HW5

小组成员：

PB20111647 鲍润晖

PB20000103 王炳勋

PB20111704 张宇昂

PB20111651 何泽昊

PB19071508 唐思渝

Paper 1

论文信息

题目：*Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace*

作者：Nicolas Christin

内容简介

本文试图对Silk Road市场（当时著名的非法网上匿名市场，及首个现代意义的黑市）进行数据分析及建模。作者认为这是对其的第一次全面的测量分析。

作者在大约六个月的时间里（2012年2月3日-2012年7月24日）收集了一组受控的测量数据，并对其进行分析。他们设计了一个（简单的）收集方法来获得公开可用的Silk Road的市场数据，分析了超过24,000个项目，以及超过180,000条反馈信息。

详细介绍

Silk Road的运作方式

用户进入Silk Road：买家B访问Tor网络（通过安装tor客户端或使用网络代理），使用伪顶级域名的URL (.onion)，连接Silk Road网站，登录后访问商品列表。存在隐形列表可供买家B选择，它们与Silk Road的其他部分不相连，需要得知URL才能访问。

购买与货物交付：使用BTC(Bitcoin)进行交易。在交易过程中，Silk Road(SR)作为第三方，托管买家B和卖家S之间的BTC，交易链为$B \to SR \to S$。也可选择"tumbler" services：交易链为$B \to I_1 \to \ldots \to I_n \to S$，I为一次性的中间人。

数据收集方法

注册Silk Road账户进行抓取。Silk Road允许通过认证cookie登录，有效期为一周，因此爬取能够绕过验证码。同时定期丢弃线路，避免相同tor节点的过度使用。

收集范围为数据商品页面的数据及买家反馈数据。对于后者，文中采用了两种记录方式：对每个时间间隔抓取网页快照，作为反馈量的下限（可能漏掉快照之后的部分反馈）；每次观察到新反馈后立即记录，作为反馈量的上限（买家对同一购买记录的多次评价会被重复计入）。

测量分析过程

该部分（原文3、4节）是文章的主要部分。

我们将原文的该部分重新分为三部分，分别作为后面提到的分析结论的依据。引号内为文中的子标题。

- "出售的是什么"：描述在Silk Road上销售的物品和卖家群体的特征。
- "谁在销售""买家满意度"：描述了所售物品和卖家人口是如何随时间演变的。
- "BTC的价格变化""交易额""佣金额"：描述作者的测量区间内的销售量。通过强制的买方反馈报告来估算Silk Road上的每日销售金额、并以此推断出Silk Road运营商收取的佣金数额。

本部分将商品、卖家、买家反馈、BTC价格等作为要素进行分析。主要内容为对数据特征的大量描述，浅显且较为冗长，在此选择不表。

测量分析结论

这些表述直接来自原文。

- Silk Road确实主要提供毒品，尽管也有其他物品。它由一个相对国际化的社区。
- 有很大一部分物品在该网站上停留的时间并不长。它有几百个卖家，活跃卖家的数量和销售量都在增加。
- 整个市场每月的销售额略高于120万美元，意味着运营商的佣金约为92,000美元/月。

评价分析

对数据分析的“分析”

统计方法

基于常规的统计方法，测量分析部分绘制了大量图表，大致分为几类：

- 数据特征随时间的变化。该数据特征（即y坐标）为商品/卖家的留存率、BTC价格、月平均交易额等。另有一种变体，即数据量随时间变化的概率，如卖家的存活率，在图中加入95%置信区间。
- 分布图。数据Y（即y坐标）在数据X中的分布及累积分布，如商品数量在各商品种类中的分布、收到的反馈数量在各卖家中的分布等。
- 表格。如商品数最多的20个分类、平台累进抽成比例等。
- 另有一种点状图，同时刻画了每个卖家（即数据点）的存活时间（x坐标）与收到反馈的数量（y坐标），以讨论卖家存活时间与收到反馈的数量的相关性。实际得出r=0.39，相关性较弱。

