三，我们展示了如何将SNARE集成到现有的黑名单中，本质上是作为第一遍过滤器。

## 1 介绍

垃圾邮件过滤系统使用两种机制来过滤垃圾邮件：内容过滤器，根据邮件内容对邮件进行分类;和发件人信誉，它维护有关发件人 IP 地址的信息作为筛选的输入。内容过滤器（例如 [22， 23]）可以阻止某些类型的不需要的电子邮件，但它们可能很脆弱且可规避，并且需要分析电子邮件的内容，这可能很昂贵。因此，垃圾邮件过滤器还依靠发件人信誉来过滤邮件;这个想法是，邮件服务器可能能够纯粹根据发件人的信誉而不是邮件内容来拒绝邮件。基于 DNS 的黑名单 （DNSBL），如 Spamhaus [7]，维护已知发送垃圾邮件的 IP 地址列表。不幸的是，这些黑名单可能不完整，对新的垃圾邮件发送者的反应也很慢[32]。这种无响应只会变得更加严重，因为僵尸网络和BGP路由劫持使垃圾邮件发送者更容易动态获取新的，未列出的IP地址[33，34]。事实上，网络管理员仍在寻找轻量级（即，它们不需要详细的消息或内容分析）和自动化（即 它们不需要手动更新、检查或验证）。

为了实现这一目标，本文介绍了SNARE（时空网络级自动信誉引擎），这是一种发件人信誉引擎，可以根据轻量级网络级特征对电子邮件发件人进行准确，自动的分类，这些功能可以在发件人历史记录的早期确定 – 有时甚至只看到一个数据包。SNARE依赖于一种直觉，即大约95%的电子邮件是垃圾邮件，其中75 – 95%可归因于僵尸网络，僵尸网络通常表现出与合法电子邮件发件人不同的不寻常发送模式。SNARE根据发送消息的方式（即流量模式）而不是发件人是谁（即他们的IP地址）对发送者进行分类。换句话说，SNARE基于这样的假设，即存在轻量级网络级功能，可以将垃圾邮件发送者和合法发件人区分开来;本文发现了这些功能，并使用它们来构建一个自动确定电子邮件发件人声誉的系统。

SNARE与其他基于网络级行为对发送者进行分类的方法有一些相似之处[12，21，

24, 27, 34]，但这些方法依赖于检查邮件内容、收集大量收件人的信息或两者兼而有之。相比之下，SNARE基于轻量级网络级功能，这可以使其更好地扩展，并在更高的流量速率下运行。此外，SNARE比以前的信誉系统更准确，后者使用网络级行为特征对发件人进行分类：例如，对于类似的检测率，SNARE的误报率比我们以前的工作[34]低一个数量级。它是第一个既能与现有静态 IP 黑名单一样准确，又能自动跟上不断变化的发件人行为的信誉系统。

尽管根据“网络级”功能自动推断发件人信誉具有优势，但仍然存在一个主要障碍：我们必须确定哪些功能可以有效且高效地区分垃圾邮件发送者和合法发件人。鉴于可能特征的巨大空间，找到一组以低误报率和低假阴率对发件人进行分类的特征具有挑战性。本文确定了 13 个这样的网络级功能，这些功能需要不同级别的发件人历史记录信息。

不同的功能会产生不同级别的开销。因此，我们首先评估可以在接收方本地精确计算的特征，没有来自其他接收方的信息，没有以前的发送历史，也没有检查消息本身。我们发现属于此类别的几个功能对于对发件人进行分类非常有效，包括：发件人的 AS、发件人的 IP 地址与收件人的 IP 地址之间的地理距离、周围 IP 地址空间中电子邮件发件人的密度以及邮件发送的时间。我们还查看了消息和接收者之间的各种聚合统计信息（例如，从单个IP地址发送的消息的平均值和标准偏差），发现虽然这些功能需要更多的计算和消息开销，但它们确实有助于区分垃圾邮件发送者和合法发件人。识别出这些特征后， 我们分析了这些功能的相对重要性，并基于RuleFit [19]集成学习算法将它们整合到自动信誉引擎中。

除了介绍第一个基于网络级特征的自动分类器外，本文还提出了一些额外的贡献。首先，我们对垃圾邮件发送者和合法发件人的各种网络级特征进行了详细研究，详细研究了每种特征如何区分垃圾邮件发送者和合法发件人，并解释了为什么这些功能可能在垃圾邮件发送者和合法发件人之间表现出差异。其次，我们使用最先进的集成学习技术来构建使用这些特征的分类器。我们的结果表明，SNARE的性能 mance 至少与基于 DNS 的静态黑名单一样好，实现了 70% 的检测率，误报率约为 0.2%。使用从单个邮件中提取的功能和这些功能的聚合可提供轻微的改进，并且添加托管最常错误分类发件人的 AS 的 AS“白名单”可将误报率降低到 0.14%。这种准确性大致相当于现有的静态IP黑名单，如SpamHaus [7]；然而，其优点是 SNARE 是自动化的，它根据发送行为而不是其 IP 地址来表征发送方，IP 地址可能会因动态寻址、新受损的主机或路由劫持而发生变化。尽管SNARE的性能仍然不完美，但我们相信其优势是显而易见的：与其他电子邮件发件人信誉系统不同，SNARE既自动化又轻量级，足以仅对网络级信息进行操作。第三，我们提供了SNARE的部署方案。 即使其他人没有完全按照我们的描述部署SNARE的算法，我们相信网络级功能本身的集合可以为其他商业和开源垃圾邮件过滤设备提供有用的输入。

本文的其余部分组织如下。第 2 节介绍了现有发件人信誉系统的背景和 SNARE 的可能部署方案，并介绍了集成学习算法。第 3 节介绍了电子邮件发件人的网络级行为属性，并测量了与垃圾邮件发送者和合法发件人的这些功能相关的一阶统计信息。第 4 节使用不同的特征子集评估 SNARE 的性能，范围从可以从单个数据包确定的特征子集到需要一定数量历史记录的子集。我们将在第 5 节中研究将分类器合并到垃圾邮件过滤系统中的可能性。第6节讨论规避和其他限制，第7节描述相关工作，第8节总结。

## 2 背景

在本节中，我们将提供有关现有发件人信誉机制的背景，介绍改进发件人信誉机制的动机（我们将在第 7 节中调查其他相关工作），并描述一种称为 RuleFit 的分类算法来构建信誉引擎。我们还介绍了 McAfee 的 TrustedSource 系统，它既是用于我们分析的数据源，也是 SNARE 的可能部署方案。

### 2.1 电子邮件发件人信誉系统

当今的垃圾邮件过滤器在基于 DNS 的黑名单（DNSBL） 中查找 IP 地址，以确定 IP 地址是否是当时已知的垃圾邮件来源

的查找。一个常用的公共黑名单是垃圾邮件屋 [7]；其他黑名单运营商包括SpamCop[6]和SORBS [5]。目前的黑名单有三个主要缺点。首先，它们仅提供 IP 地址粒度的信誉。不幸的是，正如我们早期的工作所观察到的[34]，发件人的IP地址是动态的：在任何一天中，大约有10%的垃圾邮件发件人以前没有被观察到过。这项研究还观察到，许多垃圾邮件IP地址将在数周内处于非活动状态，大概直到它们从IP黑名单中删除。这种动态性使得维护响应式 IP 黑名单成为手动、乏味和不准确的

费率过程;它们通常也是粗粒度的，将整个前缀列入黑名单（有时过于激进），而不是单个发件人。其次，IP黑名单通常是不完整的：之前的一项研究指出，在垃圾邮件陷阱中收到的垃圾邮件中，多达20%未列入任何黑名单[33]。最后，它们有时是不准确的：铁事证据充斥着合法邮件服务器的IP地址被错误列入黑名单的故事（例如，因为它们将垃圾邮件反映到邮件列表中）。为了弥补这些缺点，商业信誉系统通常会合并其他数据，如SMTP元数据或邮件指纹，以减轻这些缺点[11]。我们之前的工作引入了“行为黑名单”，并基于单个行为特征开发了一个垃圾邮件分类器：特定IP地址发送到每个收件人域的邮件数量[34]。 本文以行为黑名单为主题，寻找更好的功能，可以更早地对发件人进行分类，并且更能抵抗逃避。

## 2.2 数据和部署方案

本节介绍迈克菲的 TrustedSource 电子邮件发件人信誉系统。我们描述了如何使用来自该系统的数据来研究电子邮件发件人的网络级特征并评估SNARE的分类。我们还介绍了如何将SNARE的功能和分类算法整合到实时发件人信誉系统（如TrustedSource）中。

数据源 TrustedSource 是一个商业信誉系统，允许查找各种互联网标识符，如 IP 地址、URL、域或邮件指纹。它接收来自各种不同设备类型（如邮件网关、Web 网关和防火墙）的查询反馈。我们使用 McAfee TrustedSource 系统中的查询日志在 2007 年 10 月 22 日至 11 月 4 日的 14 天内评估了 SNARE 。 收到的每封电子邮件都会生成对 TrustedSource 数据库的查找，因此查询日志中的每个条目都表示从某个发件人发送到迈克菲某个 TrustedSource 设备的一封电子邮件。由于体积

| 字段时间戳 | *描述* |
|---|---|
| ts_server_name | UNIX 时间戳<br>处理查询的服务器的名称 |
| 得分 | 基于反垃圾邮件筛选器组合的邮件评分 数据包中的源 IP (DNS 服务器将查询中继给我们) |
| source_ip | 正在查询的 IP 邮件<br>正文的长度 收件人地址数 |
| query_ip | |
| body_length | |
| count_taddr | |

