doi:10.3969/j.issn.1000-436x.2014.09.001

第三方追踪的安全研究

张玉清，武倩如，刘奇旭，董颖
（中国科学院大学 国家计算机网络入侵防范中心，北京 101408）

摘 要：第三方追踪可以获得用户的浏览历史等隐私信息，如何保护第三方追踪带来的隐私威胁问题成为安全领域的重要研究内容。首先介绍第三方追踪的基本概念、特点及安全现状，并结合第三方应用的类型，总结了第三方追踪存在的隐私威胁。然后从有状态追踪和无状态追踪 2 个方面介绍第三方追踪的技术，并对第三方追踪防御的相关研究进行了分析和比较。最后总结第三方追踪中研究领域的开放性问题和发展方向。

关键词：信息安全；隐私；第三方追踪；Web
中图分类号：TP 393.08 **文献标识码**：A **文章编号**：1000-436X(2014)09-0001-11

Research on security of third-party tracking

ZHANG Yu-qing, WU Qian-ru, LIU Qi-xu, DONG Ying
(National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences, Beijing 101408, China)

Abstract: Third-party tracking can record users' behaviors across many websites. Increasing attention has been paid to the area of third-party tracking due to the fact that user's privacy protection is causing wider concern. The basic concept and present development were introduced. And security risks were also summarized. Then main technology of third-party tracking was intrduced from stateful tracking and stateless tracking aspects. Furthermore, several existing defenses of third-party tracking were described and compared. Finally, some hot research topics were given and the future research directions were discussed.

Key words: information security; privacy; third-party tracking; Web

1 引言

1.1 第三方追踪的概念及特点

早期的 Web 页面是由一个人或者组织设计部署的。随着多样化需求的不断增加，Web 页面中引用越来越多的第三方应用，作为广告、网站分析、社交网络等用途。如图 1 所示，用户主动浏览的网站叫做第一方网站，即在地址栏中所显示的网站（www.nytimes.com/pages/national/index.html）；嵌入在第一方网站中的，与第一方网站不属于同一域或同一个公司的网站叫做第三方应用网站，当用户浏览第一方网站的同时也会向第三方应用发送请求。图 1 中所标识出的广告、视频、社交网络都为第三方应用，其中网站分析类的第三方应用代码在页面中执行，但没有显示。除了为第一方网站提供各种各样的服务，第三方应用还具有如下特点。

1) 第一方网站主动引入第三方应用，默认第三方应用安全可信，并授权第三方应用获得网站的一些信息。

2) 同一个第三方应用会被多个第一方网站所使用。

3) 同一家公司可能拥有多种第三方应用，比如 google analysis、google adsense、doubleclick 等都是 google 公司的产品。

这些特点使得第三方应用在为第一方网站提

收稿日期：2014-07-18；修回日期：2014-08-30
基金项目：国家自然科学基金资助项目（61272481, 61303239）；北京市自然科学基金资助项目（4122089）；国家发展和改革委员会 2011 年信息安全专项基金资助项目（发改办高技[2012]1424）；中国科学院大学校长基金资助项目（Y25102HN00）
Foundation Items: The National Natural Science Foundation of China (61272481, 61303239)**;** The National Natural Science Foundation of Beijing(4122089);The National Information Security Special Projects of National Development and Reform Commission of China [(2012)1424]; The President Fund of GUCAS (Y25102HN00)

图 1 New York Time 网站上的第三方应用

供服务的同时，有能力跨多个网站追踪、记录用户的个人信息及其浏览历史。第三方应用的这种追踪行为叫做第三方追踪。第三方追踪造成的隐私威胁问题日益严重，如何保护第三方追踪带来的隐私威胁问题成为安全领域的重要研究内容。

1.2 第三方追踪安全现状

第三方追踪可以获得用户浏览历史，从而得到用户的位置、兴趣、购买过的商品、就业状况，性取向、财务状况、医疗状况[1,2]等用户的隐私信息，使用户无隐私可言[3]。收集用户的隐私信息并不是假设的猜想，而是实际存在的。下面几个事件，反映了第三方追踪的安全隐患。

1) 2011 年中旬，一家 NAI 成员的广告公司 Epic Marketplace 追踪的用户高度敏感信息数据段被曝光，敏感信息包括：怀孕生育、更年期、修复不良信贷等信息[4]。

2) 2011 年 10 月，一家在线约会网站 OkCupid 被发现向其他商家出售用户喝酒、吸烟、吸毒的频率等信息[5]。

3) Krishnamurthy 等[6]在 10 个流行的健康网站输入查询信息，在其中 9 个网站中发现第三方应用收集用户的查询信息。

4) 2013 年 3 月，中央电视台 3.15 平台曝光几家网络公司利用第三方追踪搜集上亿条 Cookie 信息，使得网民在网络上的行为成为“裸奔”。

2 第三方应用的分类及隐私威胁

本节首先介绍第三方应用的分类及其隐私威胁模型，接着总结有追踪行为的第三方应用的隐私威胁。

2.1 第三方应用的分类

第三方应用通常以 JavaScript 脚本、iframe、Web bug 或媒体等方式存在于第一方网站中[7, 8]。按照所提供服务的内容不同，第三方应用可以划分为 6 类[3]：分析服务（analysis service）、社交网络（social networks）、在线广告（online advertising）、内容提供商（content providers）、前端服务（frontend services）、托管平台（hosting platforms），并不是所有的第三方应用都有追踪行为。分析服务、社交网络、在线广告这 3 类第三方应用通常都有追踪行为，而大部分的内容提供商、前端服务和托管平台类第三方应用并不对用户进行追踪，但由于这些第三方应用都嵌插在第一方网站中，因此，都有能力追踪用户，不能完全排除其存在追踪行为的可能性。下面具体介绍这 6 类第三方应用，并介绍前 3 类第三方应用所采用的典型追踪模型。本节中追踪技术都以 http Cookies 为例，在第 3 节中，将详细介绍其他追踪技术。

1) 分析服务

分析服务类的第三方应用为第一方网站提供统计信息，使第一方网站可以更好地了解其访问者，包括用户的地域、浏览的内容、用户代理信息以及与该网站的互动信息，从而改进网站的设计。虽然分析服务类的第三方应用在实现上有很大的不同，但是几乎所有的第三方应用都采用以下 2 种商业模式之一：一些公司采用付费模式提供服务，

