在线社交网络中异常帐号检测方法研究

张玉清[1)] 吕少卿[1)] 范丹[2),3)]

[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，西安 中国 710071)
[2)](中国科学院大学 国家计算机网络入侵防范中心，北京 中国 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 中国 100093)

摘 要 异常帐号检测是在线社交网络安全研究的关键问题之一。攻击者通过异常帐号传播广告、钓鱼等恶意消息以及恶意关注、点赞等行为严重威胁到正常用户的信息安全和社交网络的信用体系，为此有大量的研究工作来检测社交网络中异常帐号。文中回顾了近年来在线社交网络中异常帐号检测的主要成果，阐述了异常帐号在不同发展阶段的表现形式以及检测异常帐号所面临的主要挑战，重点从基于行为特征、基于内容、基于图（Graph）、无监督学习四个方面总结了异常帐号检测方案，介绍了在实验过程中数据获取、数据标识以及结果验证的主要方法，并对未来异常帐号检测的研究趋势进行了展望。

关键词 在线社交网络；异常帐号；异常检测；社交网络安全；检测方案；垃圾信息
中图法分类号：TP309
论文引用格式
张玉清，吕少卿，范丹，在线社交网络中异常帐号检测方法研究，计算机学报，2015，Vol.38：在线出版号 No.15
ZHANG Yu-qing, LV Shao-qing, FAN Dan ，Anomaly Detection in Online Social Networks，Chinese Journal of Computers,2015, Vol.38: Online Publishing No.15

Anomaly Detection in Online Social Networks

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

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

Abstract Anomaly detection is increasingly becoming a focus in the field of Online Social Networks (OSNs) security. Attackers have utilized OSNs as new platforms to conduct malicious behaviors, which have seriously threatened the privacy of users and the reputation of OSNs, including sending spam, phishing and other illicit activities such as selling followers and Page Like. Many detection techniques have been specifically developed in past years for spotting anomalies in OSNs. This paper reviews important achievements in anomaly detection in recent years. First, different behaviors of anomalies are elaborated with the grand challenges to the detection. Second, we discuss the detection techniques with respect to feature-based, content-based, graph-based and unsupervised approaches. Third, major methods of collecting datasets, labeling anomalies and validating results are introduced. Finally, we conclude the paper with an exploration of future research directions on anomaly detection in OSNs.

Key words: Online Social Networks; Anomaly; Anomaly Detection; Online Social Networks Security;

本课题得到国家自然科学基金资助项目（61272481，61402434）；中国博士后科学基金资助项目（2014M550085）；信息安全国家重点实验室开放课题基金（2014-12）资助。张玉清，男，1966 年生，博士，教授，博士生导师，主要研究领域为网络与信息安全，E-mail：zhangyq@ucas.ac.cn. 吕少卿，男，1987 年生，博士研究生，主要研究方向为在线社交网络安全与隐私，E-mail：lvsq@nipc.org.cn，手机:15600616025. 范丹，女，1982 年生，博士，主要研究领域为网络与协议安全，E-mail：fand@nipc.org.cn.

Detection Techniques; Spam

1 引言

在线社交网络（Online Social Networks, OSN）也称为社会媒体网络（Social Media Networks，SMN）或社交网站（Social Network Sites，SNS），是指为拥有共同兴趣、行为、背景的人们建立社交关系的在线网络平台[1]。随着 Web2.0 的快速发展，越来越多的社交网站如雨后春笋般出现，如国外的 Facebook、Twitter，国内的新浪微博、人人网等，这些网站聚集了大量的用户，Facebook 在 2014 年 6 月已经拥有 13 亿活跃用户[1]，国内新浪微博用户注册数也已超过 5 亿[2]。从全球权威的网站排名机构 Aleax 公布的网站排名结果来看[3]，在排名前 20 的网站中提供社交网络服务的网站约占 80%。这些统计数据都表明在线社交网络已经成为人们生活、工作、交流的重要平台。

图 1 购物网站中出售相关服务

在线社交网络在带给人们各种便利、满足人们各项需求的同时，其海量的用户数也吸引了攻击者的目光[2-4]，成为攻击者获取巨大利益的新乐园。攻击者通过创建大量的虚假帐号和盗用正常用户的帐号，在社交网站中发布广告、色情、钓鱼等恶意信息[5][6]，如 Twitter 中每天新增 300 万条垃圾信息[7]，Facebook 约 8.7%的帐号为攻击者创建的虚假帐号[4]，这些虚假帐号和被盗用的帐号统称为异常帐号。由于社交网络用户之间本身具有信任关系，其中的恶意信息比传统垃圾邮件等包含的恶意信息更加危险，研究发现 Twitter 中垃圾广告链接的点击率比垃圾邮件中链接点击率[8]要高出两个数量级。这些恶意行为对正常用户的隐私信息、帐号安全以及使用体验造成了严重威胁。同时攻击者通过异常帐号进行恶意互粉[9]、添加好友[10]、点赞[11]等行为来获取利益，如图 1 所示这些异常帐号甚至在购物网站中被明码标价能够提供关注、点赞、转发、投票等服务，严重危害到在线社交网络的信誉评价体系以及用户的信任关系。

为了解决在线社交网络中异常帐号带来的安全问题，学术界和工业界提出了大量检测方案。总体而言，异常帐号检测主要涉及 3 方面的内容：（1）异常帐号的表现。由于在线社交网络拥有海量注册帐号，这些帐号具有形态各异的表现，并且帐号的表现是一个动态过程，在不同的阶段具有不同的行为特征，因此如何确定异常帐号的表现成为首要问题；（2）检测方案的设计。在确定异常帐号表现的基础之上，面对在线社交网络中用户纷繁复杂的数据和行为，如何选择合适的特征和算法来设计既满足准确率又满足效率的检测方案，是该领域的核心问题。（3）检测方案的验证。设计的检测方案只有采用真实数据验证后才能够证明有效，而在线社交网络由于涉及商业利益和用户隐私等问题对于数据的获取和使用有严苛的条件，因此如何获取相应的实验数据、对实验结果进行验证等问题也需要重点关注。

在线社交网络中异常帐号检测需要分析用户行为模式以及网络结构对于推动社交网络分析、图挖掘(Graph Mining)等理论研究具有重要价值，对于社交网站安全、用户隐私保护等具有直接的意义和价值，另外对网络群体事件监测、网络舆论导向等也具有重要价值[12]。国内外许多大学和研究机构都在此领域展开了深入研究，如加州大学伯克利分校、卡内基梅隆大学、微软研究院、清华大学、北京大学等，一些重要的研究成果频频出现在 S&P、CCS、USS、WWW、KDD、AAAI 等国际信息安全领域和数据挖掘领域的顶级会议和期刊上。

在线社交网络中异常帐号检测涉及多个领域，

[1] Cmpany Info | Facebook Newsroom. http://newsroom.fb.com /company-info/ 2015,1,7

[2] Sina Weibo. http://en.wikipedia.org/wiki/Sina_Weibo/ 2015,1,10.

[3] Alexa. Top Sites, http://www.alexa.com/.2015,1,7.

[4] Facebook. http://en.wikipedia.org/wiki/Facebook 2015,1,9

如异常检测、图（Graph）中异常检测以及垃圾信息检测等。Chandola 等人[13]对一般性的异常检测方法进行了总结，Akoglu 等人[14]介绍了通用的图中异常检测方法，Savagea 等人[15]也是从图的角度对在线社交网络中异常检测进行了概括，Mo 等人[16]介绍了网络中垃圾信息的检测方法。为了全面的总结在线社交网络中异常帐号检测的方法以及研究成果，本文从异常帐号的表现形式出发，分析了异常帐号检测所面临的挑战，然后根据核心算法将不同的检测方案分为基于行为特征、基于内容、基于图以及无监督学习，对各个类别的关键技术和发展现状进行了分析讨论，并总结了在检测实验中数据获取、数据标识和结果验证的主要方法及其优缺点，最后探讨了在线社交网络中异常帐号检测的发展趋势和研究方向。

本文组织结构为：第 2 节介绍在线社交网络中异常帐号的表现形式以及异常帐号检测所面临的挑战；第 3 节分类探讨分析了异常帐号检测方法的关键技术和研究现状并总结了在异常帐号检测实验中数据获取、数据标识和结果验证的主要方法；第 4 节展望了未来的发展趋势；第 5 节为总结。

2 社交网络中异常帐号及挑战

2.1 异常帐号分类

在社交网络中攻击者需要通过帐号作为媒介来发动攻击，这些通过攻击者创建或盗用的异常帐号可以被用来执行各种恶意行为，严重影响正常用户的使用。根据异常帐号的不同表现形式，我们将在线社交网络中异常帐号的状态变化分为三个阶段，分别为创建阶段、发展阶段和应用阶段。

（1） 创建阶段。攻击者通过自动化工具利用虚假个人信息创建大量虚假帐号，这个阶段属于异常帐号的创建阶段；

（2） 发展阶段。社交网络帐号之间需要建立联系才能够传播消息，攻击者为了使恶意消息传播更广泛需要与其他正常帐号建立联系；同时攻击者为了快速增加异常帐号的可信程度，会与其他异常帐号建立联系。这个阶段属于异常帐号的发展阶段。

（3） 应用阶段。当异常帐号与其他正常帐号及异常帐号之间建立一定联系后，攻击者就会通过这些异常帐号执行各种恶意行为，如发布广告、钓鱼、色情消息等，或通过这些帐号来恶意增加其他帐号的信誉，如批量关注、恶意点赞等。有些攻击者还会通过盗用正常用户的帐号来执行这些恶意行为。这个阶段属于异常帐号的应用阶段。

这三个阶段紧密联系，只有经过创建阶段才能发展和应用，有些异常帐号会在创建阶段完成后就进入应用阶段，如恶意点赞等，但是这样的异常帐号容易被在线社交网络现有的检测系统检测到[17]，因此攻击者创建的异常帐号一般都遵循这三个阶段的变化。

异常帐号在不同阶段具有不同的表现形式，研究人员对异常帐号进行检测时所侧重的方面也不尽相同，在此我们根据三个不同阶段以及检测时的不同侧重将异常帐号分为以下五类，如图 2 所示：