图 1：迈克菲数据集中使用的数据说明。

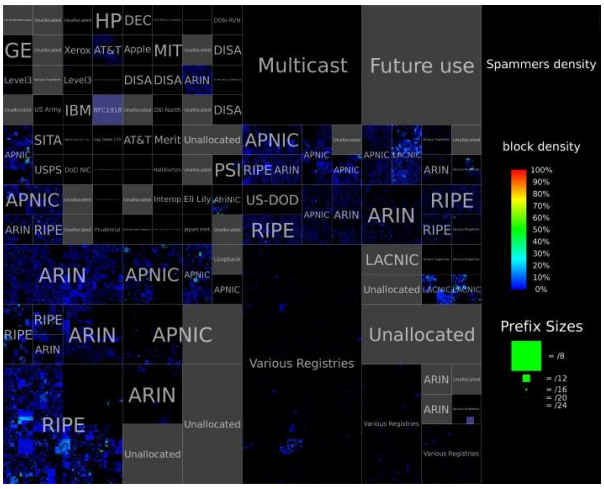



图 2：我们专题研究的一周期间（2007 年 10 月 22 日至 28 日）希尔伯特空间中发件人 IP 地址的分布。（灰色块是未使用的 IP 空间。

在完整的日志集中，我们专注于来自单个 TrustedSource 服务器的日志，该服务器反映了每天从超过 130 万个 IP 地址收到的大约 2500 万封电子邮件。这些消息来自分布在世界各地的大约 2,500 个不同的 TrustedSource 设备。虽然域和设备之间没有精确的一对一映射，并且我们没有唯一域数量的精确计数，但域的数量大致相同数量级。

日志包含许多字段，其中包含每封电子邮件的元数据;图 1 显示了我们最终用于开发和评估 SNARE 分类算法的字段子集。时间戳字段反映在某些域中的 TrustedSource 设备上接收邮件的时间;source_ip字段反映了发出 DNS 查询的计算机的源 IP（即电子邮件的收件人）。query_ip

字段是正在查询的 IP 地址（即电子邮件发件人的 IP 地址）。发送方的 IP 地址显示在希尔伯特空间中，如图 2[1]所示，其中每个像素代表一个 /24 网络前缀，强度表示每个块中观察到的 IP 密度。发件人 IP 地址的分布表明，TrustedSource 数据库在 Internet 上并置了一组具有代表性的电子邮件。我们使用图 1 中的许多其他功能作为 SNARE 分类算法的输入。

为了帮助我们在功能分析（第 3 节）和培训（第 2.3 节和第 4 节）中将发件人标记为垃圾邮件发送者或合法发件人，日志还包含每封电子邮件的分数，这些分数指示 McAfee 如何根据其当前系统对电子邮件发件人进行评分。分数字段指示 McAfee 的发件人信誉评分，我们将其分为五个标签：某些火腿、可能是火腿、某些垃圾邮件、可能是火腿和不确定。虽然这些分数不是完美的基本事实，但它们确实代表了手动分类和持续调整算法的输出，这些算法也对更重量级的特征（例如，数据包有效载荷）进行操作。我们的目标是开发一个与 TrustedSource 一样准确的全自动分类器，但 (1) 自动对发件人进行分类，(2) 仅依赖于轻量级、防规避的网络级功能。

部署和数据聚合方案 由于 SNARE 仅对电子邮件的网络级功能进行操作，因此可以部署为 TrustedSource 的一部分，也可以部署为独立的 DNSBL。SNARE 使用的一些功能依赖于聚合各种发件人的发件人行为。为了聚合这些功能，监视器可以收集有关发件人跨各种收件人域的全局行为的信息。聚合此信息是一项相当轻量级的操作：由于 SNARE 使用的功能基于简单的功能（即 IP 地址和辅助信息），因此它们可以搭载在小型控制消息或 DNS 消息中（如 McAfee 的 TrustedSource 部署）。

## 2.3 监督学习：规则拟合

**集成学习**：在过去十年中，RuleFit Learning集成一直是流行的预测学习方法之一。他们的结构模型采用以下形式

$$F(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m f_m(x) \qquad (1)$$

其中 x 是从训练数据（时空特征）派生的输入变量;$f_m$(x) 不同称为融合成员（“基础学习器”）的函数，M 是融合的大小;F（x）是预测输出（“垃圾邮件”或“火腿”的标签），它采用融合成员的线性组合。给定基本学习器，该技术通过正则化线性回归和“套索”惩罚（惩罚大系数 $a_m$）来确定学习器的参数。

弗里德曼和波佩斯库提出了RuleFit[19]，将回归和分类问题构造为简单规则的线性组合。由于在这种情况下，基本学习器的数量可能很大，因此作者建议使用决策树中的规则作为基本学习器。此外，为了提高精度，变量本身也作为基函数包括在内。此外，最小化损失函数[18]的快速算法和控制树大小的策略可以大大降低计算复杂度。

变量重要性 RuleFit 的另一个优点是解释。由于其形式简单，每个规则都易于理解。在构建预测模型后，可以评估各个变量的相对重要性。经常出现在重要规则或基本函数中的输入变量被认为更相关。变量 $x_i$的重要性是直接对应于该变量的基函数的重要性，加上涉及 $x_i$的所有其他规则的平均重要性。RuleFit论文有更多细节[19]。在第 4.3 节中，我们展示了这些功能的相对重要性。

与其他算法的比较 还有另外两个经典的分类器候选者，我们在数据集上测试了这两个分类器，并且它们都比 RuleFit 产生更差的性能（即更高的误报和更低的检测率）。支持向量机（SVM）[15]已被经验证明在手写识别、人脸检测、文本分类等各种问题上具有良好的泛化性能。另一方面，它们确实需要对参数进行大量调整才能获得最佳性能。如果训练集很大，分类器本身会占用大量存储空间，并且对新数据点进行分类将相应变慢，因为每个测试点的分类成本为 O（S），其中 S 是支持向量的数量。SVM 的计算复杂性与 SNARE 快速（以线速）做出决策的目标相冲突。决策树[30]是另一种流行的分类方法。 生成的分类器易于理解且速度更快，对新测试点的预测采用 O（log（N））， 其中 N 是训练树中的节点数。不幸的是，决策树会影响准确性：其高误报率使其不太适合我们的目的。

[1] 更大的数字可在 http://www.gtnoise.net/snare/hilbert-ip.png.

## 3 网络级功能

在本节中，我们将探讨电子邮件发件人的各种时空特征，并讨论为什么这些属性对于区分垃圾邮件发送者和合法发件人是相关且有用的。我们通过增加开销级别对分析的功能进行分类：

- 单数据包特征是指那些可以在没有先前历史记录的情况下确定的 SNARE 尝试分类的 IP 地址的特征，并且仅从相关 IP 地址中给出单个数据包的特征（第 3.1 节）。
- 可以从单个 SMTP 邮件头或电子邮件中收集单标头和单邮件功能（第 3.2 节）。
- 聚合特征可以用不同数量的历史记录（即其他特征的聚合）来计算（第 3.3 节）。

每类特征都包含那些可能纯粹是单个接收器局部特征或跨多个接收器聚合的特征;后者意味着信誉系统必须具有某种聚合网络中特征的机制。在以下各节中，我们将介绍每个类中的功能，解释选择该功能背后的直觉，并根据垃圾邮件发送者与合法发件人比较该功能。

没有一个功能需要完全区分火腿和垃圾邮件。下面的分析表明，使用单个完美功能来获得最佳分辨率是不现实的。正如我们在第 2.3 节中所描述的，SNARE 的分类算法使用这些特征的组合来构建最佳分类器。但是，我们确实使用这三种不同的特征类来评估SNARE的分类器，以了解它使用这些不同类的性能如何。具体来说，我们仅使用单数据包特征来评估SNARE的分类工作情况，以确定这种轻量级分类器的性能;然后，我们看看使用附加特征是否可以改善分类。

### 3.1 单包功能

在本节中，我们将讨论一些用于识别仅依赖于发件人 IP 地址中的单个数据包的垃圾邮件发送者的属性。在某些情况下，我们还依赖于辅助信息，例如路由表信息、从相邻 IP 地址发送历史记录等，而不仅仅是数据包本身中的信息。我们首先讨论可以从单个 IP 数据包中提取的特征：发送方和接收方之间的测地线距离、发送方邻域密度、IP 数据包到达时垃圾邮件与火腿的概率比、发送方的 AS 编号以及发送电子邮件的计算机。该分析基于 McAfee 从 2007 年 10 月 22 日至 28 日（含）的数据（7 天）。[2]

#### 3.1.1 发件人-收件人测地线距离：垃圾邮件传播得更远

最近的研究表明，通信方之间的社会结构可用于有效地隔离垃圾邮件发送者[13，20]。根据这些研究的结果，我们假设合法电子邮件的地理距离更短，而垃圾邮件传播的距离更接近随机。换句话说，垃圾邮件可能与穿越世界一样短距离传播。

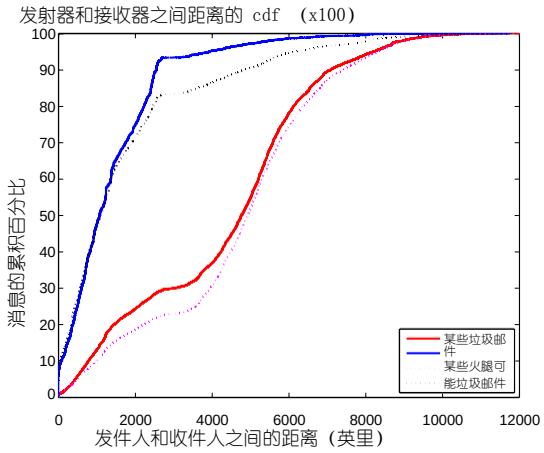
图 3（a） 显示我们的直觉大致正确：四类消息中每一类消息的发送方和目标 IP 地址之间的距离分布。这些图中使用的距离是测地线距离，即沿地球表面的距离。它的计算方法是首先使用MaxMind的GeoIP数据库[8]找到源和目标IP的物理纬度和经度，然后计算这两个点之间的距离。这些距离计算假设地球是一个完美的球体。对于某些火腿，90% 的消息传输大约 2，500 英里或更少。另一方面，对于某些垃圾邮件，只有 28% 的邮件保留在此范围内。事实上，大约 10% 的垃圾邮件传播超过 7，000 英里，这是地球赤道周长的四分之一。这些结果表明，测地线距离是区分垃圾邮件和火腿的一个有希望的指标，这也是令人鼓舞的， 因为它可以仅使用单个 IP 数据包快速计算。