它们声称有合法的权利间接获得用户的分析数据；另外一些公司提供免费的服务，它们使用搜集的用户数据来获利，比如广告定向、市场调查等。

如图 2 所示，以 Google analytics 为例分析类第三方应用典型的追踪模型如下：①用户在地址栏中输入 http://site1.com，浏览器向第一方网站 site1.com 发送请求数据分组；②第一方网站 site1.com 向用户返回 html 文件；③用户浏览器解析 html 文件，得到第三方应用的 url 链接，并向其发送请求；④第三方网站返回请求的文件内容，通常为 JavaScript 脚本文件；⑤脚本执行，读取该第一方网站的 http Cookies，获取用户 ID 等信息；⑥将这些信息作为参数传回第三方应用服务器，第三方应用可能将用户在该第一方网站的访问记录存储在其服务器中。

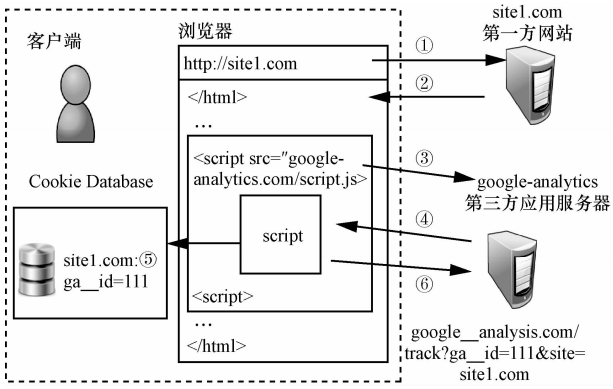


图 2　分析类第三方应用追踪模型

由于第三方应用的脚本在第一方网站中执行，根据同源策略，该脚本只能在第一方网站域下创建 http Cookies 文件，因此该类追踪模型不能跨网站追踪用户，只能追踪记录用户在 site1.com 中的浏览情况。若第三方追踪者想获得某一用户在多个网站中的浏览记录，则需要该用户的个人信息标识(PII, personally identifiable information)，来识别出哪些浏览记录是来自于该用户的。总体来说，该类追踪模型有 2 大特点[8]：①第三方应用的脚本会建立第一方网站域下的 http Cookies，将其所要获得的信息记录在该 http Cookies 中；②读取第一方网站域下 http Cookies 的内容，并将其作为 url 的参数发送给第三方。

2) 在线广告

在线广告网络模型包括 3 大元素[9]：广告发布商（publisher）、广告网络和广告商（advisers）。其主要工作流程如图 3 所示。①广告商们将广告投放到广告网络；②用户浏览广告商的页面；③获取该页面的基本 html 信息；④并向广告网络发送广告请求；⑤广告网络根据跨多个不相关网站追踪到用户的浏览历史，推断用户的喜好，这些兴趣爱好被用于选择广告发送给用户[10,11]。行为定位是广告网络用来增强其广告投放有效性的一种技术，在广告市场中有十分重要的作用。Yan 等[12]证明了行为定位技术可以增强 Bing 搜索引擎中广告的有效性。

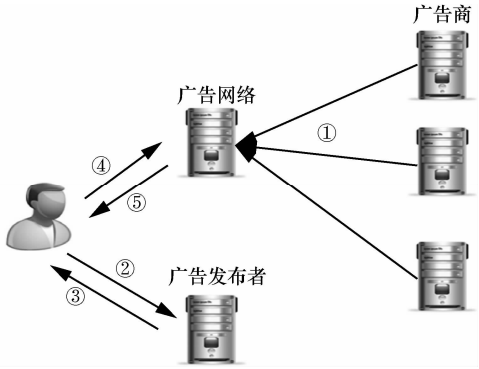


图 3　在线广告工作流程

如图 4 所示，以 Admeld 广告网络为例，分析第三方应用典型的追踪过程如下。①用户在地址栏中输入 http://site1.com，浏览器向第一方网站 site1.com 发送请求数据分组；②第一方网站 site1.com 向用户返回 html 文件；③用户浏览器解析 html 文件，得到广告网络 Admeld.com 的 url 链接，该第三方应用以 iframe 标签的形式插入在第一方网站中，向该 url 发送其请求；④广告网络 Admeld.com 返回网页文件；⑤Admeld.com 读取其域下的 http Cookies，该 http Cookies 存放了用户在多个网站的浏览记录等信息，Admeld.com 分析该用户信息，结合竞价排名等机制，选择适合该用户的广告商 trun.com；⑥向广告商 trun.com 发送广告请求信息和用户的信息，广告商根据用户所访问的网站 site1.com 及用户的信息，选择合适的广告投放到 site1.com 的页面中。

由于第三方应用以 iframe 标签的形式存在，因此该第三方应用可以创建自己域下的 http Cookies，用户浏览包含该第三方应用的第一方网站，其浏览记录都会被该第三方 http Cookies 所记录，因此可以进行跨域追踪。总体来说，该类追踪模型的特点在于[8]：一个第三方应用并不是直接插入在第一方网站中，而是被另一个第三方应用引入的。

