# Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?

Zi Chu, Steven Gianvecchio, Haining Wang, *Senior Member*, *IEEE*, and Sushil Jajodia, *Senior Member*, *IEEE*

**Abstract**—Twitter is a new web application playing dual roles of online social networking and microblogging. Users communicate with each other by publishing text-based posts. The popularity and open structure of Twitter have attracted a large number of automated programs, known as bots, which appear to be a double-edged sword to Twitter. Legitimate bots generate a large amount of benign tweets delivering news and updating feeds, while malicious bots spread spam or malicious contents. More interestingly, in the middle between human and bot, there has emerged cyborg referred to either bot-assisted human or human-assisted bot. To assist human users in identifying who they are interacting with, this paper focuses on the classification of human, bot, and cyborg accounts on Twitter. We first conduct a set of large-scale measurements with a collection of over 500,000 accounts. We observe the difference among human, bot, and cyborg in terms of tweeting behavior, tweet content, and account properties. Based on the measurement results, we propose a classification system that includes the following four parts: 1) an entropy-based component, 2) a spam detection component, 3) an account properties component, and 4) a decision maker. It uses the combination of features extracted from an unknown user to determine the likelihood of being a human, bot, or cyborg. Our experimental evaluation demonstrates the efficacy of the proposed classification system.

**Index Terms**—Automatic identification, bot, cyborg, Twitter, social networks

## 1 Introduction

Twitter is a popular online social networking and microblogging tool, which was released in 2006. Remarkable simplicity is its distinctive feature. Its community interacts via publishing text-based posts, known as *tweets*. The tweet size is limited to 140 characters. Hashtag, namely words or phrases prefixed with a # symbol, can group tweets by topic. For example, #Justin Bieber and #Women's World Cup are the two trending hashtags on Twitter in 2011 [1]. Symbol @ followed by a username in a tweet enables the direct delivery of the tweet to that user. Unlike most online social networking sites (i.e., Facebook and MySpace), Twitter's user relationship is directed and consists of two ends, friend and follower. In the case where the user A adds B as a friend, A is a *follower* of B while B is a *friend* of A. In Twitter terms, A follows B (namely, the following relationship is unidirectional from A to B). B can also add A as his friend (namely, following back or returning the follow), but is not required. When A and B follow each other, the relationship becomes bidirectional. From the standpoint of information flow, tweets flow from the source (author) to subscribers (followers). More specifically, when a user posts tweets, these tweets are displayed on both the author's homepage and those of his followers.

As reported in August 2011, Twitter has attracted 200 million users and generated 8.3 million Tweets per hour [2]. It ranks the 10th on the top 500 site list according to Alexa in December 2011 [3]. In November 2009, Twitter emphasized its value as a news and information network by changing the question above the tweet input dialog box from "What are you doing" to "What's happening." To some extent, Twitter has transformed from a personal microblogging site to an information publishing venue. Many traditional industries have used Twitter as a new media channel. We have witnessed successful Twitter applications in business promotion [4], customer service [5], political campaigning [6], and emergency communication [7], [8].

The growing user population and open nature of Twitter have made itself an ideal target of exploitation from automated programs, known as bots. Like existing bots in other web applications (i.e., Internet chat [9], blogs [10] and online games [11]), bots have been common on Twitter. Twitter does not inspect strictly on automation. It only requires the recognition of a CAPTCHA image during registration. After gaining the login information, a bot can perform most human tasks by calling Twitter APIs. More interestingly, in the middle between humans and bots have emerged cyborgs, which refer to either bot-assisted humans or human-assisted bots. Cyborgs have become common on Twitter. After a human registers an account, he may set automated programs (i.e., RSS feed/blog widgets) to post tweets during his absence. From time to time, he participates to tweet and interact with friends. Different from bots which greatly use automation, cyborgs interweave characteristics of both manual and automated behavior.

- *Z. Chu is with Twitter, Inc., 1355 Market St., Suite 900, San Francisco, CA 94103. E-mail: jchu@twitter.com.*
- *S. Gianvecchio is with the MITRE Corporation, 7515 Colshire Dr., McLean, VA 22102. E-mail: gianvecchio@mitre.org.*
- *H. Wang is with the Department of Computer Science, College of William and Mary, Williamsburg, VA 23185. E-mail: hnw@cs.wm.edu.*
- *S. Jajodia is with the Center for Secure Information Systems, George Mason University, Fairfax, VA 22030. E-mail: jajodia@gmu.edu.*

*Manuscript received 20 Dec. 2011; revised 06 June 2012; accepted 14 Aug. 2012; published online 22 Aug. 2012.*
*For information on obtaining reprints of this article, please send e-mail to: tdsc@computer.org, and reference IEEECS Log Number TDSC-2011-12-0273. Digital Object Identifier no. 10.1109/TDSC.2012.75.*

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.
1545-5971/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society
IPR2023-01160
KASPERSKY EX1033 PAGE 1

Automation is a double-edged sword to Twitter. On one hand, legitimate bots generate a large volume of benign tweets, like news and blog updates. This complies with the Twitter's goal of becoming a news and information network. On the other hand, malicious bots have been greatly exploited by spammers to spread spam. The definition of spam in this paper is spreading malicious, phishing, or unsolicited commercial content in tweets. These bots randomly add users as their friends, expecting a few users to follow back.[1] In this way, spam tweets posted by bots display on users' homepages. Enticed by the appealing text content, some users may click on links and get redirected to spam or malicious sites.[2] If human users are surrounded by malicious bots and spam tweets, their twittering experience deteriorates, and eventually the whole Twitter community will be hurt. The objective of this paper is to characterize the automation feature of Twitter accounts, and to classify them into three categories, human, bot, and cyborg, accordingly. This will help Twitter manage the community better and help human users recognize who they are tweeting with.

In the paper, we first conduct a series of measurements to characterize the differences among human, bot, and cyborg in terms of tweeting behavior, tweet content, and account properties. By crawling Twitter, we collect over 500,000 users and more than 40 million tweets posted by them. Then, we perform a detailed data analysis, and find a set of useful features to classify users into the three classes. Based on the measurement results, we propose an automated classification system that consists of four major components:

1. the entropy component uses tweeting interval as a measure of behavior complexity, and detects the periodic and regular timing that is an indicator of automation;
2. the spam detection component uses tweet content to check whether text patterns contain spam or not[3];
3. the account properties component employs useful account properties, such as tweeting device makeup, URL ration, to detect deviations from normal; and
4. the decision maker is based on Random Forest, and it uses the combination of the features generated by the above three components to categorize an unknown user as human, bot, or cyborg.

We validate the efficacy of the classification system through our test data set. We further apply the system to classify the entire data set of over 500,000 users collected, and speculate the current composition of Twitter user population based on our classification results.

The remainder of this paper is organized as follows: Section 2 covers related work on Twitter and online social networks. Section 3 details our measurements on Twitter. Section 4 describes our automatic classification system on Twitter. Section 5 presents our experimental results on classification of humans, bots, and cyborgs on Twitter. Finally, Section 6 concludes the paper.

1. Some advanced bots target potential users by keyword search.
2. Due to the tweet size limit, it is very common to use link shortening service on Twitter, which converts an original link to a short one (i.e., http://bit.ly/dtUm5Q). The link illegibility favors bots to allure users.
3. Spam is a good indicator of automation. Most spam messages are generated by bots, and very few are manually posted by humans.

## 2 Related Work

Twitter has been widely used since 2006, and there are some related literature in twittering [12], [13], [14]. To better understand microblogging usage and communities, Java et al. [12] studied over 70,000 Twitter users and categorized their posts into four main groups: daily chatter (e.g., "going out for dinner"), conversations, sharing information or URLs, and reporting news. Their work also studied 1) the growth of Twitter, showing a linear growth rate; 2) its network properties, showing the evidence that the network is scale-free like other social networks [15]; and 3) the geographical distribution of its users, showing that most Twitter users are from the US, Europe, and Japan. Krishnamurthy et al. [13] studied a group of over 100,000 Twitter users and classified their roles by follower-to-following ratios into three groups: 1) broadcasters, which have a large number of followers; 2) acquaintances, which have about the same number on either followers or following; and 3) miscreants and evangelists (e.g., spammers), which follow a large number of other users but have few followers. Wu et al. [16] studied the information diffusion on Twitter, regarding the production, flow, and consumption of information. Kwak et al. [17] conducted a thorough quantitative study on Twitter by crawling the entire Twittersphere. Their work analyzed the follower-following topology, and found nonpower-law follower distribution and low reciprocity, which all mark a deviation from known characteristics of human social networks. Kim et al. [18] analyzed Twitter lists as a potential source for discovering latent characters and interests of users. A Twitter list consists of multiple users and their tweets. Their research indicated that words extracted from each list are representative of all the members in the list even if the words are not used by the members. It is useful for targeting users with specific interests.

In addition to network-related studies, several previous works focus on sociotechnological aspects of Twitter [7], [8], [19], [20], [21], such as its use in the workplace or during major disaster events.