（1） 僵尸帐号（Social Bot），即由攻击者通过自动化工具创建的虚假帐号，能够模拟正常用户的操作如发布消息、添加好友等[18]。僵尸帐号是攻击者创建的异常帐号在创建阶段的表现，主要侧重于自动化的创建过程，而不去考虑这些帐号被创建的目的，因此针对僵尸帐号的检测也主要利用帐号创建时的特征，如帐号昵称的命名规则等。

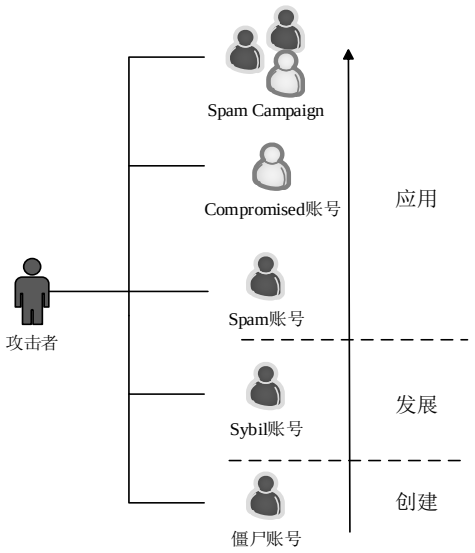

图 2　异常帐号分类

（2） Sybil 帐号，Sybil 帐号原本是应用在 P2P 等分布式网络中来描述虚假身份[19]，后来被应用到社交网络中描述在网络结构中攻击者所创建的虚假帐号[20]。Sybil 帐号相当于异常帐号在发展阶段的表现，因此针对 Sybil 帐号的检测主要通过图结构方面的异常。

（3） Spam 帐号，Spam 帐号是攻击者创建的虚假帐号在应用阶段的统称，即这些帐号主要用来发布广告、钓鱼、色情等信息，或用来恶意改变社交网络中的信誉，如恶意互粉、添加好友、点赞等行为。针对 Spam 帐号的检测主要侧重于恶意行为和恶意

内容的特征。

（4） Compromised 帐号，即被劫持帐号[21]。这些帐号原本是正常帐号，但被攻击者劫持来执行恶意行为。正常帐号拥有大量的正常用户好友，且具有正常的行为特征，所以攻击者往往通过各种方法盗取正常帐号进行恶意行为。由于 Compromised 帐号是由正常用户创建，没有帐号创建以及发展阶段的特征，因此针对 Compromised 帐号的检测主要利用帐号行为的突变来进行。

（5） Spam Campaign，即攻击者创建的大量虚假帐号以及盗用的Compromised帐号在集中时间段来传播恶意信息或执行其他恶意行为，称其为 Spam Campaign[22]。针对 Spam Campaign 的检测主要通过这些帐号在同一时间段内的群体行为，如同时发布相同消息或者同时点赞某个页面等。

这五类帐号的分类不是绝对的也不是互相排斥的，只是为了更好的标识异常帐号在不同阶段的不同表现，例如一个异常帐号在侧重网络结构方面的工作中被称为 Sybil 帐号，同时对于侧重帐号表现的工作中也可被称为 Spam 帐号。本文所涉及的异常帐号不包括用户为了保护个人隐私或为了分开工作和生活而创建的多个帐号，也不包括攻击者进行身份克隆攻击所创建的帐号[23]。

2.2 异常帐号检测主要挑战

在线社交网络海量的用户数、异常帐号的多种表现形态以及复杂的地下市场等都为异常帐号的检测带来了巨大的挑战。

（1）异常帐号的多种表现形式

如 2.1 所述，在线社交网络中异常帐号具有多种表现形式，而且不同攻击者创建的异常帐号也具有不同的行为模式和特征[5]，同时社交网络拥有海量的用户，这些用户在使用社交网络的时候也表现出不同的行为模式和特征。有经验的攻击者刻意将异常帐号的表现更加接近正常用户，使得对于异常帐号的检测更加困难，有些甚至正常用户都无法分辨清楚[24]，而且有时需要区分具体的异常帐号类型，因为社交网络服务提供商对不同类型异常帐号的处理方式不同，如对于攻击者创建的虚假帐号可以直接禁用，而对于 Compromised 帐号，需要给用户发送安全提醒或者对帐号重置密码。因此异常帐号的多种表现形式对于异常帐号的检测是一个挑战。

（2）异常帐号特征的动态变化

异常帐号检测是一个猫鼠游戏，当社交网络根据异常帐号的某些特征部署了相应的检测系统之后，攻击者在利益的驱使下总是能够很快找到绕过的方式，使异常帐号表现出新的特征，检测系统就

要重新对特征进行训练，这样社交网络对异常帐号的检测往往滞后于攻击者，使得正常用户依然受到异常帐号的危害。因此异常帐号特征的动态变化对于异常帐号的检测是一个挑战。

（3）社交网络巨大的用户数据

社交网络拥有海量的用户，如 Facebook 在 2014 年已经拥有 13 亿月活跃用户，用户每天所发布的内容以及用户的行为操作更是不可胜数，如每天 Twitter 中用户发布的消息达到 5 亿条，而异常帐号检测系统需要对每个用户的数据都进行计算，这将花费大量的时间，而异常帐号检测期望能尽早的发现异常帐号，降低对正常用户的损害。因此社交网络巨大的用户数据对于异常帐号的检测是一个挑战。

（4）网络空间的复杂性

复杂的网络空间所拥有的功能和服务为异常帐号提供了平台和便利。如众包(crowdsource)平台，能够为攻击者解决社交网络设置的验证码和手机验证[25]，并且攻击者能够在众包平台发布相应的恶意任务[26]，一些正常用户在利益的驱动下参与这些恶意任务，因此这些帐号一般表现正常，但当有恶意任务时就表现出异常行为，这样混淆了正常帐号和异常帐号的界线，使得针对异常帐号的检测更加困难。网络空间还有大量的短网址服务，攻击者通过这些短网址服务将恶意 URL 进行短网址处理，绕过了社交网络对用户发布的 URL 的检测[27]。因此复杂的网络空间对于异常帐号的检测也是一个挑战。

总而言之，尽管在线社交网络中异常帐号检测与一般数据异常检测以及图中异常检测紧密相关，但由于社交网络自身海量的数据以及用户复杂的行为方式等特性，使得社交网络中异常帐号检测面临更多的新挑战，将一般异常检测或图中异常检测的研究成果直接应用到在线社交网络中异常帐号检测无法取得令人满意的效果，因此需要对在线社交网络以及异常帐号的特性进行深入分析，并在此基础上提出有针对性的检测方案。

表 1　检测方案分类

方案	思想	特征	方法	主要检测类型
基于行为特征	分类	用户行为特征	有监督	僵尸、Spam、Spam Campaign
基于内容	分类	消息内容	有监督	Spam、Spam Campaign、Compromised
基于图	图中异常检测	图结构	无监督	Sybil、Spam
无监督学习	聚类/模型	多种特征	无监督	Spam、Spam Campaign、Compromised

3 社交网络中异常帐号检测方法

针对在线社交网络中异常帐号所带来的威胁，学术界和工业界都提出了大量的检测方案，根据这些方案所采用核心算法的不同，我们将这些方案分为四类，如表 1 所示，分别为基于行为特征的检测方案、基于内容的检测方案、基于图的检测方案和无监督学习的检测方案。基于行为特征和基于内容的检测方案将异常帐号检测看为一个分类问题，即分别利用帐号的行为特征和帐号发布的内容来区分正常帐号和异常帐号。社交网络中帐号之间的关联关系具有图的性质，基于图的检测方案是利用正常帐号和异常帐号在所形成的图中具有不同的结构模式或连接方式，将异常帐号检测问题转化为图中异常检测问题，再利用图挖掘的相关算法来区分正常帐号和异常帐号。无监督学习的方法是基于正常帐号有相同的特征或者符合一定的模型，通过特征的聚类或者建立模型来检测异常帐号。

检测方案的分类没有明确的界线，有的检测方案可能采用了多种检测技术，比如结合了基于行为特征和基于内容，或者基于行为特征和基于图等，而且在设计具体的检测方案时可以考虑结合不同的技术来达到更好的效果。

3.1 基于行为特征的检测方案

由于异常帐号的主要目的是通过恶意行为如发布广告、钓鱼、色情消息等从中获取利益，而且异常帐号往往是通过自动化工具来控制，为了获取利益的最大化，异常帐号会提高发布消息的频率或者在短时间内发出大量的好友请求等。因此异常帐号与正常帐号在某些行为特征方面必然存在差异。一些工作就是利用异常帐号与正常帐号在行为特征方面的不同来检测异常帐号。

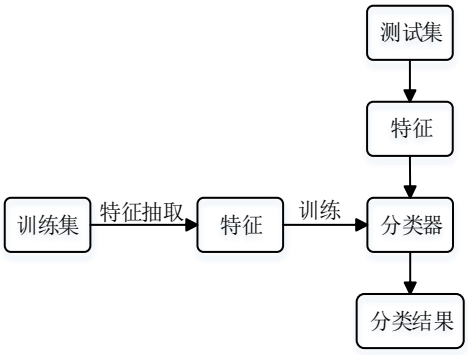

图 3 基于行为特征的检测方案

基于行为特征的检测方案将异常帐号检测看为数据挖掘中的分类问题。因此检测方案的基本流程如图 3 所示，首先在社交网络中获取数据训练集，然后从数据中抽取相应的行为特征，再利用分类算法对这些特征进行训练形成分类器，最后利用测试样本集对分类器进行测试并判断分类结果。

基于行为特征的异常帐号检测方案的技术关键主要是特征的选取以及分类算法的选择，并且在实践中发现特征的选取比分类算法的选择更加重要[15]。因为对于抽取的特征可以应用多种分类算法进行测试，然后选择效果最优的，而社交网络中帐号可利用的特征无穷无尽，无法一一测试后再选取。在表 2 中我们对相关文献中采用的特征进行了总结，按其所属分为四类，分别为用户个人信息、用户行为、好友关系以及消息内容，并列出了常见的特征。由于不同的工作所采用的数据集不同，因此对于检测结果的对比没有意义，我们只对具体的算法进行介绍。