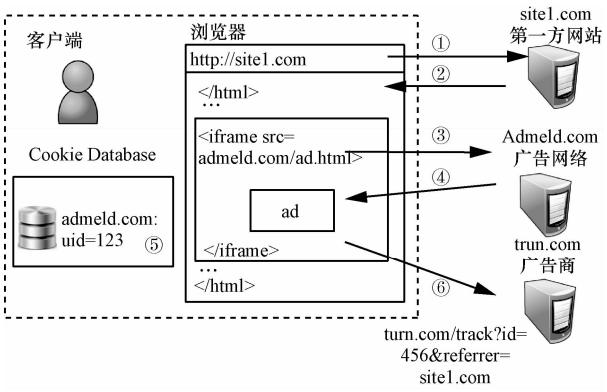


图 4　广告网络类第三方应用追踪模型

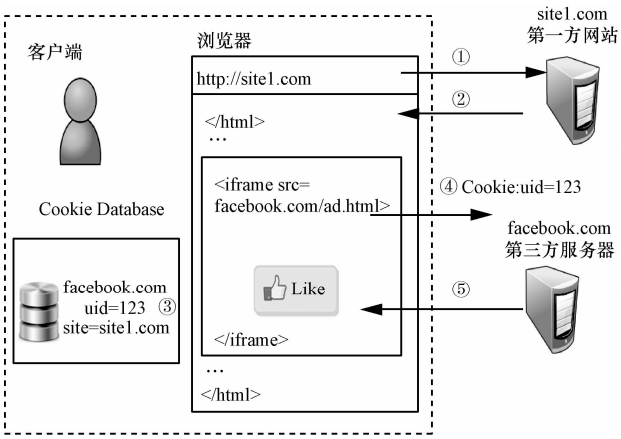


图 5　社交网络类第三方应用追踪模型

3) 社交网络

一些社交网络既是第一方网站又是第三方应用，其角色定义根据不同的场景有所不同：当用户主动浏览社交网络时，该社交网络是第一方网站；当用户浏览其他第一方网站，而社交网络嵌入在该第一方网站时，该社交网络为第三方应用。社交网络类第三方应用为用户提供个性化内容和单点登录服务。社交网络作为第三方应用时，最常见的形式为分享或点赞按钮，如 facebook 的 like 按钮，Google+1 按钮，新浪微博的分享按钮等。这些社交网络类的第三方应用免费为第一方网站提供服务来促进用户的参与和推动市场研究。一些研究[13~15]利用社交网络的数据来提供精准广告定位服务。

如图 5 所示，以 facebook like 按钮为例，用户将 facebook.com 作为第一方网站访问时，在浏览器端建立了 facebook.com 域下的 http Cookies，当用户访问其他包含该社交网站应用的第一方网站时：①用户在地址栏中输入 http://site1.com，浏览器向第一方网站 site1.com 发送请求数据分组；②第一方网站 site1.com 向用户返回 html 文件；③用户浏览器解析 html 文件，得到第三方应用的 url 链接，该第三方应用以 iframe 标签的形式插入在第一方网站，因此可以读写该第三方应用域名下的 http Cookies，在 facebook 的域名下记录浏览历史；④向解析出的 url 发送请求，并将 http Cookies 的信息作为请求分组头部发送给第三方；⑤第三方应用向用户返回 facebook like 按钮。

由于可以在第三方域下建立 http Cookies，因此，该类追踪模型可以跨网站的追踪用户，记录用户在多个第一方网站中的浏览情况。总体来说，该类追踪模型有 2 大特点[8]：①该网站作为第一方网站时，在其域下建立 http Cookies；②作为第三方应用时，利用其域下的 http Cookies 进行追踪。

4) 内容提供商

内容提供商类第三方应用管理视频、地图、新闻、天气、彩票和其他媒体类信息，如 YouTube。一些内容提供商在为第一方网站提供的同时，也通过内置广告来获取利润。

5) 前端服务

前端服务类第三方应用通常为第一方网站提供 JavaScript 类库和 API 等来丰富用户体验，增强网站的功能。例如，Google Libraries API[16]、Google Feed API[17]等。

6) 托管平台

托管平台类第三方应用帮助网站发布者发布自己的内容。常见的有博客平台（如 Wordpress.com[18]）和内容分布网络（如 Akamai[19]）。

7) 其他类

除以上 6 类第三方应用外，还有一类第三方应用并不提供任何服务，它专门用于追踪用户，获得用户浏览历史，分析用户行为特征，并将用户的数据卖给广告公司等。此类公司通常通过付费给第一方网站，并将自己的代码加入到第一方网站中。

Krishnamurthy 等[20]搜集了 2005 年到 2010 年间大约 1 200 个流行网站的页面信息，他们的报告中展现第三方应用的 2 个发展趋势。首先，平均来说，每个网站中引用第三方应用的数目逐年增多。其次，第三方应用公司扩展迅速，一些大的追踪公司，包括 Google、Adobe 和 Microsoft 都通过收购扩展了自己的市场份额。

2.2　第三方追踪的隐私威胁

有追踪行为的第三方应用被称为第三方追踪者。第三方追踪者获得的用户隐私信息主要为浏览历史和个人标识信息[21]2 大类。

1) 浏览历史

用户的浏览历史可以直接泄露用户的个人信息，如用户的位置、兴趣、性取向、财务状况、健康状态等各种高度敏感信息。与此同时，通过检测用户常浏览的页面，可以分析得到许多有关该用户的隐含信息，比如分析其行为习惯等。

当一个第一方页面嵌入第三方追踪者的内容时，该追踪者可以通过 http referrer 头部来获得第一方页面的 url。如果页面中嵌入了第三方追踪者的 JavaScript 代码，追踪者还可以通过执行代码来获得第一方网站的其他信息，如使用 document.title 代码获得页面的标题。

广告公司 Epic Marketplace 将 15 511 条网页的链接存放到一个不可见的 iframe 中，利用 JavaScript 动态加载这些 url 并判断用户是否访问过这些 url，处理过程存放在 http Cookies 中，从而来获得用户的浏览历史[4]。这些 url 除了购物类网站，还包括可以泄露用户敏感行为的网站，如医疗网站、财务网站等。

2) 个人信息标识

一些第三方追踪者获得了许多浏览记录和相关信息后，需要标识出哪些浏览记录是来自于同一个用户的，从而获得该用户尽可能多的浏览历史。可以用来识别出用户的信息叫做个人信息标识，如用户的 ID、用户名等。

一些第一方网站将用户的个人信息标识卖给第三方追踪者，并形成一种商业模式，通常以彩票或者有奖测试的形式存在。一些广告数据提供者（如 DataLlogix）购买用户的标识信息，利用一个用户的标识信息，在其离线数据库中检索出该用户的所有相关信息，并用这些信息为该用户提供广告推荐。

一些第一方网站将用户的个人信息标识无意地提供给第三方追踪者。2011 年，Krishnamurthy 等[6]在 120 个流行的非社交网站中做了一项调查，发现 56%的网站直接泄露用户的个人标识信息给第三方追踪者，如用户的电子邮件地址、姓名、性别等。

3　第三方追踪的技术

第三方追踪技术多种多样，按照是否在本地进行存储可以分为：有状态的追踪和无状态的追踪。

3.1　有状态的追踪

有状态的追踪是指第三方追踪者使用本地的存储机制来记录用户的行为、浏览历史等隐私信息。这些存储机制包括 http Cookies、Flash Cookies、HTML5 Local Storage 等。