Twitter has attracted spammers to post spam content, due to its popularity and openness. Fighting against spam on Twitter has been investigated in recent works [14], [22], [23]. Yardi et al. [14] detected spam on Twitter. According to their observations, spammers send more messages than legitimate users, and are more likely to follow other spammers than legitimate users. Thus, a high follower-to-following ratio is a sign of spamming behavior. Grier et al. [22] investigated spam on Twitter from the perspective of spam and click-through behaviors, and evaluated the effectiveness of using blacklists to prevent spam propagation. Their work found out that around 0.13 percent of spam tweets generate a visit, orders of magnitude higher than click-through rate of 0.003-0.006 percent reported for spam e-mail. Exploiting the social trust among users, social spam may achieve a much higher success rate than traditional spam methods. Thomas et al. [23] studied the behaviors of spammers on Twitter by analyzing the tweets originated from suspended users in retrospect. They found that the current marketplace for Twitter spam uses a diverse set of spamming techniques, including a variety of strategies for creating Twitter accounts, generating spam URLs, and distributing spam.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

Compared to previous measurement studies on Twitter, our work covers a relatively large group of Twitter users (more than 500,000) and differs in how we link the measurements to automation, i.e., whether posts are from humans, bots, or cyborgs. While some similar metrics are used in our work, such as follower-to-following ratio, we also introduce some metrics, including entropy of tweet intervals, which are not employed in previous research. Our work also detects spam content through Bayesian classification. However, our work focuses on determining the automation degree of Twitter accounts, and uses spam as one of the features in the classification.

Twitter is a social networking service, so our work is also related to recent studies on social networks, such as Flickr, LiveJournal, Facebook, MySpace, and YouTube [15], [24], [25]. In [15], with over 11 million users of Flickr, YouTube, LiveJournal, and Orkut, Mislove et al. analyzed link structure and uncovered the evidence of power-law, small-world, and scale-free properties. In [25], Cha et al. examined the propagation of information through the social network of Flickr. Their work shows that most pictures are propagated through the social links (i.e., links received from friends rather than through searches or external links to Flickr content) and the propagation is very slow at each hop. As a result of this slow propagation, a picture's popularity is often localized in one network and grows slowly over a period of months or even years. In [24], Cha et al. analyzed video popularity life cycles, content aliasing, and the amount of illegal content on YouTube, a popular video sharing service. While YouTube is designed to share large content, i.e., videos, Twitter is designed to share small content, i.e., text messages. Unlike other social networking services, like Facebook or YouTube, Twitter is a microcontent social network, with messages being limited to 140 characters.

As Twitter is a text-based message system, it is natural to compare it with other text-based message systems, such as instant messaging or chat services. Twitter has similar message length (140 characters) to instant messaging and chat services. However, Twitter lacks "presence" (users show up as online/offline for instant messaging services or in specific rooms for chat) but offers 1) more access methods (web, SMS, and various APIs) for reading or posting and 2) more persistent content. Similar to Twitter, instant messaging and chat services also have problems with bots and spam [9], [26]. To detect bots in online chat, Gianvecchio et al. [9] analyzed humans and bots in Yahoo! chat and developed a classification system to detect bots using entropy-based and machine-learning-based classifiers, both of which are used in our classification system as well. In addition, as Twitter is text-based, e-mail spam filtering techniques are also relevant [27], [28], [29]. However, Twitter posts are much shorter than e-mails and spaced out over longer periods of time than for instant messages, e.g., hours rather than minutes or seconds.

Twitter also differs from most other network services in that automation, e.g., message feeds, is a major feature of legitimate Twitter usage, blurring the lines between bot and human. Twitter users can be grouped into four categories: humans, bots, bot-assisted humans, and human-assisted bots. The latter two, bot-assisted humans and human-assisted bots, can be described as cyborgs, a mix between bots and humans [30].

## 3 Measurement

In this section, we first describe the data collection of over 500,000 Twitter users. Then, we detail our observation of user behaviors and account properties, which are pivotal to automatic classification.

### 3.1 Data Collection

Here, we present the methodology used to crawl the Twitter network and collect detailed user information. Twitter has released a set of API functions [31] that support user information collection. Thanks to Twitter's courtesy of including our test account to its white list, we can make API calls up to 20,000 per hour. This eases our data collection. To diversify our data sampling, we employ two methods to collect the data set covering more than 500,000 users. The first method is Depth-First Search (DFS)-based crawling. The reason we choose DFS is that it is a fast and uniformed algorithm for traversing a network. Besides, DFS traversal implicitly includes the information about network locality and clustering. Inspired by [32], [33], we randomly select five users as seeds. For each reached user, we record its follower list. Taking the following direction, the crawler continues with the depth constraint set as three. We customize our crawler with a core module of PHP cURL. Ten crawler processes work simultaneously for each seed. After a seed is finished, they move to the next. The crawl duration lasts one month, and 429,423 users are logged.

Similar to the work in [13] and [14], we also use the public timeline API to collect the information of active users, increasing the diversity of the user pool. Twitter constantly posts the 20 most recent tweets in the global scope. The crawler calls the timeline API to collect the authors of the tweets included in the timeline. Since the Twitter timeline frequently updates, the crawler can repeatedly call the timeline API. During the same time window of the DFS crawl, this method contributes 82,984 users to the data set. We totally collect 512,407 users on Twitter combining both methods.

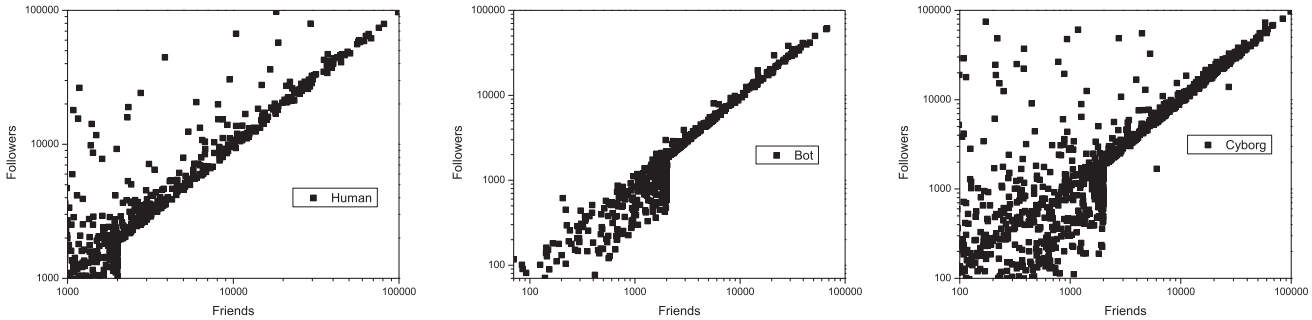
### 3.2 Ground-Truth Creation

To develop an automatic classification system, we need a ground-truth set that contains known samples of human, bot, and cyborg. Among collected data, we randomly choose different samples and classify them by manually checking their user logs and homepages. The ground-truth set includes 2,000 users per class of human, bot, and cyborg, and thus in total three are 6,000 classified samples. In summary, the data set contains 8,350,095 tweets posted by the sampled users in their account lifetime,[4] from which we can extract useful features for classification, such as tweeting behaviors and text patterns.

Our log-based classification follows the principle of the Turing test [34]. The standard Turing tester communicates with an unknown subject for 5 minutes, and decides whether it is a human or machine. Classifying Twitter users is actually more challenging than it appears to be. For many users, their tweets are less likely to form a relatively consistent context. For example, a series of successive tweets may be hardly relevant. The first tweet is the user status, like "watching a football game with my buds."

4. 4,431,923 tweets in the training set, and 3,918,172 tweets in the test set.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.




(a) Human (b) Bot (c) Cyborg

Fig. 1. Numbers of followers and friends.

The second tweet is an automatic update from his blog. The third tweet is a news report RSS feed in the format of article title followed by a shortened URL.

For every account, the following classification procedure is executed. We thoroughly observe the log, and visit the user's homepage (http://twitter.com/#!/username) if necessary. We carefully check tweet contents, visit URLs included in tweets (if any), and decide if redirected webpages are related with their original tweets and if they contain spam or malicious contents. We also check other properties, like tweeting devices, user profile, and the numbers of followers and friends. Given a long sequence of tweets (usually, we check 60 or more if needed), the user is labeled as a human if we can obtain some evidence of original, intelligent, specific and human-like contents. In particular, a human user usually records what he is doing or how he feels about something on Twitter, as he uses Twitter as a microblogging tool to display himself and interact with friends. For example, he may write a post like "I just saw Yankees lost again today. I think they have to replace the starting pitcher for tomorrow's game." The content carries intelligence and originality. Specificity means that the tweet content is expressed in relatively unambiguous words with the presence of consciousness [34]. For instance, in reply to a tweet like "How you like iPad?" a specific response made by human may be "I like its large touch screen and embedded 3G network." On the other hand, a generic reply could be "I like it."

The criteria for identifying a bot are listed as follows. The first is the lack of intelligent or original content. For example, completely retweeting tweets of others or posting adages indicates a lack of originality. The second is the excessive automation of tweeting, like automatic updates of blog entries or RSS feeds. The third is the abundant presence of spam or malicious URLs (i.e., phishing or malware) in tweets or the user profile. The fourth is repeatedly posting duplicate tweets. The fifth is posting links with unrelated tweets. For example, the topic of the redirected webpage does not match the tweet description. The last is the aggressive following behavior. In order to gain attention from human users, bots do mass following and unfollowing within a short period of time. Cyborgs are either human-assisted bots or bot-assisted humans. The criterion for classifying a cyborg is the evidence of both human and bot participation. For example, a typical cyborg account may contain very different types of tweets. A large proportion of tweets carry contents of human-like intelligence and originality, while the rest are automatic updates of RSS feeds. It represents a usage model, in which the human uses his account from time to time while the Twitter widget constantly runs on his desktop and posts RSS feeds of his favorite news channel. Lastly, the uncertain category is for non-English users and those without enough tweets to classify. The samples that are difficult and uncertain to classify fall into this category, and are discarded. Some Twitter accounts are set as "private" for privacy protection, and their webpages are only visible to their friends. We do not include such type of users in the classification either, because of their inaccessibility.

