A Study on Measurement and Classification of Twitter Accounts

N. Ashwan Kumar[1], Jayendra Kumar[2]

[1]M.Tech (CSE), Anurag Group of Institutions (Formerly CVSR College of Engineering), JNTUH, Hyderabad, Telangana, India

[2]M.Tech (CSE), Assistant Professor - CSE, Anurag Group of Institutions(Formerly CVSR College of Engineering), JNTUH, Hyderabad, Telangana, India

Abstract: ***Twitter is a new web application playing dual roles of online social networking and microblogging. 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. 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.***

Keywords: Automatic identification, Twitter, Social networks

1. Introduction

Twitter is a popular online social networking and micro-blogging tool, which was released in 2006. Remarkable simplicity is its distinctive feature. Its community interacts via publishing text-based posts, known as tweets. Social media—mobile and web-based applications that allow people to communicate and share information across multiple platforms—is experiencing rapid growth and is being adopted by many. How and why such technology diffuses is a question of current import, as it is adding new dimensions to human interaction. Internet chat is a popular application that enables real time text-based communication. Millions of people around the world use Internet chat to exchange messages and discuss a broad range of topics on-line.

Internet chat is also a unique networked application, because of its human-to-human interaction and low bandwidth consumption [9]. However, the large user base and open nature of Internet chat make it an ideal target for malicious exploitation. The abuse of chat services by automated programs known as chat bots, poses a serious threat to on-line users. Chat bots have been found on a number of chat systems, including commercial chat networks, such as AOL [15], Yahoo![19] and MSN [16],and open chat networks, such as IRC and Jabber. There are also reports of bots in some non-chat systems with chat features, including online games, such as World of War craft [7] and Second Life [2]. Chat bots exploit these on-line systems to send spam, spread malware, and mount phishing attacks. So far, the efforts to combat chat bots have focused on two different approaches: (1) keyword-based filtering and (2) human interactive proofs. The keyword-based message filters, used by third party chat clients [2,3],suffer from high false negative rates because bot makers frequently update chat bots to evade published keyword lists. The use of human interactive proofs, such as CAPTCHAs [1], is also ineffective because bot operators assist chat bots in passing the tests to log into chat rooms [5, 6]. In August 2007, Yahoo! Implemented CAPTCHA to block bots from entering chat rooms, but bots are still able to enter chat rooms in large numbers. There are online petitions against both AOL and Yahoo![8,9], requesting that the chat service providers address the growing bot problem. While on-line systems are besieged with chat bots, no systematic investigation on chat bots has been conducted. The effective detection system against chat bots is in great demand but still missing.

2. A Few Chirps about Twitter

Online social networks (OSNs) have emerged recently as the most popular application since the Web began in the early 1990s.Coincident with the growth of Web 2.0 applications (such as mashups, user generated content) and users being treated as_rst class objects, numerous social networks along with thousands of helper applications have arisen. Well known ones include Facebook, MySpace, Friendster, Bebo, hi5, and Xanga, each with over forty million[13]registered users. Many applications have been created to use the distribution platform provided by OSNs. For example, popular games like Scrabulous, allow many thousands of users on Facebook to play the game with their social network friends. A few smaller networks with supercial similarities to the larger OSNs have started recently. Some of these began as simple helper applications that work well with the larger OSNs, but then become popular in their own right. A key distinguishing factor of these smaller networks is that they provide a new means of communication. In the case of Twitter [11]it is Short Message Service (SMS [18]), a store and forward best effort delivery system for text messages. In the case of *qik*, it is streaming video from cell phones. Jaiku [10], another small OSN, allows people to share their .activity stream", while Dodge ball [6]lets users update their status along with _ne-grained geographical information, allowing the system to locate friends nearby. GyPSii[8], a Dutch OSN is aimed at the mobile market exclusively, combining geo-location of users with image uploading and works on various cell phones including Apple's iPhone. Close to Twitter, a mobile OSN that encourages constant updates is Bliin [3]. For example, Twitter messages can be received by users as a text message on their cell phone, through a Facebook application that users have added to their Facebook account to see the messages when they log in, via email, as an RSS feed, or as an Instant Message (with a choice of Jabber, Google Talk etc.). Figure 1 shows the various input and output vectors to send and receive Twitter status update messages (.tweets"). Twitter isan example of a micro-content OSN, as opposed to say, YouTube, where individual videos uploaded are much larger. Individual tweets are limited to 140 characters in Twitter. Twitter began in October 2006 and is written using Ruby on Rails [16]. Our study _ends that users from a dozen countries are heavily represented in the user population but significantly less than the U.S. Recently, Twitter has made interesting inroads into novel domains, such as help during a large-scale _re emergency [4], updates during riots in Kenya [1], and live traffic updates to track commuting delays [12].