1) http Cookies

http Cookies[22] 是存储在用户计算机中的小型文件，用来帮助网站识别用户。2.1 节已结合第三方追踪的隐私威胁模型，详细说明 http Cookies 技术的应用细节。

2) Flash Cookies

Flash Cookies[23, 24]是由 Flash player 控制的客户端共享存储技术，它具备以下特点：①类似 http Cookies，Flash Cookies 利用 SharedObject 类实现本地存储信息，SharedObject 类用于在用户计算机上读取和存储有限的数据量，共享对象提供永久存储在用户计算机上的对象之间的实时数据共享；②本地共享对象是作为一些单独的文件来存储的，文件扩展名为.SOL，尺寸默认为不超过 100 kB，并且不会过期；③本地共享对象并不是基于浏览器的，所以普通的用户不容易删除它们。如果要删掉它们的话，首先要知道这些文件所在的具体位置。这使得本地共享对象能够长时间地保留在本地系统上。

3) HTML5 Local Storage

HTML5 Local Storage[25,26]是 HTML5 提供的 API 接口。通过 JavaScript 代码调用 HTML5 Local Storage API 接口可以在客户端存储较大数据[27]。由于为存储较大数据而设计，广告商和其他第三方应用可以使用 HTML5 Local Storage 来存储用户几个星期甚至几个月的个人信息，这些信息可能包括用户的地理位置、时区、照片、购买记录、电子邮件、浏览历史等。

4) ETag

ETag 是 url 的实体，用于标识 url 对象是否改变。由于 ETag 可以生成唯一标识，即使用户删除了 http Cookies、Flash Cookies 和 HTML5 Local Storage，第三方应用仍然可以利用 ETag 来重建这些被删除的 Cookies 使追踪继续，而用户却并不知情[26]。ETag 追踪技术的最大威胁在于，即使用户使用浏览器的隐私浏览模式仍无法逃避追踪。

目前，大部分的第三方追踪者都使用 http Cookies 窃取用户的隐私记录。Good 等[28]爬取了 Quantcast 排名前 25 000 的网站，发现其中 87%的网站设置了 http Cookies，这些 http Cookies 中有 76%是第三方应用设置的 Cookies。也就是说，当用

表 1 http Cookie、Flash Cookie 和 HTML5 LocalStorage 的关键特征

特征	http Cookies	Flash Cookies	HTML5 Local Storage
存储容量	4 KB	默认 100 KB	默认 5 MB
期限	默认会话结束	默认永久	默认永久
位置	在 SQL 文件中（浏览器）	浏览器外存储	在 SQL 文件中（浏览器）
可用性	只创建它的浏览器可用	同一个机器上的多个浏览器可用	只创建它的浏览器可用

户浏览网站时，大部分的行为活动都会被第三方追踪者获取到。但是 http Cookies 本身有很多局限性：①容易被清除；②每条 http Cookie 记录的信息量小；③可存储的 http Cookies 数量有限。表 1 为 http Cookies、Flash Cookies、HTML5 Local Storage 的性能对比。越来越多的第三方追踪者开始使用 Flash Cookies、HTML5 Local Storage 等技术，甚至使用多种技术相结合来窃取用户的隐私记录[26,29]。例如，ClearSpring、Interclick、Specific Media 等在线广告公司，被发现使用 Flash Cookies 来追踪用户；在 2011 年中旬，Soltani 发现一家提供第三方分析服务应用的公司 KISSmetrics，使用 http Cookies、Flash Cookies、ETag Cookies、HTML5 Local Storage 以及指纹识别技术等相结合，能够在用户删除 Cookies 的情况下，自动重建被删除的 Cookies[26]。

3.2 无状态的追踪

有状态的追踪将用户的信息存储在本地，有可能被用户清除，而无状态的追踪中，第三方追踪者并不在本地的存储和记录用户的行为。其工作原理是通过用户的指纹信息来识别用户，从而获得该用户浏览历史等隐私信息[30, 31]。指纹信息是用户固有的状态特征信息，本不是用户的隐私信息，然而，第三方追踪者可以通过多个指纹信息组合来识别用户，形成个人信息标识，这时，可以形成个人信息标识的指纹信息组合也成为用户的隐私信息。

无状态追踪的典型过程如下：第三方应用A存在于多个第一方网站中；当用户U浏览第一方网站B时，第三方应用A获得该用户的指纹信息，为该用户创建ID，并将用户ID、指纹信息以及网站B的信息记录到服务器的数据库中；当该用户U浏览第一方网站C时，第三方应用A在获得该用户的指纹信息后，与数据库中的用户指纹信息进行对比，若对比成功，则在该用户的浏览历史中增加网站C的信息，否则为该用户在数据库中新建一条记录。从而第三方应用A可以获得其所在第一方网站的所有用户的浏览记录。

如表2所示，指纹信息可以通过2种途径获得。一是通过执行脚本或插件中的代码获得，叫做主动指纹(active fingerprinting)信息[3]。如CPU型号、时区、安装的字体、安装的插件、始终脉冲相位差、可使用的插件、支持的MIME类型、Cookies是否可用等信息；另一种是通过查询网络流获得，叫做被动指纹(passive fingerprinting)信息[3]。如IP地址、语言、http可接受头部等。一些信息既可以通过代码执行获得，也可以通过网络流获得，比如操作系统类型、用户代理等。

表 2 指纹信息

ID	指纹信息	主动指纹信息	被动指纹信息
1	CPU 型号	√	
2	用户代理	√	√
3	时区	√	
4	操作系统	√	√
5	时钟脉冲相位差	√	
6	安装的字体	√	
7	安装的插件	√	
8	可使用的插件	√	
9	支持的 MIME 类型	√	
10	Cookies 是否可使用	√	
11	IP 地址		√
12	语言		√
13	http 可接受头部		√

对于主动指纹信息，Peter Eckersley等[30]在近500 000个浏览器样本中发现有83.6%的浏览器可以被一系列主动指纹信息唯一标识。尽管浏览器的状态会经常改变，如升级等，也可通过比对算法以99%的准确率判断出原状态浏览器和现状态浏览器是否为同一浏览器。一些公司已经使用浏览器指纹识别技术，如BlueCava等[32]。

对于被动指纹信息，Ting-Fang Yen等[33]通过统

计分析大量的网站用户数据，证明在浏览器状态不改变的情况下，被动指纹信息也足够用来识别浏览器，并发现用户即使在浏览器端清除了http Cookies或使用隐私模式浏览，再次访问该网站的时候仍然可以被识别追踪。