### 3.3 Data Analysis

As mentioned before, Twitter API functions support detailed user information query, ranging from profile, follower, and friend lists to posted tweets. In the above crawl, for each user visited, we call API functions to collect abundant information related with user classification. Most information is returned in the format of XML or JSON. We develop some toolkits to extract useful information from the above well-organized data structures. Our measurement results are presented in the question-answer format.

*Q1. In terms of social relationship, do bots have more friends than followers?* A user's tweets can only be delivered to those who follow him. A common strategy shared by bots is following a large number of users (either targeted with purpose or randomly chosen), and expecting some of them will follow back. Following back is considered as a form of etiquette on Twitter. To increase follower number, some users blindly follow back all the followers including spammers, without carefully checking their profiles. Fig. 1 shows the scatter plots of the numbers of followers and friends for the three categories. For better illustration, the scale is chopped and a small amount of extraordinary points are not included. Fig. 1 contains three different groups of users: group I where the number of one's followers is clearly greater than the number of its friends; group II where the situation is reverse; and group III where the nodes stick around the diagonal.

In the human category, as shown in Fig. 1a, the majority of the nodes belong to group III, implying that the number of their followers is close to that of their friends. This result complies with [15], revealing that human relationships are

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

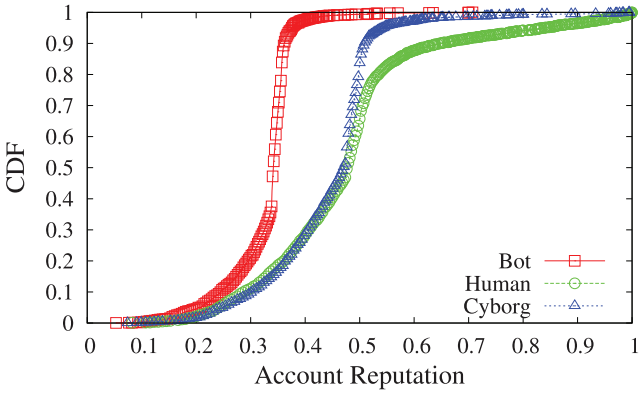



Fig. 2. CDF of account reputation.

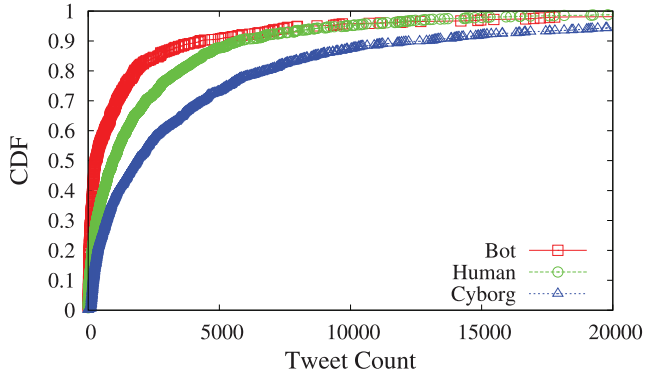



Fig. 3. CDF of tweet count.

typically reciprocal in social networks. Meanwhile, there are quite a few nodes belonging to group I with far more followers than friends. They are usually accounts of celebrities and famous organizations. They generate interesting media contents and attract numerous subscribers. For example, the singer Justin Timberlake has 1,645,675 followers and 39 friends (the ratio is 42,197-to-1).

In the bot category, many nodes belong to group II, as shown in Fig. 1b. Bots add many users as friends, but few follow them back. Unsolicited tweets make bots unpopular among the human world. However, for some bots, the number of their followers is close to that of their friends. This is due to the following reason. Twitter imposes a limit on the ratio of followers over friends to suppress bots. Thus, some more advanced bots unfollow their friends if they do not follow back within a certain amount of time. Those bots cunningly keep the ratio close to 1.

Besides, we have observed that, normal human users are more likely to follow "famous" or "reputable" accounts. We define and normalize

$$Account\ Reputation = \frac{follower_no}{follower_no + frined_no}. \quad (1)$$

A celebrity usually has many followers and few friends (such as Justin Timberlake), and his reputation is close to 1. In contrast, for a bot with few followers and many friends, its reputation is close to 0. Fig. 2 shows the cumulative distribution function (CDF) of account reputation for three categories. The human category has the largest account reputation, closely followed by cyborg. However, bot's value is much lower. Around 60 percent of bots have fewer followers than friends, causing account reputation less than 0.5.

Inspired by account reputation, we define *account taste* as average account reputation of all the friends of the account. Intuitively, the user freely chooses whom to follow (namely, friends), and this reflects his "taste." If the account follows spammers, its "taste" is bad. By doing this, it helps spread spam to more users, making itself a "supporter" of spammers. We have observed "spammer clusters" in our data set where spam accounts tend to follow each other. The CDF result of account taste is similar with Fig. 2, and is not presented due to the space limit.

*Q2. Does automation generate more tweets?* To answer this question, we measure the number of tweets posted in a user's lifetime.[5] Fig. 3 shows the CDF of the tweet counts, corresponding to the human, bot, and cyborg category. It is clear that cyborg posts more tweets than human and bot. A large proportion of cyborg accounts are registered by commercial companies and websites as a new type of media channel and customer service. Most tweets are posted by automated tools (i.e., RSS feed widgets, Web 2.0 integrators), and the volume of such tweets is considerable. Meanwhile, those accounts are usually maintained by some employees who communicate with customers from time to time. Thus, the high tweet count in the cyborg category is attributed to the combination of both automatic and human behaviors in a cyborg. It is surprising that bot generates fewer tweets than human. We check the bot accounts, and find out the following fact. In its active period, bot tweets more frequently than human. However, bots tend to take long-term hibernation. Some are either suspended by Twitter due to extreme or aggressive activities, while the others are in incubation and can be activated to form bot legions.

*Q3. Does automation generate higher tweeting frequency?* Extended from the previous question, here we examine account's tweeting frequency in active status. Fig. 4 plots the interarrival timing distribution of three categories. Due to space limit, each category contains 100 accounts. Tweets posted by an account are sorted on timestamp, and the timestamp of the first tweet is set as 0. The tweeting interarrival sequence of each account is denoted as a vertical strip in the figure, and each of its tweets is denoted as a tiny segment in the strip. We observe the wide existence of burstiness (namely, a block of intensive tweets) in bot, whereas human exhibits less intensive interarrival distribution. Automated programs used by bot accounts can constantly operate in the background and intensively post tweets. Most human users tweet with large interarrivals (such as hours, and even days for some inactive users), and manual behavior cannot generate tweeting frequency as high as bot. The statistics in Table 1 may present a better insight into interarrival distributions of three categories. The average and median of interarrivals of bot are the smallest among three categories. This fact matches the frequent appearance of bursts in Fig. 4b. In contrast, interarrivals of

5. It is the duration from the time when his account was created to the time when our crawler visited it.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

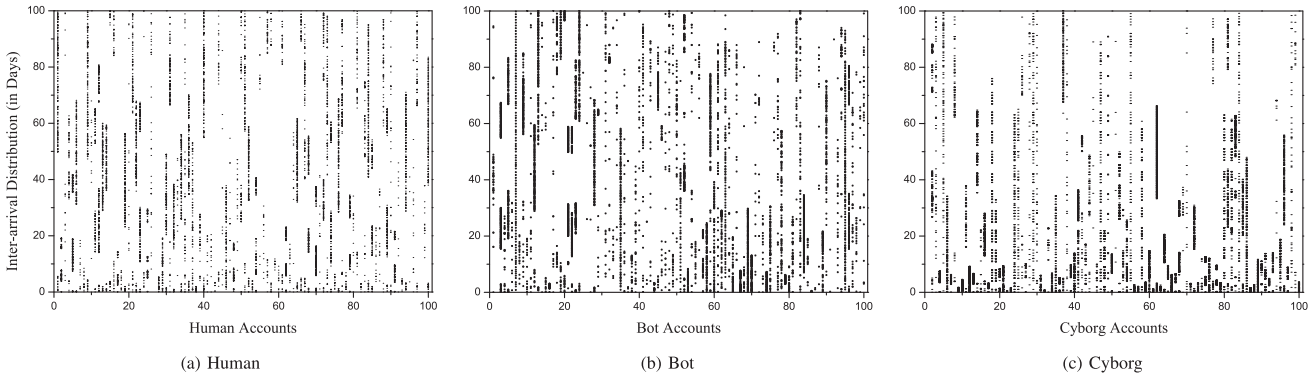



Fig. 4. Interarrival timing distribution of accounts.

human are relatively sparser, and the large standard deviation suggests the irregularity of human behavior.