选取不同的特征能够检测不同类别的异常帐号。不同类别的异常帐号在社交网络中有不同的行为表现，因此利用不同的行为特征能够检测不同类型的异常帐号。利用帐号注册时的特征如帐号的命名规则、注册时间、注册时的表单提交、点击时间等特征可用来检测僵尸帐号[5][28][29]。利用用户行为特征以及消息内容的特征如帐号创建时间、每天发布消息数、消息评论数、消息相似度等来检测 Spam 帐号[30][31]。Amleshwaram 等人[32]新提出了 15 项消息内容的特征来检测 Twitter 中异常帐号，然后采用 k-means 算法对这些异常帐号进行聚类来检测 Spam Campaigns。Stringhini 等人[33]利用好友关系特征以及消息内容特征分别对 Facebook、Twitter 中 Spam 帐号进行检测，然后根据帐号所发布 URL 的最终地址对 Spam 帐号进行聚合来检测 Spam Campaign。Yang 等人[34]利用好友关系的特征如发出好友请求数以及好友请求被接受数等特征来检测人人网中 Sybil 帐号。

不同的特征对于检测结果的贡献程度也不同。Lee 等人[35]对比了消息内容特征跟好友关系特征对检测 Twitter 中 Spam 帐号的贡献，发现利用消息内容特征更加有效。Benevenuto[36]等人采用了 39 项消息内容的特征以及 23 项用户个人信息和用户行为的特征来检测 Twitter 中异常帐号，并根据对检测结果的贡献程度将特征进行了分组，每组 10 个特征，然后分别对比每组特征的检测效果，发现利用贡献程度最高的 10 个特征的检测结果与利用所有特征的检测结果相同，并且即使是贡献程度最低的 10 个特征对于异常帐号也具有区分度。Lee 等人[37]采用结合用户个人信息、好友关系、消息内容共 16 项特征来对异常帐号进行检测，并分析了利用不同类别特征进行检测的准确率，发现采用结合这三项特征的检测准确率最高。Wang 等人[38]利用好友关系以及消息内容的特征对 Twitter 中异常帐号进行检测并分析了每项特征具有区分度的原因。Ahmed 等人[39]通过每次删除一个特征来计算每个特征对检测结果的贡献程度。

表 2 特征列表

类别	特征
用户个人信息	用户名长度、用户简介长度、帐号注册时间、用户名命名规则、被访问次数
用户行为	消息发布时间间隔、评论回复时间、帐号注册流程、帐号点击顺序、点赞数
好友关系	好友数、粉丝数、关注数、好友请求/好友数、关注数/粉丝数、好友网络聚类系数、二阶好友数、二阶好友消息数
消息内容	消息中 URL 比率、#比率、@比率、消息相似度、消息单词数、消息字符数、评论数、Spam 关键词数、消息来源、消息数量、消息转发次数、

不同的特征具有不同的鲁棒性，有些特征攻击者能够轻易绕过。Yang 等人[40]对常见的 24 个特征的鲁棒性进行了实验性分析，发现这些特征很容易被攻击者绕过并给出了攻击者可能的绕过策略。Hu 等人[41]采用在线学习的方法能够及时更新模型，提高了特征的鲁棒性，通过有向拉普拉斯图对社交网络的关系信息进行建模，通过矩阵分解框架来表达用户信息，因此即使攻击者改变了 Spam 帐号的行

为方式，也能够及时训练模型。

基于行为特征的检测方案的基础是异常帐号与正常帐号在一些行为特征上表现不同，因此基于行为特征的检测方案的重点是寻找这样的特征，而且这样的特征需要有较强的鲁棒性才能够防止攻击者绕过，只要利用这样的特征形成分类器就能够对社交网络中其他帐号进行检测。基于行为特征的检测方案检测速度快，但是需要提前在线下对样本数据进行训练，同时需要正常帐号和异常帐号的样本，当特征数量比较多时线下训练的时间较长。基于行为特征的检测方案利用的是异常帐号的行为特征，只有异常行为发生之后才能够检测到，因此无法做到对恶意行为的实时检测[42]。

3.2 基于内容的检测方案

异常帐号通过发布广告、钓鱼、色情等消息来获取利益，因此在发布的消息内容方面异常帐号与正常帐号之间存在区别。基于内容的检测方案是利用异常帐号所发布内容与正常帐号所发布内容的不同来进行检测，因此检测的重点放在判断用户发布的消息是否为恶意消息。基于内容的检测方案能够在用户发布消息时即可判断该消息是否为恶意消息，与基于行为特征的检测方案相比检测更加及时。基于行为特征的检测方案中也涉及到部分消息内容的特征，但是那些特征是消息所附加的特征如消息中 URL 的比率等，而不是消息内容本身，如文本内容或者所嵌入的 URL 等。

根据不同的消息内容利用对象，我们将基于内容的检测方案分为以下两类，一类为利用单个帐号的内容特征，另一类为利用群体帐号的内容特征。我们将分别介绍这两类方法。

（1） 利用单个帐号内容特征。

利用单个帐号内容特征是根据单个异常帐号发布的消息内容如消息中嵌入的 URL 以及发布消息的行为方式与正常帐号的区别等来检测异常帐号。

通过判断消息内容中嵌入的 URL 的最终指向来进行检测。异常帐号在发布的恶意信息中嵌入了指向广告、钓鱼、色情、恶意软件下载等网址的 URL，可以通过判断消息内容中的 URL 是否恶意来判断发布消息的帐号是否异常。虽然互联网中存在大量标识钓鱼、恶意网站等黑名单列表，如 Google Safe Browsing[1]，Phishing Tank[2]、URIBL[3]、SURBL[4]、Spamhaus[5]、APWG[6]、MalwarePatrol[7]等，但这些黑名单列表往往存在延迟，在将异常帐号发布的恶意 URL 列入黑名单之前，90%的点击已经发生[27]。因此 Thomas 等人[43]提出了实时的 URL 检测方案，通过访问社交网络帐号提交的每一个 URL，对这些 URL 页面抽取相应的特征，再根据线下训练的分类器来判断是否为恶意 URL，其中使用的特征包括 URL 的特征、URL 跳转的特征、HTML 内容、HTTP 头部、JavaScript 事件、DNS 等，平均检测一个页面的时间只需要 5.54 秒。

通过对单个帐号的消息内容特征的变化来检测异常。对于被攻击者劫持的 Compromised 帐号，由于其被劫持前后所发布消息的内容和行为有巨大的变化，因此可以通过消息内容特征的变化来检测 Compromised 帐号。Egele 等人[44]通过对比帐号发布的消息是否存在突变来判断帐号是否被劫持，首先通过时间、消息来源、消息语言、消息主题、消息中的链接、直接用户交互、邻近性这 7 个特征对帐号所发布消息建模，然后判断帐号之后所发布的消息是否与创建的模型有偏离，如果偏离超过了阈值，就判断为可疑。但单个用户的偏差并不能确定一定是帐号被劫持，也许可能仅仅是换了新的设备，因此再对这些可疑帐号的消息内容进行聚类，通过消息内容以及所包含 URL 判断这些消息是否相似，如果大量帐号通过这些消息聚为一类，那么就判断为 Compromised 帐号。

（2） 利用群体帐号内容特征。

攻击者为了扩大恶意消息的传播范围来获取更大的利益会控制大量的异常帐号发布相同或相似的恶意消息，因此可以利用群体帐号的消息内容特征来检测异常帐号。

通过判断消息是否相似来检测异常。Gao 等人[45]发现 Twitter 中 63%的垃圾消息的文本内容是基于模板产生，因此设计了 Tangram，将已知的恶意消息分割为字段，用字段来生成匹配模板，然后用生成的模板去检测更多的恶意消息。

---

[1] Google Safe Browsing API. https://developers.google.com/safe-browsing/ 2015,1,7

[2] PhishTank- Join the fight against phishing. http://www.phishtank.com/ 2015,1,7

[3] Uribl. http://uribl.com/ 2015,1,12

[4] SURBL. http://www.surbl.org/lists 20151,12

[5] The Spamhaus Project. http://www.spamhaus.org/ 2015,1,10

[6] Unifying the Global Response to Cybercrime. http://www.antiphishing.org/ 2015,1,7

[7] MalwarePatrol- Malware is everywhere. http://www.malware.com.br/ 2015,1,8

通过判断消息中包含的 URL 是否相似来检测异常。Gao 等人[22]对 Facebook 用户所发布的消息进行聚类，即找到发布大量相同或相似消息的群组，然后通过黑名单列表等方式判断这些群组中用户所发布的消息是否为恶意，如果所发送的消息为恶意消息，那么这些帐号就是 Spam Campaign 帐号。与此类似，Chu 等人[46]则是对 Twitter 中 Spam Campaign 进行检测，先通过 Twitter 微博内容中嵌入的 URL 的最终跳转地址对微博聚类，然后利用 URL 的最终跳转地址以及微博内容判断类内帐号是否为恶意。Lee 等人[47]发现 Twitter 中恶意 URL 经过短网址的多次处理，会根据访问来源的不同跳转到不同的最终地址，对于正常用户的访问会跳转到恶意网页而通过自动化爬虫访问将跳转到正常网页，这样就造成利用 URL 最终地址来检测 URL 是否为恶意的结果不准确。鉴于此，他们利用恶意 URL 的跳转只有有限的中转资源这一特点，首先将具有相同 IP 地址的 URL 的域名替换为 IP 地址；其次找到这些 URL 在跳转中的入口地址；然后根据入口地址重建 URL 的跳转路径，最后抽取跳转路径以及微博内容的特征对 URL 进行检测。

上述基于消息文本、消息中 URL 等内容的检测方案，根据所利用的不同内容能够检测不同类型的异常帐号（单个帐号内容特征能够检测异常帐号个体如 Spam 帐号或 Compromised 帐号，群体帐号内容特征能够检测异常帐号群体，如 Spam Campaign），虽然能够在异常帐号发布恶意消息的时候及时检测到异常[48]，但是由于只是利用了帐号发布的消息，因此无法对其他恶意行为的异常帐号进行检测，如虚假粉丝、虚假点赞等不发布恶意消息的异常帐号。基于内容的检测方案的核心算法还是有监督学习的方法，即需要根据现有的恶意内容进行训练，形成分类器，然后再进行检测。因此攻击者可以通过修改恶意页面的特征或者改变发布消息的模板来绕过检测，如攻击者可以利用 Spinbot 形成大量文本不同但是意思相同的内容，就能够绕过这些检测[40]。