#### 3.1.2 发件人 IP 邻域密度：垃圾邮件发送者被其他垃圾邮件发送者包围

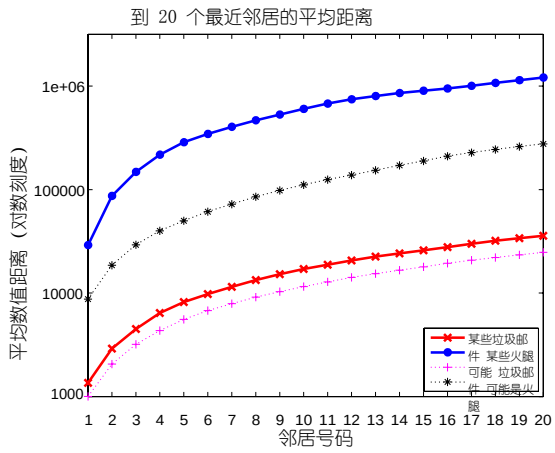
今天大多数垃圾邮件都是由僵尸网络生成的[33，37]。对于来自同一僵尸网络的邮件，受感染的 IP 地址可能在数字空间中彼此靠近，通常甚至位于同一子网中。检测 IP 地址是否属于僵尸网络的一种方法是查看过去的历史记录，并确定是否从与当前发件人位于同一子网中的其他 IP 接收了邮件，其中子网大小可以通过实验确定。如果来自同一子网的许多不同 IP 正在发送电子邮件，则整个子网被机器人感染的可能性很高。

仅使用子网密度的问题在于参考框架不会超越子网

[2]第 4 节中的评估使用 2007 年 10 月 22 日至 11 月 4 日（14 天）的数据，其中一些数据不包括在用于测量研究的数据跟踪中。




(a) 发件人和收件人地理位置之间的测地线距离。




(b) 到 IP 空间中 20 个最近邻的数值距离的平均值。

图 3：垃圾邮件发送者和合法发件人之间的空间差异。

边界。IP 邻域中电子邮件发件人密度的更灵活度量是到其 k 个最近邻居的距离。到k个最近邻的距离可以通过将IP视为从0到$2^{32}-1$（对于IPv4）的数字集并在此一维空间中找到最近邻来计算。我们可以预期这些距离对垃圾邮件和火腿表现出不同的模式。如果社区拥挤，这些邻居距离将很小，表明可能存在僵尸网络。在正常情况下，在较小的 IP 地址空间范围内看到大量 IP 地址发送电子邮件是不寻常的（一个例外可能是出站邮件服务器集群，因此选择适当的阈值很重要，运营商可能需要评估哪个阈值在运行 SNARE 的特定网络上效果最好）。

到发件人的 20 个最近邻居的平均距离如图 3（b）所示。x 轴表示我们在 IP 空间中考虑的最近邻数，y 轴显示样本中到该多邻域的平均距离。该图反映了这样一个事实，即来自主机的绝大多数垃圾邮件在给定 IP 区域中具有较高的电子邮件发件人密度。垃圾邮件到$k^{th}$最近邻的距离平均往往比合法发件人短得多，这表明垃圾邮件发送者通常居住在电子邮件发件人密度较高的区域（就IP地址空间而言）。

### 3.1.3 时间：垃圾邮件发送者根据机器关闭/打开模式发送邮件

可以使用单个数据包中的信息提取的另一个功能是发送消息的时间。我们使用发件人实际位置的当地时间，而不是协调

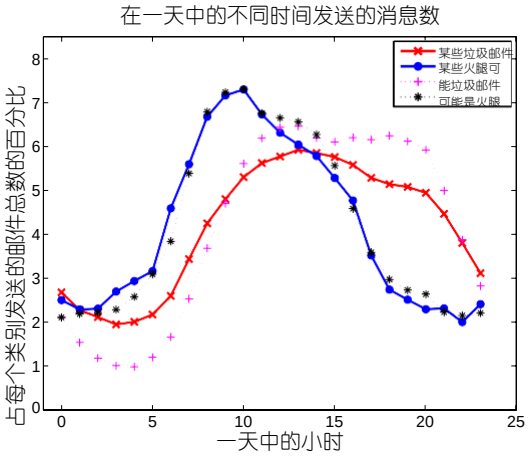



图 4：垃圾邮件发送者和合法发件人的昼夜发送模式的差异。

世界时 （UTC）。此功能背后的直觉是，与垃圾邮件发送模式相比，本地合法电子邮件发送模式可以更密切地跟踪“常规”昼夜模式。

图 4 显示了每种类型的消息在一天中不同时间的相对百分比。合法发件人和垃圾邮件发件人显示不同的昼夜模式。一天中的两个时间特别引人注目：火腿的相对数量往往在工作日开始时迅速增加，并在清晨达到顶峰。在工作日结束时，交易量也会相对较快地减少。另一方面，垃圾邮件以更慢、更稳定的速度增加，可能是因为机器在早上打开。垃圾邮件量全天保持稳定，并在晚上 9：00 左右开始下降，可能是在机器再次关闭时。总之，合法

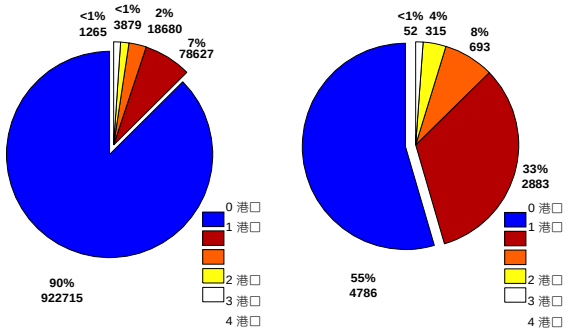
发件人倾向于遵循工作日周期，垃圾邮件发送者倾向于遵循机器电源周期。

为了使用时间戳作为一项功能，我们计算在收到消息的一天中垃圾邮件与火腿的概率比。首先，我们计算在
第 T 天，$s,t$ = 纳秒/$n_s$，其中 $n_{s,t}$ 是数量
如 p 吨
在 T 小时内收到的垃圾邮件，$n_s$是全天收到的垃圾邮件数。我们可以以类似的方式计算某个小时的先验概率$t$，$p_{h,t}$。概率比 $r_t$则简单地为 $p_{s,t}/p_{h,t}$。收到新消息时，
预先计算的发件人时区一天中相应小时的垃圾邮件与业余概率比 $r_t$可用作一项功能;这个比率可以每天重新计算。

### 3.1.4 发件人数量：少量 AS 发送大量垃圾邮件

如前所述，由于多种原因，使用 IP 地址识别垃圾邮件发送者的效率降低。首先，发件人的 IP 地址通常是暂时的。受感染的计算机可能来自拨号用户，这取决于动态 IP 分配。如果垃圾邮件来自移动设备（如笔记本电脑），一旦人们将设备带到不同的地方，IP 地址就会更改。此外，众所周知，垃圾邮件发送者采用隐蔽的垃圾邮件策略，其中每个机器人仅向单个目标域发送多个垃圾邮件，但总体而言，僵尸网络可以向许多域发送大量垃圾邮件[33]。低排放率和分布式攻击需要跨域共享信息以进行检测。

另一方面，我们之前的研究表明，很大一部分垃圾邮件发送者来自相对较小的AS集合[33]。更重要的是，负责垃圾邮件的 AS 与发送合法电子邮件的 AS 不同。因此，电子邮件发件人的AS数量可能是评估发件人声誉的一个有前途的功能。在我们的追踪的七天中，超过 10% 的唯一垃圾邮件 IP（发送某些垃圾邮件的 IP）仅来自 3 个 AS;前 20 个 AS 托管 42% 的垃圾邮件 IP。尽管我们之前的工作注意到少数ASes发起了大部分垃圾邮件[33]，但我们认为这是第一个建议使用电子邮件发件人的AS编号作为发件人声誉自动分类器的输入的工作。




(a) 垃圾邮件　(b) 合法邮件

图 5：发送垃圾邮件和合法邮件的主机上开放端口数量的分布。

### 3.1.5 服务端口的状态：合法邮件往往来自具有开放端口的计算机

我们假设合法的邮件发件人也可能侦听 SMTP 端口以外的其他端口，而机器人可能不会;我们的直觉是，机器人通常将垃圾邮件直接发送到受害者域的邮件服务器，而合法电子邮件则从其他域的MSA（邮件提交代理）移交。反向DNS（rDNS）和正向确认反向DNS（FCrDNS）技术已被广泛用于检查电子邮件是否来自拨号用户或动态分配的地址，邮件服务器将拒绝来自此类来源的电子邮件[1]。