*Q4. Is tweeting behavior regular or complex?* In our measurement, we have observed that, many bots are driven by timers to post tweets at fixed interarrivals, and thus exhibit regular behavior. In contrast, human behavior carries the inherent complexity [9], [11], [35]. We use entropy rate to measure periodic or regular timing of account's posting behavior. More theoretical details of entropy are presented in Section 4.1. For normalization, we define relative entropy as the entropy rate of an account over the maximum entropy rate in the ground-truth set. Fig. 5 demonstrates the CDF of relative entropy of the three categories. Entropy clearly separates bot from human. High entropy indicates irregularity, a sign of manual behavior, whereas low entropy indicates regularity, a sign of automation.

*Q5. How do users post tweets? manually or via auto piloted tools?* Twitter supports a variety of channels to post tweets. The device name appears below a tweet prefixed by "from." Our whole data set includes 41,991,545 tweets posted by 3,648 distinct devices. The devices can be roughly divided into the following four categories:

1. Web, a user logs into Twitter and posts tweets via the website.
2. Mobile devices, there are some programs exclusively running on mobile devices to post tweets, like Txt for text messages, Mobile web for web browsers on handheld devices, TwitterBerry for BlackBerry, and twidroid for Android mobile OS.
3. Registered third-party applications, many third parties have developed their own applications using Twitter APIs to tweet, and registered them with Twitter. From the application standpoint, we can further categorize this group into subgroups including website integrators (twitpic, bit.ly, Facebook), browser extensions (Tweetbar and Twitterfox for Firefox), desktop clients (TweetDeck and Seesmic Desktop), and RSS feeds/blog widgets (twitterfeed and Twitter for Wordpress).
4. APIs, for those third-party applications not registered or certificated by Twitter, they appear as "API" in Twitter.

Fig. 6 shows the makeup of the above tweeting device categories. Among them, the website of Twitter is the most widely used and generates nearly half of the tweets (46.78 percent), followed by third-party devices (40.18 percent). Mobile devices and unregistered API tools contribute 6.81 and 6.23 percent, respectively. Table 2 lists the top 10

TABLE 1
Interarrival Distribution Statistics

| Category | Avg(day) | Median(minute) | Std dev |
|---|---|---|---|
| Human | 1.10 | 71.0 | 10.92 |
| Cyborg | 0.64 | 56.6 | 2.77 |
| Bot | 0.36 | 30.2 | 1.89 |

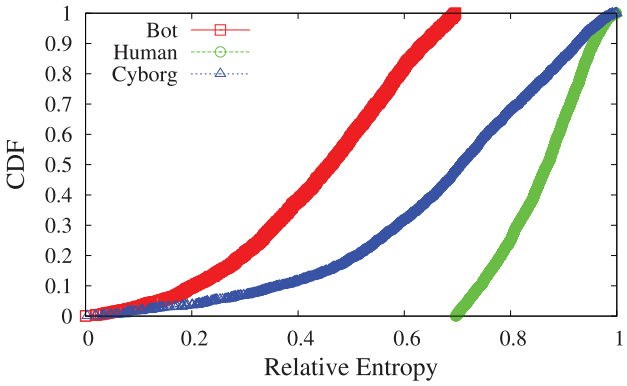



Fig. 5. CDF of account's relative entropy.

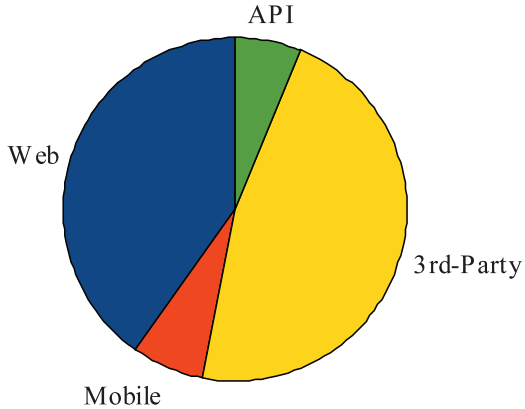



Fig. 6. Tweeting device makeup.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Top 10 Tweeting Devices

| Rank | Human | Bot | Cyborg | All |
|---|---|---|---|---|
| #1 | Web (50.53%) | API (42.39%) | Twitterfeed (31.29%) | Web (46.78%) |
| #2 | TweetDeck (9.19%) | Twitterfeed (26.11%) | Web (23.00%) | TweetDeck (9.26%) |
| #3 | Tweetie (6.23%) | twitRobot (13.11%) | API (6.94%) | Twitterfeed (7.83%) |
| #4 | UberTwitter (3.64%) | RSS2Twitter (2.66%) | Assetize (5.74%) | API (6.23%) |
| #5 | Mobile web (3.02%) | Twitter Tools (1.24%) | HootSuite (5.22%) | Echofon (2.80%) |
| #6 | Txt (2.56%) | Assetize (1.17%) | WP to Twitter (2.40%) | Tweetie (2.50%) |
| #7 | Echofon (2.22%) | Proxifeed (1.08%) | TweetDeck (1.54%) | Txt (2.13%) |
| #8 | TwitterBerry (2.10%) | TweetDeck (0.99%) | UberTwitter (1.19%) | HootSuite (2.10%) |
| #9 | Twitterrific (1.93%) | bit.ly (0.91%) | RSS2Twitter (1.18%) | UberTwitter (1.71%) |
| #10 | Seesmic(1.64%) | Twitme for WordPress (0.84%) | Twaitter (0.86%) | Mobile web (1.53%) |

devices used by the human, bot, and cyborg categories, and the whole data set.[6]

More than half of the human tweets are manually posted via the Twitter website. The rest of top devices are mobile applications (Tweetie, UberTwitter, Mobile web, Txt, TwitterBerry) and desktop clients (TweetDeck, Echofon, and Seesmic). In general, tweeting via such devices requires human participation. In contrast, the top tools used by bots are mainly auto piloted, and 42.39 percent of bot tweets are generated via unregistered API-based tools. Bots can abuse APIs to do almost everything they want on Twitter, like targeting users with keywords, following users, unfollowing those who do not follow back, or posting prepared tweets. Twitterfeed, RSS2Twitter, and Proxifeed are RSS feed widgets that automatically pipeline information (usually in the format of the page title followed by the URL) to Twitter via RSS feeds. Twitter Tools and Twitme for WordPress are popular WordPress plug-ins that integrate blog updates to Twitter. Assetize is an advertising syndicator mainly targeting at Twitter, and twitRobot is a bot tool that automatically follows other users and posts tweets. All these tools only require minimum human participation (like importing Twitter account information, or setting RSS feeds and update frequency), and thus indicate great automation.

Overall, humans tend to tweet manually and bots are more likely to use auto piloted tools. Cyborgs employ the typical human and bot tools. The cyborg group includes many human users who access their Twitter accounts from time to time. For most of the time when they are absent, they leave their accounts to auto piloted tools for management.

*Q6. Do bots include more external URLs than humans?* In our measurement, we find out that, most bots tend to include URLs in tweets to redirect visitors to external webpages. For example, spam bots are created to spread unsolicited commercial information. Their topics are similar to those in e-mail spam, including online marketing and affiliate programs, working at home, selling fake luxury brands or pharmaceutical products.[7] However, the tweet size is up to 140 characters, which is rather limited for spammers to express enough text information to allure users. Basically, a spam tweet contains an appealing title followed by an external URL. Fig. 7 shows the external URL ratios (namely, the number of external URLs included in tweets over the number of tweets posted by an account) for the three categories, among which the URL ratio of bot is highest. Some tweets by bots even have more than one URL.[8] The URL ratio of cyborg is very close to the bot's level. A large number of cyborgs integrate RSS feeds and blog updates, which take the style of webpage titles followed by page links. The URL ratio of human is much lower, on average it is only 29 percent. When a human tweets what is he doing or what is happening around him, he mainly uses text and does not often link to web pages.

*Q7. Are there any other temporal properties of Twitter users helpful for differentiation among human, bot, and cyborg?* Many research works like [36] and [37] have shown the weekly and diurnal access patterns of humans in the Internet. Figs. 9a and 9b present the tweeting percentages of the three different categories on daily and hourly bases, respectively. The weekly behavior of Twitter users shows clear differences among the three categories. While humans are more active during the regular workdays, from Monday to Friday, and less active during the weekend, Saturday and Sunday, bots have roughly the same activity level every day of the week. Interestingly, cyborgs are the most active ones on Monday and then slowly decrease their tweeting activities during the week; on Saturday, cyborgs reach their lowest active point but somehow bounce back a bit on Sunday. Such a cyborg activity trend is mainly caused by their message feeds and the high level of news and blog activities at the start of a week. Similarly, the hourly behavior of human is more active during the daytime, which mostly overlaps with office hours. The bot activity is nearly even except a little drop in the deep of night. Some more advanced bots have the setting of "only tweet from a time point to another," which helps save API calls [38]. Thus, they can tweet more in the daytime to better draw the attention of humans.

Fig. 8 shows account registration dates grouped by quarter. We have two observations from the figure. First, the majority of accounts (80.0 percent of humans, 94.8 percent of bots, and 71.1 percent of cyborgs) were registered in 2009. It confirms the dramatic growth of Twitter in 2009. Second, we do not find any bot or cyborg in our ground-truth data set earlier than March 2007. However, human registration has continued increasing since Twitter was founded in 2006. Thus, old accounts are less likely to be bots.

6. The whole data set contains around 500,000 users, and the human, bot, and cyborg categories equally contain 1,000 users in the training data set.

