doi:10.11959/j.issn.1000-436x.2015144

分享式 Spam 攻击的轻量级检测方案

吕少卿[1]，范丹[2,3]，张玉清[1,2]

（1. 西安电子科技大学 综合业务网理论及关键技术国家重点实验室，陕西 西安 710071；
2. 中国科学院大学 国家计算机网络入侵防范中心，北京 100190；
3. 中国科学院信息工程研究所 信息安全国家重点实验室，北京 100093）

摘 要：Spam 攻击是针对社交网络最主要的攻击方式，分享式 Spam 攻击具有 Spam 内容的存储与传播分离的新特性，目前没有有效的检测方案。针对这一问题分析了其攻击过程和特征，利用分享式 Spam 攻击传播和存储的特征设计了轻量级迭代检测算法 LIDA，通过目标筛选和内容检测 2 个步骤实现对分享式 Spam 的检测。同时，轻量级算法避免了传统算法对每个用户都做深度检测的问题，更具实用性。通过人人网的 4 次迭代实验，共检测到 9 568 个 Spam 账号、30 732 个 Spam 相册以及 2 626 780 条 Spam URL，表明所提的检测算法对于分享式 Spam 攻击是行之有效的。

关键词：社交网络；分享式 Spam；Spam 检测；人人网

中图分类号：TP393.08 **文献标识码**：A

Lightweight detection system of shared spam attacks

LV Shao-qing[1], FAN Dan[2,3], ZHANG Yu-qing[1,2]

（1. Information Security Research Center of State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China;
2. National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences, Beijing 100190, China;
3. State Key Laboratory of Information Security Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

Abstract: Spam is one of the most serious attacks against online social networks (OSN). Recently a new type of spam attack occurs, which named shared spam attack. The shared spam attack can separate the storage and dissemination of spam content, making the existing detection systems no longer effective. To address this problem, an empirical analysis of the process and properties of this new spam attack is performed. A novel lightweight iterative detection algorithm (LIDA) is proposed to detect spam accounts in OSN with these properties. LIDA contains two steps: target filter and content detection. It also noteworthy that LIDA is a lightweight algorithm to infer more spam accounts by exploiting spam accounts' sharing instead of scanning or analyzing all accounts. Experimental results in RenRen, which has successfully detected 9 568 spam accounts, 30 732 spam albums and 2 626 780 spam URL in four round iterations, indicate that LIDA is effective and efficiency in detecting shared spam accounts.

Key words: online social networks; Spam based on sharing; Spam detection; RenRen

1 引言

随着社交网络的快速发展，越来越多的互联网用户通过社交平台进行交流沟通，以 Facebook 为例，2014 年月活跃用户数达到 12.8 亿[1]；国内人人网 2012 年用户数也已达到 2.2 亿[2]。这些社交平台已经深入地影响了人们的生活、工作、学习以及交流的方式。

社交网络的快速发展也吸引了一些攻击者的目光，他们将社交网络作为获取利益的新平台[3]。攻击者在社交网络中创建大量的虚假账号来发布广告信息、钓鱼信息以及 Drive for Download[4]。Spam 攻击

收稿日期：2014-07-10；**修回日期**：2014-11-20

基金项目：国家自然科学基金资助项目（61272481）；信息安全国家重点实验室开放课题基金资助项目

Foundation Items: The National Natural Science Foundation of China (61272481); Open Fund of State Key Laboratory of Information Security

已经成为社交网络受到的最主要的攻击方式，以 2008 年的研究为例，83%的社交网络用户在当年接收到至少一条 Spam 消息[5]。因此对 Spam 攻击的检测引起了学术界和工业界的广泛关注。如文献[6~10]采用基于行为的检测方案，利用社交网络中 Spam 账号的行为特征进行检测。文献[11~13]采用基于内容的检测方案，针对社交网络中用户发布的 URL 进行检测，判断是否为 Spam URL。虽然这些工作能够以较高的准确率检测出 Spam 账号或 Spam URL，但它们针对的都是传统的 Spam 攻击方式，即 Spam 账号通过社交网络的状态、微博、回复、评论等功能发送、传播大量包含恶意内容的文本消息，这些文本消息在每次发送或传播的过程中都携带有恶意 URL。

Spam 攻击与检测是一个交替进行的过程，Spam 攻击者在面对社交网络中的检测机制时能够很快找到绕过的策略[14]。当前在社交网络中出现了一种新的 Spam 攻击方式[15]，攻击者利用社交网络的分享功能传播包含有 Spam 信息的相册，当正常用户访问到该分享相册，在浏览照片时就会在照片描述中显示这些 Spam 信息，称其为分享式 Spam 攻击。与传统 Spam 攻击相比，分享式 Spam 攻击将 Spam 信息的存储与传播割裂为 2 个独立的部分，在传播过程中只表现为相册或照片的分享，不直接携带有恶意内容。因此现有的检测算法，无论是基于内容的检测，还是基于行为的检测都不再适用。

针对新型的分享式 Spam 账号的检测，目前只有 Wang 等 [15]利用正常账号与 Spam 账号在鼠标点击方面的不同来进行，但他们的工作只能够检测 Spam 相册传播者，因为只有传播者会有大量分享操作，而 Spam 相册上传者只执行一次上传操作。同时他们采用的是对社交网络中所有用户都进行深度检测，随着用户数量的不断增加，对于用户数量巨大的社交网络这种完全的检测算法是不现实的[16]。因此本文主要关注点是在有限时间、有限资源情况下检测出更多的分享式 Spam 账号。

本文针对分享式 Spam 攻击，分析了其具体攻击过程和特征，根据这种攻击方式的特征，设计了轻量级迭代检测算法（LIDA, lightweight iterative detection algorithm），利用本文所提的检测算法通过对人人网的 4 次迭代实验，共检测到 9 568 个 Spam 账号、30 732 个 Spam 相册以及 2 626 780 条 Spam URL。

本文的主要贡献和创新点如下。

1) 针对分享式 Spam 攻击的检测：本文检测算法充分利用分享式 Spam 攻击中 Spam 信息存储与传播分离的特征，能够有效检测分享式 Spam 攻击账号，从人人网的实验结果来看，比之前的工作更加有效。

2) 轻量级算法：本文针对分享式 Spam 攻击的检测算法是一种轻量级迭代检测算法，避免了传统算法对每个用户都做深度检测，能够在有限时间和资源情况下检测到更多 Spam 账号。