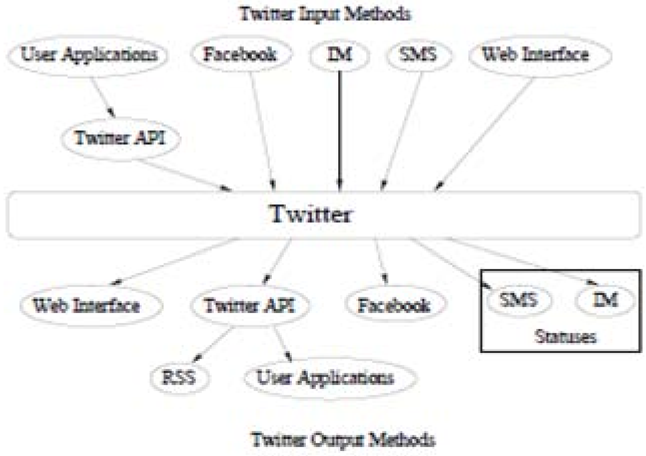
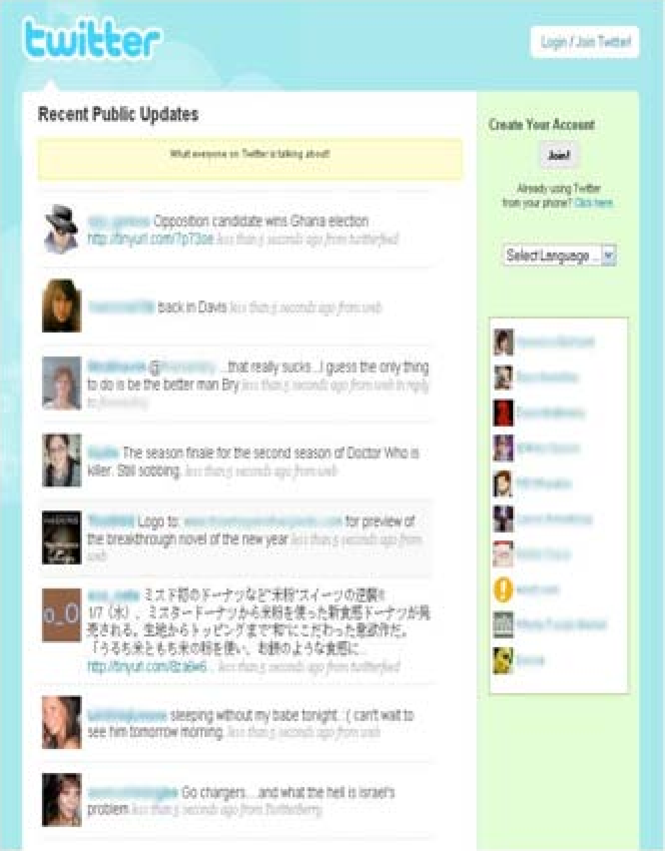


Figure 1: Twitter input and output methods

3. Research Background

3.1 Data Collection

Using the Twitter search API6 we collected publicly available tweets during the four events of study. As a security feature Twitter users can choose to make their profile either public or private. All tweets sent by a public profile are publicly available for anyone to view, even those without an account. These public tweets are also aggregated into a tweet stream called the public timeline (see Figure 2 for an example), which lets anyone view what people are tweeting about at a given time. If a user marks their profile as private, their tweets can only be viewed by other users that they have given permission to follow them, so these tweets are not ones we could sample.



Opposition candidate wins Ghana election http://tinyurl.com/7p73oe

back in Davis

...that really sucks...I guess the only thing to do is be the better man Bry

The season finale for the second season of Doctor Who is killer. Still sobbing.