7. A new topic is attracting more followers on Twitter. It follows the style of pyramid sales by asking newly joined users to follow existing users in the spam network.

8. Many such accounts belong to a type of bot that always appends a spam link to tweets it retweets.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

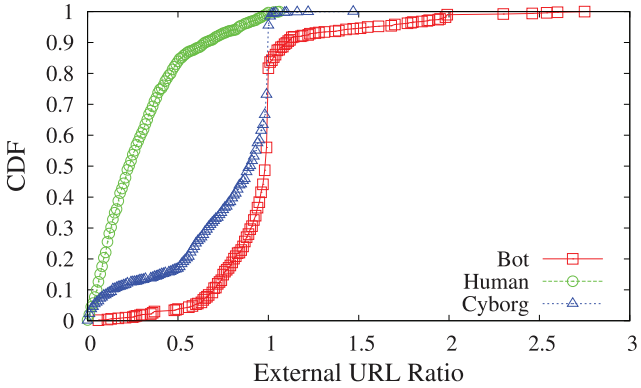



Fig. 7. External URL ratio in tweets.

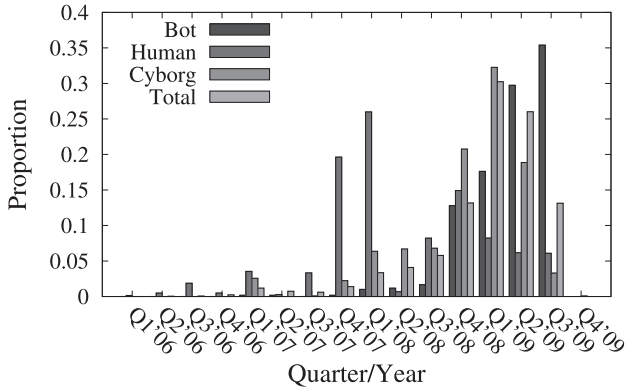



Fig. 8. Account registration date (grouped by quarter).

*Q8. Are users aware of privacy and identity protection on Twitter?* Twitter provides a `protected` option to protect user privacy. If it is set as true, the user's homepage is only visible to his friends. However, the option is set as false by default. In our data set of over 500,000 users, only 4.9 percent of them are protected users. Twitter also verifies some accounts to authenticate users' real identities. More and more celebrities and famous organizations have applied for verified accounts. For example, Bill Gates has his verified Twitter account at http://twitter.com/billgates. However, in our data set, only 1.8 percent of users have verified accounts.

## 4 Classification

This section describes our automated system for classification of Twitter users. The system classifies Twitter users into three categories: human, bot, and cyborg. The system consists of several components: the entropy component, the spam detection component, the account properties component, and the decision maker. The high-level design of our Twitter user classification system is shown in Fig. 10.

The entropy component uses corrected conditional entropy to detect periodic or regular timing, which is a sign of automation. The spam detection component uses a variant of Bayesian classification to detect text patterns of known spam on Twitter. The account properties component uses account-related properties to catch bot deviation from the normal human distribution. Lastly, the decision maker based on Random Forest algorithm analyzes the features identified by the other three components and makes a decision: human, cyborg, or bot.

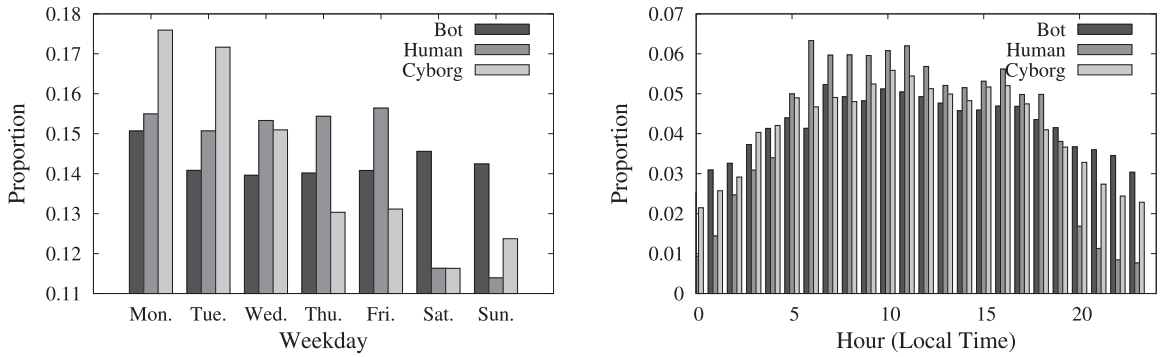
### 4.1 Entropy Component

The entropy component detects periodic or regular timing of the messages posted by a Twitter user. On one hand, if the entropy or corrected conditional entropy is low for the intertweet delays, it indicates periodic or regular behavior, a sign of automation. More specifically, some of the messages are posted via automation, i.e., the user may be a potential bot or cyborg. On the other hand, a high entropy indicates irregularity, a sign of human participation.

#### 4.1.1 Entropy Measures

The entropy rate is a measure of the complexity of a process [39]. The behavior of bots is often less complex than that of humans [40], [41], which can be measured by entropy rate. A low entropy rate indicates a regular process, whereas a high entropy rate indicates a random process. A medium entropy rate indicates a complex process, i.e., a mix of order and disorder [42].

The entropy rate is defined as either the average entropy per random variable for an infinite sequence or as the conditional entropy of an infinite sequence. Thus, as real data sets are finite, the conditional entropy of finite sequences is often used to estimate the entropy rate. To estimate the entropy rate, we use the corrected conditional entropy [43]. The corrected conditional entropy is defined as follows:




(a) Tweets by Day of Week (b) Hourly Tweets

Fig. 9. Tweets posted.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

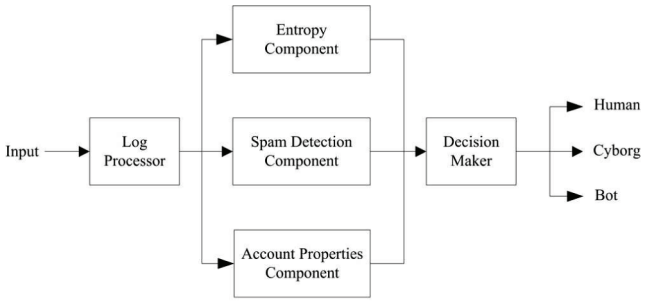



Fig. 10. Classification system.

A random process $X = \{X_i\}$ is defined as a sequence of random variables. The entropy of such a sequence of random variables is defined as

$$H(X_1, \ldots, X_m) = -\sum_{i=1}^{m} P(x_i) \log P(x_i), \tag{2}$$

where $P(x_i)$ is the probability $P(X_i = x_i)$.

The conditional entropy of a random variable given a previous sequence of random variables is

$$\begin{aligned} &H(X_m \mid X_1, \ldots, X_{m-1}) \\ &\quad = H(X_1, \ldots, X_m) - H(X_1, \ldots, X_{m-1}). \end{aligned} \tag{3}$$

Then, based on the conditional entropy, the entropy rate of a random process is defined as

$$\overline{H}(X) = \lim_{m \to \infty} H(X_m \mid X_1, \ldots, X_{m-1}). \tag{4}$$

The corrected conditional entropy is computed as a modification of (4). First, the joint probabilities, $P(X_1 = x_1, \ldots, X_m = x_m)$ are replaced with empirically derived probabilities. The data are binned into $Q$ bins, i.e., values are converted to bin numbers from 1 to $Q$. The empirically derived probabilities are then determined by the proportions of bin number sequences in the data. The entropy estimate and conditional entropy estimate, based on empirically derived probabilities, are denoted as $EN$ and $CE$, respectively. Second, a corrective term, $perc(X_m) \cdot EN(X_1)$, is added to adjust for the limited number of sequences for increasing values of $m$ [43]. The corrected conditional entropy, denoted as $CCE$, is computed as

$$\begin{aligned} &CCE(X_m \mid X_1, \ldots, X_{m-1}) \\ &\quad = CE(X_m \mid X_1, \ldots, X_{m-1}) + perc(X_m) \cdot EN(X_1), \end{aligned} \tag{5}$$

where $perc(X_m)$ is the percentage of unique sequences of length $m$ and $EN(X_1)$ is the entropy with $m$ fixed at 1 or the first-order entropy.

The estimate of the entropy rate is the minimum of the corrected conditional entropy over different values of $m$. The minimum of the corrected conditional entropy is considered to be the best estimate of the entropy rate from the limited number of sequences.

### 4.2 Spam Detection Component

The spam detection component examines the content of tweets to detect spam. We have observed that most spam tweets are generated by bots and only very few of them are manually posted by humans. Thus, the presence of spam patterns usually indicates automation. Since tweets are text, determining if their content is spam can be reduced to a text classification problem. The text classification problem is formalized as $f : T \times C \to \{0, 1\}$, where $f$ is the classifier, $T = \{t_1, t_2, \ldots, t_n\}$ are the texts to be classified, and $C = \{c_1, c_2, \ldots, c_k\}$ are the classes [44]. A value of 1 for $f(t_i, c_j)$ indicates that text $t_i$ belongs to class $c_j$, whereas a value of 0 indicates it does not belong to that class. Bayesian classifiers are very effective in text classification, especially for e-mail spam detection, so we employ Bayesian classification for our machine learning text classification component.