正如文章第7部分所言，本文的这些研究手段与测量网络犯罪/网上商场的交易的工作（以及毒品政策领域的研究）类似，没有特殊之处。

但是，本文的侧重点有所不同。与测量网络犯罪的工作描述网络攻击的形式相比，本文侧重于对市场的准确描述。利用用户反馈的数据，对网上商场的测量工作用于反映卖家声誉，而本文注重于反映销售量。而从结果上看，Silk Road处于以上两种领域的交叉，统计结果将同时具备两者的特点。

局限性

本文研究较为全面，但仍然对Silk Road买家的活跃程度研究有限，只通过卖家情况与交易情况来侧面反映。例如，对于活跃买家用户数量的变化、地区分布等特征，本文没有展开调查。

这种侧面研究也影响了准确性。以下这段源自原文的表述可作为例子：对商品数的研究只能通过交易量反映，即默认一次交易对应一件商品。当一次交易涵盖多种商品，或商品走隐形列表来完成交易时，这些商品就无法检测和反映。

有其他许多影响测量精度的因素，如在上文的详细介绍-数据收集方法中，买家反馈数据的两种记录方式都有误差。其余的因素在原文6.1中已经具体讨论了。

本文之后Silk Road的历史

作为本文相对次要的部分，作者在6.3讨论了对Silk Road的可能的干预政策：破坏全部tor网络，攻击Bitcoin的金融基础，攻击线下的交易模式，或放任不管，做一些预防毒品的工作。这些讨论没有什么实际价值：都只是一些间接手段，且没有多少可行性。

实际上，这些干预可能根本不存在。尽管文中怀疑"如果没有操作员的出错，是否有可能获得诸如Silk Road这样的隐藏服务的确切位置的确凿证据，仍然是一个有待解决的问题"，但事实上FBI通过许多非人为的漏洞，最终直接取缔了Silk Road。Silk Road在本文初版发表（2012年5月）之后的历史如下：2013年5月，Silk Road被持续的DDoS攻击短暂瘫痪。一名调查员潜入该网站并成为管理员，从而获得了有关网站运作的内部信息。 FBI称通过网站验证码直接泄露的数据发现Silk Road服务器的真实IP地址，并在冰岛查获了一台Silk Road服务器，获得了大量的聊

天记录。最后，创立者 Ulbricht的名字与一个论坛帖子的发帖者建立了联系。他在2013年10月于旧金山公共图书馆格伦公园分馆被联邦调查局逮捕。[3]

但是，与Silk Road类似的网站仍然难以根除。Silk Road被取缔之后一个月，原网站的管理员就创立了Silk Road 2.0，并在一个月后被重新取缔。类似的网站也陆续出现。

Paper 2

论文信息

题目：*Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*

作者：Kai Greshake，Sahar Abdelnabi，Shailesh Mishra，Christoph Endres，Thorsten Holz，Mario Fritz

背景简介

以Chat GPT，Mircrosoft365为例的**LLMs(Large Language Models)**在最近几个月里正越来越多地整合到各类应用中，同时带来了诸多的安全隐患，其中重要的一种是**PI(Prompt Injection)攻击**。

输入到LLM的提示性文字称为**Prompt**，以使得LLM实现功能扩展，因此可视作一种"指令"。但prompt可以是任意一段描述或者要求，因此格式比代码形式的指令更加自由，也能更轻易地对LLM产生影响。（有时，仅仅一句话就能产生可观效果。）而PI攻击就是通过输入恶意的Prompt，对LLM产生负面影响的攻击手段。在此篇文章之前，所讨论的PI攻击是"dierct"的：恶意的Prompt由恶意用户直接输入。

主要贡献

以下是作者在文中声称的研究贡献。

- 引入了有别于PI攻击的、针对LLM的**IPI(Indirect Prompt Injection)攻击**："完全没被研究过的攻击载体"。Prompt不由用户直接输入，对LLM产生影响。
- 首次对与LLM应用中与IPI相关的威胁情况进行了分类和系统分析。
- 针对上述分类法中的每种威胁，展示了这些攻击在现实世界和合成系统中的实际可行性，并提供了开源示例。