3) 基于检测结果的特征分析：分析了检测到的 Spam 账号、Spam 相册和 Spam URL 的特征，这些特征可作为下一步 Spam 检测工作的基础。

2　背景介绍

2.1　人人网

人人网是国内类似于美国 Facebook 的实名制社交网络，与 Twitter、新浪微博等非实名制社交网络不同，在人人网中用户之间的好友关系需要 2 个用户共同同意才能够建立。用户好友的动态信息将在用户的新鲜事中显示，这些动态信息包括发布以及分享状态、日志、视频和照片(相册)。用户点击这些动态信息就能够访问到好友上传或分享的日志、视频和照片的具体内容。在人人网中，用户上传的照片都属于某个相册。分享照片后，其他用户通过点击此分享即可访问到被分享的照片，然后点击下一张即可访问到该相册内的其他照片，分享式 Spam 攻击就利用了分享照片的功能。

2.2　传统 Spam 攻击

在实名制社交网络中，传统的 Spam 攻击方式是攻击者通过创建大量的虚假账号，以发布状态、评论、回复等方式在发布的文本中嵌入恶意 URL 诱使正常用户点击[11]。对于非实名制社交网络，攻击者通过发布微博、转发微博、回复、私信、@等方式传播带有恶意 URL 的 Spam 信息[17]。传统 Spam 攻击方式中每一个 Spam 账号都相对独立，在攻击过程中所承担的功能都相同，而且每次传播 Spam 信息的操作本身都携带有恶意 URL[18]。

3　分享式 Spam 攻击

当前社交网络中存在一种新的 Spam 攻击方式[15]，称其为分享式 Spam 攻击。分享式 Spam 攻击利用社交网络中分享照片的功能，通过分享对 Spam 信息进行传播。具体的攻击流程如图 1 所示。在分享式 Spam 攻击的过程中，有 2 类作用不同的 Spam

账号参与，其中一类负责上传夹杂有 Spam 信息的相册（Spam 相册），称其为上传者，另一类负责传播 Spam 相册，即对相册进行分享，称其为传播者。整个 Spam 攻击分为 2 步。

1) 上传者将 Spam 相册上传到社交网络中，即 Spam 信息存储在上传者的账号中。

2) 传播者对 Spam 相册进行分享，传播者的好友（正常用户）在新鲜事中就会接收到该分享信息。

这 2 步不需要有密切的关联，当上传者上传 Spam 相册后，传播者就能够在任何时间对 Spam 相册进行分享，即进行 Spam 攻击。有些账号会同时承担上传者与传播者的功能，称其为分享者。

为了吸引正常用户的访问，上传者在上传 Spam 相册时会以吸引用户的相册名称以及相关照片作为蓝本，然后在其中插入 Spam 照片和 Spam URL。当正常用户点击传播者的分享信息，浏览该相册内其他照片时，就会接收到上传者添加的 Spam 照片和 Spam URL。

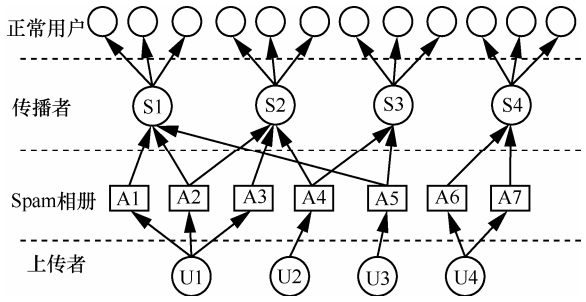

图 1　分享式 Spam 攻击模式

如图 2 所示，上传者在正常相册中插入了 Spam 照片，其中粗线框中的照片即为 Spam 照片，并且在照片的描述部分加入了 Spam URL。图 3 为该相册内的 Spam 照片，当用户浏览 Spam 相册时就会被迫接收到 Spam 信息，其中包含 Spam URL，即图中粗线框中部分。

与传统 Spam 攻击相比，分享式 Spam 攻击具有如下特点。

1) 检测更难。传统的基于内容或基于行为的检测方法是针对用户发布信息的文本内容或者用户添加好友、发布信息等行为的特征进行检测。但对于分享式 Spam 攻击，整个攻击过程被割裂为 2 个部分，Spam 信息只存储在上传者的账号中，而 Spam 信息的传播是通过传播者的分享操作，并不直接携带有 Spam 信息，因此传统的检测算法不再适用。

图 2　Spam 相册

图 3　Spam 相册中的照片以及 Spam URL

2) 危害更大。传统的 Spam 攻击中，Spam 信息主要出现在用户的新鲜事或者微博中[6]，用户能够直接忽略这些 Spam 信息。而在分享式 Spam 中，攻击者会利用吸引用户的相册名和照片创建夹杂有 Spam 照片和 URL 的相册，属于干扰式 Spam。用户在浏览此相册的过程中注意力不断被打断，严重危害正常用户的访问体验。

4　LIDA：轻量级迭代检测算法

针对分享式 Spam 攻击传播与存储分离的特性，并且考虑到传统的完全检测算法在用户数巨大的社交网络中对每个用户都做深度检测是不现实的情况[16]，本文提出了用于检测分享式 Spam 攻击的轻量级迭代检测算法。分享式 Spam 攻击是以相册作为基本的传播单位，因此 LIDA 也以相册作为基本的检测单元，利用分享式 Spam 攻击的传播特征获取可疑相册，然后利用分享式 Spam 攻击的存储特征检测可疑相册的内容，避免对社交网络中所

有用户都进行深度检测。

具体算法如图 4 所示，LIDA 在一次迭代中主要分为 2 个步骤：1) 目标筛选，基于分享式 Spam 攻击的传播特征，利用初始 Spam 账号种子筛选出最有可能是 Spam 的可疑相册；2) 内容检测，基于分享式 Spam 攻击的存储特征对筛选出的可疑相册进行基于内容的检测，判断可疑相册是否的确为 Spam 相册，以及 Spam 相册的拥有者是否为 Spam 账号。在新一次迭代过程中将上一次迭代结果中检测到的 Spam 账号作为新的 Spam 账号种子。

4.1　目标筛选

分享式 Spam 攻击割裂了 Spam 信息的存储与传播，在传播的过程中 Spam 信息没有直接参与。因此 Spam 攻击的效果与被分享的照片是否吸引正常用户有很大关系，只有正常用户访问该分享后才能接收到 Spam 信息。基于此特征，提出针对分享式 Spam 攻击的目标筛选算法，能够通过已知 Spam 内容找到最有可能是 Spam 的其他可疑内容。