Logo to: for preview of the breakthrough novel of the new year

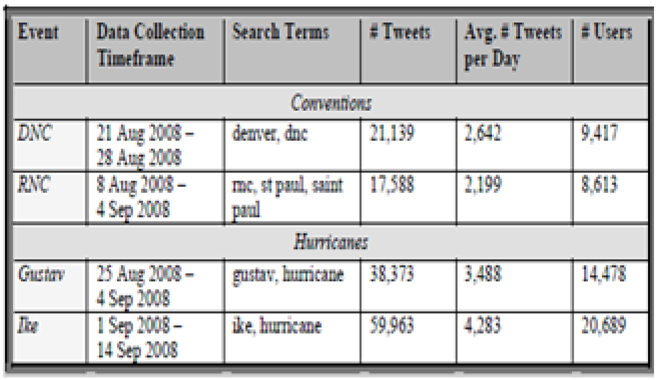
ミスド初のドーナツなど"米粉"スイーツの逆襲!! 1/7（水）、ミスタードーナツから米粉を使った新食感ドーナツが発売される。生地からトッピングまで粉にこだわった意欲作だ。「うるち米ともち米の粉を使い、お餅のような食感に... http://tinyurl.com/8za6w6

sleeping without my babe tonight. :(can't wait to see him tomorrow morning.

Go chargers....and what the hell is israel's problem

Figure 2: Example of the Twitter public timeline

Data collection timeframes (see Table 1) for each event were determined by the nature of the event. Both the DNC and the RNC started on a Monday and ended on a Thursday. However, there were many pre-convention activities and so data capture began the Thursday before continuing until the last day of the convention, rendering eight consecutive days of data collection for each event. For the two hurricanes, data collection began the day each hurricane was officially named and continued until the hurricane was declared over. Table 1 also describes how many tweets were captured in each data set, including the number of unique Twitter users sending these tweets [19].



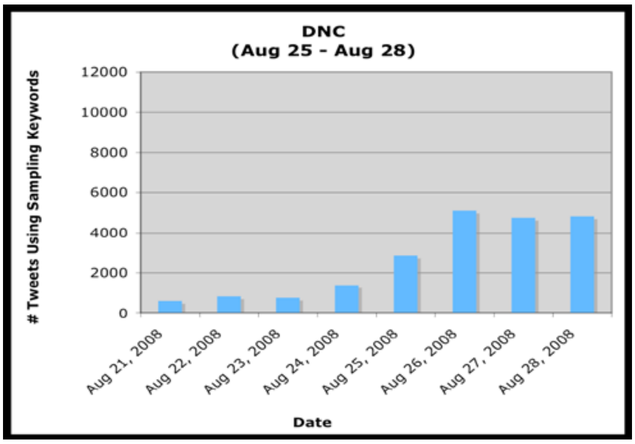
Event	Data Collection Timeframe	Search Terms	# Tweets	Avg. # Tweets per Day	# Users
		Conventions			
DNC	21 Aug 2008 – 28 Aug 2008	denver, dnc	21,139	2,642	9,417
RNC	8 Aug 2008 – 4 Sep 2008	rnc, st paul, saint paul	17,588	2,199	8,613
		Hurricanes			
Gustav	25 Aug 2008 – 4 Sep 2008	gustav, hurricane	38,373	3,488	14,478
Ike	1 Sep 2008 – 14 Sep 2008	ike, hurricane	59,963	4,283	20,689

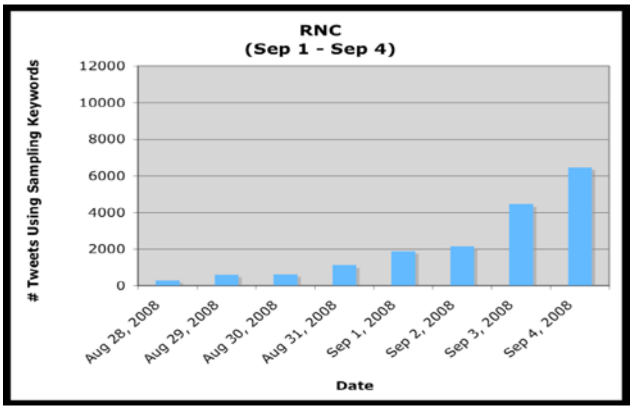
Table 1. Description of the collection criteria and the data collected for the four events.