3.3 基于图的检测方案

社交网络的一个重要特性就是帐号之间存在联系，如好友关系、关注、粉丝等，而且也只有两个帐号之间存在联系时才能够进行信息交流。因此一方面攻击者为了更广泛传播恶意信息从而试图在短时间内与大量正常帐号建立联系，另一方面攻击者通过有偿与其他用户建立联系，从中获取利益，如出售粉丝等。因此异常帐号与正常帐号在所组成的图结构中存在区别，基于图的检测方案就是利用这种区别来检测异常帐号。基于图的检测方案的本质是异常帐号与正常帐号在组成的图中具有不同的结构或者连接方式，因此基于图的检测方案关键是构造一个图，在图中异常帐号与正常帐号具有不同的结构或者连接方式，然后利用图挖掘的相关算法找到图中具体的异常结构或者异常节点[49]。

在社交网络中存在众多的图结构，除了显性的好友关系图（如 Facebook 中好友关系组成的图、Twitter 中关注关系组成的图），还存在利用其他关系建立的隐性图结构，如访问关系、分享关系、URL 共享关系等，根据组成图所利用的不同关系，我们将现有的工作分为好友关系图和其他关系图。

表 3 Sybil 检测方案比对

	假设	算法	可证明	引入 Sybil 节点数	分布式设计	实验数据集
SybilGuard[20]	假设 1,2,3	随机游走	√	$O(\sqrt{n}\log n)$	√	Kleinberg[57]
SybilLimit[50]	假设 1,2,3	多重随机游走	√	$O(\log n)$	√	Friendster,LiveJournal , DBLP ,Kleinberg
SybilInfer[51]	假设 1,2,4	随机游走/贝叶斯推理	X	/	X	LiveJournal
SumUp[52]	假设 1,2,3,4	自适应最大流	√	1+O(1)	X	YouTube[58],Flickr[59]
Gatekeeper[53]	假设 1,2,3	随机游走/广度优先搜索	√	$O(\log k)$	√	YouTube[59],Digg，
SybilDefender[54]	假设 1,2,3,4,5	随机游走	X	/	X	Orkut[59], Facebook[60]
SybilRank[55]	假设 1, 2,3	随机游走/幂次迭代	√	$O(\log n)$	X	Facebook[61]
SybilBelief[56]	假设 1,2,3,4	马尔可夫随机场	X	/	X	Facebook[62]，Email[63]

（1）好友关系图

由帐号之间好友关系组成的图是在线社交网络中最基础也是最明显的图结构。目前利用好友关系图检测异常帐号的主要工作是检测 Sybil 帐号，有大量的工作对社交网络中 Sybil 帐号检测进行了总结[64-70]。

对 Sybil 帐号的检测需要基于一定的网络结构模型，基本的网络结构模型如图 4 所示，正常帐号和异常帐号分别形成较密集的结构，而正常帐号与异常帐号之间存在稀疏的连接（攻击边）。不同的 Sybil 检测方案会对网络结构和攻击模型提出不同假设，主要的假设条件有[71]：

1) 正常帐号形成的网络结构是快速融合的（fast-mixing）。所谓的快速融合就是在网络中随机游走算法经过 O(log*n*)（*n* 为网络中节点总数，下同）步后到达的最终节点[1]符合平稳分布，即最终节点与随机游走算法开始的节点无关。

2) 已知至少一个正常节点。大部分 Sybil 检测方案都是从正常节点出发，推测其他节点是否正常。

3) 攻击边（attack edges）的数目是有限的。如图 4 所示，攻击边是 Sybil 节点与正常节点的连接边。即使攻击者能够创建大量的 Sybil 节点，但这些 Sybil 节点与正常节点之间的连接是有限的，即在正常节点组成的网络和 Sybil 节点组成的网络之间存在最小割（small cut）。最小割增加了整个网络的融合时间。

4) 整个网络的拓扑结构已知。类似 Facebook 等在线社交网络都符合这一假设，而 DHT（Distributed Hash Table）网络等不符合这条假设。

5) Sybil 节点组成的子网远远小于正常节点组成的子网。对于在线社交网络，正常帐号数都达到千万甚至亿级别，攻击者无法创建如此多的 Sybil 节点，因此都符合这条假设。

不同的检测算法对网络结构和攻击模型有着不同假设，在表 3 中我们列出了目前主要的 Sybil 检测算法所需的假设、核心算法、是否结果可证明以及证明结果、是否为分布式设计以及算法实验所使用的数据集。Sybil 检测方案只利用了网络结构，而 Sybil 节点与正常节点之间存在连接，因此必然存在 Sybil 节点被误判为正常节点，所以 Sybil 检测方案的一个主要衡量标准就是每条攻击边引入的 Sybil 节点数的上界，部分检测方案能够基于假设和算法给出相应的数学证明，即表 3 中“引入 Sybil 节点数”。

[1] 用“节点”来表示抽象图结构中的组成单元，与在线社交网络中的“帐号”相对应。

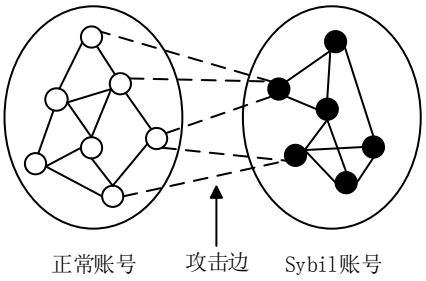

图 4 Sybil 帐号网络结构模型

根据 Sybil 检测方案所采用的核心算法我们将现有的检测方案分为以下几类：

1) 随机游走

现有的 Sybil 检测方案都主要基于假设 1，即正常节点组成的网络是快速融合的，并且正常节点与 Sybil 节点之间存在最小割，从一个正常节点出发经过若干步后到达的节点依然是正常节点的概率较大。因此可以利用随机游走算法，通过计算节点与已知的正常节点之间的关系来判断帐号是否为异常。SybilGuard[20]通过修改的随机游走算法计算每个节点的转移路径，根据节点的路径与给定的正常节点的路径是否有交集来判断是否为正常节点，但是 SybilGuard 的误报率很高，每条攻击边引入的 Sybil 节点数是 $O(\sqrt{n}\log n)$，因此 Hu 等人又提出了 SybilLimit[50]通过判断节点路径的最后一条边与正常节点的最后一条边是否相同来判断是否为正常节点，使得每条攻击边引入的 Sybil 节点数降为 O(log*n*)。SybilInfer[51]给出了一种基于随机游走和贝叶斯推理的检测方案，首先计算随机游走形成的迹 T，再构建概率模型描述迹 T 是由正常节点形成的似然概率，随后利用贝叶斯理论计算给定迹 T 时，节点是正常节点的概率即 P=(X=正常节点|T)。Gatekeeper[53]结合广度优先搜索和随机游走算法，通过随机游走算法选取一部分节点作为根节点，然后进行广度优先搜索，如果一个节点被搜索到的次数大于一个阈值，那么就判定为正常节点，否则为 Sybil 节点，通过这种方法使得每条攻击边引入的 Sybil 节点数为 O(log*k*)（*k* 为攻击边数目，下同）。SybilDefender[54]从可信节点出发执行 f 次随机游走，游走的路径长度为 log*n*，然后将这 f+1 个节点作为判断节点进行 R 次随机游走并记录节点被访问到的次数，最后通过计算节点被访问到次数的均值和标准差来判断节点是否为正常节点。SybilRank[55]利用随机游走算法通过 O(log*n*)幂次迭代将已知的多个正常节点的可信值分配给其他节点，并将节点的可信值与节点的度进行标准化，根据标准化后的

结果进行排序，其中可信值较小的节点认为是可疑节点。

2) 社区发现

现有的 Sybil 检测算法本质上都是计算每个节点与已知正常节点的联系紧密程度来进行检测，因此可以利用现有的社区发现的算法来检测 Sybil 帐号，Viswanath 等人[64]提出了基于 Mislove[72]社区发现的检测方案，能够获得比 SybilGuard 和 SybilInfer 更好的效果。

Alvisi 等人[65]认为基于随机游走算法的 Sybil 检测方案对于网络结构和攻击模型的假设在现实网络中并不成立。现实网络结构中正常节点并不是快速融合的而是形成多个社区结构，社区内是紧密联系而社区间存在稀疏的割边[73]，而且攻击者创建的 Sybil 节点能够与其他正常节点建立大量的连接[34]，因此利用随机游走算法的方案存在较大的误检率。作者基于 Andersen,Chung,Lang 的社区发现算法[74]提出了 ACL 检测方案，将 Sybil 检测方案由区分正常节点与 Sybil 节点转变为判断正常节点所在社区内其他节点的正常概率，并给出了数学证明[75]，且实验结果表明 ACL 的检测效果比 SybilLimit 更好。

3) 其他

除了利用随机游走和社区发现的方案，还有利用其他算法的 Sybil 检测方案。SumUp[52]是检测在投票情景下的 Sybil 节点，首先设立一个投票收集者（Vote Collector），用最大流（max flow）理论给不同的节点分配不同的票数，通过这样的方法来收集大部分正常（可信）节点的投票，拒绝 Sybil 节点的投票，对于有多次违规行为的节点就判定为 Sybil 节点。SybilBelief[55]利用马尔科夫随机场来检测 Sybil 帐号，首先给网络中未知的节点都分配一个随机值来标识是正常节点还是 Sybil 节点，然后通过马尔科夫随机场来计算每个节点的后验概率，即节点是正常节点的概率。