类似文献[16]，将从 Spam 攻击者的开销与收益的角度来考虑。由于在整个 Spam 攻击过程中，攻击者的时间开销主要分为：上传者账号的创建、Spam 相册的创建、传播者账号的创建与维持、传播者的分享操作。攻击者的收益即为正常用户点击 Spam URL 的次数，正常用户对 Spam URL 点击次数越多，Spam 攻击者的收益越多。与传统的直接发送 Spam 信息相比较，分享式 Spam 攻击增加了 Spam 相册创建的开销。在分享式 Spam 攻击传播过程中 Spam 信息没有直接参与，正常用户只能接收到传播者分享了某个 Spam 相册名称的信息，因此攻击者需要精心构建相册名称以及部分正常相册的内容，然后将 Spam 照片和信息夹杂在正常相册中，这将会花费攻击者大量的时间。因此，攻击者为了利益最大化，会做如下操作：为了降低上传者创建的开销，攻击者会在一个上传者中上传多个 Spam 相册；为了降低 Spam 相册创建的开销，Spam 相册内容会被重复利用，多个上传者会上传同一个 Spam 相册。

基于以上分析，提出以下 2 条假设。

假设 1　一个上传者会上传多个 Spam 相册。

假设 2　一个 Spam 相册的内容会被多个上传者上传。

在实验部分将根据实验结果验证所做的假设是合理且有效的。

基于以上 2 条假设，提出自己的目标筛选算法。

1) 获取 Spam 用户其他相册。利用初始的 Spam 账号种子，根据假设 1，上传者会上传多个 Spam 相册，因此获取 Spam 账号的其他相册作为可疑相册。

2) 搜索同名 Spam 相册。根据假设 2，一个 Spam 相册会被重复使用，通过 Spam 相册名称搜索同名相册分享，将被分享相册作为可疑相册。

上述这 2 步就能够通过已知的 Spam 账号找到更多可疑相册，然后对可疑相册进行基于内容的检测，而不需要对所有用户的相册都进行检测，即完成对目标的筛选。

4.2　内容检测

当前存在大量基于其他信息的 Spam 账号检测算法，如基于用户个人信息、网络结构、用户行为等。但是社交网络中 Spam 账号有着明显的内容特征即发送 Spam 信息，因此通过内容检测算法对可疑相册进行检测，判断是否为 Spam，这样就能从根本上判断一个账号是否为 Spam 账号。

4.2.1　URL 获取

Spam 攻击者会在照片描述中嵌入 URL 并进行混淆，针对不同的 URL 混淆采用不同的获取方式来提取文本中的 URL。

普通 URL。直接通过对以“http://”开始的字符串进行匹配。

混淆 URL。被混淆的 URL，一般在 URL 中间加入汉字，或者将 http://头去掉，针对此类型的 URL，采用匹配字符串，然后将其组合为正常的可访问的 URL。

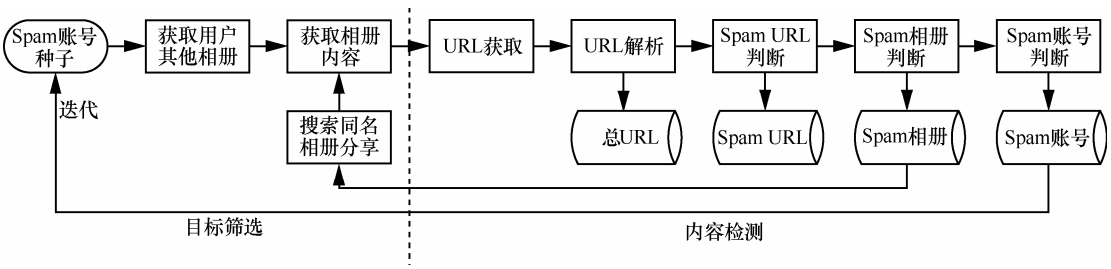

图 4　检测算法原理

URL 在评论。一些攻击者为了防止在照片描述中的 URL 被检测到，会通过对该 Spam 照片评论的方式发布 Spam URL，同时会在照片描述部分添加相关“地址在一楼”等提示信息，因此针对此类型的 URL，先匹配关键字符，然后获取 URL。

4.2.2 URL 解析

社交网络中为了减少字符的长度，对用户发布的 URL 都采用短网址的形式。攻击者为了防止社交网络自身检测系统的检测会利用其他的短网址服务将 Spam URL 转为短网址，这样对于社交网络来说，攻击者只是发布了一个短网址，但不知道该网址真实的地址[19]。短网址采用的跳转方式主要有 30x、JavaScript、meta 标签等。通过对获取到的 URL 进行逐层解析，对于每一个短网址域名构建<*domain, redirectType*>数据结构，根据这个数据表，就能够采用不同的方式对短网址进行解析。

4.2.3 Spam URL 判断

由于分享式 Spam 攻击的主要内容为广告信息，即 Spam URL 的最终地址主要指向一些购物网站，所以一般的 Google Safe Browsing[20]、Spamhaus DBL[21]、Wepawet[22]等 Spam 检测服务或列表并不能用来判断 URL 是否为恶意。本文采用了类似文献[11]的 URL 判断方法，对每个 URL 解析获得的最终地址采用多种方式进行判断。

1) 最终跳转地址。通过 URL 解析获得的最终跳转地址，如果该地址是指向一些购物网页，那么就判断为 Spam URL。

2) 短网址域名判断。存在专门为 Spam 攻击者提供短网址服务的网站或域名。对于使用此类恶意短网址域名的 URL，通过域名将其判断为 Spam。

3) 照片描述。对于一些最终地址已经不能被访问到的 URL，结合其使用的短网址以及照片描述内容的关键词来确定是否为 Spam URL。

4) 手动确定。一些最终跳转地址为人人网内部相册、博客地址、微博等 URL，通过手动访问该 URL，根据其内容是否是商品推广网页来判断是否为 Spam URL。

4.2.4 Spam 相册判断

对于可疑相册 Sus_Album_i，通过该相册内照片描述中嵌入 Spam URL 的数量 Num_i 和相册的总照片数 $Size_i$ 的比值，利用式(1)对相册是否为 Spam 相册进行判断。其中，Q 为相应阈值。

$$Sus\_Album_i=\begin{cases}Spam, \dfrac{Num_i}{Size_i}\geqslant Q\\ 正常, \dfrac{Num_i}{Size_i}<Q\end{cases} \tag{1}$$