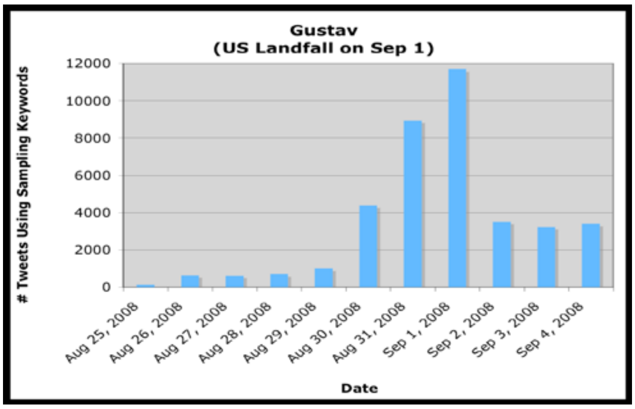
Tweets were selected using high-level, case-insensitive search terms (see Table 1). Ideally we would have included searches based on location but, unfortunately, the location field on a user profile is an editable field that is only specified or updated if the user chooses to do so. We found that inclusion of a location search returned too many irrelevant tweets and so we did not use this information in the data collection.

3.2 Daily Twitter Activity

Twitter activity varied over the days of each event, with the graphs of this activity (see Figure 3) corresponding with the significant happenings of the events they reflect. For example, both the DNC and RNC show the number of tweets, according to our sampling method, was highest on the designated days of each convention—August 25-28, 2008 and September 1-4, 2008 respectively. Hurricane Gustav experienced the highest number of tweets according to our sampling method on September 1, 2008, the day it hit landfall in the US. For Hurricane like two spikes in activity appear, one when it made landfall in Cuba on September 8, and another when it model and fall in the US on September 13, 2008.







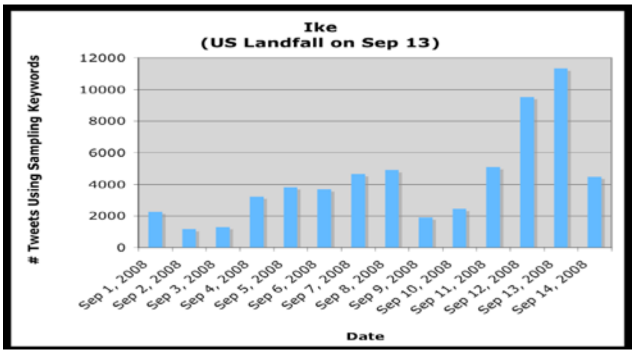


Figure 3: Graphs of the number of daily tweets our research sampled using specific keywords.

Similarly, the number of tweets collected for each event corresponds with the size and impact of each event (see Table 1). Tweets collected for the DNC, the larger of the two conventions studied, outnumbered those collected for the RNC by more than 20%. Hurricane tweet collection totals were far more than any of the convention totals due to the larger geographical impact of the Hurricanes. Comparison of the two hurricanes, shows that Hurricane Ike which had the larger impact, financially speaking—estimated $27 billion in damages (Masters,2008b)—had much higher tweet activity than Hurricane Gustav—estimated $4-14 billion in damages (Masters,2008a). Because we cannot be sure our search selection yielded completely comparable samples, we can only speculate that there is a correlation here. But these preliminary results suggest that the quantity of Twitter activity measured correlates to both size and significance of happenings.

3.3 Number of Tweets per User

To understand how many tweets each user in our data contributed to the Twitter conversation around each event, we determined the tweet count for each user. Users within each data set were then sorted according to their tweet count, after which we calculated the percentage of users who contributed one tweet for each event. We then performed the same percentage calculation for those who contributed two tweets up to seven tweets. We chose a limit of seven tweets because over 95% of the users in each of the four data sets contributed seven or less tweets to the Twitter conversation around each corresponding event. Somewhat surprisingly, we found the percentage of users who sent a certain number of tweets to be consistent across events, which can be seen more clearly in Figure 4. This suggests similar patterns of macro Twitter behavior: that the number of Twitter senders decreases as the number of messages sent increases. This supports—but does not prove—the idea that people serve as "information hubs" (Palen and Liu, 2007) to collect and deploy information, but that many others "participate" in the event in a more peripheral fashion.

