**论文名字** SiamHaN: IPv6 Address Correlation Attacks on TLS Encrypted Traffic via Siamese Heterogeneous Graph Attention Network

**所属会议** USENIX Security

**发表年份** 2021

**1、研究意义是什么？为什么要研究这个事情？这个研究为什么值得做？**

随着IPv6使用量的增长，人们越来越关注其安全和隐私问题，对IPv6网络的测量和监管的需求逐渐提升。其中，一个特别关注点是用户活动关联。在这种关联方法中，即使用户流量被传输层安全协议 (Transport Layer Security, TLS) 有效加密，网络管理员仍可以识别和跟踪用户。由于不像IPv4网络，IPv6网络很少部署NAT，导致一个IPv6地址通常对应一个单一用户而不是一个用户组，因此针对IPv6的关联方法将可以长期有效地被应用于用户追踪等任务。

**2、这个问题现在是否有人在做，现有研究都是怎么解决的？存在哪些不足？**

IPv6 流量的活动关联可以分为两类——基于地址的关联和流量特征关联：

**基于地址的关联**允许网络管理员依靠 IPv6 地址来关联用户的活动，尤其是当 IPv6 地址处于弱配置时。例如，用户可能会配置一个恒定的接口标识符，通过它管理员可以从多个通信上下文中精确定位用户的活动。然而，为了消除这种地址关联问题以保护用户隐私，RFC 4291标准要求网络运营商将接口标识符视为语义不透明。RFC 4941 标准扩展了无状态地址自动配置 (Stateless Address Auto-Configuration, SLAAC) 以允许 IPv6 用户使用临时地址。这些方法导致客户端地址频繁变换且具有伪随机的接口标识符，因而很难进行有效的追踪。

与基于地址的关联不同，**流量特征关联**通过分析加密流量中的模式将流量与用户的活动相关联。随着TLS的广泛应用，许多方法也已成功在TLS流量下实现用户活动关联任务，例如TLS指纹。虽然该方法在流量与用户的关联方面表现出很高的有效性，但由于贫乏的知识描述和不可靠的相似性学习导致误报，因此使用该技术的方法只能关联选定用户子集的已知用户的流量。

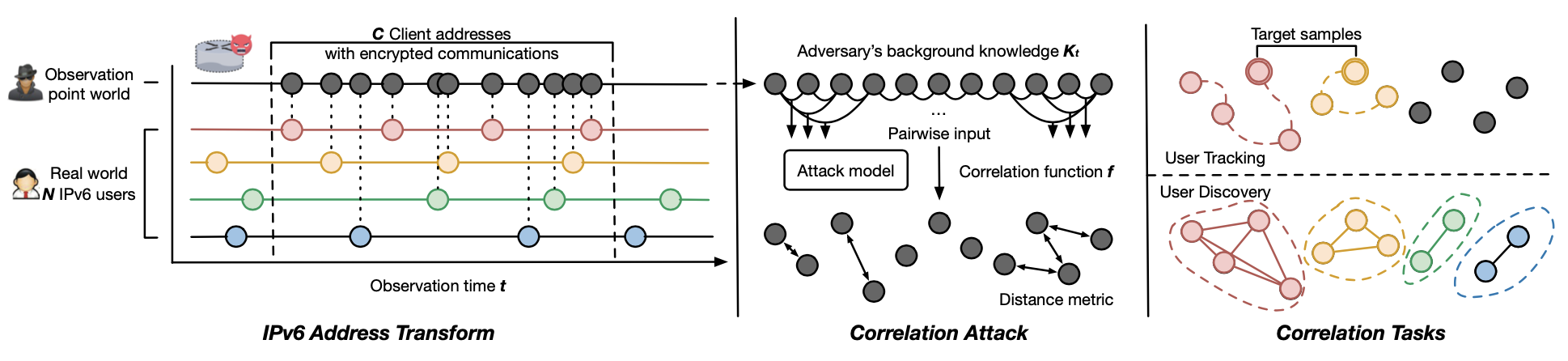
**3、本文的创新是什么？**

提出了一种更先进的方法以克服用户活动关联的诸多限制，从而实现IPv6用户的长期活动关联工作。特别是，论文介绍了一种从 TLS 加密流量中学习关联函数的方法。使用该函数，攻击者可以确定两个任意地址是否属于同一用户。与之前的工作不同，该方法可以实现大规模的用户活动关联。论文提出的方法包括以下两个步骤：首先，攻击者在有利位置监听 TLS 加密流量，然后为每个客户端地址构建知识图。其次，通过利用图神经网络和孪生网络，论文引入了孪生异质图注意力网络，它采用多级注意力和度量学习来捕获两个IPv6地址的加密流量之间的关系。