4.2.5 Spam 账号判断

对于被检测账号是否为 Spam 账号，通过式(2)进行判断。其中，*Count* 为可疑账号所上传的 Spam 相册数，P 为相应的阈值。

$$Sus\_Account=\begin{cases}Spam, Count\geqslant P\\ 正常, Count<p\end{cases} \tag{2}$$

4.2.6 迭代控制

LIDA 采用迭代算法，随着迭代过程的进行会获取到大量采用系统命名或字符数较少的 Spam 相册，这些 Spam 相册作为目标筛选的相册名将降低可疑相册中 Spam 相册的比例，而本文算法的目的是在有限时间、有限资源情况下检测到更多 Spam 相册。因此通过式(3)判断是否结束迭代。

$$Iteration=\begin{cases}\text{True}, \dfrac{Count(Spam\_album_i)}{Count(Sus\_album_i)}\geqslant T\\ \text{False}, \dfrac{Count(Spam\_album_i)}{Count(Sus\_album_i)}<T\end{cases} \tag{3}$$

对于第 i 次迭代，Spam 相册 $Spam\_album_i$ 的数量与可疑相册 Sus_album_i 的数量的比值如果超过阈值 T，则继续迭代，否则，结束迭代。

5 针对人人网的实验

根据所提出的检测算法，本文从 2013 年 12 月到 2014 年 4 月对人人网做了检测分享式 Spam 攻击实验。

5.1 实验数据获取

与文献[23]类似，通过蜜罐账号来获取初始 Spam 账号。2013 年 12 月 14 日在人人网中创建了 20 个蜜罐账号，为了防止正常用户发送好友申请并降低实验数据偏差，蜜罐账号的个人信息采用虚假个人信息，并采用不同的性别、学校、年龄、网络等。到 2014 年 1 月 14 日对向蜜罐账号发送好友请求的可疑账号进行人工检测，以及利用账号搜索功能，共获得了 76 个 Spam 账号，将其作为初始 Spam 账号种子。

根据人人网的隐私保护策略[24]，用户的相册列表默认不能直接访问，这对获取 Spam 账号的其他相册造成了一定的困难。通过对人人网的分析，最

终利用人人网 VIP 会员中心的访问控制策略，能够获取 Spam 账号的 6 个相册。

人人网提供了开放搜索功能[25]，能够搜索同名相册的分享。为了防止 Spam 攻击者在上传相册时对相册名所做的混淆（在实验过程中发现攻击者会改变 Spam 相册名的部分关键词的顺序），利用 NLPIR 汉语分词系统[26]对相册名进行分词，然后对关键词进行搜索。

对于 Spam 相册判断的阈值 Q，发表于 2012 年 WWW 会议中的文献[16]选取 URL 与 tweets 的比值 0.1 作为 social promoters 的判断阈值取得较好的结果，“ We extract those social promoters whose URL ratios(the ratio of the number of URLs to the number of tweets) are higher than 0.1”，即在一个群体中如果某类物体所占比例超过 10%，就认为这个群体属于另一个状态，会认为这样的判断阈值在 Spam 相册判断中比较合理，因此，选取 Q=0.1 作为 Spam 相册判断的阈值。

对于 Spam 账号判断的阈值 P，由于大量 Spam 上传者只上传一个 Spam 相册，在实验中，69%的上传者只上传了一个 Spam 相册，因此，采用 P=1，即只要用户上传了一个 Spam 相册，那么就判断其为 Spam 账号。

由于在社交网络中 Spam 账号的比例约为 3%~5%[11]，在具体实践中，如果 LIDA 检出率降到 5%，则算法中目标筛选完全失效。为了保证本文算法 LIDA 的高效，并且降低数据获取数量，采用 3 倍 Spam 账号比例作为迭代控制的阈值 T，即 15%，如果低于此阈值就认为现有的 Spam 账号种子中已经存在较大的误差，因此停止迭代，然后选取新的 Spam 账号种子开始新一轮检测。

5.2　实验结果

利用初始的 76 个 Spam 账号种子，经过 4 次迭代共获得了 126 930 个相册（如表 1 所示），其中，包含有 2 710 361 条 URL，经过去重后有 940 200 条独立 URL（如表 2 所示）。通过不同的 URL 判断方式，共确定了 915 922 条 Spam URL (97.42%)、30 732 个 Spam 相册，以及 9 568 个 Spam 账号。将向人人网提供所检测到的 Spam 内容，并协助清除这些 Spam 账号。

从 4 次迭代的结果发现，第 1 次迭代所获得的 Spam 相册比例最高，这是因为初始作为 Spam 账号种子都是经过人工过滤选择，因此，搜索后的结果中 Spam 相册的比例达到 65.44%。而在第 2 次以及第 4 次迭代过程中，作为 Spam 账号种子的相册中包含有“手机相册”、“应用相册”等系统在用户创建过程中产生的相册名称，对实验结果的影响较大。在第 4 次迭代后 Spam 相册的比例已经降到 14.28%，低于所设置的阈值 T=0.15，结束迭代过程。

表 1　迭代次数与 Spam 相册、Spam 账号

迭代次数	获取相册数	Spam 相册数	Spam 相册比	新增 Spam 账号
1	1 389	909	65.44%	349
2	12 989	4 880	37.57%	1 709
3	23 473	12 225	52.08%	3 237
4	89 079	12 718	14.28%	4 273
总数	126 930	30 732	24.21%	9 568

表 2　Spam URL 判断方式与所判断 URL 数

判断依据	独立 URL 数（所占比例）	总 URL 数（所占比例）
最终跳转地址	670 564(71.32%)	1 616 438(59.64%)
域名判断	255 496(27.17%)	1 057 363(39.01%)
照片描述	9 241(0.98%)	24 313(0.90%)
手动确定	4 899(0.53%)	12 247(0.45%)
Spam URL	915 922(97.42%)	2 626 780(96.92%)
正常 URL	24 278(2.58%)	83 581(3.08%)
总数	940 200	2 710 361

5.3　实验结果评估

与文献[27]类似，本文的检测算法是轻量级检测算法而不是一个完全的检测算法，因此不考虑漏报率和误报率，而是以命中次数(hit count)或检出率作为衡量的标准。在实验中共检测了 126 930 个相册，包含 50 785 个用户，其中检测出 9 568 个 Spam 账号，检出率为 18.84%。由于目前无法获得与文献[15]中同样的数据集，无法在相同的环境或标准下与其进行对比。不过从算法思想的角度出发，WANG 等[15]的算法是完全检测算法，即需要对社交网络中所有用户都进行检测。而本文的算法先进行目标筛选，筛选出最有可能是 Spam 的可疑相册，然后进行检测。当本文算法中目标筛选完全失效时，本文的检测率与 WANG 等[15]的相同。而在现实环境中面对社交网络中数十亿用户，在有限时间有限资源情况下，由于本文的算法先进行目标筛选，因此，所需检测的账号更少，单位时间内的检出率会更高、更有效，能够在较短时间内检测到大