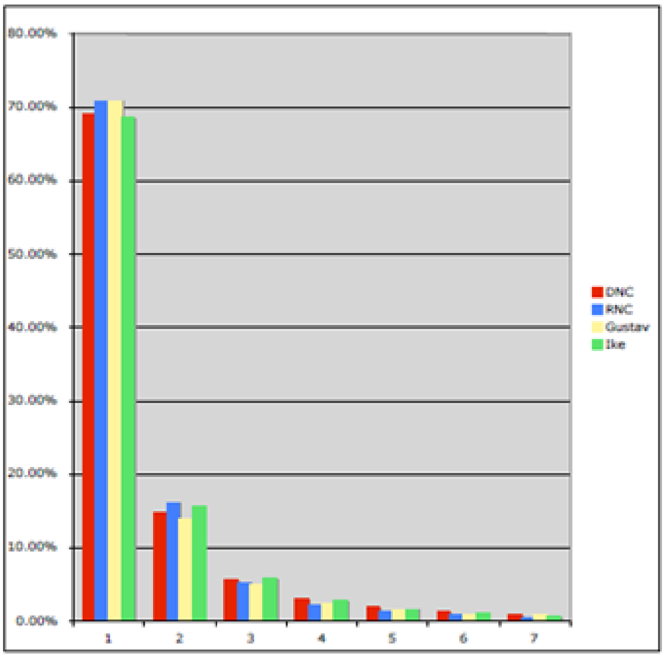
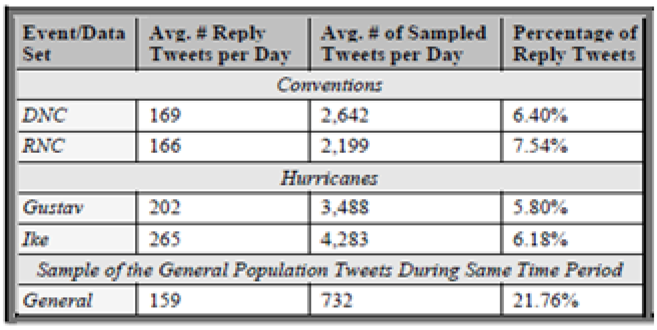


Figure 4. Graph of the percentage of users who sent a total of 1-7 tweets containing the search keywords.

3.4 Reply Tweets

Within the world of Twitter, a norm has evolved such that a sender can designate a tweet as a specific reply to another user, even when the tweets are publicly broadcast. Users begin these reply messages with the "@"symbol directly followed by the username of the person being replied to (ie.@KCTV5). The message is then typed after this reply signifier. Reply messages are a way of getting the attention of a specific user by directing a public tweet message that anyone can read to a specified recipient. We compared how many reply tweets occur in our data sets with the number of reply tweets contained in a random sample of all Twitter tweets (including those around our events of study) during our entire data collection timeframe, August 21, 2008 – September 14, 2008 (see Table 2), to see if there were any differences. To begin gathering a random sample of all tweets, we discovered that approximately 27 million tweets were sent during the designated timeframe [18]. Therefore, we set up a random sampling method designed to obtain a data set of roughly 27,000 tweets. However, when making requests, not all tweets are publicly readable. In fact, we found that roughly 30% of all the tweets we tried to collect are marked private; consequently, the actual sample is 18,308 tweets, despite making requests to obtain around 27,000 tweets.



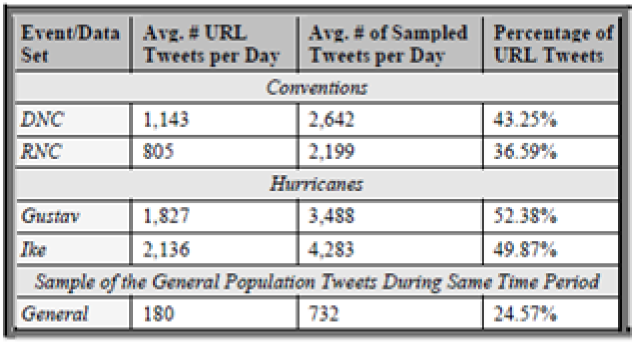
Event/Data Set	Avg. # Reply Tweets per Day	Avg. # of Sampled Tweets per Day	Percentage of Reply Tweets
Conventions			
DNC	169	2,642	6.40%
RNC	166	2,199	7.54%
Hurricanes			
Gustav	202	3,488	5.80%
Ike	265	4,283	6.18%
Sample of the General Population Tweets During Same Time Period			
General	159	732	21.76%

Table 2. Percentage of tweets in each data set that are reply tweets.