Sybil 检测方案都是基于一定的假设，但是有些工作表明在一些现实网络中这些假设并不一定成立[23][76]。现有的 Sybil 检测方案都是基于假设 1 即正常节点组成的网络是快速融合的，而且攻击边是有限的，但是研究发现在现实网络中攻击者一方面能够通过发出大量建立联系的请求来增加与正常帐号的联系[77]；另一方面正常帐号很容易被欺骗与攻击者创建的 Sybil 帐号建立联系[18]。Yang 等人[34]的实验也发现 Sybil 帐号之间并不存在紧密的联系，反而与正常帐号之间联系紧密，Sybil 帐号 75%的连接是与正常帐号。Koll 等人[78]对现有的检测方案在假设不成立情况下的效果进行了验证，发现在假设不成立的情况下，现有检测方案的效果并不理想。针对在实验过程中现实网络并不是快速融合的问题，Hu 等人提出在实验过程中需要删掉度数小于一定阈值的节点，剩下的网络就符合快速融合[66]。

虽然现有的 Sybil 检测算法在线下的实验中取得了较好的效果，但 Alvisi 等人[65]表明对于经验丰富的攻击者来说目前的 Sybil 检测方案并不实用，因此建议结合其他检测方式或社交网络中的其他信息来更准确的描述用户之间的关系。Boshmaf 等人[79]通过结合用户行为特征的有监督学习，首先判断帐号是受害者（Victim）的概率，即与 Sybil 帐号连接的帐号，然后利用修改的随机游走算法，通过降低对受害者帐号的转移概率来减少信任值到达 Sybil 区域的概率，对于 k 条攻击边最多引入 $O(\mathrm{vol}(k)\log n)$ 个 Sybil 节点，其中 $\mathrm{vol}(k)$ 是 k 条攻击边的权值之和，通过在西班牙最大的社交网站 Tuenti 中大规模部署，检测出大量 Sybil 帐号。

（2）其他关系图

社交网络中除了显性的好友关系图，还存在大量利用其他关系组成的隐性关系图。一些工作利用隐性关系组成的图来检测异常帐号。

Xue 等人[10]通过帐号发出的好友请求以及好友请求是否被接受来构成好友请求图和好友接受图。利用异常帐号发出大量的好友请求但是好友接受率很低的特征，通过初始的正常帐号种子采用类似 PageRank 的算法对图中其他帐号进行评价，从而检测异常帐号。Tan 等人[80]利用 URL 链接组成的图结合好友关系图来检测异常帐号，首先根据帐号发布的消息中嵌入的 URL 建立 URL 链接图，即发布相同 URL 的帐号之间建立联系，然后根据帐号之间的好友关系建立好友关系图，对好友关系图采用图挖掘的方法找到正常帐号，将这些帐号发布的 URL 作为白名单，并在 URL 链接图中将白名单中 URL 所建立的联系去掉，最后在 URL 链接图中剩下的度数较高的帐号即为异常帐号。Beutel 等人[81]将 Facebook 中帐号的点赞行为抽象为一个二分图，图的一部分为帐号，另一部分为被点赞的页面，如果帐号对某个页面点赞，那么该帐号与这个页面就有一条边相连接。由于攻击者所控制的异常帐号是有限的，为了获取更大的利益，异常帐号会集中对某

些页面进行点赞，组成了二分图的核，因此通过在二分图中寻找核来检测异常帐号。

基于图的检测方案利用异常帐号与正常帐号在组成的图结构中的不同来进行检测，与基于行为特征或者基于内容的检测方案相比，基于图的检测方案本质上是无监督学习方案，因此不需要提前对样本进行训练，但是单纯的利用图结构来检测异常帐号存在较大的误检率，而且只能检测能够组成图结构的异常行为。同时有些基于图的检测方案需要基于一定的假设，但由于现实中社交网络的复杂性，不同的社交网络中异常帐号具有不同的特征，因此在应用基于图的检测方案时需要对检测方案的假设进行严格的验证。在具体的检测方案设计时可以考虑结合其他检测方案来提高准确率和检测范围。

3.4 无监督学习的检测方案

基于行为特征和基于内容的检测方案都是有监督学习的方案，即对分类器的训练需要提前对帐号是否异常进行标记，因此有监督学习的方法需要花费大量的时间来标记异常帐号，而且标记的样本数量与质量对于检测结果有较大的影响。基于图的检测方案尽管是无监督学习的，但是需要构建图结构。无监督学习的检测方案不需要提前对数据进行标记，因此能够更快的形成检测系统。根据具体的算法我们将无监督学习的方案分为两类：基于聚类和基于模型。

（1）基于聚类

基于聚类的方案将异常帐号检测看为数据挖掘中聚类问题。如图 5 所示，通过对帐号的某些特征进行聚类，将正常帐号聚为一类，而不在类中的帐号即为异常帐号；或正常帐号聚为一类同时异常帐号也聚为一类，通过对类中帐号进行抽样验证就能够判断该类内的其他帐号是否为异常。因此不需要提前对样本数据进行标识。

基于聚类的检测方案的关键是选择合适的特征对帐号进行聚类。Miller 等人[82]对 Twitter 中用户个人信息特征以及微博文本内容的特征进行聚类，利用 StreamKM++与 DenStream 结合的数据流聚类算法将正常帐号聚为一类，类之外的即异常帐号，实验结果能够达到 100%的召回率以及 97.8%的准确率，但是所采用的训练集和测试集只有 1,500 左右，而且利用的特征主要是微博文本内容的特征，只能够检测发送恶意消息的 Spam 帐号，无法检测其他类型的异常帐号。

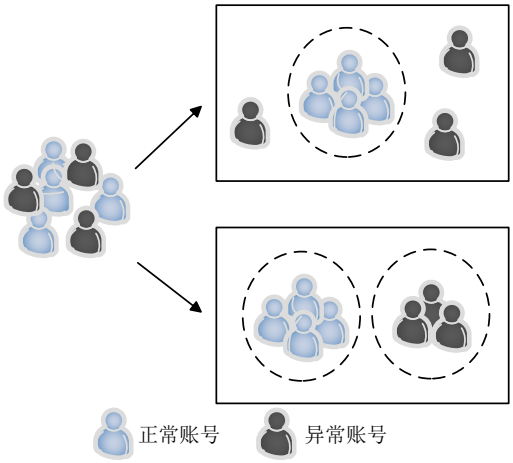

图 5 基于聚类的检测方案

Wang 等人[83]利用用户访问社交网络时的 HTTP 请求序列进行聚类。通过用户的点击序列和时间间隔构建用户的点击行为图结构，再利用图聚类的方法对点击行为图进行聚类，根据类中是否有初始设置的正常帐号种子判断类中其他帐号是否为异常帐号。针对人人网以及 LinkedIn 的真实数据实验表明该方案能够检测未知的异常帐号，但该方案需要对大量帐号长期的点击序列进行训练，如果训练样本中时间序列较短就会有较高的漏检率。

Cao 等人[84]认为异常用户的行为在社交网络中表现出松散的同步行为，而正常用户在同一时间段内表现出离散的行为，因此通过对帐号的行为进行聚类来检测异常帐号。首先利用 Jaccard 计算每两个用户在某个时间段内行为的相似度，然后根据相似度对用户进行聚类，对于类内成员数超过阈值的类就判断该类内的帐号为异常帐号。通过在 Facebook 中的部署测试，该方案能够达到将近 100%的准确率，而且能够检测未知的异常行为，但该方案在计算用户相似度时需要结合 IP 地址的约束，对相同登陆 IP 的帐号进行聚类，而攻击者可以通过设置代理等方式绕过这项约束。

（2）基于模型

基于模型的检测方案的基础是认为正常帐号的行为符合某种模型，而异常帐号的行为不符合这种模型，因此基于模型的方案的关键是抽取合适的特征对正常帐号进行训练，形成相应的模型，然后根据其他帐号是否与模型匹配来判断是否为异常帐号。

表 4 检测方案对比

方案	优点	缺点

有监督学习方法	基于行为特征	1. 检测速度快，效率高 2. 准确率高 3. 算法设计、部署相对成熟	1. 需要提前训练分类器 2. 检测未知攻击方式能较差 3. 与恶意行为存在延迟 4. 易被攻击者绕过
	基于内容	1. 准确率高 2. 实时性高 3. 检测速度快 4. 算法设计、实施相对成熟	1. 只能检测发布恶意信息的异常帐号 2. 易被攻击者绕过 3. 检测未知异常行为能力差
无监督学习方法	基于图	1. 只需要图结构 2. 不需要提前训练 3. 能够检测未知异常行为	1. 准确率相对较低 2. 不同类型的社交网络存在差异 3. 只能检测具有图结构的异常行为 4. 与恶意行为存在延迟 5. 理论研究较多，在现实中部署较少
	无监督	1. 不需要提前训练 2. 能够检测未知异常行为 3. 不易被攻击者绕过	1. 准确率相对较低 2. 与恶意行为存在延迟 3. 算法设计、实施较复杂

Viswanath 等人[85]发现社交网络中帐号的一些行为（时域、空域、时-空域行为）能够用低维数据表达，因此通过主成分分析算法对正常帐号的行为建立模型，然后根据其他帐号与模型之间的偏离程度来判断是否为异常，对 Facebook 中点赞行为的实验能够发现样本中 94.3%的异常帐号。Jiang 等人[86][87]对社交网络中异常行为定义了两个衡量指标：（1）同步值，即社交网络中异常帐号经常具有相同的行为，比如共同关注一些帐号等；（2）异常值，即这些帐号的行为与大部分其他帐号的行为不同。通过计算这两个值，对于低于阈值的帐号判断为异常帐号。针对 Twitter 以及腾讯微博的实验表明准确率能够达到 99.8%，但是只适用于具有图结构的异常行为如恶意关注等。

无监督学习的检测方案是目前异常帐号检测的新方向。无监督学习方案不需要提前对样本进行标识，因此能够检测到未知的恶意行为。但目前的无监督检测方案也有各自不足，如采用文本内容特征聚类只能检测传播恶意信息的异常帐号，而采用帐号的点击序列聚类，需要对大量数据进行训练，不适合在拥有海量用户的社交网络中部署。Cao 等人[84]对帐号的行为进行聚类需要 IP 地址的约束，而攻击者能够通过代理更改不同异常帐号的登陆 IP 或增加异常帐号的行为时间间隔来绕过约束。Viswanath 等人[85]也是对帐号的行为进行建模，但是对 Compromised 帐号的检测效果较差，而且攻击者可以通过执行大量的正常行为进行伪装通过检测。Jiang 等人[86]的工作只能针对形成图结构的恶意行为。