量 Spam 账号，有效降低社交网络中 Spam 账号的比例。

在此对所提出的假设进行验证。对于假设 1，通过对实验结果中 Spam 账号的 Spam 相册数分布进行分析，如图 5 所示，31%的 Spam 用户上传了 1 个以上 Spam 相册，11%的 Spam 用户上传了 5 个以上 Spam 相册。其中，在获取到的数据集中最多上传 Spam 相册数是 125，通过这些上传多于 1 个 Spam 相册的 Spam 账号，能够获得更多 Spam 相册。即一个 Spam 账号会上传多个 Spam 相册，说明假设 1 是有效的。

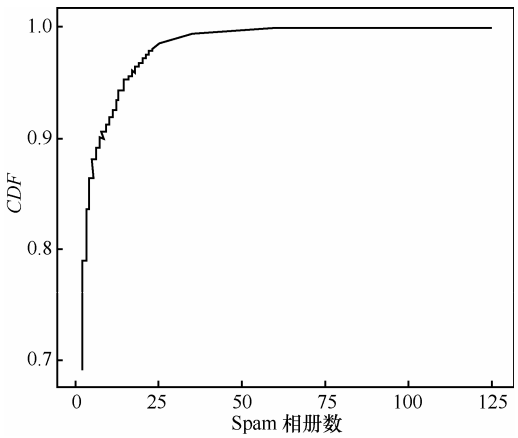

图 5 Spam 相册数分布

在实验中只能获得一个账号最近创建的 6 个相册，还包括头像相册、手机相册等系统自动创建的相册，这会造成 Spam 账号上传的 Spam 相册数减少。

对于假设 2，通过对同名 Spam 相册数的分析，如图 6 所示，72.14%的 Spam 相册存在同名相册，即一个 Spam 相册内容会被多个 Spam 账号重复使用，说明本文的假设 2 是有效的。

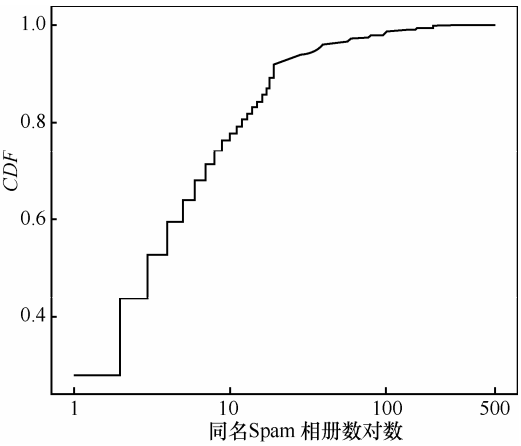

图 6 同名 Spam 相册数对数累积分布

6 Spam 分析

针对所获得的 30 732 个 Spam 相册、9 568 个 Spam 账号，以及 2 710 361 条 URL，分别对 Spam 相册、Spam 账号以及 Spam URL 进行了分析。

6.1 Spam 相册分析

为了分析 Spam 相册随时间的变化，对 Spam 相册的创建时间进行了分析。

从图 7 中可以看出实验中所检测到的 Spam 相册主要为 2012 年创建的相册（63.99%）。这是因为一方面在初始 Spam 账号种子中存在较多 2012 年创建的相册，目标筛选通过相册名获取可疑相册，而相册名具有时效性，因此在目标筛选中获取了大量 2012 年创建的可疑相册；另一方面从 2013 年以后人人网部署了基于文献[15]的检测系统，能够在一定程度降低分享式 Spam 相册的传播。但该系统只能够检测传播者，而且 Spam 攻击者通过降低分享速率就能够绕过该系统的检测，事实证明依然存在大量新创建的分享式 Spam 账号。

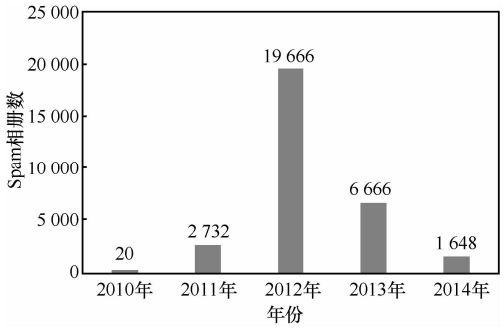

图 7 Spam 相册创建年份时间分布

本文分析了不同创建时间的 Spam 相册中包含 Spam URL 的照片数 *Num* 与整个相册内照片数 *Size* 比值的累积分布，如图 8 所示，可以看出 75%的 2014 年创建的 Spam 相册 *Num/Size* 值小于 50%，即包含 Spam URL 的照片数占总照片数的比值较小；2013 年创建的相册中 50%的 Spam 相册 *Num/Size* 的值小于 50%。而对于 2012 年以及 2011 年之前 92%的 Spam 相册中全部照片都包含有 Spam URL。

通过手动抽样分析不同年份创建的 Spam 相册，发现对于 2012 年以前的 Spam 相册，一般是推广性质的相册，即整个相册都包含 Spam URL；而对于 2013 年之后创建的 Spam 相册是在正常相册中

插入 Spam 照片以及 Spam URL，伪装为正常相册来吸引用户的浏览。这表明随着时间的推移以及社交网络检测算法的改变，Spam 攻击者的攻击方式也在变化。

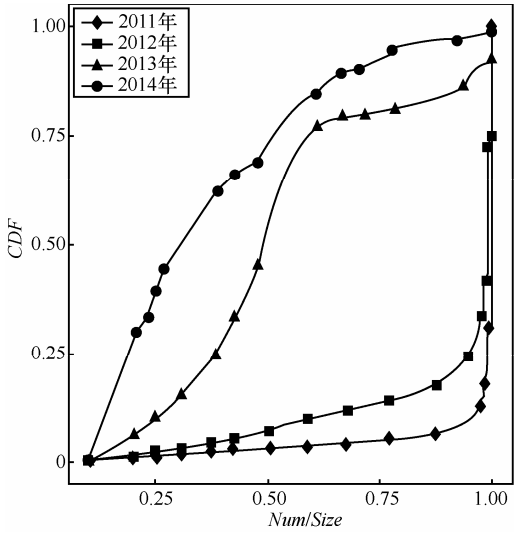

图 8　*Num/Size* 随年份变化

6.2　Spam 用户分析

针对检测到的 9 568 个 Spam 账号，获取了这些账号的基本信息，如好友数、来访数、总的分享数以及相册分享数。其中 671 个账号已经被封禁。对剩余 8 897 个能够访问的 Spam 账号，通过获取的数据集，发现其中 1 027 个 Spam 账号同时承担传播者的功能，即属于分享者。