**4、本文是模型创新？还是测量分析创新？还是应用创新？详细阐述**

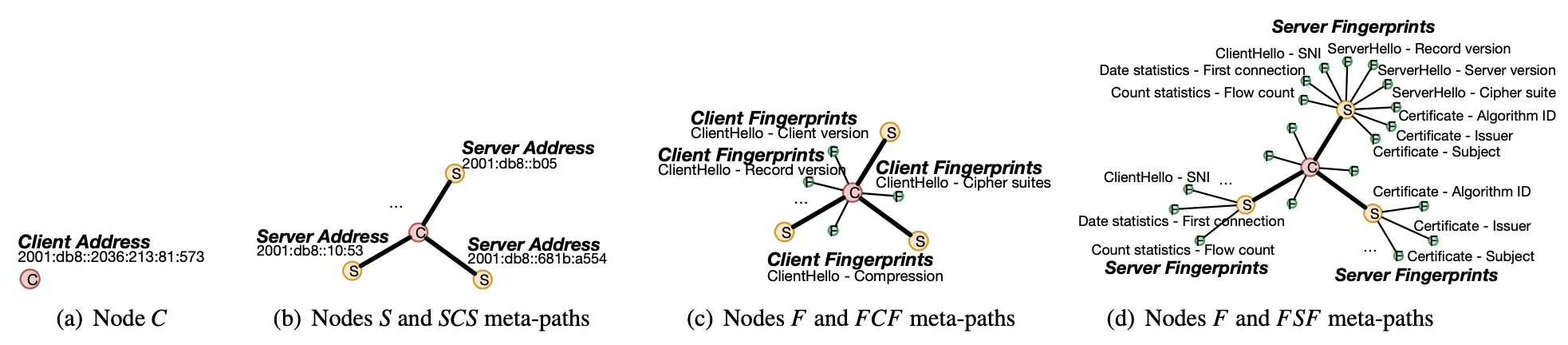
论文既有**应用创新**又有**模型创新**。

论文在IPv6网络客户端地址频繁变更场景下提出了一种新的地址关联手段，从而实现了IPv6用户追踪领域的应用，开辟了全新的IPv6地址变换场景下的地址关联任务。如图1所示，通过利用关联模型，攻击者可以衡量每两个地址间的相似度从而实现IPv6用户追踪与用户发现工作。同时，论文明确了攻击者使用关联攻击带来的严重安全威胁，并进行了对抗实验以阐述可能的防御手段。