3.5 检测方案对比

各类检测方案各有优劣，有各自的应用场景，如表 4 所示，我们列出了不同检测算法各自的优点和缺点。基于行为特征和基于内容的检测方案是有监督学习方案，优点是只要训练形成了分类器，就能够对异常帐号进行检测而且能够区分不同类别的异常帐号，检测准确率较高。但由于是有监督学习方法，需要提前对样本数据进行标记，这将花费大量的时间和人力成本，而且只能够检测已知的攻击类型，在攻击者改变攻击方式后就无法检测[88]。基于内容的检测方案能够做到对异常帐号实时检测，但是只能够检测发布恶意消息的异常帐号。基于行为特征和基于内容的检测方案目前比较成熟，不单单有理论的研究还有现实中大规模的部署。基于图的检测方案利用了图结构特征，具有较强的鲁棒性，但是需要建立相应的图结构，而且只能检测与其他帐号有联系的恶意行为，准确率较低，目前只是理论研究较多，在现实中大规模部署较少。无监督学习方案不需要提前对样本数据进行标记，能够较快形成检测系统，同时能够检测未知的攻击行为，且不易被攻击者绕过，但是这种检测方案不容易区分不同类型的异常帐号，而且需要对大量数据进行学习。因此社交网站可以结合多种不同的检测方案，从不同的层次对异常帐号进行检测，如可以先采用无监督学习的方案检测未知的攻击行为，然后对攻击行为抽取特征，再利用有监督学习的方案进行检测。

3.6 实验方法总结

在线社交网络中异常帐号检测方案都需要相应的实验进行验证，因此需要获取大量的帐号信息，但由于社交网站商业保密以及对用户隐私的保护，获取大量帐号的信息比较困难。同时有监督学习检测方案需要对获取的数据进行标识才能够进行训练。在实验的最后也需要对检测结果的效果进行分析，因此需要对检测结果进行验证。这些实验方法的选择对于检测方案的设计和验证至关重要。根据目前的相关工作，我们分别总结了在实验过程中数据获取方式、数据标识方式以及结果验证方式。

3.6.1 数据获取方式

检测方案的实现与验证都需要大量的真实数据，数据的获取方式主要有以下几种：

（1） 爬虫获取。社交网站都提供了相应的API，能够直接利用爬虫调用 API 获取帐号信息，但是社交网络对 API 的使用有一定的限制，可以要求社交网络服务提供商将相应的 IP 地址加入白名单来突破限制。也可利用网络爬虫直接获取，但是这种方式只能获取帐号的公开数据而且需要选择合适的抽样算法对社交网络中用户进行爬取，因为有些抽样算法无法准确描述社交网络的整体特性[89]。

（2） 公开数据集。一些学者公开了自己所获取的数据集，有些机构汇总了相关数据，如 SNAP[1]等，Xu 等人[90]对社交网络相关的公开数据集进行了总结。因此可以利用公开数据集进行实验。

（3） 与社交网站合作。社交网站拥有全部的用户数据，因此可以通过与社交网站合作，由社交网站提供所需的数据。

表 5 数据获取方式

方法	工作
爬虫获取	[21][22][37][38][42][88]
公开数据集	[47][51][81][86]
与社交网站合作	[30][34][41][80][81][82]

表 6 数据标识方式

方法	工作
人工标识	[21][34][37][38][80]
URL 黑名单检测工具	[33][34] [40][42] [88]
蜜罐系统	[35][43]
地下市场购买	[5][81]
社交网络自身功能	[37][43] [47]

如表 5 所示，我们列出了利用各种数据获取方式的相关工作。通过爬虫获取数据是最常见的数据获取方式，而且能够根据实验需求获取指定数据，但是需要耗费一定的人力成本来编写相应的爬虫以及一定的机器时间来爬取数据。利用公开的数据集可以节约时间成本，而且能够在相同数据集上与其他工作进行对比，但是公开的数据集与实验所需的数据内容不一定完全符合，会对实验结果有一定的影响。通过与社交网站合作能够获取大量自己所需的数据，但是与社交网站建立联系比较困难。

3.6.2 数据标识方式

有监督学习的检测方案需要提前对获取的数据标识为正常或者异常，而且在整个社交网络中异常帐号占的总体比例较少，因此对于分类器的训练有时需要标识额外的异常帐号。对于无监督学习的检测方案也需要在获取的数据中将异常数据剔除。因此需要对获取的数据进行标识，现有的工作主要用到以下方法：

（1） 人工标识。通过访问帐号的主页等其他内容，由人工手动判断是否为异常帐号。

（2） URL 黑名单检测工具。通过 URL 黑名单列表来确定帐号发布消息中的 URL 是否为恶意URL，进而确定帐号是否为异常帐号。一些公开的URL 黑名单在本文 3.2 节已经介绍。

（3） 蜜罐系统。在社交网络中建立蜜罐系统，能够获取主动添加好友以及发送恶意消息的异常帐号。

（4） 地下市场购买。通过在地下市场中直接购买所需的服务，如虚假粉丝或者点赞，根据购买到的服务来确定异常帐号。

（5） 社交网络自身功能。Twitter 自身提供了@spam 的功能，用户能够向 Twitter 报告可疑的异常帐号。通过获取@spam 的相关内容就能够获取用户报告的大量异常帐号。

如表 6 所示，我们列出了采用各种数据标识方式的相关工作。这些方式有各自的优势也有各自的不足。人工标识准确率高，但需要花费大量的时间和人力，只适用于少量数据集。URL 黑名单检测的方式尽管时间成本不高，但是存在一定的延时。蜜罐系统需要建立多个蜜罐，而且需要长时间运行才能够获取到足够的异常帐号。通过地下市场购买能够在短时间内获取大量异常帐号，但只能获取部分形式的异常帐号，无法获取攻击者自己使用的异常帐号。利用社交网络自身功能会存在一定的误报，并且需要爬虫进行实时获取，否则帐号被禁用之后就无法获取其具体内容。因此在实验过程中可以考虑综合使用这些方法，根据具体检测方案的目的结合使用不同的标识方式。

3.6.3 结果验证方式

异常帐号检测方案的实验结果需要进行验证才能够证明检测方案是否有效以及具体的检测效果，即需要判断检测结果中的异常帐号是否的确为异常帐号。主要的结果验证方式有：

（1） 人工验证。通过手动检查检测结果中异常帐号的行为或者消息内容，人工判断是否为异常帐号。

（2） 与社交网站合作。通过与社交网站合作，

[1] Stanford Large Network Dataset Collection http://snap.stanford.edu/data/index.html. 2015,3,25.

将检测结果提供给社交网站，让社交网站的安全团队对这些帐号进行验证。

（3） 社交网站自身功能。利用社交网站自身的检测系统，通过在一段时间间隔后再访问所检测到的异常帐号，根据其是否被禁用来判断帐号是否正常。

表 7 结果验证方式

方法	工作
人工验证	[22][30][33][41][82]
与社交网站合作	[10][34][81]
社交网站自身功能	[21][32]

如表 7 所示，我们汇总了具体方案的结果验证方式，这些结果验证方式各有优劣，人工验证方式是最常见也是最准确的方式，但是需要花费大量的人力和时间成本。与社交网站合作能够节约时间和人力成本，但可能会由于商业保密等原因无法获取精确的检测效果。利用网站自身功能误差较大，而且有些异常帐号会存在很长时间[91]。在实验过程中可以根据具体的检测方案，结合使用不同的结果验证方式，如可以抽样一部分检测结果进行人工验证，获取样本的精确检测效果，然后将整体的检测结果提交社交网站获取整体的检测效果。

4 研究展望

在线社交网络中异常帐号检测是一个不断交替发展的过程，如前所述当前的研究工作面临着巨大的挑战，尽管现有的检测方案针对这些挑战提出了一些解决方案，如无监督学习方案能够检测未知攻击行为，以及专门针对众包平台、短网址服务等的研究工作，但是这些工作依然存在不足。展望未来，我们认为以下方向将是未来的研究核心：

（1） 发现并检测新的攻击方式

通过分析社交网络中异常帐号攻击形式的发展，我们发现社交网络总会面临新的攻击方式。首先攻击者在利益的驱动下，面对社交网络的检测系统会改变原有的攻击方式，形成新的攻击方式，如攻击者为了绕过检测由原来直接发布文本型广告消息发展为在照片中嵌入恶意广告进行传播。其次社交网络推出的新功能也会带来新的攻击方式，如对页面的点赞功能出现了专门的恶意点赞异常帐号，以及利用热门话题的功能出现了专门针对热门话题的垃圾信息[92]。最后新的网络资源也会带来新的攻击方式，攻击者会利用各种网络资源来辅助进行恶意行为，如短网址、众包平台等网络资源的出现为攻击者提供了新的攻击方式。新的攻击方式具有新的特征，利用异常帐号之前的行为特征训练的检测系统对于新的攻击方式不再适用。尽管无监督学习检测方案的出现能够检测一定的新形式攻击，但是现有的无监督学习检测方案有各自的不足。因此需要时刻关注异常帐号的发展，及时发现攻击者在社交网络中新的攻击方式并从中提取相应的特征进行检测，降低对社交网络正常用户的危害。

（2） 结合多特征多维度的检测模型

目前检测方案所采用的特征大都为静态特征而且特征类别比较单一，这种检测方案只能够检测到有限类别的异常帐号，而且这些特征的鲁棒性不足，攻击者能够采用各种方式绕过检测，因此需要深入分析正常帐号和异常帐号行为方式的区别，挖掘鲁棒性较高的特征来进行检测。同时异常帐号在社交网络中的操作具有时间维度，每一次恶意行为都具有时间属性，而且攻击者为了获取更多利益，对异常帐号的操作会与正常帐号存在差异，如操作行为频繁、操作时间与正常帐号不同等，而且这种时间序列的特征难以绕过，因此可以在用户行为特征中加入时间维度，设计结合时间维度的多特征检测模型。

（3） 高效的轻量级、分布式并行检测算法

社交网络拥有海量用户，现有的检测方案，无论是有监督学习还是无监督学习方案都需要对每个帐号都进行处理，这样会浪费大量时间，而异常帐号在社交网络中只占一小部分，因此可以设计轻量级检测方案，利用异常帐号之间的联系根据已经检测到的异常帐号发现更多未知的异常帐号，而不需要对每个帐号都进行检测。同时无论检测模型的线下训练还是检测系统的线上实时处理，都需要及时性才能够有效降低对正常用户的损害，因此需要设计高效的检测算法，能够适用于各种分布式并行计算平台，从而提高检测效率，降低检测时间。

5 总结

随着社交网络的飞速发展，越来越多的攻击者将目光集中在社交网络中，其中异常帐号带来的各种危害严重威胁到正常用户的信息安全以及社交网站的安全发展，为此学术界和工业界都提出了大量的检测方案。我们介绍了社交网络中异常帐号的表现形式以及当前检测工作所面临的巨大挑战，对目前的检测方案分为基于行为特征、基于内容、基于图、无监督学习四类，并分别对具体方案做了总

结、对比，随后汇总了检测实验中数据获取、数据标识以及结果验证的主要方法，最后探讨了未来的研究趋势。

参考文献：

[1] Ellison N B. Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communication, 2007, 13(1): 210-230.