我们提出了一个与 DNSBL 或 rDNS 检查正交的附加功能。发送邮件服务器打开特定端口以接受用户的连接，而机器人是受感染的主机，其中众所周知的服务端口已关闭（需要root权限才能打开）。当数据包到达邮件服务器时，服务器会发出发送到源主机的活动探测，以扫描以下四个通常用于传出邮件服务的端口：25 （SMTP）、465 （SSL SMTP）、80 （HTTP） 和 443 （HTTPS），它们与传出邮件服务相关联。由于当前邮件服务器和 McAfee 的数据均不提供电子邮件发件人的端口信息，因此我们需要探测发件人的 IP 以检查可能打开的服务端口。调查过程是在 2008 年 10 月和 2009 年 1 月期间执行的，远远晚于收到电子邮件的时间。尽管有这种延迟，开放端口的状态仍然暴露了合法发件人和垃圾邮件发送者之间的显着差异。 图 5 分别显示了垃圾邮件和火腿类别的百分比和打开端口的数量。统计数据是根据我们在第 4 节（2007 年 10 月 22 日至 28 日）中使用的评估数据集中的发件人 IP 计算的。在垃圾邮件情况下，90% 的垃圾邮件 IP 地址没有打开任何标准邮件服务端口;相比之下，

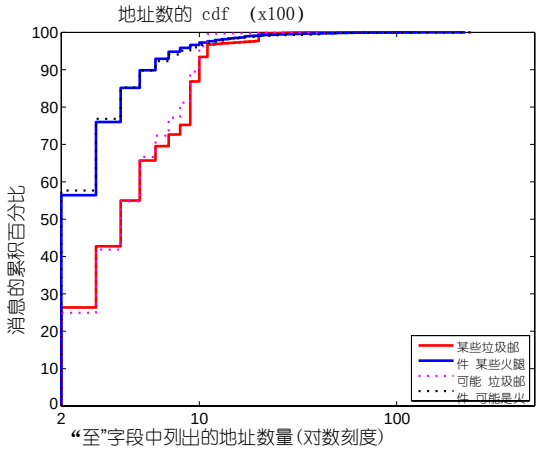



图 6：每个类别的“收件人”字段上列出的地址数分布（忽略单个收件人邮件）。

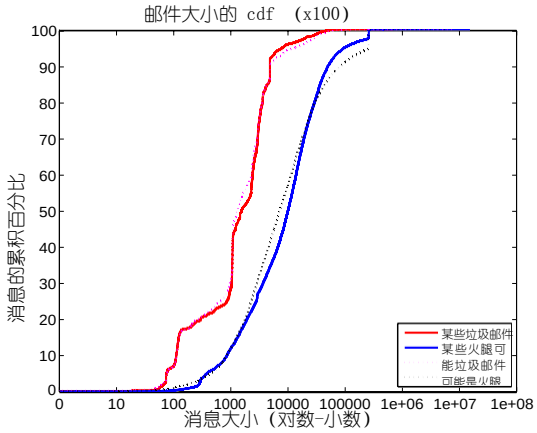



图 7：不同类别消息的消息大小（以字节为单位）的分布。

一半的合法电子邮件来自侦听至少一个邮件服务端口的计算机。尽管防火墙可能会阻止探测尝试（这会导致合法邮件服务器显示没有端口侦听），但与电子邮件相关的端口的状态似乎仍与发件人的区别高度相关。当将此功能作为分类器的输入提供时，我们将其表示为位图（4 位），其中每个位指示发送方 IP 是否正在侦听特定端口。

## 3.2 单头 和 单消息功能

在本节中，我们将讨论可以从单个 SMTP 标头或邮件中提取的其他功能：邮件中的收件人数和邮件的长度。我们将这些功能与上一节中的功能区分开来，因为提取这些功能实际上需要打开 SMTP 连接、接受邮件或两者兼而有之。一旦接受连接，就会收到 SMTP 标头和随后的竞争消息。此时，垃圾邮件过滤器可以提取其他非内容功能。

### 3.2.1 收件人数量：垃圾邮件往往有更多的收件人

到目前为止讨论的功能可以从任何给定特定 IP 地址的单个 IP 数据包中提取，并结合来自其他 IP 的消息的一些历史知识。无需查看内容即可使用的另一个功能是标题“收件人”字段中的地址数量。此功能可以在收到整个 SMTP 标头，但在接受邮件正文之前。但是，大多数邮件只列出了一个地址。超过 94% 的垃圾邮件和 96% 的合法电子邮件发送给单个收件人。图 6 显示了发送给多个收件人的所有电子邮件的每个邮件类别的“收件人”字段中的地址数分布。x 轴在对数刻度上，以将绘图集中在较小的值上。根据此图并查看实际值，如果“收件人”字段上有非常多的收件人（100 或更多），则此度量值的不同类型的发件人之间似乎没有显着差异。围绕 2 到 10 个地址的明显差异表明，一般来说，ham 的收件人较少（接近 2 个），而垃圾邮件发送到多个地址（接近 10 个）。（我们承认此功能可能是可以规避的，并在第 6.1 节中更详细地讨论）。

### 3.2.2 邮件大小：合法邮件的邮件大小可变；垃圾邮件往往很小

收到整封邮件后，电子邮件正文大小（以字节为单位）也是已知的。由于给定的垃圾邮件发件人将主要在所有邮件中发送相同或相似的内容，因此可以预期垃圾邮件发送者发送的邮件大小的差异将小于合法发件人发送的邮件之间的差异。为了保持有效，垃圾邮件机器人还需要保持较小的邮件大小，以便它们可以最大化可以发送的邮件数量。因此，垃圾邮件可能会偏向较小的大小。图 7 显示了每个类别的消息分布。垃圾邮件都聚集在 1–10KB 范围内，而合法发件人的邮件大小分布更均匀。因此，梅斯

—

Sage 正文大小是邮件的另一个属性，可能有助于区分垃圾邮件发送者和合法发件人。

### 3.3 聚合要素

到目前为止讨论的行为属性都可以使用单个消息（带有辅助或邻域信息）来构建。如果 IP 中的某些历史记录可用，则还可以构造一些聚合 IP 级功能。给定有关来自单个 IP 地址的多封邮件的信息，可以使用以下平均值和方差的组合来捕获以下度量的总体分布：（1） 发件人和收件人之间的测地线距离，（2） SMTP 标头的“收件人”字段中的收件人数，以及 （3） 邮件正文长度（以字节为单位）。通过汇总多个消息和一段时间内的行为，这些聚合特征可能会产生更可靠的预测。另一方面，计算这些功能是以增加延迟为代价的，因为我们需要在计算这些消息之前收集大量消息。有时，收集汇总信息甚至需要跨域协作。通过对多条消息求平均值，这些特征还可以平滑特征空间的结构，使边缘情况更难分类。

## 4 评估信誉引擎

在本节中，我们使用不同的特征集评估SNARE的RuleFit分类算法的性能：仅来自单个数据包的特征，来自单个标头或消息的特征以及聚合特征。

### 4.1 设置

在本次评估中，我们使用了从2007年10月22日到2007年11月4日的14天跟踪数据，其中部分数据与第3节中的分析数据不同。换句话说，整个数据跟踪分为两部分：前半部分用于测量研究，后半部分用于评估SNARE的性能。此设置的目的是验证我们发现的特征统计量随着时间的推移会保持相同分布的假设，并确保特征提取不会干扰我们对预测的评估。

训练 我们首先为跟踪的子集收集了每条消息的特征。然后，我们平均每天随机抽取 100 万封邮件，其中垃圾邮件与火腿的数量比与原始数据相同（即 5% 火腿和 95% 垃圾邮件;目前，我们只考虑“某些火腿”和“某些垃圾邮件”类别中的邮件以获得更准确的基本事实）。只我们的评估基于此采样数据集，而不是第 3 节中的特征分析，因此这些特征的选择不应受到抽样的影响。然后，我们有意抽样等量的垃圾邮件作为火腿数据（每天每个类别中 30，000 封邮件）来训练分类器，因为训练要求每个类具有相同数量的样本。实际上，垃圾邮件量巨大，在进入 SNARE 引擎之前可能会丢弃大量垃圾邮件，因此对垃圾邮件进行抽样进行训练是合理的。

验证 我们使用时间交叉验证评估分类器，这是通过沿时间序列将数据集拆分为子集、在时间窗口中对数据子集进行训练、使用下一个子集进行测试以及将时间窗口向前移动来完成的。此过程重复十次（2007 年 10 月 26 日至 2007 年 11 月 4 日测试），每个子集考虑一天的数据，时间窗口设置为 3 天（表示不需要长周期历史记录）。对于每一轮，我们分别计算检测率和误报率，其中检测率（“真阳性”率）是发现的垃圾邮件与整个垃圾邮件语料库的比率，误报率反映了错误分类的火腿与所有火腿实例的比例。最终评估反映了所有试验的平均值。由于我们用于此实验的高采样率， 我们对几项试验重复了上述实验，以确保各试验的结果一致。如本节中的结果所示，即使分类器仅基于单数据包特征，检测率约为 70%，误报率约为 0.4%。假阳性降至 0.2% 以下，检测率为 70%，因为分类器包含其他功能。尽管这种误报率可能仍然太高，SNARE无法涵盖所有其他垃圾邮件过滤技术，但我们相信其性能可能足以与其他方法结合使用，可能作为早期分类器，或作为传统IP信誉系统（例如SpamHaus）的替代品。

### 4.2 信誉引擎的准确性

在本节中，我们将评估SNARE在三组不同特征上的准确性。令人惊讶的是，我们发现，即使仅依靠单数据包功能，SNARE也可以自动区分垃圾邮件发送者和合法发件人。添加基于单标头或单消息的其他功能，或基于 24 小时历史记录的这些功能的聚合，可进一步提高准确性。

(a) 单个数据包

| | 分类为 垃圾邮件 | 他 |
|---|---|---|
| 垃圾邮件 | **70%** | 30% |
| 他 | **0.44%** | 99.56% |

(b) 单个标头/消息

| | 分类为 垃圾邮件 | 他 |
|---|---|---|
| 垃圾邮件 | **70%** | 30% |
| 他 | **0.29%** | 99.71% |

(c) 24+ 小时历史

| | 分类为 垃圾邮件 | 他 |
|---|---|---|
| 垃圾邮件 | **70%** | 30% |
| 他 | **0.30%** | 99.99% |