During the creation of ground-truth set in Section 3.2, when the human inspector encounters spam content in tweets, or tweets contain URLs caught by the blacklists, their content (URLs are excluded if any) is added to the spam content set. In contrast, the nonspam content set contains ham tweets posted by human. As a conservative measure, the set does not contain content posted by bot or cyborg.

In Bayesian classification, deciding if a message belongs to a class, e.g., spam, is done by computing the corresponding probability based on its content, e.g., $P(C = spam|M)$, where $M$ is a message and $C$ is a class. If the probability is over a certain threshold, then the message is from that class.

The probability that a message $M$ is spam, $P(spam|M)$, is computed from Bayes theorem:

$$\begin{aligned} P(spam|M) &= \frac{P(M|spam)P(spam)}{P(M)} \\ &= \frac{P(M|spam)P(spam)}{P(M|spam)P(bot) + P(M|not\ spam)P(not\ spam)}. \end{aligned} \tag{6}$$

The message $M$ is represented as a feature vector $\langle f_1, f_2, \ldots, f_n \rangle$, where each feature $f$ is one or more words in the message and each feature is assumed to be conditionally independent

$$\begin{aligned} P(spam|M) = &\left\{ P(spam) \prod_{i=1}^{n} P(f_i|spam) \right\} \Big/ \\ &\left\{ P(spam) \prod_{i=1}^{n} P(f_i|spam) + P(not\ spam) \right. \\ &\left. \prod_{i=1}^{n} P(f_i|not\ spam) \right\}. \end{aligned} \tag{7}$$

The calculation of $P(spam|M)$ varies in different implementations of Bayesian classification. The implementation used for our machine learning component is CRM114 [45]. CRM114 is a powerful text classification system that offers a variety of different classifiers. The default classifier for CRM114 is Orthogonal Sparse Bigram (OSB), a variant of Bayesian classification, which has been shown to perform well for e-mail spam filtering. OSB differs from other Bayesian classifiers in that it treats pairs of words as features. OSB first chops the whole input into multiple basic units with five consecutive words in each unit. Then, it extracts four word pairs from each unit to construct features, and derives their probabilities. Finally, OSB applies Bayes theorem to compute the overall probability that the text belongs to one class or another.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

### 4.3 Account Properties Component

Besides intertweet delay and tweet content, some Twitter account-related properties are very helpful for the user classification. As shown in Section 3.3, obvious difference exists between the human and bot categories. The first property is the URL ratio. The ratio indicates how often a user includes external URLs in its posted tweets. External URLs appear very often in tweets posted by a bot. Our measure shows, on average the ratio of bot is 97 percent, while that of human is much lower at 29 percent. Thus, a high ratio (e.g., close to 1) suggests a bot and a low ratio implies a human. The second property is tweeting device makeup. According to Table 2, about 70 percent tweets of human are posted via web and mobile devices (referred as manual devices), whereas about 87 percent tweets of bot are posted via API and other auto-piloted programs (referred as auto devices). The third property is the followers to friends ratio.

The fourth property is link safety, i.e., to decide whether external links in tweets are malicious/phishing URLs or not. We run a batch script to check a URL in five blacklists: Google Safe Browsing, PhishingTank, URIBL, SURBL, and Spamhaus [46], [47], [48], [49], [50]. Google Safe Browsing checks URLs against Google's constantly updated lists of suspected phishing and malware pages. PhishingTank focuses on phishing websites. The mechanisms of URIBL, SURBL, and Spamhaus are similar. They contain those suspicious websites that have appeared in spam e-mails, primarily Unsolicited Bulk/Commercial E-mail (UBE/UCE). If the URL appears in any of the blacklists, the feature of link safety is set as false.

The fifth property is whether a Twitter account is verified. No bot in our ground-truth data set is verified. The account verification suggests a human. The sixth property is the account registration date. According to Fig. 8, 94.8 percent of bots were registered in 2009. The last two properties are the hashtag ratio and mention ratio. Hashtag ratio of an account is defined as the number of hashtags included in the tweets over the number of tweets posted by the account. Mention ratio is defined similarly.

The account properties component extracts these properties from the user log, and sends them to the decision maker. It assists the entropy component and the spam detection component to improve the classification accuracy.

### 4.4 Decision Maker

Our classification problem can be formulated as follows: Given an unknown user $U$ represented by the feature vector, the decision maker determines the class $C$ to which $U$ belongs to. Namely,

$$U = <f1, f2, \ldots, fn> \rightarrow C = \{human, bot, cyborg\}.$$

We select Random Forest [51] as the machine learning algorithm, and implement the decision maker based on it.

Random Forest creates an ensemble classifier consisting of a set of decision trees. The algorithm applies the random feature selection in [51] and bagging idea in [52] to construct a "collective forest" of decision trees with controlled variation. The decision tree contains two types of nodes, the leaf node labeled as a class, and the interior node associated with a feature. We denote the number of features in the data set as $M$, and the number of features used to make the decision at a node of the tree as $m(\ll M)$. Each decision tree is built top-down in a recursive manner. For every node in the construction path, $m$ features are randomly selected to reach a decision at the node. The node is then associated with the feature that is the most informative. Entropy is used to calculate the information gain contributed by each of the $m$ features (namely, how informative a feature is). In other words, the recursive algorithm applies a greedy search by selecting the candidate feature that maximizes the heuristic splitting criterion.

We denote $D$ as the data set of labeled samples, and $C$ as the class with $k$ values, $C = \{C_1, C_2, \ldots, C_k\}$. The information required to identify the class of a sample in $D$ is denoted as *Info(D) = Entropy(P)*, where $P$, as the probability distribution of $C$, is

$$P = \left\{\frac{|C_1|}{|D|}, \frac{|C_2|}{|D|}, \ldots, \frac{|C_k|}{|D|}\right\}.$$

If we partition $D$ based on the value of a feature $F$ into subsets $\{D_1, D_2, \ldots, D_n\}$,

$$Info(F, D) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} Info(D_i). \tag{8}$$

After the value of feature $F$ is obtained, the corresponding gain in information due to $F$ is denoted as

$$Gain(F, D) = Info(D) - Info(F, D), \tag{9}$$

As *Gain* favors features that have a large number of values, to compensate for this *GainRatio* is defined as

$$GainRatio(F, D) = \frac{Gain(F, D)}{SplitInfo(F, D)}, \tag{10}$$

where *SplitInfo(F,D)* is the information due to the splitting of $D$ based on the value of attribute $F$. Thus,

$$SplitInfo(A, D) = Entropy\left(\frac{|D_1|}{|D|}, \frac{|D_2|}{|D|}, \ldots, \frac{|D_n|}{|D|}\right). \tag{11}$$

More details of decision tree learning can be found in [53]. To classify an unknown sample, it is push downwards in the tree, and assigned with the class of the leaf node with which it ends up. Every decision tree determines a classification decision on the sample. Random Forest applies the majority voting of all the individual decisions to reach the final decision.

## 5 Evaluation

In this section, we first evaluate the accuracy of our classification system based on the ground-truth set. Then, we apply the system to classify the entire data set of over 500,000 users collected. With the classification results, we further speculate the current composition of Twitter user population. Finally, we discuss the robustness of the proposed classification system against possible evasions.

### 5.1 Methodology

As shown in Fig. 10, the components of the classification system collaborate in the following way. The entropy

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

component calculates the entropy (and corrected conditional entropy) of intertweet delays of a Twitter user. The entropy component only processes logs with more than 100 tweets.[9] This limit helps reduce noise in detecting automation. A lower entropy indicates periodic or regular timing of tweeting behavior, a sign of automation, whereas a higher entropy implies irregular behavior, a sign of human participation. The spam detection component determines if the tweet content is either spam or not, based on the text patterns it has learned. The content feature value is set to 1 for spam but 0 for nonspam. The account properties component checks all the properties mentioned in Section 4.3, and generates a real-number-type value for each property. Given a Twitter user, the above three components generate a set of features and input them into the decision maker. For each class, namely human, bot, and cyborg, the decision maker computes a classification score for the user, and classifies it into the class with the highest score. The training of the classification system and cross validation of its accuracy are detailed as follows.

### 5.2 Classification System Training

The spam detection component of the classification system requires training before being used. It is trained on spam and nonspam data sets. The spam data set consists of spam tweets and spam external URLs, which are detected during the creation of the ground-truth set. Some advanced spam bots intentionally inject nonspam tweets (usually in the format of pure text without URLs, such as adages[10]) to confuse human users. Thus, we do not include such vague tweets without external URLs. The nonspam data set consists of all human tweets and cyborg tweets without external URLs. Most human tweets do not carry spam. Cyborg tweets with links are hard to determine without checking linked webpages. They can be either spam or nonspam. Thus, we do not include this type of tweets in either data set. Training the component with up-to-date spam text patterns on Twitter helps improve the accuracy. In addition, we create a list of spam words with high frequency on Twitter to help the Bayesian classifier capture spam content.

### 5.3 Cross Validation of Accuracy

We use Weka, a machine learning tool [54], to implement the Random Forest-based classifier. We apply cross validation with 10-folds to train and test the classifier over the ground-truth set [55]. The data set is randomly partitioned into 10 complementary subsets with equal size. In each round, one out of 10 subsets is retained as the test set to validate the classifier, while the remaining nine subsets are used as the training set to train the classifier. At the beginning of a round, the classifier is reset and retrained. Thus, each round is an independent classification procedure, and does not affect subsequent ones. The individual results from 10 rounds are averaged to generate the final estimation. The advantage of cross validation is that, all samples in the data set are used for both training and validation, while each sample is validated exactly once. The confusion matrix listed in Table 3 demonstrates the classification results.