[2] Gao H, Hu J, Huang T, et al. Security issues in online social networks. IEEE Internet Computing, 2011, 15(4): 56-63.

[3] Fire M, Goldschmidt R, Elovici Y. Online social networks: threats and solutions survey. IEEE Communications Surveys and Tutorials, 2013, 16(4): 2019-2036.

[4] Caviglione L, Coccoli M, Merlo A. A taxonomy–based model of security and privacy in online social networks. International Journal of Computational Science and Engineering, 2014, 9(4): 325-338.

[5] Thomas K, McCoy D, Grier C, et al. Trafficking fraudulent accounts: the role of the underground market in twitter spam and abuse//Proceedings of the 22rd USENIX Security Symposium. Washington, USA, 2013: 195-210.

[6] Huang T K, Rahman M S, Madhyastha H V, et al. An analysis of socware cascades in online social networks//Proceedings of the 22nd international conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 619-630.

[7] Chu Z, Gianvecchio S, Wang H, et al. Who is tweeting on Twitter: human, bot, or cyborg?//Proceedings of the 26th annual computer security applications conference. Austin, USA, 2010: 21-30.

[8] Kanich C, Kreibich C, Levchenko K, et al. Spamalytics: An empirical analysis of spam marketing conversion//Proceedings of the 15th ACM conference on Computer and communications security. Alexandria, USA, 2008: 3-14.

[9] Stringhini G, Wang G, Egele M, et al. Follow the green: growth and dynamics in twitter follower markets//Proceedings of the 2013 conference on Internet measurement conference. Barcelona, Spain, 2013: 163-176.

[10] Xue J, Yang Z, Yang X, et al. VoteTrust: leveraging friend invitation graph to defend against social network sybils//Proceedings of the 32nd IEEE International Conference on Computer Communications. Turin, Italy, 2013: 2400-2408.

[11] De Cristofaro E, Friedman A, Jourjon G, et al. Paying for likes?: understanding facebook like fraud using honeypots//Proceedings of the 2014 Conference on Internet Measurement Conference. Vancouver, Canada, 2014: 129-136.

[12] Thomas K, Grier C, Paxson V. Adapting social spam infrastructure for political censorship//Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats. Berkeley, USA, 2012: 13-13.

[13] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Computing Surveys, 2009, 41(3): 15.

[14] Akoglu L, Tong H, Koutra D. Graph-based anomaly detection and description: a survey. Data Mining and Knowledge Discovery, 2014, 5:1-17.

[15] Savage D, Zhang X, Yu X, et al. Anomaly detection in online social networks. Social Networks, 2014, 39: 62-70.

[16] Mo Qian, Yang Ke. Overview of web spammer detection. Ruan Jian Xue Bao/Journal of Software, 2014, 25(7):1505-1526(in Chinese)
(莫倩，杨珂．网络水军识别研究．软件学报，2014, 25(7): 1505-1526)

[17] Stein T, Chen E, Mangla K. Facebook immune system//Proceedings of the 4th Workshop on Social Network Systems. Salzburg, Austria, 2011: 8.

[18] Boshmaf Y, Muslukhov I, Beznosov K, et al. The socialbot network: when bots socialize for fame and money//Proceedings of the 27th Annual Computer Security Applications Conference. Orlando, USA, 2011: 93-102.

[19] Douceur J R. Peer-to-peer systems: The sybil attack. Berlin Heidelberg: Springer, 2002.

[20] Yu H, Kaminsky M, Gibbons P B, et al. Sybilguard: defending against sybil attacks via social networks. IEEE Transactions on Networking, 2008, 16(3): 576-589.

[21] Zangerle E, Specht G. “Sorry, I was hacked” a classification of compromised twitter accounts//Proceedings of the 29th Annual ACM Symposium on Applied Computing. Gyeongju, Korea, 2014: 587-593

[22] Gao H, Hu J, Wilson C, et al. Detecting and characterizing social spam campaigns//Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. Melbourne, Australia, 2010: 35-47.

[23] Bilge L, Strufe T, Balzarotti D, et al. All your contacts are belong to us: automated identity theft attacks on social networks//Proceedings of the 18th international conference on World wide web. 2009: 551-560.

[24] Wang G, Mohanlal M, Wilson C, et al. Social turing tests: crowdsourcing sybil detection//Proceedings of the 19th Annual Network & Distributed System Security Symposium. San Diego, USA, 2012.

[25] Motoyama M, McCoy D, Levchenko K, et al. Dirty jobs: The role of freelance labor in web service abuse//Proceedings of the 20th USENIX conference on Security. San Francisco, USA, 2011: 14-14.

[26] Wang G, Wilson C, Zhao X, et al. Serf and turf: crowdturfing for fun and profit//Proceedings of the 21st international conference on World

Wide Web. Lyon, France, 2012: 679-688.

[27] Maggi F, Frossi A, Zanero S, et al. Two years of short urls internet measurement: security threats and countermeasures//Proceedings of the 22nd international conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 861-872.

[28] Lee S, Kim J. Early filtering of ephemeral malicious accounts on Twitter. Computer Communications, 2014, 54: 48-57.

[29] Ni Ping, Zhang Yu-Qing, Wen Guan-Xing, et al. Detection of socialbot networks based on population characteristics. Journal of University of Chinese Academy of Sciences, 2014, 31(5): 691-700. (in Chinese)
(倪平, 张玉清, 闻观行, 等. 基于群体特征的社交僵尸网络检测方法. 中国科学院大学学报, 2014, 31(5): 691-700.)

[30] Rahman M S, Huang T K, Madhyastha H V, et al. Efficient and scalable socware detection in online social networks//Proceedings of 10th USENIX Symposium on Networked Systems Design and Implementation. San Jose, USA, 2012: 663-678.

[31] Zheng X, Zeng Z, Chen Z, et al. Detecting spammers on social Networks. Neurocomputing, 2015, 159(2): 27-34.

[32] Amleshwaram A A, Reddy N, Yadav S, et al. CATS: characterizing automation of twitter spammers//Proceedings of the Fifth International Conference on Communication Systems and Networks. Bangalore, India, 2013: 1-10.

[33] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks//Proceedings of the 26th annual computer security applications conference. Austin, USA, 2010: 1-9.

[34] Yang Z, Wilson C, Wang X, et al. Uncovering social network sybils in the wild. ACM Transactions on Knowledge Discovery from Data, 2014, 8(1): 2.

[35] Lee K, Caverlee J, Webb S. Uncovering social spammers: social honeypots+ machine learning//Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. Geneva, Switzerland, 2010: 435-442.

[36] Benevenuto F, Magno G, Rodrigues T, et al. Detecting spammers on twitter//Proceedings of the Collaboration, electronic messaging, anti-abuse and spam conference. Redmond, USA, 2010, 6: 12.

[37] Lee K, Eoff B D, Caverlee J. Seven months with the devils: a long-term study of content polluters on twitter//Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 185-192.

[38] Wang A H. Don't follow me: spam detection in twitter//Proceedings of the 2010 International Conference on Security and Cryptography. Athens, Greece, 2010: 1-10.

[39] Ahmed F, Abulaish M. A generic statistical approach for spam detection in Online Social Networks. Computer Communications, 2013, 36(10): 1120-1129.

[40] Yang C, Harkreader R C, Gu G. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers//Proceedings of the 14th international conference on Recent Advances in Intrusion Detection. California, USA, 2011: 318-337.

[41] Hu X, Tang J, Liu H. Online social spammer detection//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Quebec, Canada, 2014: 59-65.

[42] Grier C, Thomas K, Paxson V, et al. @ spam: the underground on 140 characters or less//Proceedings of the 17th ACM conference on Computer and communications security. Chicago, USA, 2010: 27-37.

[43] Thomas K, Grier C, Ma J, et al. Design and evaluation of a real-time url spam filtering service//Proceedings of the Symposium on Security and Privacy. Oakland, USA, 2011: 447-462.

[44] Egele M, Stringhini G, Kruegel C, et al. COMPA: detecting compromised accounts on social networks//Proceedings of the 20th Annual Network & Distributed System Security Symposium. San Diego, USA. 2013.

[45] Gao H, Yang Y, Bu K, et al. Spam ain't as diverse as it seems: throttling OSN spam with templates underneath//Proceedings of the 30th Annual Computer Security Applications Conference. New Orleans, USA, 2014: 76-85.

[46] Chu Z, Widjaja I, Wang H. Detecting social spam campaigns on twitter//Proceedings of the Applied Cryptography and Network Security. Singapore, 2012: 455-472.

[47] Lee S, Kim J. WarningBird: detecting suspicious urls in twitter stream//Proceedings of the 19th Annual Network & Distributed System Security Symposium. San Diego, USA, 2012.

[48] Song J, Lee S, Kim J. Spam filtering in twitter using sender-receiver relationship//Proceedings of the 14th international conference on Recent Advances in Intrusion Detection. California, USA, 2011: 301-317.