表 1：使用规则拟合对使用协变移位的不同特征集的 SNARE 性能。检测率和误报率以粗体显示。（根据今天的 DNSBL [10]，检测率固定为 70% 进行比较）。

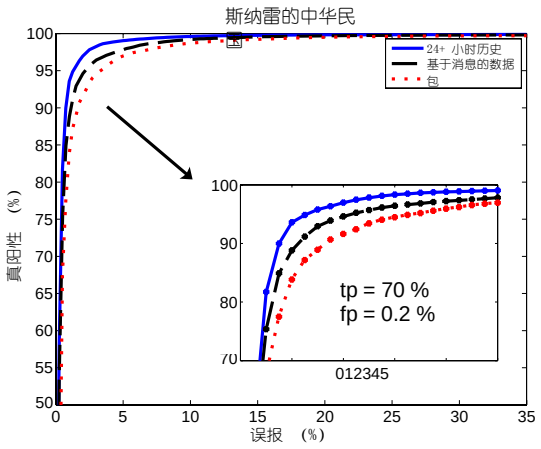



图 8：SNARE 中的 ROC。

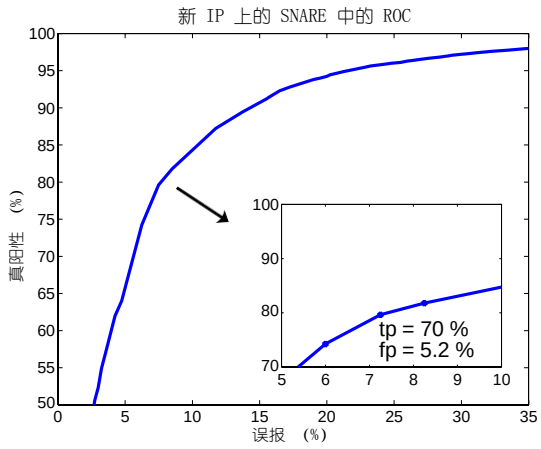



图 9：SNARE 中新 IP 上的 ROC。

#### 4.2.1 单包功能

当邮件服务器收到新的连接请求时，服务器可以仅根据 TCP SYN 数据包向 SNARE 提供发件人和收件人的 IP 地址以及时间戳。回想第 3 节 即使SNARE以前从未见过此IP地址，它仍然可以将此信息与其他电子邮件服务器的近期行为历史记录相结合，并构建以下特征：（1）发件人和收件人之间的测地线距离，（2）日志中到发件人20个最近邻居的平均距离，（3） 请求连接时垃圾邮件与 HAM 的概率比（4） 发件人 IP 的 AS 编号，以及 （5） 发件人上电子邮件服务端口的状态。

为了评估这些功能的有效性，我们对 RuleFit 进行了这些功能的培训。图 8 中的虚线点曲线展示了 SNARE 信誉引擎的 ROC 曲线。fp = 0.2% 和 tp = 70% 统计数据是指具有 24 小时历史记录（实线）的曲线，稍后将讨论。我们检查假阳性给定固定的真阳性，70%。混淆矩阵如表1（a）所示。超过0.44%的合法电子邮件被标记为垃圾邮件。此结果非常重要，因为它依赖于由有限数据量构建的特征并且仅来自候选 IP 的单个 IP 数据包。发件人信誉系统将与包括基于内容的过滤在内的其他技术的组合一起部署。因此，作为第一道防线，该系统将非常有效地消除许多不需要的发件人。事实上，一旦发件人被识别为垃圾邮件发送者，邮件服务器甚至不需要接受连接请求，从而节省了网络带宽和计算资源。我们在下面描述的功能进一步提高了准确性。

**4.2.2 单报头和单邮件功能** 单报件功能允许SNARE在查看邮件头之前快速识别和丢弃垃圾邮件发送者的连接。一旦邮件服务器确认了连接并检查了整个邮件，SNARE就可以通过查看一组附加功能来更有信心地确定发件人信誉。如第 3.2 节所述，这些功能包括收件人数和邮件正文长度。表1（b）显示了我们将单数据包特征（即前一节的特征）与这些附加特征相结合时的预测准确性。如第 3 节的结果所示，添加消息正文长度和

特征集的接收者数量进一步提高了SNARE的检测率和误报率。

值得一提的是，“收件人”字段中列出的收件人数量可能在某种程度上是可以回避的：发件人可以在“抄送”和“密件抄送”字段中列出目标电子邮件地址。此外，如果垃圾邮件发送者始终在“收件人”字段中放置单个收件人地址，则此值将与绝大多数合法邮件相同。由于除了“收件人”字段上的电子邮件地址计数之外，SMTP标头中没有其他字段的日志，因此我们无法评估考虑“抄送”和“密件抄送”标头下列出的收件人数量是否值得。

| 排 | 功能说明 |
|---|---|
| 1 | 发件人 IP 的 AS 编号 |
| 2 | 过去 24 小时内消息长度的平均值 |
| 3 | 到日志中发件人的 20 个最近 IP 邻居的平均距离 |
| 4 | 过去 24 小时内消息长度的标准偏差 |
| 5 | 发件人上电子邮件服务端口的状态 |
| 6 | 发件人和收件人之间的测地线距离 |
| 7 | 收件人数量 |
| 8 | 过去 24 小时内的平均测地线距离 |
| 9 | 过去 24 小时内的平均收件人数量 |
| 10 | 收到邮件时垃圾邮件与火腿的概率比 |
| 11 | 过去 24 小时内收件人编号的标准偏差 |
| 12 | 消息正文的长度 |
| 13 | 过去 24 小时内测地线距离的标准偏差 |

表2：SNARE中特征重要性的排名。

#### 4.2.3 聚合要素

如果来自一个发件人的多条消息可用，则可以计算以下特征：测地线距离、邮件正文长度和收件人数的平均值和方差。我们评估一个分类器，该分类器是根据过去 24 小时内的汇总统计信息以及前几节中的特征训练的。

表 1（c） 显示了具有这些聚合特征的 RuleFit 的性能，ROC 曲线绘制为图 8 中的实线。应用聚合特征可进一步降低错误率：70% 的垃圾邮件被正确识别，而误报率仅为 0.20%。基于内容的过滤对识别垃圾邮件非常有效，但不能满足专业要求。为大型邮件服务器处理大量邮件。RuleFit 的预测阶段更快，其中查询从决策树的根目录遍历到底部标签。鉴于误报率较低，SNARE将是完美的第一道防线，可疑邮件将被丢弃或重新路由到服务器场进行进一步分析。

### 4.3 其他注意事项

检测“新鲜”垃圾邮件发送者 我们检查了数据跟踪，提取了上一个训练窗口中未显示的IP地址，并进一步研究了具有SNARE所有功能的“新鲜”垃圾邮件发送者的检测准确性。如果将真阳性固定为 70%，则误报将增加到 5.2%，如图 9 所示。与图 8 相比，关于新合法用户的决策变得更糟，但仍然可以识别大多数新的垃圾邮件发送者，这验证了 SNARE 能够自动分类“新鲜”垃圾邮件发送者。

单个特征的相对重要性 我们使用 RuleFit 可以评估我们在第 3 节中检查的特征的相对重要性这一事实。表 2 对所有时空要素进行了排名（最重要的要素位于顶部）。前三大功能 — AS num、平均长度和 neig 密度 — 在区分垃圾邮件发送者和良好发件人方面发挥着重要作用。这个结果非常有希望，因为大多数这些功能都是轻量级的：更好的是，在仅从发送方收到一个数据包的情况下，可以计算出这三个功能中的两个。正如我们将在第 6 节中讨论的那样，它们也相对抗逃避。

特征之间的相关性分析 我们使用互信息来研究特征耦合的紧密程度，以及它们可能包含冗余信息的程度。给定两个随机变量，互信息衡量一个变量在了解另一个变量（即它们共享的信息）后减少了多少不确定性。对于离散变量，X 和 Y 的互信息计算如下：$I(X, Y) = \sum_{x,y} p(x, y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$。当对数底 2 为使用时，数量反映了在给定一个变量的情况下可以删除多少位来编码另一个变量。表 3 显示了一天训练数据的特征对之间的相互信息（2007 年 10 月 23 日）。我们不显示其他日期的统计数据，但这些日期的特征反映了相互信息的相似数量。通过将值范围划分为 4，000 个箱（产生良好的离散近似），将具有连续值的特征（例如，发送方和接收方之间的测地线距离）转换为离散变量;我们计算离散概率上的互信息。表中特征的索引与表 2 中的排名相同;基于数据包的功能用黑色圆圈标记。我们还计算每个特征的熵，并将它们显示在表 3 中的索引旁边。互信息的解释仅在单个列或行内是一致的， 因为没有任何公共变量的互信息比较是没有意义的。当然，该表需要额外的分析，但显示了一些有趣的观察结果。排名靠前的特征 AS 编号与其他几个特征共享高互助信息（以粗体显示），尤其是特征 6，发件人和收件人之间的测地线距离。一阶统计的聚合特征-

| | 0 (8.68) | 2 (7.29) | @ (2.42) | 4 (6.92) | 0 (1.20) | 0 (10.5) | 7 (0.46) | 8 (9.29) | 9 (2.98) | ® (4.45) | 11 (3.00) | 12 (6.20) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 (7.29) | **4.04** | | | | | | | | | | | |
| @ (2.42) | **1.64** | 1.18 | | | | | | | | | | |
| 4 (6.92) | **3.87** | 4.79 | 1.23 | | | | | | | | | |
| 0 (1.20) | **0.65** | 0.40 | 0.11 | 0.43 | | | | | | | | |
| 0 (10.5) | **5.20** | 3.42 | 0.88 | 3.20 | 0.35 | | | | | | | |
| 7 (0.46) | **0.11** | 0.08 | 0.02 | 0.08 | 0.004 | 0.15 | | | | | | |
| 8 (9.29) | **5.27** | 5.06 | 1.20 | 4.79 | 0.46 | 5.16 | 0.13 | | | | | |
| 9 (2.98) | **1.54** | 1.95 | 0.53 | 2.03 | 0.09 | 1.17 | 0.10 | 2.08 | | | | |
| ® (4.45) | **0.66** | 0.46 | 0.07 | 0.49 | 0.02 | 0.87 | 0.006 | 0.85 | 0.13 | | | |
| 11 (3.00) | **1.87** | 1.87 | 0.75 | 2.04 | 0.16 | 1.55 | 0.09 | 2.06 | 1.87 | 0.20 | | |
| 12 (6.20) | **2.34** | 2.53 | 0.49 | 2.12 | 0.20 | 2.34 | 0.07 | 2.30 | 0.52 | 0.31 | 0.73 | |
| 13 (8.89) | **4.84** | 4.78 | 1.15 | 4.69 | 0.41 | 4.77 | 0.11 | 6.47 | 1.98 | 0.69 | 2.04 | 2.13 |