对于指纹追踪的防御，Acar 等[34]根据指纹识别技术中“字体”的识别，分析 JavaScript 代码和 flash 脚本，从而编写了一个识别采用这种指纹技术的追踪者的工具。一些用户使用代理服务器或浏览器插件等工具来隐藏指纹信息。然而 Nick Nikiforakis 等[2]通过分析 3 大提供指纹追踪技术的商业广告公司的代码，发现一些公司使用用户安装的代理服务器信息和浏览器插件信息作为指纹信息，可以更准确地识别出用户。

4　第三方追踪防御及有效性

第三方追踪引起的隐私问题受到越来越多的关注，为减缓第三方追踪的隐私威胁，越来越多的组织、研究机构以及 IT 企业投入到第三方追踪的防御工作中。美国和欧盟都针对第三方追踪制定了相关的政策及标准。然而这些政策标准目前都不能有效地防御第三方追踪，因此一些研究机构和 IT 企业也投入到对第三方追踪防御工作的研究中。国内对于第三方追踪的研究很少，国际上对于第三方追踪的防御工作的研究主要分为 2 大方向：1）平衡用户隐私和第三方追踪（特别是广告厂商），既能保护用户的隐私，又可以完成广告的定制推送服务；2）在客户端采取措施，使用户完全控制自己的信息是否可被第三方应用获取。本节首先介绍关于第三方追踪的相关政策，然后从平衡用户隐私和第三方追踪与用户完全控制 2 个方向介绍第三方追踪防御的相关研究和发展。

4.1　相关政策及标准

美国联邦贸易委员会（FTC, federal trade commission）是最主要的联邦消费者保护监管和执法机构。2007 年，几个消费者团体对 FTC 提出倡议，为在线广告创建一个“Do Not Track”列表[35]。2010 年 12 月，FTC 在发布的隐私报告中要求设计一个“Do Not Track”系统可以使用户能控制自己在网络上的隐私信息。2012 年，白宫发布了一份与美国商务部合作的在线隐私报告[36]，该报告提出了一个隐私保护的框架和基本的隐私立法。

欧盟在 2002 的电子隐私指示文件（ePrivacy directive）2002/58/EC 中指出，网站需要赋予用户“选择不同意”（opting-out）的权利，使用户可以选择不允许网站在用户本地浏览器存储信息，然而这项文件并没有起到任何效果[37]。2009 年，欧盟又在其修订文件 2009/136/EC 中出使用“选择同意”(opting-in)原则来替代“选择不同意”原则，大部分的成员国认为“Do Not Track”机制可以满足该指示文件的要求。2012 年，欧盟委员会（European Commission）在欧盟数据保护法[38]（EU data protection law）中增加了一条规定，该规定明确要求非欧盟企业非法追踪欧盟公民将被严厉处理，其罚款金额高达其公司利润的 2%。

W3C（world wide Web）近年来一直在标准化和实现“Do Not Track”机制[29]：当用户提出“Do Not Track”请求时，具有“Do Not Track”功能的浏览器在 http 数据传输中添加一个头信息，这个头信息向第三方应用表明用户不希望被追踪，这样，遵守该规则的第三方应用就不会追踪用户的个人信息来用于更精准的在线广告。目前，几乎所有的主流浏览器都采用了“Do Not Track”，如 IE[39]、Firefox[40]、Google Chrome、Safari[41]等。然而，一些第一方网站可以忽略“Do Not Track”头部继续追踪用户。

4.2　平衡用户隐私和第三方追踪

一些研究希望在保证用户隐私的情况下，同时也保证第三方应用的利益。这一类研究的总体原则为：既保证目前第三方应用可以正常地为用户提供服务，又使得第三方应用无法获得用户真正的隐私信息。目前这些研究主要针对于广告类第三方应用和网站分析类第三方应用。

对于广告类第三方应用，一些研究者试图将广告信息存放在客户端。由于客户端中存放了用户所有的浏览历史，因此可以为用户提供更精准的广告推荐服务。

Toubiana 等[42]提出了一个在不影响用户隐私前提下保证行为广告的使用系统——Adnostic。该系统的工作原理是：广告网络根据少量的用户信息，选出一定数量的广告发给客户端；由于用户的浏览历史等隐私信息存储在浏览器，行为分析过程则在浏览器端完成，并根据行为分析的结果选择推荐广告显示在广告发布商的网站上；对于广告付费的模块，该系统使用高效的加密算法，从而保证用户的隐私安全。然而，该系统存在以下几个问题：1) 由于浏览器中只存储一定数量的广告，广告定位

的精准性可能会受到影响；2) 该系统会增加带宽和延时，削弱用户体验；3) 没有采取匿名化的措施，广告商可以通过用户的指纹信息识别出用户，从而获得用户的浏览历史等隐私信息；4) 并没有解决广告模式中竞价机制。Reznichenko 等[43]提出一种算法，能够在尽量保证用户隐私数据的情况下，允许广告审计方进行广告排名，从而解决 Adnostic 系统中的竞价机制问题。

Guha 等[44]也提出了一个与 Adnostic 类似的系统——Privad，将行为分析和目标广告选择在用户浏览器中执行。不同的是，这个系统更为复杂。为解决 Adnostic 系统中存在的匿名化的问题，该系统在现有的网络框架中的 3 大要素广告发布商，广告网络和广告商的基础上又增加了 2 个要素终裁者（dealer）和监督者（monitor）。然而，该系统中新添加的 2 个元素会改变广告模型的架构，因此不利于整个系统的推广使用。

对于网站分析类的第三方应用，由于第三方应用需要获得用户的一些信息才可以为第一方网站提供分析统计服务。因此广告类的平衡架构不适用于网站分析类的第三方应用。一些研究者试图在用户信息中加入噪音信息来保护用户的隐私。

Akkus 等[45]提出了一种不需要追踪也可以进行 Web 分析的系统。该系统假设网站分析类第三方应用只需要第一方网站用户的统计信息，而非每个用户的个人信息，通过差分隐私算法，为这些统计信息添加噪音，从而保护了用户的隐私行为。然而该系统的限制条件较多，不能真正在实际中使用。