Notably, the percentage of reply tweets found in the random tweets data sample was much higher than that of our convention and hurricane data samples (see Table 2). We hypothesize this could be for several reasons. The first is that more broadcast-based information sharing activities happen during mass convergence and crisis events, where the user is pushing information out to many users and not directing it toward one specific user. A second reason is that a reply implies that there is some prior context between the user sending the reply tweet and the user the tweet is directed to. In this case, the user sending the reply tweet may not repeat key contextual words like "dnc" or "denver" because the user they are directing their message to would already know they are in Denver at the DNC. Our Twitter search methods would not pick up a reply tweet like this one, which may have contributed to the lower event reply tweet percentages found in Table 2.

3.5 URL Tweets

Twitter allows users to include URLs in their tweets. This is useful for multiple reasons. Sometimes the 140character limit for Twitter messages can be too constricting when a user wants to convey large amounts of information. Other times, tweets serve as pointers to resources that followers might find interesting or important. Readers of the tweet can then follow the URL to a website with a click on the link. Again, we wanted to compare how many tweets in our data sets contain URLs with the number of tweets containing URLs found in a random sample of all tweets appearing in Twitter during our collection time frame. Using the same sample of random tweets we collected in the last section we were able to make this comparison (see Table 3).



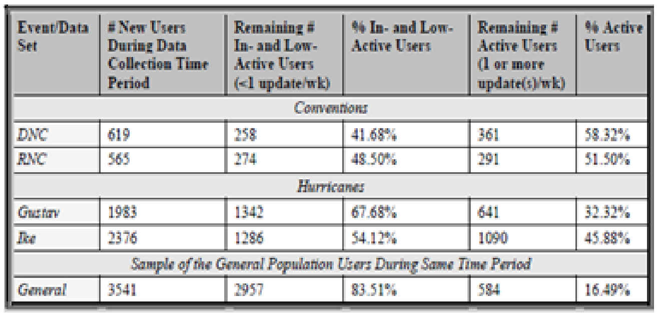
Event/Data Set	Avg. # URL Tweets per Day	Avg. # of Sampled Tweets per Day	Percentage of URL Tweets
Conventions			
DNC	1,143	2,642	43.25%
RNC	805	2,199	36.59%
Hurricanes			
Gustav	1,827	3,488	52.38%
Ike	2,136	4,283	49.87%
Sample of the General Population Tweets During Same Time Period			
General	180	732	24.57%

Table 3. Percentage of Tweets in each data set that contain an URL.

We found the percentage of tweets containing URLs to be notably lower in the general sample than that of our convention and hurricane data samples (see Table 3). This observed behavior supports the idea that users are serving as information brokers, and distributing web-based information resources to others during times of non-routine events. Also notable is the difference in percentage of URL tweets between the two conventions and the two hurricanes. Roughly 40% of the convention tweets contained URLs, while around 50% of the hurricane tweets contained URLs. What could explain this difference is that emergency events have higher information demands than mass convergence but non-emergency events.

3.6 Adoption of Twitter

To better understand Twitter adoption, we collected information about all the new users in each data set. New users are those user accounts that were created during the data collection timeframe for each event. We compared the hurricane-based or convention-based new user data to the general pool of Twitter users, with a random sample of all new Twitter users from August 21, 2008 to September 14, 2008.We examined how many tweets each new user has sent since the time of the original data collection to understand the adoption patterns of these users. To do this, we queried Twitter to find out what the updated tweet count for each new user was on January 8-9, 2009. Using these recent tweet counts we could determine how many of these new users could be considered *active users*. By active users, we mean those users who have contributed one or more tweets every week since the events took place. The elapsed time since the end of the original data collection period (September 14, 2008) to the point of retrieval of updated tweet counts (January 8-9, 2009) is about a period of 17 weeks. Therefore, those users who have a tweet count of 17 or more we call active users. Conversely, low-active or inactive users are those users who have contributed less than one tweet every week, a new adoptees during the hurricane and convention events with less than 17 status updates in the 17weeks since that time.