贡献分析

以下对上述3点贡献一一进行分析，子标题为每种贡献的简述。对于第一点——IPI概念的提出，放在最后以展示人们IPI攻击的认识的发展。

分类法：

文章第3节基于下图的威胁分类法，详细介绍了IPI的各种威胁。

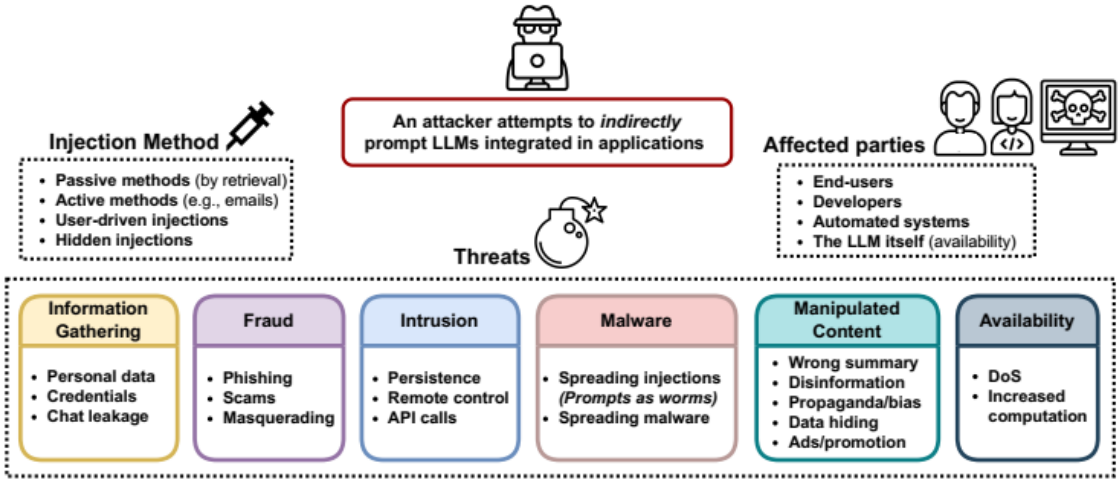

Figure 2: A high-level overview of new indirect prompt injection threats to LLM-integrated applications, how the prompts can be injected, and who can be targeted by these attacks.

Injection method（注入方式）：

- Passive methods：攻击者对用户输入习惯和LLM的检索行为进行推断，将prompt设置在LLM可能达到的检索位置上。随着LLM越来越多地调用API、与其他应用整合在一起，以增强自身的输出能力，恶意的prompt更容易被放置和检索。
- Active methods/User-driven methods：用户无意识复制含有prompt的文字，并输入到LLM中。
- Hidden injections：编码注入及多阶段利用。

Threats（威胁）：

- Information gathering：攻击者通过诱导用户直接/间接泄露数据，或直接对LLM攻击，来收集隐私信息。
- Fraud：利用LLM对用户进行诈骗。
- Malware：利用LLM向用户推荐Malware（恶意软件）。
- Intrusion：集成在系统中的LLM可能成为攻击者攻击系统的媒介。
- Manipulated content：操纵LLM的输出，即向用户展示不客观的信息，达成某种目标（eg:隐藏某些事实、展示广告）。
- Availability：影响LLM的可用性（eg:增加计算时长、破坏API的输入输出）

评价：

该分类法基于威胁而非攻击技术，对攻击的结果进行分类。这样的分类较为全面，且强调攻击的实际影响。这当然也模糊了其他攻击要素，如对具体攻击手段的描述。实际上，罗列的许多威胁之间在攻击方式和对LLM的影响（而非对用户/系统的影响）等方面上有相似之处，因此存在重复讨论的情形。例如，Malware可以视为fraud的特例，都属于利用用户对LLM的信任，向用户进行别有用心的推荐。