本文分析了 2 种类型 Spam 账号的好友数，如图 9 所示，49.51%上传者好友数不超过 10 个。只有 7.01%分享者的好友数小于 10，而 59.21%的分享者好友数超过了 100。即攻击者通过新创建的 Spam 账号作为主要上传者，利用好友数较多的 Spam 账号作为分享者。

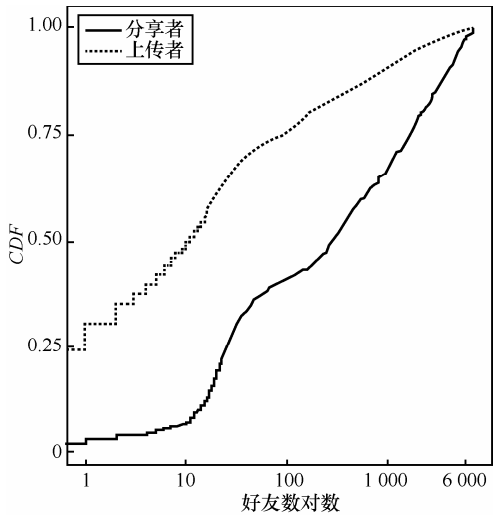

图 9　Spam 用户好友数对数累积分布

针对上传者与分享者在分享方面的区别，如图 10 所示，58.33%上传者的相册分享数与总分享数的比值小于 0.25；53.94%分享者的相册分享数与总分享数的比值超过 0.75，有 31.16%的分享者(320 位)的相册分享数与总分享数的比值超过 0.9。即对于分享者，在其整个分享内容中，对相册的分享占到了主要部分。这是因为分享者承担着传播 Spam 相册的功能，攻击者通过不断分享 Spam 相册来获得更大的收益。

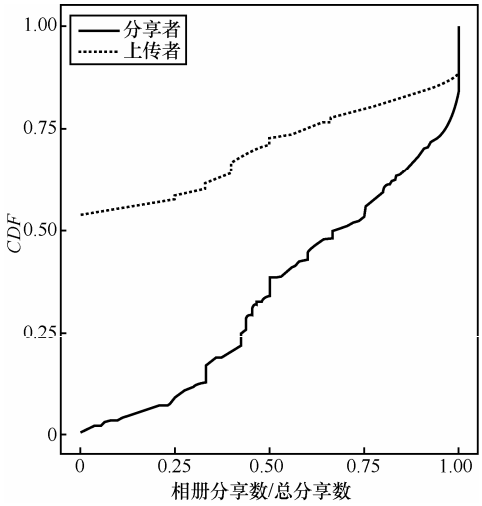

图 10　相册分享数与总分享数比值累积分布

6.3　Spam URL 分析

针对人人网的实验中共获取了 126 930 个相册，这些相册共包括 7 072 685 张照片，其中 2 710 361 张照片包含有 URL。通过对这些 URL 进行去重，共有 940 200 条独立 URL。本文对 URL 的重复数进行了分析，如图 11 所示，虽然 76.37%的 Spam URL 只出现了一次，但依然有 1.95%(1 782 条)独立 Spam URL 的重复出现超过 100 次，占整个 Spam URL 的 23.82%(625 775 条)。与 Spam URL 相比，正常 URL 重复次数主要小于 10 次，而 Spam URL 的重复次数在 10 到 100 之间的比例要超过正常 URL。

本文统计了 Spam URL 以及正常 URL 的短网址使用情况，如表 3 所示。

表 3　　短网址使用分布

类别	独立 URL 数（所占比例）	总 URL 数（所占比例）
Spam URL	845 751(92.34%)	1 795 548(68.36%)
正常 URL	8 742(36.01%)	11 140(13.33%)

与正常 URL 相比，Spam URL 使用短网址服务的比率更高，并且在使用短网址的正常 URL 中 86.68%是使用新浪微博的短网址服务 t.cn，这些照片的内容都来自新浪微博。

对 Spam URL 所使用的短网址域名进行分析，列出了使用频率最多的前 10 个域名，如表 4 所示。

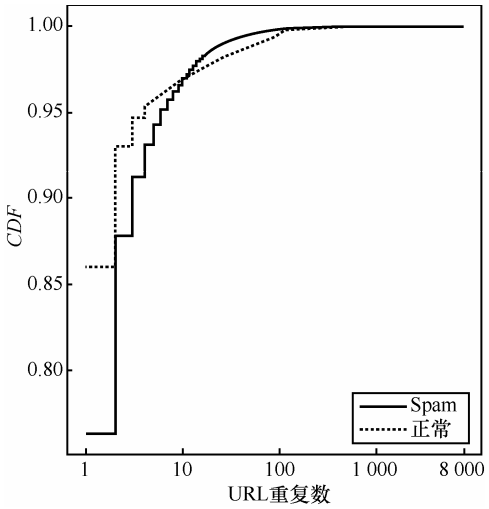

图 11 Spam URL 重复数对数累积分布

表 4 Spam URL 短网址域名分布

短网址域名	独立 Spam URL 数（所占比例）	总 Spam URL 数（所占比例）
url7.me	182 157(19.89%)	503 982(19.19%)
t.cn	167 667(18.31%)	272 568(10.38%)
bit.ly	92 927(10.15%)	115 847(4.41%)
sinaapp.com	48 411(5.29%)	145 619(5.54%)
dwz.cn	44 098(4.81%)	81 219(3.09%)
taourl.com	38 328(4.18%)	44 843(1.71%)
feiyiban.cn	27 665(3.02%)	101 933(3.88%)
126.am	27 586(3.01%)	45 724(1.74%)
url.cn	15 800(1.73%)	19 772(0.75%)
tinyurl.com	13 932(1.52%)	16 988(0.65%)