4.3 用户完全控制

目前，所有平衡用户与第三方应用的利益的研究都侧重于某一类型的第三方应用。无论对于单独一种类型的第三方应用，还是对于所有类型的第三方应用，都不存在可以推广使用的系统架构。然而，第三方追踪对用户隐私安全造成的威胁亟待解决，因此一些技术方法将对第三方追踪的防御完全作用在客户端，使用户来决定自己的隐私信息是否可以被第三方应用获取。这些技术可以分为如下几类。

1) 阻止 Cookies 或 Flash Cookies

如第 3 节所述，Cookies 和 Flash Cookies 被用来记录用户的信息。这种方法主要由浏览器和浏览器插件提供，用来阻止第三方应用在用户浏览器中建立 Cookies 或 Flash Cookie。“No Cookie for Google search”[46]是一款浏览器插件，用来阻止 Google 搜索建立的 Cookies，以此防止 Google 追踪用户的搜索记录。BetterPrivacy[47]提供一些方法用户可以处理来自 Google、YouTube、Ebay 等的 Flash Cookies。然而，一些第三方应用仍然可以使用指纹信息来对用户进行追踪。

2) 阻止脚本执行

脚本代码在第三方追踪有着重要的作用：在有状态的追踪技术中，脚本代码可以用来设置 http Cookie、html Local Storage 与 Flash Cookie 进行交互等；在无状态的追踪技术中，脚本代码可以用来获取主动指纹信息。因此一些工具采取阻止脚本执行的方式来防御第三方追踪，如 NoScript[48]。然而，阻止脚本执行的这种保护方法，会使页面没有办法正常加载和工作，影响用户应用体验。

3) Opting out Cookies

Opting out Cookies 是一些网站在用户的浏览器文件夹中创建的 Cookies，当用户浏览这些网站时，网站若检测出用户安装了 Opting out Cookies，则将停止该用户在网站中继续安装 Cookies。Opting out Cookies 用来告诉第三方应用不要在用户的浏览器上安装 Cookies，如 Keep My Opt-Outs[49]和 Targeted Advertsing Cookie Opt-Out[50]都用来提供 Opting out Cookies 的功能，然而，一些第三方应用可以忽略这条规则继续追踪。Leon 等[51]证明了 Opting out Cookies 机制的无效性。

4) 过滤协议头部

通过过滤 http 协议头部中的信息来保护用户的隐私安全，比如，一些工具被用来修改或移除 Referer 头部。然而，http 协议头部的一些信息在网络安全的其他方面有着重要的作用，如 Referer 头部在对抗跨站请求伪造攻击中有着重要的作用，因此移除或修改头部会对其他方面的安全造成影响。与此同时，头部过滤只能保护用户部分隐私信息。

5) 黑名单

一些浏览器插件通过阻止向黑名单中的第三方应用发送请求来防御第三方追踪，如 DoNotTrackMe[52]、Ghostery[53]、Adblock Plus[54]等。当用户浏览第一方网站时，这些浏览器插件将截获并检查数据分组，若存在向黑名单中的第三方应用发送的请求数据分组，则将这些数据分组丢弃，从而黑名单中的第三方应用将无法获得用户的任何信息，有状态的追踪和无状态的追踪都无法实现。目前，这种防御方

式被认为是最有效的防御第三方追踪的措施[3,47,55]然而，黑名单需要人工来建立和维护，且现有黑名单中的有追踪行为的第三方应用的数目很有限，仍然有大量未知的有追踪行为的第三方应用没有被发现。因此只能被防御黑名单中存在的第三方应用。

4.4 小结

综上所述，平衡用户与第三方应用的利益的研究尚处于起步阶段，且现有的作用在客户端的防御措施都不能很好地对抗第三方追踪技术。表 3 列出了现有的防御措施针对第三方追踪技术的有效性[3,6,7]。除了可以防御主动指纹追踪，阻止脚本执行在一定情况下对有状态追踪较为有效：当第三方应用通过执行脚本来获取 http Cookies，Flash Cookies, html Local Storage 信息时，阻止脚本执行可以阻止这些有状态的追踪，而当第三方应用通过 http 头部获取 http Cookies 时，则无法进行防御；阻止 http Cookies 或 Flash Cookies 可以防御有状态的追踪；Opting out Cookies 可以防御 HTTP Cookies；过滤协议头部可以防御某些被动指纹信息被获取；黑名单防御对黑名单中的第三方应用防御有效。从表 3 中可以看出，尚未存在一种有效的防御措施可以完全防御第三方追踪问题。

黑名单作为一种最为有效的防御措施，但防御的有效性取决于黑名单的覆盖率。由于并不是所有的第三方应用都会对用户的行为进行追踪，如何判断一个第三方应用是否为第三方追踪者需要专家进行多方面的判断。因此，现有的工具中，黑名单大多通过人工建立和维护。人工建立和维护需要的工作量和资源较大，第三方追踪者的数量不断增多，黑名单需要定期的维护和更新，如何准确且自动化地获取黑名单成为亟待解决的问题。目前，通过机器学习的方法，根据现有的黑名单中追踪者的特征建立分类器。该分类器可以准确提取出使用脚本追踪的第三方应用的名单。

5 思考与讨论

随着 Web 多样性的发展，第三方应用被越来越多的第一方网站使用，第三方追踪引起的隐私问题受到越来越多的组织、研究机构以及 IT 企业的关注。越来越多的新技术用于第三方追踪，因此第三方追踪的防御工作也面临着更多的挑战。

从第三方追踪技术方面考虑，追踪技术的发展趋势如下。

1) 多种有状态追踪技术配合使用[25,26]。当用户删除一种追踪信息后，其他追踪信息会立刻进行复制恢复。比如，当用户删除 http Cookies 后，存放在 Flash Cookies 中的数据会恢复 http Cookies 的值。

2) 着重发展无状态追踪技术[30,32,33]。相对于有状态的追踪技术，无状态追踪技术更难防御。现有的许多防御措施都针对于有状态的追踪技术。因此，越来越多的第三方应用采用无状态追踪技术来进行追踪。

3) 无状态追踪技术与有状态追踪技术结合使用[26,29,32]。无状态追踪技术与有状态追踪技术各有其优势。有状态追踪更易追踪用户，直接且准确地获取用户的信息，且数据存放在客户端，减少服务器负载。但有状态的追踪技术易于防御，只需在客户端定期删除 http Cookies、Flash Cookies 或 HTML5 Local Storage 中的信息记录即可；无状态追踪技术获取信息后，还需要指纹识别算法才能得到用户的浏览历史，且存在一定误差。然而，无状态的追踪技术难以防御。因此，多种无状态追踪技术与有状态追踪技术的结合使用，将成为未来追踪技术发展的主要方向。