Event/Data Set	# New Users During Data Collection Time Period	Remaining # In- and Low-Active Users (<1 update/wk)	% In- and Low-Active Users	Remaining # Active Users (1 or more update(s)/wk)	% Active Users
Conventions					
DNC	619	258	41.68%	361	58.32%
RNC	565	274	48.50%	291	51.50%
Hurricanes					
Gustav	1983	1342	67.68%	641	32.32%
Ike	2376	1286	54.12%	1090	45.88%
Sample of the General Population Users During Same Time Period					
General	3541	2957	83.51%	584	16.49%

Table 4. Percentage of new users who have become low-active/inactive and active users.

The percentage of active and inactive users in each data set appears in Table 4. Our collected data shows that there are more accounts who became active users in our hurricane- and convention-event data sets than there are in the general sample. If we define "active user status" as adoption of Twitter technology, then we can see that more users in our data sets (who specifically sent at least one twitter about one or more of the events) adopted Twitter, than a general sample of the new users to Twitter during the same time period. This suggests that when faced with a need and having important and direct experience of usefulness with it, people are more likely adopt a new technology for the long term.

4. Conclusion and Future Scope

In this paper, we have studied the problem of automation by bots and cyborgs on Twitter. The problem of bots on Twitter is further complicated by the key role that automation plays in everyday Twitter usage. We have collected one month of data with good number of 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. Lastly, we have discovered that certain account properties, like external URL ratio and tweeting device makeup, are very helpful on detecting automation. In the future, there is a possibility to block the automated tweets by using any engineering method and also there is a scope to extend this work, which restricts dumping of huge data into

5. Acknowledgements

We are very thankful to all the authors in the references to make this paper in a right way.

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 FrontLines: Emerging Use of Social Media in the 2007 SouthernCalifornia Wildfires," Proc. Int'l ISCRAM Conf., May 2008.

[8] A.L. Hughes and L. Palen, "Twitter Adoption and Use in MassConvergence and Emergency Events," Proc. Sixth Int'l ISCRAMConf., May 2009.

[9] S. Gianvecchio, M. Xie, Z. Wu, and H. Wang, "Measurement andClassification of Humans and Bots in Internet Chat," Proc. 17thUSENIX 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 MyBotnet: 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 ObservationalProofs," Proc. 16th ACM Conf. Computer and Comm. Security, 2009.

[12] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter:UnderstandingMicroblogging Usage and Communities," Proc.NinthWebKDD and First SNA-KDD Workshop Web Mining andSocial Network Analysis, 2007.

[13] B. Krishnamurthy, P. Gill, and M. Arlitt, "A Few Chirps aboutTwitter," Proc.First Workshop Online Social Networks, 2008.

[14] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "DetectingSpam 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 SocialNetworks," Proc. Seventh ACM SIGCOMM Conf. Internet Measurement,2007.

[16] S. Wu, J.M. Hofman, W.A. Mason, and D.J. Watts, "Who SaysWhat 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 SocialNetwork 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 Listsas a Potential Source for Discovering Latent Characteristics ofUsers," 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: TheRole that Micro-Blogging Plays in Informal Communication atWork," Proc. ACM Int'l Conf. Supporting Group Work, 2009.

Author Profile

N. Ashwan Kumar received the B.Tech Degree in Computer Science and Engineering from JNTUH, Hyderabad in 2012 and pursuing M.Tech Degree in Computer Science and Engineering from Anurag Group of Institutions (Formerly CVSR College of Engineering) JNTUH, Hyderabad.

Jayendra Kumar is working as an Assistant Professor in Computer Science Engineering from Anurag Group of Institutions (Formerly CVSR College of Engineering) Hyderabad, Telangana, India. He has received the B.E in Computer Science and Engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya University in 2001, Bhopal, Madhya Pradesh, India and M.Tech in Computer Science and Engineering from JNTUH Hyderabad in 2012.