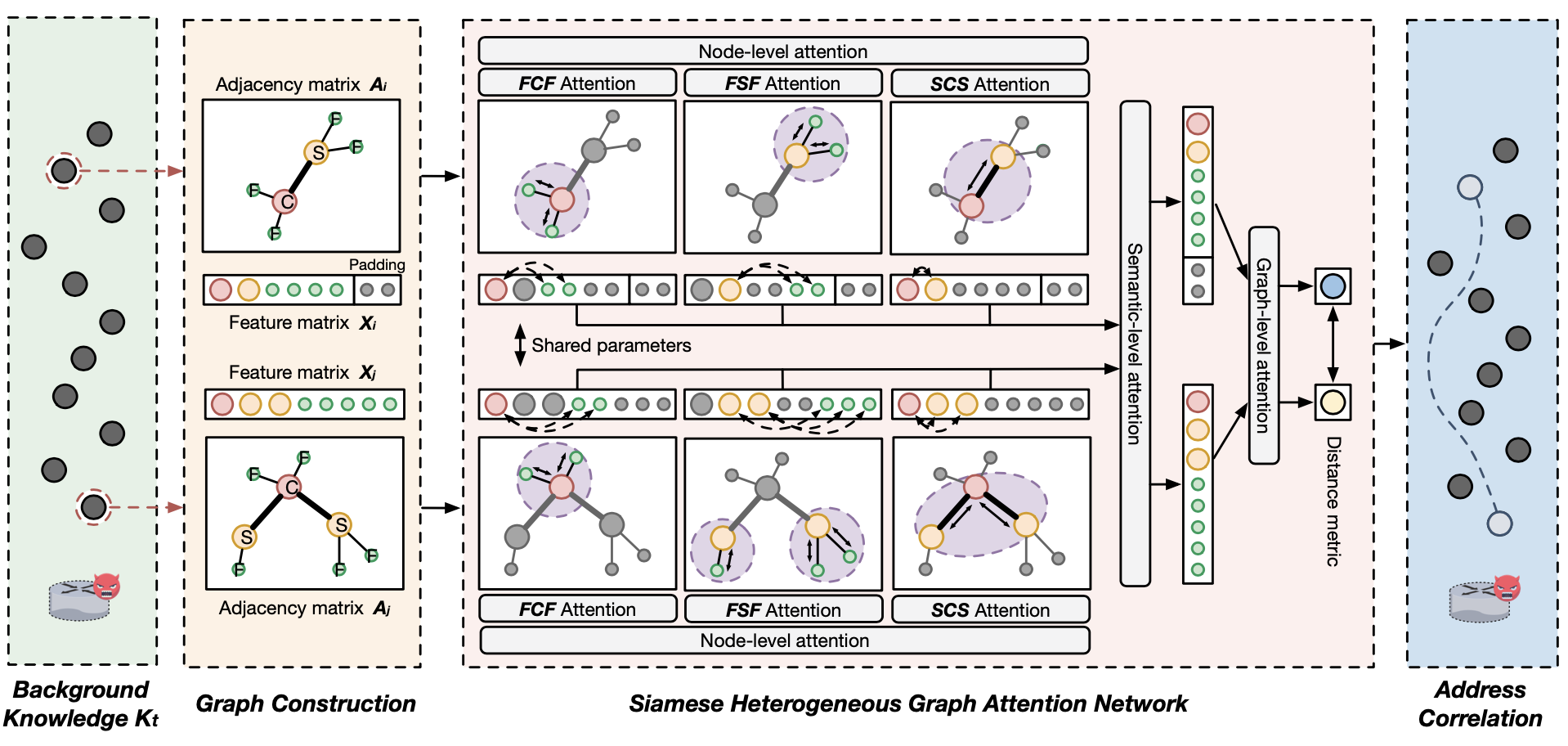


**图1 IPv6地址关联攻击的威胁模型**

论文的模型创新表现在利用**知识图谱**实现了网络安全领域的知识建模，如图2所示，主要通过利用流量中提取的TLS指纹和统计特征形成客户端地址的知识图谱。同时，论文提出了**孪生异质图注意力网络SiamHAN**来学习地址的知识图谱，如图3所示，模型通过多级注意力学习异质图中的实体属性与实体间关系知识并最终形成图嵌入表示，最后通过孪生网络完成相似度学习与预测两地址是否属于同一用户。



**图2 IPv6客户端地址知识图谱构建过程**



**图3 SiamHAN模型架构**

**5、实验结果/测量结果，详细阐述**

**（1）数据集**

论文在中国科技网 (CSTNET) 上被动监测了 2018 年 3 月至 7 月的 IPv6 用户流量，并收集了大量用户数据集用于关联实验。通过利用持久化 HTTP 明文 cookie 来标记泄漏这些 cookie 的地址的 TLS 流量。论文利用用户在流量监控过程中访问部署了 HTTPS 的网站的同时使用相同的地址访问没有部署 HTTPS 的网站的现象，广泛收集HTTP与TLS流量实现了数据标注工作。

**（2）评估标准**

评估指标包括真阳性率、假阳性率、ROC 曲线下面积和准确率。

**（3）对比方法**

（i）**User IP Profiling**。User IP Profiling 是通过使用客户端地址的所有目标 IP 构建用户画像，并使用贝叶斯分类器来识别封闭世界数据集中的已知用户。