TABLE 3
Confusion Matrix

| | | Classified | | | Total | True Pos |
|---|---|---|---|---|---|---|
| | | Human | Cyborg | Bot | | |
| Actual | Human | 1972 | 27 | 1 | 2000 | 98.6% |
| | Cyborg | 65 | 1833 | 102 | 2000 | 91.7% |
| | Bot | 2 | 46 | 1952 | 2000 | 97.6% |
| | | | | | Avg | 96.0% |

The "Actual" rows in Table 3 denote the actual classes of the users, and the "Classified" columns denote the classes of the users as decided by the classification system. For example, the cell in the junction of the "Human" row and column means that 1,972 humans are classified (correctly) as humans, whereas the cell of "Human" row and "Cyborg" column indicates that 27 humans are classified (incorrectly) as cyborgs. There is no misclassification between human and bot.

We examine the logs of those users being classified by mistake, and analyze each category as follows:

- For the human category, 1.4 percent of human users are classified as cyborg by mistake. One reason is that the overall scores of some human users are lowered by spam content penalty. The tweet size is up to 140 characters. Some patterns and phrases are used by both human and bot, such as "I post my online marketing experience at my blog at http://bit.ly/xT6klM. Please ReTweet it." Another reason is that the tweeting interval distribution of some human users is slightly lower than the entropy means, and they are penalized for that.
- For the bot category, 2.3 percent of bots is wrongly categorized as cyborg. The main reason is that, most of them escape the spam penalty from the spam detection component. Some spam tweets have very obscure text content, like "you should check it out since it's really awesome. http://bit.ly/xT6klM". Without checking the spam link, the component cannot determine if the tweet is spam merely based on the text.
- For the cyborg category, 3.3 percent of cyborgs are misclassified as human, and 5.1 percent of them are misclassified as bot. In analyzing those samples misclassified as human, we find out a common fact that, some owners of cyborg accounts interact with followers from time to time, and use manual devices to reply or retweet to followers. Besides, the manual behavior of owners increases the entropy of tweeting interarrivals. The two factors tend to influence the classifier to make decisions in favor of human. The difficulty here is that, a cyborg can be either a human-assisted bot or a bot-assisted human. A strict policy could categorize cyborg as bot, while a loose one may categorize it as human.

9. The intertweet span could be wild on Twitter. An account may be inactive for months, but suddenly tweets at an intensive frequency for a short-term, and then enters hibernation again. It generates noise to the entropy component. Thus, the entropy component does not process logs with less than 100 tweets. Besides, in practice, it is nearly impossible to determine automation based on a very limited number of tweets.

10. A typical content pattern is listed as follows: Tweet 1, A friend in need is a friend in deed. Tweet 2, Danger is next neighbor to security. Tweet 3, Work home and make $3k per month. Check out how, http://tinyurl.com/bF234T.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

TABLE 4
Feature Weights

| Feature | Accuracy (%) |
|---|---|
| Entropy | 82.8 |
| URL Ratio | 74.9 |
| Automated Device % | 71.0 |
| Bayesian Spam Detection | 69.5 |
| Manual Device % | 69.2 |
| Registration Date | 62.9 |
| Mention Ratio | 56.2 |
| Link Safety | 49.3 |
| Hashtag Ratio | 47.0 |
| Followers to Friends Ratio | 45.3 |
| Account Verification | 35.0 |

- There is negligible misclassification between human and bot. The classifier clearly separates these two classes.

Overall, our classification system can accurately differentiate human from bot. However, it is much more challenging for a classification system to distinguish cyborg from human or bot. After averaging the true positive rates of the three classes with equal sample size, the overall system accuracy can be viewed as 96.0 percent.

Among the set of features used in classification, some play a more important role than others. Now, we evaluate the discrimination weight of each feature. In every test, we only use one feature to independently cross validate the ground-truth set. Table 4 presents the results sorted on accuracy. The entropy feature has the highest accuracy at 82.8 percent. It effectively captures the timing difference between regularity of automated behavior and complexity of manual behavior. Limited by tweet size, bot usually relies on URLs to redirect users to external websites. This fact makes the URL ratio feature have a relatively high accuracy at 74.9 percent. Recognizing the tweeting device makeup (manual or automated) and detecting spam content also help the classification. By comparing the collective performance in Table 3 and individual performance in Table 4, we observe that, no single feature works perfectly well, and the combination of multiple features improves the classification accuracy.

### 5.4 Twitter Composition

We further use the classification system to automatically classify our whole data set of over 500,000 users. We can speculate the current composition of Twitter user population based on the classification results. The system classifies 53.2 percent of the users as human, 36.2 percent as cyborg, and 10.5 percent as bot. Thus, we speculate the population proportion of human, cyborg and bot category roughly as 5:4:1 on Twitter.

### 5.5 Resistance to Evasion

Now, we discuss the resistance of the classification system to possible evasion attempts made by bots. Bots may deceive certain features, such as the followers to friends ratio as mentioned before. However, our system has two critical features that are very hard for bots to evade. The first feature is tweeting device makeup, which corresponds to the manual/auto device percentage in Table 4. Manual device refers to web and mobile devices, while auto device refers to API and other auto-piloted programs (see Section 3.3, Q5). Tweeting via web requires a user to log in and manually post via the Twitter website in a browser. Posting via HTTP form is considered by Twitter as API. Furthermore, currently it is impractical or expensive to run a bot on a mobile device to frequently tweet. As long as Twitter can correctly identify different tweeting platforms, device makeup is an effective metric for bot detection. The second feature is URL ratio. Considering the limited tweet length that is up to 140 characters, most bots have to include a URL to redirect users to external sites. Thus, a high URL ratio is another effective metric for bot detection. If we exclude the features of URL ratio and tweeting device makeup, and retrain the classifier, the overall classification accuracy drops to 88.9 percent. Bot may try to bypass some features when it knows our detection strategy. For timing entropy, bot could mimic human behaviors but at the cost of much reduced tweeting frequency. For spam content, bot could intermix spam with ham tweets to dilute spam density. We will continue to explore new features emerging with the Twitter development for more effective bot detection in the future.

## 6 Conclusion

In this paper, we have studied the problem of automation by bots and cyborgs on Twitter. As a popular web application, Twitter has become a unique platform for information sharing with a large user base. However, its popularity and very open nature have made Twitter a very tempting target for exploitation by automated programs, i.e., bots. The problem of bots on Twitter is further complicated by the key role that automation plays in everyday Twitter usage.

To better understand the role of automation on Twitter, we have measured and characterized the behaviors of humans, bots, and cyborgs on Twitter. By crawling Twitter, we have collected one month of data with over 500,000 Twitter users with more than 40 million tweets. Based on the data, we have identified features that can differentiate humans, bots, and cyborgs on Twitter. Using entropy measures, we have determined that humans have complex timing behavior, i.e., high entropy, whereas bots and cyborgs are often given away by their regular or periodic timing, i.e., low entropy. In examining the text of tweets, we have observed that a high proportion of bot tweets contain spam content. Lastly, we have discovered that certain account properties, like external URL ratio and tweeting device makeup, are very helpful on detecting automation.

Based on our measurements and characterization, we have designed an automated classification system that consists of four main parts: the entropy component, the spam detection component, the account properties component, and the decision maker. The entropy component checks for periodic or regular tweet timing patterns; the spam detection component checks for spam content; and the account properties component checks for abnormal values of Twitter-account-related properties. The decision maker summarizes the identified features and decides whether the user is a human, bot, or cyborg. The

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

effectiveness of the classification system is evaluated through the test data set. Moreover, we have applied the system to classify the entire data set of over 500,000 users collected, and speculated the current composition of Twitter user population based on the classification results.

## ACKNOWLEDGMENTS

The authors are grateful to the anonymous referees for their insightful feedback. This work was partially supported by ARO grant W911NF-11-1-0149 and US National Science Foundation (NSF) grant 0901537. The work of Sushil Jajodia was partially supported by the Air Force Office of Scientific Research under grant FA9550-09-1-0421 and by the Army Research Office under MURI grant W911NF-09-1-0525 and DURIP grant W911NF-11-1-0340.