基于威胁进行分类的另一个缺点是，无法强调哪些特性是IPI攻击**特有的**。直接的PI攻击仍然可以进行系统侵入，而IPI对用户的攻击理论上也能由用户自身进行（尽管在动机上没有必要）。因此，也许更好的解读方式是将IPI攻击视作PI攻击的子集（而不是强调IPI有别于PI），并将这些威胁视为PI攻击所共有的威胁。

展示威胁的可行性：

研究工具：

- 合成应用：研究者构建了一个聊天应用程序，使用OpenAI的API构建。prompt通过一个名为LangChain的库向LLM发挥作用。应用程序具有搜索、查看当前网站等一系列子工具的接口，利用初始prompt向LLM介绍这些工具的使用方法。
- 攻击对象：bing chat。利用edge浏览器启用侧边栏启用的bing chat的功能，在本地HTML的注释中插入prompt。此外也对代码完成器Github Copilot进行了简单攻击实验。

研究内容：

研究者进行了对各种威胁的演示。关于Intrusion，介绍了远程控制（来自攻击者服务器的命令能够通过LLM传递给系统），或将prompt写入持久存储中的示例。关于Availability，介绍了使LLM静默、功能失效、扰乱搜索结果和搜索查询的示例。而其余威胁的攻击效果较为直观。另外，简要展示了Hidden injections的成功实例。多阶段利用中，微小的injection可能触发LLM自主获取payloads（更大的诱导信息）。

研究展现了IPI攻击的强大能力：LLM经历一次搜索查询，就可能被突破安全防线；攻击者编写简短的prompt，就能产生攻击效果。

作者的讨论：

由于LLM本身存在发展迅速等特点，github上的演示不保证可重复性。研究者对测试结果的评估难以量化，应对措施的研究也尚且欠缺。

评价：

由于沿用了IPI攻击分类法来讨论，这部分演示的叙述也有所侧重。作者主要叙述设计prompt的技巧，进行了大量的prompt内容及会话结果的演示，而隐藏了注入prompt方式的具体细节。例如，在信息收集的介绍中，只提及prompt通过设计的"注入程序"来指示LLM，而攻击方式的展示则留到了4.3小节专门介绍。这样的安排进一步突出了IPI攻击对用户的影响。

本节仍然存在重复讨论的情形：除了入侵和可用性之外的其他四种威胁，其攻击目标彼此类似，即使用prompt引导LLM输出特定的内容，只是这些特定的内容所期望的达成目标不同。实际上如果其中一种威胁被成功验证，其他三种威胁的prompt设计和会话构造便容易效仿，有水字数的嫌疑。

至于示例效果，作者采用概念验证，以避免对实际应用造成影响。实验主要使用了edge浏览器侧边栏的bing chat来测试合成应用和本地HTML的攻击。尽管作者声称这些攻击原则上对实际应用也可行，并拿出twitter上用户的反馈来佐证，但并非所有的威胁都一定实际有效。例如，代码完成的介绍中，作者针对Github Copilot的攻击的手段是在代码注释中隐藏prompt，但这种注入对环境十分敏感，缺乏稳定性。

IPI概念：

本文第一版发表于2023年2月，首次研究了IPI攻击这一攻击载体，将它从PI攻击中区别开来，并进行了详细讨论。在本文之外，2023年5月的文章显示[4]，IPI攻击已经被视为PI攻击的重要的组成部分，结合XSS攻击、检索结果放置攻击等手段。另外，OWASP网站将PI攻击列为LLM的十大弱点之首[5]，对此的描述为"绕过过滤器或使用精心制作的提示来操纵LLM，使模型忽略以前的指令或执行非预期的行动"，实质上也包含了IPI攻击。可见在过去几个月里，IPI攻击的重要性也得到了公认。

参考文献：

1. Traveling the silk road: a measurement analysis of a large anonymous online marketplace
2. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection
3. Wikipedia-Silk Road (marketplace)
4. Prompt Injection Attacks: A New Frontier in Cybersecurity
5. OWASP Top 10 List for Large Language Models version 0.1