表3：SNARE中特征之间的相互信息；基于数据包的特征以黑眼圈中的数字显示。（索引是表 2 中的功能排名。

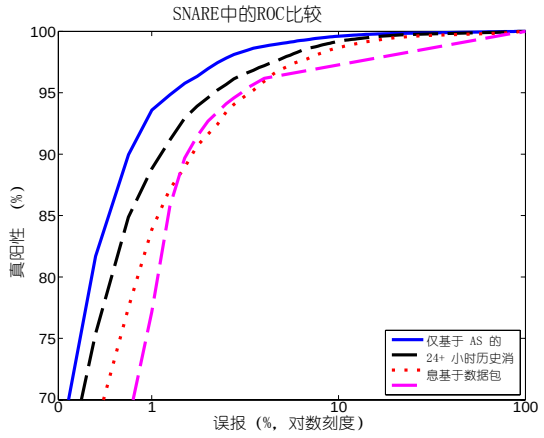



图 10：ROC 与仅 AS 案例的比较。

抽搐（例如，特征 2、4、8）彼此之间也具有较高的值。由于垃圾邮件发送者可能会在每封邮件中表现出一个或多个这些功能，因此随着时间的推移，跨多封邮件聚合这些功能表明，随着时间的推移观察垃圾邮件发送者将揭示其中的许多特征，尽管不一定在任何邮件或单个邮件组上。因此，聚合功能可能会与垃圾邮件发送者常见的其他功能共享高度相互信息。

聚合特征彼此之间具有高度互助信息的一个可能原因是，随着时间的推移，跨多个消息聚合特征会包含 IP 地址的历史记录，该地址可能会随着时间的推移表现出许多这些特征。

仅基于 AS 编号的性能 由于根据 RuleFit 的说法，AS 编号是最具影响力的功能，并且与许多其他功能共享高度互助信息，因此我们调查了仅此功能在多大程度上区分垃圾邮件发送者和合法发件人。我们将 AS 特征馈送到预测模型中，并将 ROC 绘制为图 10 中的下虚线。要制作一个

仔细比较，还显示了“基于数据包”、“基于消息”和“基于历史记录”的 ROC（与图 8 中的相同），并且误报以对数刻度显示。分类器在 70% 的检测率下获得 0.76% 的误报。从表1中回想一下，具有“基于数据包”功能的误报率几乎是一半，即0.44%，而具有“基于历史”功能的误报率将进一步降低到0.20%，这表明其他功能有助于提高性能。我们还注意到，仅使用 AS 号码作为区分功能可能会导致大量合法电子邮件被错误分类，并且如果垃圾邮件发送者决定使用伪造的 ori- 宣布路由，则可能会被规避

gin AS（这是一种易于安装的攻击，并且有些常见）[2，26，39]。

## 5 垃圾邮件过滤系统

本节介绍如何将SNARE的信誉引擎集成到整个垃圾邮件过滤系统中，该系统包括白名单和不断对标记数据（例如，来自垃圾邮件陷阱，用户收件箱等）重新训练分类器的机会。由于SNARE的信誉引擎仍然具有非零误报率，因此我们展示了如何将其与有助于进一步提高其准确性的机制相结合，并且即使在某些误报的情况下也可以防止丢弃合法邮件。我们提出了系统的概述，并评估了这两个功能对整体系统精度的好处。

### 5.1 系统概述

图 11 显示了整个系统框架。系统不需要驻留在单个服务器上。大型公共电子邮件提供商可能会运行自己的 SNARE 实例，因为它们拥有大量的电子邮件数据和处理资源。较小的邮件服务器可能会查询远程

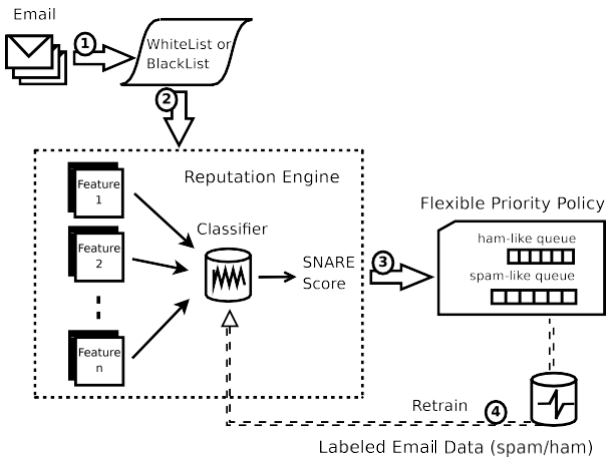



图 11：SNARE 框架。

SNARE服务器。我们设想SNARE可以通过以下方式集成到工作流程中：

1. 电子邮件到达。收到第一个数据包后，邮件服务器向 SNARE 服务器提交查询（仅源和目标 IP）。邮件服务器可以选择在获取SMTP标头或整封邮件后向SNARE发送更多信息。对单个数据包或消息发送查询是在检测准确性和电子邮件处理时间之间进行权衡（即，提前发送请求将使邮件服务器尽早获得响应）。收到的查询中的消息统计信息将用于构建 SNARE 分类器。
2. 白名单。未在白名单中列出的查询将在进行任何垃圾邮件过滤检查或基于内容的分析之前传递到 SNARE 的信誉引擎（在第 2.3 节中提供）。输出是一个分数，默认情况下，正值表示可能是垃圾邮件，负值表示可能是火腿;绝对值表示分类的置信度。管理员可以设置不同的分数阈值，以便在误报和检测率之间进行权衡。我们在 5.2.1 节中评估了白名单的好处。
3. 灰名单和基于内容的检测。一旦信誉引擎计算出分数，电子邮件将被传递到不同的队列中。此时可以应用对资源更敏感且耗时更耗时的检测方法（例如，基于内容的检测）。当邮件服务器能够接收电子邮件时，类 ham-queue 中的邮件具有更高的处理优先级，而类垃圾邮件队列中的邮件将获得较少的资源。此策略允许服务器加速处理 SNARE 归类为垃圾邮件的邮件。这种分层检测方案的优点是合法电子邮件将更快地传递到用户的收件箱。类似垃圾邮件队列中的邮件在最终被丢弃之前可能会分流到资源密集型垃圾邮件过滤器。[3]
4. 再培训 IP 地址是否在该连接中发送垃圾邮件或合法邮件在请求时未知，但在垃圾邮件筛选器处理邮件后已知。SNARE依赖于准确标记的训练数据。电子邮件最终将收到更仔细的检查（如图 11 中的“重新训练”所示）。这些滤波器的结果被视为基本事实，可用作动态调整 SNARE 阈值的反馈。例如，当邮件服务器有备用资源或类似垃圾邮件的队列中的大量电子邮件稍后被认为是合法的时，SNARE系统将被要求采取更慷慨的行动，以将电子邮件评分为可能的火腿;另一方面，如果邮件服务器不堪重负或类似火腿的队列有太多不正确的标签，SNARE将不太可能将电子邮件放入类似火腿的队列中。第 5.2.2 节评估了不同间隔再训练的好处。

## 5.2 评估

在本节中，我们将评估这两个附加功能（白名单和再训练）如何提高 SNARE 的整体准确性。

### 5.2.1 白名单的好处

我们相信白名单可以帮助降低 SNARE 的整体误报率。为了评估这种白名单的影响，我们检查了与误报相关的特征，并确定一天内 SNARE 的所有误报中有 43% 来自仅 10 个 AS。我们检查了不同日期的这一特征，发现任何给定日期的 30% 到 40% 的误报来自前 10 个 AS。然而，不幸的是，这些排名前 10 位的 AS 并不是每天都保持不变，因此白名单可能需要定期重新训练。也可能是除了源的 AS 编号之外的其他特性为白名单提供了更好的机会。我们将细化白名单的细节留给以后的工作。