## REFERENCES

[1] "Top Trending Twitter Topics for 2011 from What the Trend," http://blog.hootsuite.com/top-twitter-trends-2011/, Dec. 2011.
[2] "Twitter Blog: Your World, More Connected," http://blog.twitter.com/2011/08/your-world-more-connected.html, Aug. 2011.
[3] Alexa, "The Top 500 Sites on the Web by Alexa," http://www.alexa.com/topsites, Dec. 2011.
[4] "Amazon Comes to Twitter," http://www.readwriteweb.com/archives/amazon_comes_to_twitter.php, Dec. 2009.
[5] "Best Buy Goes All Twitter Crazy with @Twelpforce," http://twitter.com/in_social_media/status/2756927865, Dec. 2009.
[6] "Barack Obama Uses Twitter in 2008 Presidential Campaign," http://twitter.com/BarackObama/, Dec. 2009.
[7] J. Sutton, L. Palen, and I. Shlovski, "Back-Channels on the Front Lines: Emerging Use of Social Media in the 2007 Southern California Wildfires," *Proc. Int'l ISCRAM Conf.*, May 2008.
[8] A.L. Hughes and L. Palen, "Twitter Adoption and Use in Mass Convergence and Emergency Events," *Proc. Sixth Int'l ISCRAM Conf.*, May 2009.
[9] S. Gianvecchio, M. Xie, Z. Wu, and H. Wang, "Measurement and Classification of Humans and Bots in Internet Chat," *Proc. 17th USENIX Security Symp.*, 2008.
[10] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your Botnet Is My Botnet: Analysis of a Botnet Takeover," *Proc. 16th ACM Conf. Computer and Comm. Security*, 2009.
[11] S. Gianvecchio, Z. Wu, M. Xie, and H. Wang, "Battle of Botcraft: Fighting Bots in Online Games with Human Observational Proofs," *Proc. 16th ACM Conf. Computer and Comm. Security*, 2009.
[12] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," *Proc. Ninth WebKDD and First SNA-KDD Workshop Web Mining and Social Network Analysis*, 2007.
[13] B. Krishnamurthy, P. Gill, and M. Arlitt, "A Few Chirps about Twitter," *Proc. First Workshop Online Social Networks*, 2008.
[14] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting Spam in a Twitter Network," *First Monday*, vol. 15, no. 1, Jan. 2010.
[15] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," *Proc. Seventh ACM SIGCOMM Conf. Internet Measurement*, 2007.
[16] S. Wu, J.M. Hofman, W.A. Mason, and D.J. Watts, "Who Says What to Whom on Twitter," *Proc. 20th Int'l Conf. World Wide Web*, pp. 705-714, 2011.
[17] H. Kwak, C. Lee, H. Park, and S. Moon, "What Is Twitter, a Social Network or a News Media?" *Proc. 19th Int'l Conf. World Wide Web*, pp. 591-600, 2010.
[18] I.-C.M. Dongwoo Kim, Y. Jo, and A. Oh, "Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users," *Proc. CHI Workshop Microblogging: What and How Can We Learn From It?*, 2010.
[19] D. Zhao and M.B. Rosson, "How and Why People Twitter: The Role that Micro-Blogging Plays in Informal Communication at Work," *Proc. ACM Int'l Conf. Supporting Group Work*, 2009.
[20] K. Starbird, L. Palen, A. Hughes, and S. Vieweg, "Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged Information," *Proc. ACM Conf. Computer Supported Cooperative Work*, Feb. 2010.
[21] B.J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth," *Am. Soc. for Information Science and Technology*, vol. 60, no. 11, pp. 2169-2188, 2009.
[22] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: The Underground on 140 Characters or Less," *Proc. 17th ACM Conf. Computer and Comm. Security*, pp. 27-37, 2010.
[23] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended Accounts in Retrospect: An Analysis of Twitter Spam," *Proc. ACM SIGCOMM Conf. Internet Measurement Conf.*, pp. 243-258, 2011.
[24] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," *Proc. Seventh ACM SIGCOMM Conf. Internet Measurement*, 2007.
[25] M. Cha, A. Mislove, and K.P. Gummadi, "A Measurement-Driven Analysis of Information Propagation in the Flickr Social Network," *Proc. 18th Int'l Conf. World Wide Web*, 2009.
[26] M. Xie, Z. Wu, and H. Wang, "Honeyim: Fast Detection and Suppression of Instant Messaging Malware in Enterprise-Like Networks,," *Proc. 23rd Ann. Computer Security Applications Conf.*, 2007.
[27] P. Graham, "A Plan for Spam," http://www.paulgraham.com/spam.html, Jan. 2008.
[28] J.A. Zdziarski, *Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification.* No Starch Press, 2005.
[29] M. Xie, H. Yin, and H. Wang, "An Effective Defense Against Email Spam Laundering," *Proc. 13th ACM Conf. Computer and Comm. Security*, 2006.
[30] J. Yan, "Bot, Cyborg and Automated Turing Test," *Proc. 14th Int'l Workshop Security Protocols*, Mar. 2006.
[31] Twitter, "Twitter api Wiki," http://apiwiki.twitter.com/, Feb. 2010.
[32] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of Osns," *Proc. 27th IEEE Int'l Conf. Computer Comm.*, Mar. 2010.
[33] M.R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On Near-Uniform Url Sampling," *Proc. Ninth Int'l World Wide Web Conf. Computer Networks*, May 2000.
[34] A.M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, pp. 433-460, 1950.
[35] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. Gnecchi-Ruscone, A. Malliani, and S. Cerutti, "Measuring Regularity by Means of a Corrected Conditional Entropy in Sympathetic Outflow," *Biological Cybernetics*, vol. 78, no. 1, pp. 71-78, 1998.
[36] H.J. Fowler and W.E. Leland, "Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management," *IEEE J. Selected Areas in Comm.*, vol. 9, no. 7, pp. 1139-1149, Sept. 1991.
[37] M. Dischinger, A. Haeberlen, K.P. Gummadi, and S. Saroiu, "Characterizing Residential Broadband Networks," *Proc. Seventh ACM SIGCOMM Conf. Internet Measurement*, 2007.
[38] Tweetadder, "Automatic Twitter Software," http://www.tweetadder.com/, Feb. 2010.
[39] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* Wiley-Interscience, 2006.
[40] S. Gianvecchio and H. Wang, "Detecting Covert Timing Channels: An Entropy-Based Approach," *Proc. ACM Conf. Computer and Comm. Security*, Oct./Nov. 2007.
[41] H. Husna, S. Phithakkitnukoon, and R. Dantu, "Traffic Shaping of Spam Botnets," *Proc. Fifth IEEE Conf. Consumer Comm. and Networking*, Jan. 2008.
[42] B.A. Huberman and T. Hogg, "Complexity and Adaptation," *Physics D*, vol. 2, nos. 1-3, pp. 376-384, 1986.
[43] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. Gnecchi-Ruscone, A. Malliani, and S. Cerutti, "Measuring Regularity by Means of a Corrected Conditional Entropy in Sympathetic Outflow," *Biological Cybernetics*, vol. 78, no. 1, pp. 71-78, Jan. 1998.
[44] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.

[45] B. Yerazunis, "CRM114 - the Controllable Regex Mutilator," http://crm114.sourceforge.net, Sept. 2009.
[46] Google, "Google Safe Browsing API," http://code.google.com/apis/safebrowsing/, Feb. 2010.
[47] "Phishtank, Join the Fight Against Phishing," http://www.phishtank.com/, Aug. 2011.
[48] "Uribl, Realtime Uri Blacklist," http://http://www.uribl.com/about.shtml, Aug. 2011.
[49] "Surbl," http://www.surbl.org/lists, Aug. 2011.
[50] "The Spamhaus Project," http://www.spamhaus.org/, Aug. 2011.
[51] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.
[52] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 832-844, Aug. 1998.
[53] R. Kohavi and R. Quinlan, "Decision Tree Discovery," *Handbook of Data Mining and Knowledge Discovery,* pp. 267-276, Univ. Press, 1999.
[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The Weka Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter,* vol. 11, pp. 10-18, 2009.
[55] G. McLachlan, K. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data.* Wiley, 2004.

**Zi Chu** received the PhD degree in computer science from the College of William and Mary, Virginia, in 2012. He is a software engineer with Twitter, Inc., San Francisco, California. His research interests include web technologies, machine learning, data mining, Internet anomaly detection, and security privacy.

**Steven Gianvecchio** received the PhD degree in computer science from the College of William and Mary in 2010. He is a senior scientist and principal Investigator at the MITRE Corporation in McLean, Virginia. His research interests include networks, distributed systems, network monitoring, intrusion detection, traffic modeling, and covert channels.

**Haining Wang** received the PhD degree in computer science and engineering from the University of Michigan at Ann Arbor in 2003. He is an associate professor of computer science at the College of William and Mary, Williamsburg, Virginia. His research interests lie in the area of security, networking system, and distributed computing. He is a senior member of the IEEE.

**Sushil Jajodia** received the PhD degree from the University of Oregon, Eugene. He is a university professor, BDM International professor, and the director of Center for Secure Information Systems in the Volgenau School of Engineering at the George Mason University, Fairfax, Virginia. The scope of his current research interests encompasses information secrecy, privacy, integrity, and availability. He has authored or coauthored six books, edited 38 books, and conference proceedings, and published more than 400 technical papers in the refereed journals and conference proceedings. He is also a holder of 10 patents and has several patent applications pending. He has supervised 26 doctoral dissertations. Nine of these graduates hold tenured positions at the US universities; four are US National Science Foundation (NSF) CAREER awardees and one is US Department of Energy (DoE) Young Investigator awardee. Two additional students are tenured at foreign universities. He received the 1996 IFIP TC 11 Kristian Beckman award, 2000 Volgenau School of Engineering Outstanding Research Faculty Award, 2008 ACM SIGSAC Outstanding Contributions Award, and 2011 IFIP WG 11.3 Outstanding Research Contributions Award. He was recognized for the most accepted papers at the 30th anniversary of the IEEE Symposium on Security and Privacy. His h-index is 71 and Erdos number is 2. He is a senior member of the IEEE. The URL for his webpage is http://csis.gmu.edu/jajodia.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Authorized licensed use limited to: Erin Block. Downloaded on June 23,2023 at 14:56:49 UTC from IEEE Xplore. Restrictions apply.