排名前 10 的短网址域名占整个 Spam URL 的 51.34%。在前 10 的短网址域名中既有专业的短网址服务提供商(url7.me、bit.ly、tinyurl.com)，也有互联网企业提供的短网址服务（t.cn、dwz.cn、126.am、url.cn）以及专门为购物网站推广者提供的短网址服务（taourl.com、feiyiban.cn），而且有些短网址利用了云计算平台（sinaapp.com）。可见 Spam 攻击者能够充分利用现有的 Web 服务为 Spam 攻击提供便利。虽然一些短网址服务部署了 Spam 检测系统[28]，但由于 Spam 攻击者发布的网址一般是指向购物网站，混淆了与正常商品推广的界线[29]，因此，短网址服务依然被滥用。

7 讨论

本节将对文中没有考虑的问题进行讨论。

1) 传播者的检测

LIDA 主要是针对上传者进行检测，因为对于分享式 Spam 攻击，Spam 信息只存储在上传者的相册内，只要此 Spam 相册被清除，正常用户就不会接收到 Spam 信息。并且之前已经有相关工作对传播者进行检测[15]。

2) 初始 Spam 账号种子

由于本文提出的算法需要初始 Spam 账号作为种子，并且通过对表 1 的分析，初始 Spam 账号种子与下一次迭代的检测效率有很大关系，如果初始 Spam 账号种子中包含有大量以系统命名的 Spam 相册，将会在一定程度降低 LIDA 的检测效率。在实验中第 1 次迭代的检出率是 65.44%，而第 4 次迭代的检出率是 14.28%，就是因为在 Spam 账号种子中包含大量以系统相册命名的 Spam 相册。但是由于本文的算法会对 Spam 相册的内容进行检测，因此初始 Spam 账号种子的选取并不会影响检测结果的准确率。由于本文算法的目的是在有限时间和有限资源情况下检测到更多 Spam 账号，因此只利用人工获得的 Spam 账号作为初始种子进行检测，没有进行不同初始 Spam 账号种子的对比实验。考虑在下一步工作中，分析不同初始 Spam 账号种子（不同的 Spam 账号数量、不同的 Spam 相册数量、不同的 Spam 相册创建时间等）对检测结果以及检测效率的影响情况，进一步提高成果的水平。

3) 不同类型 Spam 相册

通过图 8，发现不同时期创建的 Spam 相册所包含的 Spam URL 数量有很大的区别，在 2012 年之前创建的相册主要为推广性质的 Spam 相册，而在 2013 年之后的 Spam 相册主要是伪装成正常相册吸引正常用户的点击。由于这 2 种相册都包含有 Spam URL，因此没有做进一步的区分，都将其判定为 Spam 相册。而且通过手动检测，这 2 种 Spam 相册的上传者有很大一部分是重复的，即这些相册的拥有者都是 Spam 账号，只是在不同创建时间所创建的 Spam 相册形式不同。

4) 优势与不足

本文首次对分享式 Spam 攻击的过程和特征进

行了分析，利用此特征设计了专门针对分享式 Spam 攻击的轻量级迭代检测算法 LIDA，由于采用轻量级设计，有效避免了对社交网络中所有用户都进行深度检测的问题，通过人人网的实验结果表明，本文的算法与之前的工作相比能够在有限时间和资源下检测到更多 Spam 账号。但本文所提到的算法主要是利用了分享式 Spam 攻击中上传者的特征来进行目标筛选，没有利用传播者的特征，这是下一步的工作。

8　相关工作

当前学术界在社交网络中 Spam 攻击检测方面已经有很多相关工作。在此简要介绍并与本文的工作进行比较，具体如表 5 所示。

现有的检测算法从方法上主要分为基于行为的检测算法和基于内容的检测算法。

基于行为的检测算法主要有 Egele 等[6]通过检测用户之后的行为是否违反了之前建立的模型来判断账号是否被劫持。文献[7]中作者利用人人网中用户的好友请求以及网络结构等 4 个特征构造分类器检测虚假账号。Zhu 等[8]通过有监督的机器学习对用户的行为如访问相册、分享、发状态等进行建模来检测 Spam 账号。Stringhini 等[9]利用蜜罐账号收集 Spam 账号种子，通过分析这些 Spam 账号种子的特征来检测 Spam 账号。Thomas 等[10]通过购买 120 019 个 Twitter 虚假账号，根据这些账号在注册时的命名规则以及注册过程中的特征来检测 Spam 账号。但是这些基于用户行为特征的检测算法主要针对传统的 Spam 攻击方式，在分享式 Spam 攻击中，上传者只负责存储 Spam 相册，没有其他的操作；而传播者只分享相册，因此，这些算法无法检测分享式 Spam 攻击账号。

基于内容的检测算法有 Gao 等[11]针对 Facebook 用户新鲜事中的 URL，通过判断 URL 是否为 Spam URL，并对 Spam URL 进行聚簇来检测 Facebook 中大规模 Spam 攻击。Lee 等[12]根据 Twitter 中攻击者只有有限资源，所生成的短网址在跳转过程中会有重复出现的特点，利用短网址跳转以及 tweets 文本内容的特征来检测 Spam 信息。Kurt 等[13]根据 Spam URL 在 HTTP 头信息、JavaScript 事件、跳转行为等特点对 Twitter 用户发布的 URL 进行检测。但是这些工作只是针对社交网络中新鲜事或者微博内容中出现的 URL 进行检测，而对于分享式 Spam 攻击，在传播的过程中并不直接带有 URL，因此，这些方法不能够直接用来检测分享式 Spam 攻击。

目前只有 Wang 等[15]能够检测分享式 Spam 攻击。作者利用虚假账号与正常账号在使用社交网络时鼠标点击事件的区别来检测虚假账号。通过 3 个不同角度分析鼠标点击事件：Session-level、Activities、Click Transitions。随后利用 SVM 对这

表 5　　**相关工作比较**

相关工作	检测方法	特征	分享式 Spam	轻量级
文献[6]	基于行为	用户发送消息特征	×	×
文献[7]		好友关系网络	×	×
文献[8]		用户的活动行为	×	×
文献[9]		好友关系、文本特征	×	×
文献[10]		用户名命名规则、注册时行为	×	×
文献[14]		用户好友网络结构、语义相似性	×	√
文献[15]		用户鼠标点击行为	√	×
文献[11]	基于内容	用户发送消息中 URL	×	×
文献[12]		URL 重定向链	×	×
文献[13]		URL 的内容和行为、IP 地址	×	×
LIDA		照片描述中 URL	√	√

注：表中√表示是；×表示否。

些特征进行区分，最后对人人网中 100 万用户进行测试，发现 22 000 个可疑账号。虽然此工作跟本文的工作都检测分享式 Spam 攻击，但是 LIDA 是轻量级检测算法，不需要对每个账号都进行检测，而且作者利用的特征是鼠标点击，只能够检测 Spam 攻击中的传播者，因为只有传播者会进行大量的分享操作，而对于上传者只需要进行一次上传操作。本文主要侧重于上传者的检测。