从第三方追踪防御方面考虑，第三方追踪的防御研究重点如下。

1) 平衡用户的隐私和第三方追踪，既能保护用户的隐私又能保证第三方应用的利益。现有的研究

表 3　现有的防御措施的有效性

ID	防御措施	有状态的追踪			无状态的追踪	
		http Cookies	Flash Cookies	HTML5 Local Storage	主动指纹	被动指纹
1	阻止脚本执行	√	√	√	√	
2	阻止 Cookies 或 Flash Cookies	√	√			
3	Opting out Cookies	√				
4	过滤协议头部					√
5	黑名单	√	√	√	√	√

仍处于初级阶段[42~45]：只针对于广告或网站分析类的第三方应用进行了研究，且假设条件过多，需要改变现有的商业模式，无法真正地扩展应用。这方面的研究，还有许多问题需要克服。

2) 无状态追踪的防御[34]。无状态追踪技术发展迅速，哪些指纹信息可以被用来识别用户，第三方应用是否使用了无状态追踪都很难确认，从而防御工作很难进行。如何对无状态追踪进行防御，值得深入研究。

3) 黑名单的自动化获取。黑名单作为目前最为有效的防御方法[55]，如何准确且自动化地获取黑名单成为亟待解决的问题。目前，这方面的工作仍然很缺乏，值得深入研究。

6 结束语

本文首先介绍了第三方应用的类型及发展趋势，越来越多的网站使用第三方应用，平均每个网站中引用第三方应用的数目也逐年增多。接着介绍了现有的第三方追踪技术。按照是否在本地进行存储可以分为有状态的追踪和无状态的追踪。有状态的追踪技术包括：http Cookies、Flash Cookies 和 HTML5 Local Storage 等。相对于有状态的追踪技术，无状态的追踪技术采用指纹信息来识别用户，获取用户的浏览历史，因此更难防御。现有的防御工作主要分为 2 个方向。一是如何平衡用户隐私与第三方追踪，既能保护用户的隐私又能保证第三方应用的利益。这个方向的研究目前仍处于初级阶段，然而，第三方追踪对用户隐私安全造成的威胁亟待解决。因此另一种防御工作的方向将对第三方追踪的防御完全作用在客户端，以保证用户隐私为首要目的。在现有防御方法中黑名单最为有效，如何准确且自动化地获取黑名单成为亟待解决的问题。

参考文献：

[1] LIBERT T. Privacy implications of health information seeking on the Web[EB/OL].http://papers.ssm.com/sol3/papers.cfm?abstractid=2423006.2014.

[2] NIKIFORAKIS N, KAPRAVELOS A, JOOSEN W, *et al*. Cookieless monster: exploring the ecosystem of Web-based device fingerprinting[A]. 2013 IEEE Symposium on Security and Privacy[C]. 2013.541-555.

[3] MAYER J R, MITCHELL J C. Third-party Web tracking: policy and technology[A]. IEEE Symposium on Security and Privacy (SP)[C]. 2012.413-427.

[4] MAYER J. Tracking the trackers: to catch a history thief[EB/OL]. http://cyberlaw.stanford.edu/node/6695, 2011.

[5] MAYER J. Tracking the trackers: where everybody knows your username[EB/OL].http://cyberlaw.stanford.edu/blog/2011/10/tracking-trackers-where-everybody-knows-your-username, 2011.

[6] KRISHNAMURTHY B, NARYSHKIN K, WILLS C. Privacy leakage vs. protection measures: the growing disconnect[A]. Web 2.0 Security and Privacy Workshop[C]. 2011.1-10.

[7] MALANDRINO D, SCARANO V. Privacy leakage on the Web: diffusion and countermeasures[J]. Computer Networks, 2013,57: 2833-2855.

[8] ROESNER F, KOHNO T, WETHERALL D. Detecting and defending against third-party tracking on the Web[A]. Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation[C]. 2012.12-12.

[9] LI Z, ZHANG K, XIE Y, *et al*. Knowing your enemy: understanding and detecting malicious web advertising[A]. Proceedings of the 2012 ACM Conference on Computer and Communications Security[C]. 2012.674-686.

[10] BARFORD P, CANADI I, KRUSHEVSKAJA D, *et al*. Adscape: harvesting and analyzing online display ads[A]. Proceedings of the 23rd International Conference on World Wide Web[C]. 2014.597-608.

[11] SMIT E G, VAN N G, VOORVELD H A M. Understanding online behavioural advertising: user knowledge, privacy concerns and online coping behaviour in Europe[J]. Computers in Human Behavior, 2014, 32: 15-22.

[12] YAN J, LIU N, WANG G, *et al*. How much can behavioral targeting help online advertising[A]. Proceedings of the 18th International Conference on World Wide Web[C]. 2009.261-270.

[13] KOROLOVA A. Privacy violations using microtargeted ads: a case study[A]. IEEE International Conference on Data Mining Workshops (ICDMW)[C]. 2010.474-482.

[14] STUTZMAN F, GROSS R, ACQUISTI A. Silent listeners: the evolution of privacy and disclosure on facebook[J]. Journal of Privacy and Confidentiality, 2013,4(2):2.

[15] RADER E. Awareness of behavioral tracking and information privacy concern in facebook and Google[A]. Symposium on Usable Privacy and Security (SOUPS)[C]. 2014.

[16] Google Libraries API [EB/OL]. https://developers.google.com/speed/libraries/?hl=zh-CN.2014.

[17] Google Feed API[EB/OL]. https://developers.google.com/feed/? hl=zh-cn, 2014.

[18] Wordpress[EB/OL]. https://wordpress.com/, 2014.

[19] Akamai[EB/OL]. http://www.akamai.cn/enzs/, 2014.

[20] KRISHNAMURTHY B. Privacy leakage on the Internet[EB/OL]. http://www.ietf.org/proceedings/77/slides/plenaryt-5.pdf, 2010.

[21] KRISHNAMURTHY B, WILLS C E. On the leakage of personally identifiable information via online social networks[A]. Proceedings of the 2nd ACM workshop on Online social networks[C]. 2009.7-12.