（ii）**User SNI Profiling**。User SNI Profiling 是使用来自客户端的所有 TLS ClientHello 消息中的 SNI 作为用户兴趣标识并构建用户画像。

（iii）**Client Fingerprinting**。Client Fingerprinting 是提取 TLS ClientHello 消息的特定字段作为用户的客户端指纹，并利用随机森林来学习任何两个配对指纹的关联性。

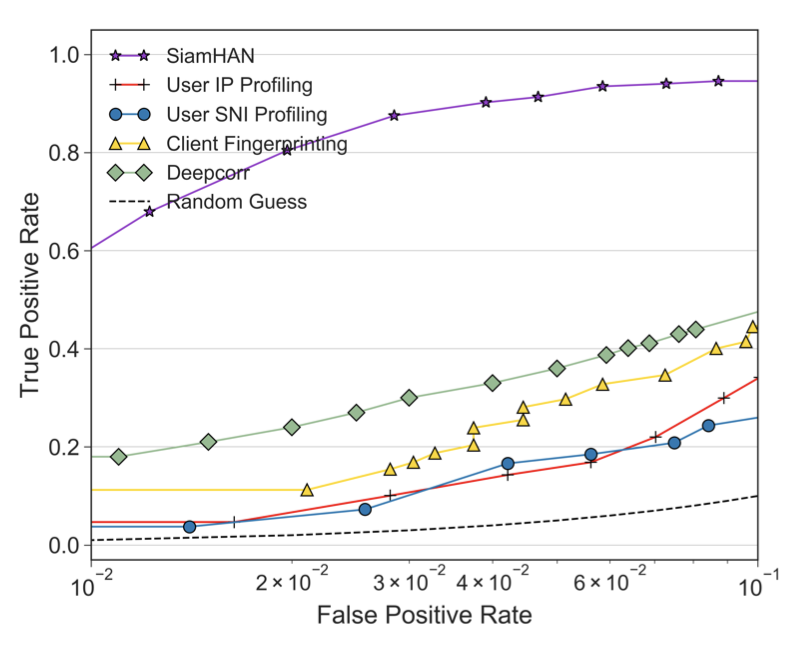
（iv）**Deepcorr**。Deepcorr 使用流序列特征来实现多场景下的关联任务。

**（4）实验结果**

论文包括众多的实验结果以论证地址关联攻击的有效性，这里挑选了几个比较关键的实验结果进行阐述。

（i）地址关联

论文首先测量关联模型在任意地址对上实现的关联性能。在这个实验环境中，攻击者可以基于背景知识对任意成对地址进行关联性测试。论文通过构建训练对样本和测试对样本来综合评估关联模型在成对地址关联任务上的表现。如图4所示，对于目标 FPR = 4 × 10−2，论文方法SiamHAN实现了 0.90 的 TPR，而所有对比方法提供的 TPR 都小于 0.40。