图 12 显示了当我们将每天对错误分类最多的 ham 负责的前 50 个 AS 列入白名单时的平均 ROC 曲线。这个白名单减少了最好的

[3]尽管 SNARE 的误报率很低，但一些运营商可能认为合法邮件或发件人可能被错误分类的任何非零可能性都需要至少第二次通过更严格的过滤器。

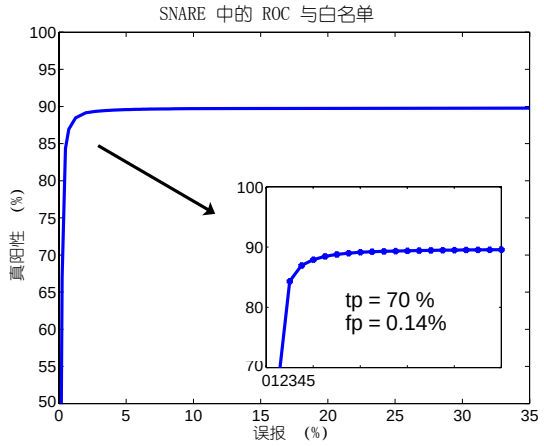



图 12：在 AS 上带有白名单的 SNARE 中的 ROC。

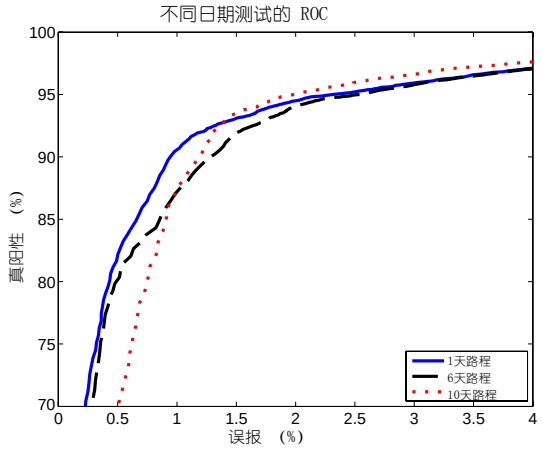



图 13：ROC 使用先前的训练规则对未来的消息进行分类。

可能的检测率相当高（有效是因为大约 11% 的垃圾邮件来自这些 AS）。然而，对于 70% 的检测率，此白名单也将误报率降低到约 0.14%。更激进的白名单或其他功能的白名单可能会导致更低的误报率。

### 5.2.2 再培训的好处

设置 由于电子邮件发件人的行为是动态的，因此使用较早时间段的数据训练 SNARE 最终可能会变得陈旧。为了检查定期重新训练分类器的要求，我们根据前 3 天（截至 2007 年 10 月 23 日至 25 日）的数据训练 SNARE，并在接下来的 10 天进行测试。和以前一样，我们每天使用 100 万条随机抽样的垃圾邮件和非垃圾邮件来测试分类器。

结果 图 13 显示了未来 3 天的假阳性和真阳性，分别是 2007 年 10 月 26 日、10 月 31 日和 11 月 4 日。随着时间的流逝，对未来日子的预测将变得更加不准确。例如，在 11 月 4 日（训练后十天），假设 ROC 曲线上有相同的真阳性，假阳性率下降了。这一结果表明，对于此跟踪中的垃圾邮件发送者行为，每天重新训练 SNARE 的分类算法应该足以保持准确性。 （我们预计重新训练的需求可能因不同的数据集而异。）

## 6 讨论和限制

在本节中，我们将讨论 SNARE 的各个方面，这些方面可能会引起实际问题。我们首先讨论攻击者可能能够规避各种功能的程度，以及这些功能的程度特征可能因时间和数据集而异。然后，我们讨论 SNARE 的生产部署可能存在的可扩展性问题，以及各种可能的解决方法。

### 6.1 逃避抵抗和鲁棒性

在本节中，我们将讨论构成 SNARE 分类算法输入的各种网络级特征的逃避阻力。在某种程度上，这些功能中的每一个都是可以避免的。尽管如此，SNARE 提高了标准，使垃圾邮件发送者在不改变他们用来发送垃圾邮件的技术的情况下更难逃避检测。尽管垃圾邮件发送者可能会适应以规避以下某些功能，但我们认为垃圾邮件发送者很难调整所有功能以通过 SNARE，尤其是在不降低垃圾邮件僵尸网络的有效性的情况下。我们依次调查表 2 中的每个功能。

AS 编号 AS 编号与发件人身份的关联比 IP 地址更持久，原因有二：(1) 垃圾邮件服务器可能设置在特定的 AS 中，而网络管理员并未将其关闭。 (2) Bot 倾向于在 AS 内聚集，因为同一 AS 中的机器很可能具有相同的漏洞。垃圾邮件发送者很难将邮件服务器或机器人大军移动到不同的 AS；因此，AS 编号可用于指示恶意主机。

消息长度 在我们的分析中，我们发现合法电子邮件消息的大小往往比垃圾邮件的多变（可能是因为垃圾邮件发送者经常使用模板发送大量邮件 [25]）。了解此功能后，垃圾邮件发送者可能会开始随机化其电子邮件的长度 -

圣贤；这种攻击并不难发动，但它可能会限制垃圾邮件发送者可以发送的邮件类型，或者使协调具有类似邮件的大规模垃圾邮件活动变得稍微困难一些。

最近邻距离最近邻距离是另一个难以修改的特征。到 k 个最近邻居的距离有效地隔离了一小段 IP 地址中异常大量的电子邮件服务器的存在。如果垃圾邮件发送者试图改变他们的邻居密度，他们将无法在受感染的子网中使用太多机器将垃圾邮件发送到同一组目的地。尽管僵尸网络控制器可以指示同一子网上的僵尸程序以不同的目标集为目标，但这种规避确实需要更多的协调，并且在某些情况下，可能会限制每个垃圾邮件僵尸程序在选择其目标目的地时的灵活性。

电子邮件服务端口的状态 某些限制可能会使主动探测失败，例如，外发邮件服务器使用自己的协议来减轻消息（例如 Google 邮件）或防火墙阻止来自域外的连接。但僵尸不会打开这些端口的概率很高，攻击者需要获得 root 权限才能启用这些端口（这需要更复杂的方法和资源）。基本思想是找出发件人是否是合法的邮件服务器。虽然我们在 SNARE 中使用了主动探测，但其他方法也可以方便测试，例如域名检查或邮件服务器身份验证。

发件人-接收方测地距离 垃圾邮件发送者的物理位置与其目标 IP 位置之间的测地距离分布是垃圾邮件要求在尽可能短的时间内到达尽可能多的目标邮箱的结果。即使在地理分布广泛的大型僵尸网络中，要求每个僵尸网络偏向收件人域以规避此功能也可能会限制僵尸网络用于发送垃圾邮件的方式的灵活性。虽然也可以通过调整每个机器人的收件人域来规避此功能，但如果机器人只向附近的收件人发送垃圾邮件，则僵尸网络的灵活性也会受到一定程度的限制：例如，不可能针对来自完全分布式垃圾邮件僵尸网络的特定区域。

收件人数量 我们发现垃圾邮件的收件人往往多于合法邮件；垃圾邮件发送者可能会通过减少每封邮件的收件人数量来规避此功能，但这可能会降低发送邮件的效率，并且可能会以其他方式改变发件人的行为，从而使垃圾邮件发送者更加引人注目（例如，强迫垃圾邮件发送者以打开更多连接）。

一天中的时间 此功能可能比其他功能更难逃避。话虽如此，垃圾邮件僵尸网络的昼夜模式是由受感染机器打开时产生的。为了让僵尸网络在一天中修改它们的昼夜消息量以匹配合法的消息模式，它们将不得不在晚上降低它们的垃圾邮件量，尤其是下午 3:00 到下午 3:00 之间。和晚上 9:00并减少下午的电子邮件量。这将再次降低僵尸网络发送大量电子邮件的能力。

## 6.2 其他限制

我们简要讨论 SNARE 当前的其他局限性，包括它扩展到大量收件人的能力以及它对发送垃圾邮件和合法邮件的 IP 地址进行分类的能力。

规模 SNARE 最终必须扩展到数千个域并每天处理数亿个电子邮件地址。不幸的是，即使是最先进的机器学习算法也无法很好地处理这么大的数据集；此外，将数据发送给中央协调员进行培训可能会消耗大量带宽。尽管我们的评估表明 SNARE 的分类对训练数据的采样相对稳健，但我们打算进一步研究采样训练数据的最佳方法，或者甚至可能执行网络内分类。

双用途 IP 地址 我们与大型邮件提供商的对话表明，最大的新兴威胁之一是从基于 Web 的电子邮件帐户发送垃圾邮件的“网络机器人”[35]。随着这些类型的攻击的发展，越来越多的垃圾邮件可能会从同时发送大量合法邮件的 IP 地址发送。在这些情况下，IP 地址既不好也不坏，将需要更复杂的分类器和功能，可能涉及基于时间序列的功能。

# 7 相关工作

我们调查了以前关于表征电子邮件发件人、电子邮件发件人信誉系统和其他不基于内容的电子邮件过滤系统的网络级属性和行为的工作。

特征研究 最近的特征研究提供了越来越多的证据表明垃圾邮件发送者具有不同的网络级行为模式。拉马钱德兰等人。 [34] 显示垃圾邮件发送者利用瞬态僵尸网络以低速率从任何特定 IP 向任何域发送垃圾邮件。谢等。 [38] 发现一个巨大的

大多数在动态 IP 地址上运行的邮件服务器仅用于发送垃圾邮件。在他们最近发表的研究 [37] 中，他们展示了一种通过使用从垃圾邮件中的 URL 构建的签名来识别机器人的技术。与 SNARE 不同，他们基于签名的僵尸网络识别在分析消息内容方面有很大不同。其他人也检查了僵尸网络的相关行为，主要是为了表征而不是检测 [25、31]。帕塔克等人。[29] 部署了一个中继坑来收集来自多个发往多个域的垃圾邮件发件人的数据。他们使用这些数据来演示垃圾邮件发送者如何利用受损的中继服务器来逃避检测；这项研究从多个有利的角度审视了垃圾邮件发送者，但主要侧重于描述垃圾邮件发送者的特征，而不是开发新的检测机制。牛等。分析网络垃圾邮件发送者的网络级行为（例如， URL 重定向和“门口”页面）并提议使用基于上下文的分析来防御网络垃圾邮件 [28]。