在文献[16]中，Yang 等分析了 Twitter 中的网络犯罪生态系统（cyber criminal ecosystem），将虚假账号分为 2 类：Criminal 账号和 Criminal 支持账号，然后分析这 2 种账号内部以及两者之间的网络结构。最后作者利用这种网络结构和语义相似性设计了 CIA 检测算法来检测 Twitter 中的虚假账号。与本文类似，作者提出的 CIA 检测算法也是轻量级检测算法，但只是利用 Twitter 中的 following- follower 关系以及相邻 2 个账号之间所发布内容的语义相似度，而利用了分享式 Spam 攻击的特征并结合了基于内容的检测。

9 结束语

随着社交网络的飞速发展，Spam 攻击者也将其作为获取利益的乐土。针对社交网络中出现的分享式 Spam 攻击，首次分析其特点并据此提出了轻量级迭代检测算法 LIDA。通过人人网的实验表明，本文算法对分享式 Spam 攻击的检测是有效的，并且与现有工作相比，本文算法检测所需的基数更少且能检测上传者，能够作为社交网络现有检测算法的补充。通过对检测到的 Spam 相册、账号、URL 的分析，发现 Spam 攻击者能够有效利用各种 Web 资源辅助针对社交网络的 Spam 攻击。下一步工作，将考虑根据传播者的分享结构来获取更多 Spam 账号，并对 Facebook 中的分享式 Spam 攻击进行检测。

参考文献：

[1] FaceBook[EB/OL]. http://en.wikipedia.org/wiki/Facebook.2014.

[2] RenRen[EB/OL]. http://en.wikipedia.org/wiki/Renren. 2013.

[3] MICHAEL F, ROY G, YUVAL E. Online social networks: threates and solutions[J]. IEEE Communications Surveys & Tutorials, 2013, 11(4): 1-19.

[4] WANG A. Don't follow me: spam detection in twitter[A]. International Conference on Security and Cryptography (SECRYPT)[C]. Athens, Greece, 2010. 142-151.

[5] HARRIS. A Study of Social Networks Scams Interactive[R]. Public Relations Research, 2008.

[6] EGELE M, STRINGHINI G, KRUEGEL C. COMPA: detecting compromised account on social networks[A]. Network & Distributed System Security Symposium[C]. San Diego, CA, USA ,2013.

[7] YANG Z, WILSON C, WANG X. Uncovering social network sybils in the wild[A]. Proceedings of the ACM SIGCOMM Conference on Internet Measurement[C]. Berlin, Germany, 2011. 259-268.

[8] ZHU Y, WANG X, ZHONG E. Discovering spammers in social networks[A]. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence[C]. Toronto, Canada, 2012. 171-177.

[9] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks[A]. Annual Computer Security Applications Conference[C]. Austin, Texas, USA, 2010.1-9.

[10] THOMAS K, MCCOY D, GRIER C. Trafficking fraudulent accounts: the role of the underground market in twitter spam and abuse[A]. USENIX Security Conferences[C]. Washington D C, USA, 2013.

[11] GAO H, HU J, WILSON C. Detecting and characterizing social spam campaigns[A]. Proceedings of the ACM SIGCOMM Conference on Internet Measurement[C]. New York, USA, 2010.35-47.

[12] LEE S, KIM J. WARNINGBIRD: detecting suspicious URLs in twitter stream[A]. Network & Distributed System Security Symposium[C]. San Diego, California, USA. 2012.

[13] THOMAS K, GRIER C, MA J. Design and evaluation of a real-time URL spam filtering service[A]. IEEE Symposium on Security & Privacy[C]. Oakland, California, USA, 2011.447-462.

[14] YANG C, HARKREADER R, GU G. Empirical evaluation and new design for fighting evolving twitter spammers[J]. IEEE Transactions on Information Forensics and Security, 2013,8 (8): 1280-1293.

[15] WANG G, KONOLIGE T, WILSON C. You are how you click: clickstream analysis for sybil detection[A]. USENIX Security Conferences[C]. Washington D C, USA, 2013.241-256.

[16] YANG C, HARKREADER R, ZHANG J. Analyzing spammer's social networks for fun and profit[A]. Proceedings of the 21th International Conference on World Wide Web[C]. Lyon, France, 2012.71-80.

[17] GRIER C, THOMAS K, PAXSON V, *et al*. @spam: the underground on 140 characters or less[A]. Proceedings of the 17th ACM Conference on Computer and Communications Security[C]. Chicago, USA, 2010.27-37.

[18] LUPHER A, ENGLE C, XIN R. Detecting Spam on Social Networking Sites: Related Work[R]. University of California Berkeley, 2012.

[19] FLORIAN K, STROHMAIER M. Short links under attack: geographical analysis of spam in a URL shortener network[A]. Proceedings of the 23rd ACM Conference on Hypertext and Social media[C]. Milwaukee, WI, USA, 2012. 83-88.

[20] Google safe browsing[EB/OL]. https://developers.google.com/ safe -browsing/. 2014.

[21] DBL - the spamhaus project[EB/OL]. http://www.spamhaus.org/ dbl/. 2014.

[22] Wepawet[EB/OL]. http://wepawet.iseclab.org/. 2014.

[23] LEE K, EOFF B, CAVERLEE J. Seven months with the devils: a long-term study of content polluters on twitter[A]. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media[C]. Barcelona, Spain, 2011. 185-192.

[24] RenRen privacy statement[EB/OL]. http://www.renren.com/siteinfo /privacy. 2013.

[25] RenRen search[EB/OL]. http://browse.renren.com/. 2013.

[26] [EB/OL]. http://ictclas.nlpir.org/, 2014.

[27] ZHANG J, PORRRAS P, ULLRICH J. Highly predictive blacklisting[A]. USENIX Security Conferences[C]. San Jose, CA, USA, 2008.

[28] Spam and malware protection[EB/OL]. http://blog.bitly.com/post/ 263859706/spam-and- malware -protectio, 2014.

[29] THOMAS K, GRIER C, PAXSON V. Suspended accounts in retrospect: an analysis of twitter spam[A]. Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference[C]. Berlin, Germany, 2011. 243-258.

作者简介：

吕少卿（1987-），男，山西五寨人，西安电子科技大学博士生，主要研究方向为在线社交网络安全。

范丹（1982-），女，河北定州人，博士，主要研究方向为网络和协议安全。

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。