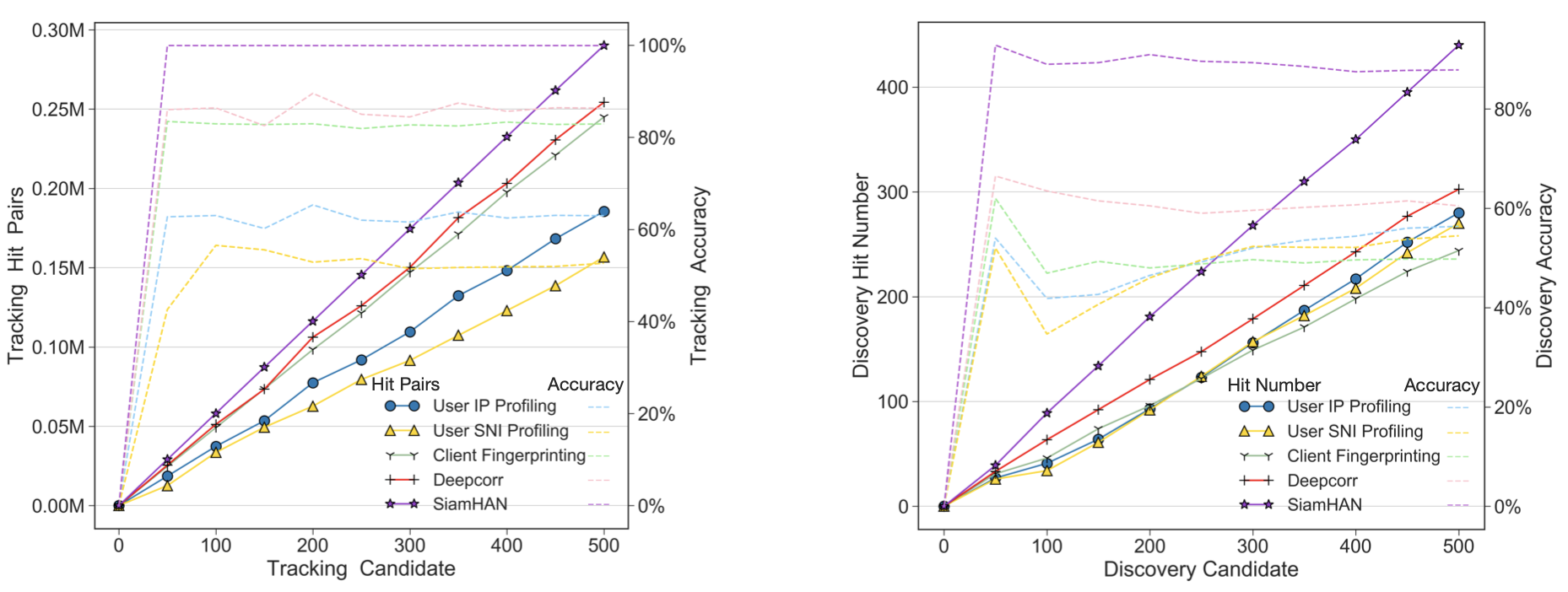
**图4 地址关联任务性能对比**

（ii）用户追踪

获得预先训练好的关联模型后，IPv6 地址关联攻击就可以应用于长期的用户关联任务。论文在基于时间的数据切分训练集上训练关联模型和所有比较方法，并从测试用户中选择目标用户地址以测量测试数据集上的跟踪性能。如图5左所示，SiamHAN可以正确识别出与目标用户样本相关或不相关的地址对，并且是最先进的关联系统 Deepcorr 的 1.10∼1.19 倍。与 Deepcorr 在用户跟踪任务上的 85% 准确率相比，关联模型以 99% 的准确率优于现有的关联技术。

（iii）用户发现

用户发现是关联模型应用的第二项具有挑战性的任务，它可以通过关联不断迭代划分地址集合以发现大规模加密流量上的唯一 IPv6 用户数量。如图5右所示，SiamHAN提供了 88% 的发现准确率，而使用相同设置的最先进系统 Deepcorr 的发现准确度为 60%，本发明的关联模型的发现准确率将是 Deepcorr 的 1.40 ∼ 1.54 倍。SiamHAN的极高准确性确保了在广阔未知流量中发现活跃 IPv6 用户的实用性。



**图5 用户追踪与用户发现性能对比**

**6、文章存在什么不足？**

论文仅是IPv6地址关联与用户追踪探索方向实践的第一步，未来还有许多工作需要实施和完善。（1）论文用于地址关联的用户数据集是十分难获取的，需要依靠有利节点的长期HTTP、TLS流量观测实现工作量庞大。（2）论文实现的关联模型只能每次输入两个地址以判断其是否关联，在海量客户端地址场景下可能会形成难以承受的时间复杂度，尽管在用户追踪场景下需要定向追踪的用户仅为一小部分，但该问题没有得到根本性的解决，仍然需要进一步的工作以降低时间复杂度。

**7、文章给你带来了哪些收获？（数据集、方法启发、写作优点等）**

论文公开了其关联模型源码与部分匿名数据集，为该领域的进一步实验提供了研究基础，对我未来在IPv6领域的研究工作提供了非常大的便利。