基于网络级行为的发件人信誉 SpamTracker [34] 与 SNARE 关系最密切；它使用跨多个域聚合的数据中的网络级行为特征来推断发件人的信誉。虽然这项工作引发了行为黑名单的想法，但我们发现了许多其他功能，这些功能比那篇论文中使用的单一功能更轻量级、更不易逃避。 Beverly 和 Sollins 基于传输级特征（例如，往返时间、拥塞窗口）[12] 构建了一个类似的分类器，但他们的分类器都是重量级的，因为它依赖于 SVM，并且还需要接受消息以收集特点。唐等。通过分析查询模式观察到的 IP 地址行为，探索了垃圾邮件发件人的检测 [36]。他们的工作重点是与查询来源相关的消息量的广度和周期性。 之前的各种工作还尝试根据收件人组对电子邮件发件人进行聚类，通常着眼于垃圾邮件过滤 [21、24、27]，这在本质上类似于 SNARE 的测地线距离功能；然而，这些以前的技术通常需要对大量收件人或两者的消息内容进行分析，而 SNARE 可以在更轻量级的功能上运行。 McAfee 的 TrustedSource [4] 和 Cisco IronPort [3] 将垃圾邮件过滤设备部署到成百上千个域，然后向中央服务器查询发件人信誉，并提供有关它们收到的消息的元数据；我们正在与 McAfee 合作，将 SNARE 部署为 TrustedSource 的一部分。

非内容垃圾邮件过滤 Trinity [14] 是一种分布式、无内容的垃圾邮件检测系统，用于检测来自依赖于消息量的僵尸网络的消息。 SpamHINTS 项目 [9] 也有既定目标使用网络流量模式分析而不是邮件内容来过滤垃圾邮件。 Clayton 早期在挤压检测方面的工作涉及监控本地 ISP [16] 和远程 ISP [17] 的服务器日志以检测垃圾邮件发送者。这项工作与我们的目标相似，但所提出的方法更多地关注与仅来自单个发件人的 SMTP 会话相关的属性。

## 8 结论

尽管在基于内容的垃圾邮件过滤方面取得了很大进展，但最先进的发件人信誉系统（例如 DNSBL）相对反应迟钝、不完整且粒度较粗。为了改善这种情况，本文介绍了 SNARE，这是一种发件人信誉系统，可以根据发件人历史早期确定的特征准确自动地对电子邮件发件人进行分类——有时在只看到一个 IP 数据包之后。

未来工作的几个领域仍然存在。也许最未知的领域是使用时间特征来提高准确性。 SNARE 的所有特征本质上都是离散变量，但我们从经验中知道垃圾邮件发送者和合法发件人也表现出不同的时间模式。在 SNARE 的未来版本中，我们的目标是将此类时间特征合并到分类引擎中。另一个需要改进的领域是让 SNARE 更难逃避。尽管我们认为垃圾邮件发送者很难规避 SNARE 的功能并仍然保持有效，但设计更强大的分类器以应对主动规避和错误训练分类器的尝试可能是未来工作的一个有前途的领域。

## 致谢

我们感谢我们的牧羊人 Vern Paxson，他提供了许多有用的建议，包括查看特征之间相互信息的建议以及对分析和表示的其他几项改进。我们还感谢 Wenke Lee、Anirudh Ramachandran 和 Mukarram bin Tariq 对本文提出的有益评论。这项工作由 NSF 职业奖 CNS-0643974 和 NSF 奖 CNS-0716278 和 CNS-0721581 资助。

## 参考文献


[1] FCrDNS 查找测试。 http://ipadmin.junkemailfilter.com/rdns.php。

[2] 互联网警报登记处。 http://iar.cs.unm.edu/.

[3] 铁港。http://www.ironport.com.

[4] 迈克菲安全计算。 http://www.securecomputing.com.

[5] SORBS：垃圾邮件和开放中继阻止系统。 http://www.au.sorbs.net/.

[6] 垃圾邮件警察。http://www.spamcop.net/bl.shtml.

[7] 垃圾邮件豪斯 IP 阻止列表。 http://www.spamhaus.org.

[8] 地理知识产权接口。麦克斯明有限责任公司。 http://www.maxmind.com/app/api, 2007.

[9] 垃圾邮件提示：令人高兴的是，它不一样。 http://www.spamhints.org/, 2007.

[10] DNSBL 资源：统计中心。 http://stats.dnsbl.com/, 2008.

[11] *ALPEROVITCH, D., JUDGE, P., AND KRASSER, S. 电子邮件信誉系统的分类法。首届大规模分布式计算系统（TRAM）信任和声誉管理国际研讨会（2007年）。*

[12] BEVERLY, R. 和 SOLLINS, K. 利用垃圾邮件的传输级特征。第五届电子邮件和反垃圾邮件会议（CEAS）（2008年）。

[13] BOYKIN, P.和ROYCHOWDHURY, V.个人电子邮件网络：一种有效的反垃圾邮件工具。IEEE Computer 38, 4 (2005), 61–68.

[14] 布罗茨基和布罗茨基。一种独立于内容的分布式垃圾邮件检测方法。在理解僵尸网络（HotBots）的热门话题的第一研讨会（2007）。

[15] 伯吉斯，C.关于模式识别的支持向量机的教程。Data Mining and Knowledge Discovery 2, 2 (1998), 121–167.

[16] 克莱顿，R.通过挤压检测阻止垃圾邮件。第一届电子邮件和反垃圾邮件会议（CEAS）（2004年）。

[17] Clayton, R. 通过检查传入的服务器日志来阻止传出垃圾邮件。第二届电子邮件和反垃圾邮件会议（CEAS）（2005年）。

[18] 弗里德曼，J.和波佩斯库，B.梯度定向正则化。斯坦福大学，技术报告（2003年）。

[19] 弗里德曼，J.和波佩斯库，B.通过规则集成进行预测学习。《应用统计年鉴》（待出版）（2008年）。

[20] GOLBECK, J. 和 HENDLER, J. 电子邮件过滤的信誉网络分析。第一届电子邮件和反垃圾邮件会议（CEAS）（2004年）。

[21] GOMES, L. H., CASTRO, F. D. O., ALMEIDA, R. B., BETTENCOURT, L. M. A., ALMEIDA, V. A. F. 和 ALMEIDA, J. M. 改进基于结构相似性的垃圾邮件检测。减少互联网上不需要的流量的步骤研讨会（SRUTI）（2005年）的会议记录。

[22] GOODMAN, J., CORMACK, G. 和 HECKERMAN, D. 垃圾邮件和正在进行的收件箱之战。ACM 50的通信，2（2007），24-33。

[23] *HULTON, E. 和 GOODMAN, J. 垃圾邮件过滤教程。第21届机器学习国际会议（ICML）教程（2004）。*

[24] 约翰森，L.，罗威尔，M.，巴特勒，K.和麦克丹尼尔，P. 向感兴趣的社区发送电子邮件。第四届电子邮件和反垃圾邮件会议（CEAS）（2007年）。

[25] KANICH, C., KREIBICH, C., LEVCHENKO, K., ENRIGHT, B., PAXSON, V., VOELKER, G.M. 和 Savage, S. Spamalytics：垃圾邮件营销转换的实证分析。第 15 届 ACM 计算机和通信安全会议（CCS）会议记录（2008）。

[26] KARLIN, J., FORREST, S.和REXFORD, J.自治系统的自治安全。计算机网络 52, 15 (2008), 2908–2923.

[27] LAM, H., AND YEUNG, D.一种基于社交网络的垃圾邮件检测学习方法。第四届电子邮件和反垃圾邮件会议（CEAS）（2007年）。

[28] NIU, Y., WANG, Y.-M., CHEN, H., MA, M., AND HSU, F.使用基于上下文的分析对论坛垃圾邮件进行定量研究。第 14 届年度网络和分布式系统安全研讨会（NDSS）论文集（2007）。

[29] PATHAK, A., HU, C., Y. AND MAO, Z., M. 从一个独特的有利位置窥视垃圾邮件发送者的行为。在第一届USENIX大规模漏洞利用和紧急威胁研讨会（LEET）（2008年）。

[30] 昆兰，J.决策树的诱导。机器学习 1, 1 (1986), 81–106.

[31] RAJAB, M., ZARFOSS, J., MONROSE, F. 和 TERZIS, A.了解僵尸网络现象的多方面方法。第六届ACM SIGCOMM互联网测量会议（IMC）会议记录（2006）。

[32] 拉马钱德兰，A.，达贡，D.和费姆斯特，N。DNSBL 能否跟上机器人的步伐？第三届电子邮件和反垃圾邮件会议（CEAS）（2006年）。

[33] RAMACHANDRAN, A. 和 FEAMSTER, N. 了解垃圾邮件发送者的网络级行为。ACM SIGCOMM会议记录（2006）。

[34] RAMACHANDRAN, A., FEAMSTER, N.和VEMPALA, S.使用行为黑名单过滤垃圾邮件。ACM计算机和通信安全会议（CCS）（2007）。

[35] 与马克·里舍的私人对话，雅虎邮箱，2008年。

[36] TANG, Y. C., KRASSER, S., JUDGE, P., AND ZHANG, Y.-Q. 使用精细的 SVM 进行快速有效的垃圾邮件 IP 检测，用于对高度不平衡的频谱邮件服务器行为数据进行垃圾邮件过滤。第二届协作计算国际会议（CollaborationCom）（2006年）。

[37] XIE, Y., YU, F., ACHAN, K., PANIGRAHY, R., HULTEN, G. 和 OSIPKOV, I. 垃圾邮件机器人：签名和特征。ACM SIGCOMM（2008）论文集。

[38] XIE, Y., YU, F., ACHAN, K., GILUM, E., GOLDSZMIDT, M., AND WOBBER, T.IP 地址的动态程度。ACM SIGCOMM论文集（2007）。

[39] 赵晓、裴丹、王林、梅西、曼金、吴顺芳、张林BGP 多源 AS（MOAS）冲突分析。第一届ACM SIGCOMM互联网测量研讨会（IMW）会议记录（2001年）。