[22] HTTP cookie [EB/OL]. http://en.wikipedia.org/wiki/HTTP_cookie. 2014.

[23] SOLTANI A, CANTY S, MAYO Q, *et al*. Flash Cookies and privacy[A]. AAAI Spring Symposium: Intelligent Information Privacy Management[C]. 2010.

[24] COSTANTE E, DEN HARTOG J, PETKOVIĆ M. What Web Sites Know About You Data Privacy Management and Autonomous Spontaneous Security[M]. Springer Berlin Heidelberg, 2013.146-159.

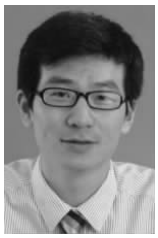
[25] PRINCE J D. HTML5: not just a substitute for flash[J]. Journal of

Electronic Resources in Medical Libraries, 2013, 10(2):108-112.
[26] AYENSON M, WAMBACH D J, SOLTANI A, *et al*. Flash cookies and privacy II: now with HTML5 and etag respawning[EBOL]. http://ssrn.com/abstract=1898390.2011.
[27] LAWSON B, SHARP R. Introducing html5[M]. New Riders, 2011.
[28] CHRIS J, HOOFNAGLE N G. The Web privacy census[EB/OL]. http://www.law.berkeley.edu/privacycensus.htm, 2012.
[29] RUIZ-MARTÍNEZ A. A survey on solutions and main free tools for privacy enhancing Web communications[J]. Journal of Network and Computer Applications, 2012,35(5):1473-1492.
[30] ECKERSLEY P. How unique is your Web browser?[A]. Privacy Enhancing Technologies[C]. 2010.1-18.
[31] CARRASCAL J P, RIEDERER C, ERRAMILLI V, *et al*. Your browsing behavior for a big mac: Economics of personal information online[A]. Proceedings of the 22nd international conference on World Wide Web[C]. 2013.189-200.
[32] BlueCava [EB/OL]. http://www.bluecava.com/, 2014.
[33] YEN T F, XIE Y, YU F, *et al*. Host fingerprinting and tracking on the web: Privacy and security implications[A]. Proceedings of NDSS[C]. 2012.
[34] ACAR G, JUAREZ M, NIKIFORAKIS N, *et al*. FPDetective: Dusting the web for fingerprinters[A]. Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security[C]. 2013.1129-1140.
[35] The history of the do not track header[EB/OL].https://www.cdt.org/privacy/20071031consumerprotectionsbehavioral.pdf, 2007.
[36] Consumer data privacy in a networked world[EB/OL]. http://whitehouse.gov/sites/default/files/privacy-final.pdf, 2012.
[37] Letter to the online advertising industry [EB/OL]. http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2011/20110803_letter_to_oba_annexes.pdf, 2011.
[38] Commission proposes a comprehensive reform of the data protection rules[EB/OL].http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm, 2012.
[39] IE9 and Privacy: Introducing Tracking Protection[EB/OL]. http://blogs.msdn.com/b/ie/archive/2010/12/07/ie9-and-privacy-introducing-tracking-protection-v8.aspx,2010.
[40] Web tool on firefox to deter tracking[EB/OL]. http://allthingsd.com/20110124/web-tool-on-firefox-to-deter-tracking/,2011.
[41] WINGFIELD N. Apple adds do-not-track tool to new browser [EB/OL]. http://online.wsj.com/news/articles/SB10001424052748703551304576261272308358858, 2011.
[42] TOUBIANA V, NARAYANAN A, BONEH D, *et al*. Adnostic: privacy preserving targeted advertising[A]. NDSS[C]. 2010.
[43] REZNICHENKO A, GUHA S, FRANCIS P. Auctions in do-not-track compliant internet advertising[A]. Proceedings of the 18th ACM Conference on Computer and Communications Security[C]. 2011,667-676.
[44] GUHA S, CHENG B, FRANCIS P. Privad: practical privacy in online advertising[A]. Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation[C]. 2011.169-182.
[45] AKKUS I E, CHEN R, HARDT M, *et al*. Non-tracking web analytics[A]. Proceedings of the 2012 ACM Conference on Computer and Communications Security[C]. 2012.687-698.
[46] No cookie for Google search [EB/OL]. https://addons.mozilla.org/zh-cn/firefox/addon/no-cookie-for-google-search/, 2014.
[47] BetterPrivacy[EB/OL].https://addons.mozilla.org/zh-cn/firefox/addon/betterprivacy/, 2014.
[48] NoScript[EB/OL].https://addons.mozilla.org/zh-cn/firefox/addon/noscript/, 2014.
[49] Keep My Opt-Outs[EB/OL]. https://chrome.google.com/webstore/detail/keep-my-opt-outs/hhnjdplhmcnkiecampfdgfjilccfpfoe, 2014.
[50] Targeted advertising Cookie Opt-Out [EB/OL]. https://addons.mozilla.org/zh-cn/firefox/addon/targeted-advertising-cookie-op/,2014.
[51] LEON P, UR B, SHAY R, *et al*. Why Johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising[A]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems[C]. 2012.589-598.
[52] DoNotTrackMe: online privacy protection[EB/OL].https://addons.mozilla.org/zh-cn/firefox/addon/donottrackplus/,2014.
[53] Ghostery[EB/OL].https://addons.mozilla.org/zh-cn/firefox/addon/ghostery/, 2014.
[54] Adblock plus[EB/OL]. https://addons.mozilla.org/zh-cn/firefox/addon/adblock-plus/, 2014.
[55] BAU J, MAYER J, PASKOV H, *et al*. A promising direction for Web tracking countermeasures[A]. Web 2.0 Security & Privacy[C]. 2013.

作者简介：

张玉清（1966-），男，陕西宝鸡人，博士，中国科学院大学教授、博士生导师，主要研究方向为网络与信息系统安全。

武倩如（1988-），女，满族，辽宁凤城人，中国科学院大学博士生，主要研究方向为 Web 隐私安全和第三方追踪。



刘奇旭（1984-），男，江苏徐州人，博士，中国科学院大学讲师，主要研究方向为网络与信息系统安全，包括漏洞挖掘、漏洞评估、漏洞管理、应急响应等。

董颖（1991-），女，陕西渭南人，中国科学院大学硕士生，主要研究方向为信息安全。