[49] Jiang M, Cui P, Beutel A, et al. Inferring strange behavior from connectivity pattern in social networks//Proceedings of the 18th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Taiwan, China, 2014: 126-138.

[50] Yu H, Gibbons P B, Kaminsky M, et al. Sybillimit: a near-optimal social network defense against sybil attacks//Proceedings of the Symposium on Security and Privacy. Oakland, California, USA, 2008: 3-17.

[51] Danezis G, Mittal P. SybilInfer: detecting sybil nodes using social networks//Proceedings of the 16th Annual Network & Distributed System Security Symposium. San Diego, CA, USA, 2009.

[52] Tran D N, Min B, Li J, et al. Sybil-resilient online content voting//

Proceedings of the.6th USENIX Symposium on Networked Systems Design and Implementation. Boston, USA, 2009, 9(1): 15-28.

[53] Tran N, Li J, Subramanian L, et al. Optimal sybil-resilient node admission control//Proceedings of the 30th IEEE International Conference on Computer Communications. Shanghai, China, 2011:3218-3226.

[54] Wei W, Xu F, Tan C C, et al. Sybildefender: defend against sybil attacks in large social networks// Proceedings of the 31th IEEE International Conference on Computer Communications. Orlando, USA, 2012:1951-1959

[55] Cao Q, Sirivianos M, Yang X, et al. Aiding the detection of fake accounts in large scale social online services//Proceedings of 10th USENIX Symposium on Networked Systms Design and Implementation. San Jose, USA, 2012: 197-210.

[56] Gong N, Frank M, Mittal P. SybilBelief: a semi-supervised learning approach for structure-based sybil detection. IEEE Transactions on Information Forensics and Security, 2014, 9(6): 976-987

[57] Kleinberg J. The small-world phenomenon: an algorithmic perspective//Proceedings of the thirty-second annual ACM symposium on Theory of computing. New York, USA, 2000: 163-170.

[58] Mislove A, Post A, Druschel P, et al. Ostra: leveraging trust to thwart unwanted communication//Proceedings of the 5th Symposium on Networked Systems Design and Implementation. San Francisco, USA, 2008, 8: 15-30.

[59] Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks//Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.. San Diego, USA, 2007: 29-42

[60] Wilson C, Boe B, Sala A, et al. User interactions in social networks and their implications//Proceedings of the 4th ACM European conference on Computer systems. Nuremberg, Germany, 2009: 205-218.

[61] Gjoka M, Kurant M, Butts C T, et al. Walking in facebook: A case study of unbiased sampling of OSNs//Proceedings of the 29th IEEE International Conference on Computer Communications. San Diego, USA, 2010: 1-9

[62] Viswanath B, Mislove A, Cha M, et al. On the evolution of user interaction in facebook//Proceedings of the 2nd ACM workshop on Online social networks. Barcelona, Spain, 2009: 37-42.

[63] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: densification laws, shrinking diameters and possible explanations//Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. Chicago, USA, 2005: 177-187

[64] Viswanath B, Post A, Gummadi K P, et al. An analysis of social network-based sybil defenses. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 363-374.

[65] Alvisi L, Clement A, Epasto A, et al. Sok: the evolution of sybil defense via social networks//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2013: 382-396.

[66] Yu H. Sybil defenses via social networks: a tutorial and survey. ACM SIGACT News, 2011, 42(3): 80-101.

[67] Boshmaf Y, Beznosov K, Ripeanu M. Graph-based sybil detection in social and information systems//Proceedings of the ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara, Canada, 2013: 466-473.

[68] Li F, Liu B, Xiao Z, et al. Detecting and defending against sybil attacks in social networks: an overview//Proceedings of the Ninth International Conference on Broadband and Wireless Computing, Communication and Applications. Guangdong, China, 2014: 104-112.

[69] Koll D, Li J, Stein J, et al. On the state of OSN-based Sybil defenses// Proceedings of the IEEE 13th Networking Conference. Trondheim, Norway, 2014: 1-9.

[70] Zhang K, Liang X, Lu R, et al. Sybil attacks and their defenses in the internet of things. IEEE Internet of Things, 2014, 1(5): 372-383

[71] Viswanath B, Mondal M, Clement A, et al. Exploring the design space of social network-based sybil defenses//Proceedings of the Fourth International Conference on Communication Systems and Networks. Bangalore, India, 2012: 1-8.

[72] Mislove A, Viswanath B, Gummadi K P, et al. You are who you know: inferring user profiles in online social networks//Proceedings of the 3th ACM international conference on Web search and data mining. New York, USA, 2010: 251-260

[73] Mohaisen A, Hopper N, Kim Y. Keep your friends close: incorporating trust into social network-based sybil defenses//Proceedings of the 30th IEEE International Conference on Computer Communications. Shanghai, China, 2011: 1943-1951

[74] Andersen R, Chung F, Lang K. Local graph partitioning using pagerank vectors//Proceedings of the 47th IEEE Symposium on Foundations of Computer Science. Washington, USA, 2006: 475-486.

[75] Alvisi L, Clement A, Epasto A, et al. Communities, random walks, and social sybil defense. Internet Mathematics, 2014, 10(3-4): 360-420.

[76] Mohaisen A, Yun A, Kim Y. Measuring the mixing time of social graphs//Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. Melbourne, Australia, 2010: 383-389.

[77] Ghosh S, Viswanath B, Kooti F, et al. Understanding and combating link farming in the twitter social network//Proceedings of the 21st international conference on World Wide Web. Lyon, France, 2012: 61-70.

[78] Koll D, Li J, Stein J, et al. On the effectiveness of sybil defenses based on online social networks//Proceedings of the 21st IEEE International

Conference on Network Protocols. Göttingen, Germany, 2013:147-158.

[79] Boshmaf Y, Logothetis D, Siganos G, et al. Íntegro: leveraging victim prediction for robust fake account detection in OSNs//Proceedings of the 22th Annual Network & Distributed System Security Symposium. San Diego, USA, 2015.

[80] Tan E, Guo L, Chen S, et al. UNIK: unsupervised social network spam detection//Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. San Francisco, USA, 2013: 479-488.

[81] Beutel A, Xu W, Guruswami V, et al. CopyCatch: stopping group attacks by spotting lockstep behavior in social networks//Proceedings of the 22nd international conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 119-130.

[82] Miller Z, Dickinson B, Deitrick W, et al. Twitter spammer detection using data stream clustering. Information Sciences, 2014, 260: 64-73.

[83] Wang G, Konolige T, Wilson C, et al. You are how you click: clickstream analysis for sybil detection//Proceedings of the 22rd USENIX Security Symposium. Washington, USA, 2013: 241-256.

[84] Cao Q, Yang X, Yu J, et al. Uncovering large groups of active malicious accounts in online social networks//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale, USA, 2014: 477-488.

[85] Viswanath B, Bashir M A, Crovella M, et al. Towards detecting anomalous user behavior in online social networks//Proceedings of the 23rd USENIX Security Symposium. San Diego, USA, 2014: 223-238.

[86] Jiang M, Cui P, Beutel A, et al. CatchSync: catching synchronized behavior in large directed graphs//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2014: 941-950.

[87] Jiang M, Cui P, Beutel A, et al. Detecting suspicious following behavior in multimillion-node social networks//Proceedings of the companion publication of the 23rd international conference on World wide web companion. Seoul, Korea, 2014: 305-306.

[88] Zhu Y, Wang X, Zhong E, et al. Discovering spammers in social networks//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 171-177.

[89] Cui Ying-An, Li Xue, Wang Zhi-Xiao, et al. A comparison on methodologies of sampling online social media. Jisuanji Xuebao/Chinese Journal of Computers, 2014, 37(8): 1859-1876. (in Chinese)
(崔颖安, 李雪, 王志晓, 等. 在线社交媒体数据抽样方法的比较研究. 计算机学报, 2014, 37(8): 1859-1876.)

[90] Xu Ke, Zhang Sai., Chen Hao, Li Hai-Tao. Measurement and analysis of online social networks. Jisuanji Xuebao/Chinese Journal of Computers, 2014, 37(2):165-188.(in Chinese)
(徐恪, 张赛, 陈昊, 李海涛. 在线社会网络的测量与分析. 计算机学报, 2014, 37(2):165-188)

[91] Lin P C, Huang P M. A study of effective features for detecting long-surviving Twitter spam accounts//Proceedings of the 15th International Conference on Advanced Communication Technology. PyeongChang, Korea, 2013: 841-846.

[92] Martinez-Romo J, Araujo L. Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 2013, 40(8): 2992-3000.

Zhang Yu-Qing, born in 1966, PH.D. , Professor, Ph.D. supervisor. His research interests include network and information security.

Lv Shao-Qing, born in 1987, Ph.D. candidate. His main research interests include security and privacy of online social networks.

Fan Dan, born in 1982, Ph.D. . Her research interests include network and. protocol security.

Background

With the rapid growth of online social networks (OSNs) for communicating, sharing, storing and managing significant information, it is attracting attackers who misuse the OSNs to exploit vulnerabilities for their illicit benefits. Both fake accounts and compromised accounts are used as tools to attack OSNs, and hard to identify. Many papers have been published on the detection of anomaly in OSNs. Since anomaly detection is playing an increasingly important role in the security of OSNs, the purpose of this paper is to survey existing techniques, and to outline the types of challenges that can be addressed.

This paper reviewed important achievements made by computer scientists in anomaly detection in recent years. The authors elaborated the behaviors of different type of anomalies and the grand challenges towards detection. Then we grouped existing techniques into four categories based on the underlying approaches adopted by each techniques which are feature-based, content-based, graph-based and unsupervised approaches. For each category, we identify the key assumptions, which are used by the techniques to distinguish between normal and anomalies. We also identify the advantages and disadvantages of the techniques in each category. Further, major methods of collecting datasets, labeling anomalies and validating results are introduced. Finally, we discussed future research directions of anomaly detection in OSNs.

This work is supported by the Natural Science Foundation of China under Grant No.61272481, No.61402434, China Postdoctoral Science Foundation Funded Project under Grant No.2014M550085, Open Fund of State Key Laboratory of Information Security.