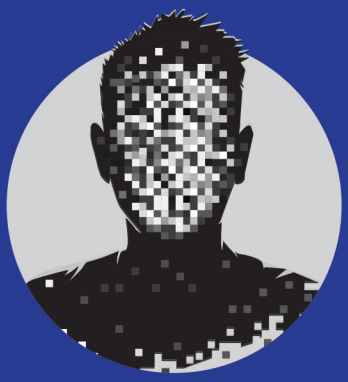
fakespace



The fake account problem on social media platforms

Margaret Gratian, PhD

Email or Phone

Password

Log In

[Photo credit: iStock.com/ne2pi]

Introduction

Every day, major online social media platforms purge millions of accounts from their sites for engaging in inauthentic or deceptive behaviors [1, 2]. Facebook alone reports detecting and banning close to 13.7 billion active fake accounts from their site between October 2017 and June 2020 [1]. Regardless of these actions, social media platforms are plagued by accounts, behavior, and content that are fake or manipulative. There are well-documented and far-reaching consequences; consider, for example, the networks of fake accounts used by nation-state actors for global election tampering [3, 4]. Left unchecked, fake accounts can be used to distribute spam and malware; influence and shape public opinion; defame or impersonate real people; propagate hate and violence; cultivate mass fear, panic, and distrust; and more.

In fact, the COVID-19 global pandemic has reintroduced the phrase *infodemic*—first coined in 2003 to describe the spread of false information during the SARS outbreak in Asia [5]—into public discourse. The reintroduction of the term highlights the importance of countering false information about the virus (e.g., its origins, how it spreads, and how it can be prevented) to encourage the public to take virus precautions seriously. While the research on COVID-19 misinformation (i.e., accidentally misleading information) and disinformation (i.e., intentionally misleading information) is still emerging (at the time this article was written), research suggests that fake accounts—specifically automated bot accounts on Twitter—are largely responsible for promoting political conspiracies about the virus in the United States [6]. In the United Kingdom, the National Health Service has reportedly worked to shut down fake Twitter accounts purporting to be hospital accounts and using that identity to spread falsehoods as a trusted source [7]. Facebook's July 2020 Coordinated Inauthentic Behavior Report details entire networks of accounts taken down for COVID-19 misinformation and disinformation, such as a coordinated group of 303 Facebook accounts and 31 Instagram accounts operating across Asia, Europe, and the United States [8].

The ability to rapidly detect and remove inauthentic or fake accounts is therefore not only crucial for maintaining the integrity of online platforms and protecting users from abuse and manipulation but also has serious implications on public health.

Though current headlines may suggest otherwise, detecting and mitigating deceptive or fraudulent accounts and behaviors is an old problem that emerges in new forms with new challenges. In the early 2000s, researchers focused on countering email spam by developing techniques and rules to identify low-reputation IP addresses and domain names [9]. Later, e-commerce platforms battled similar problems by developing reputation systems to mitigate the effects of dishonest buyers and sellers [10]. Although we face a host of new problems today, it is worth noting that the old ones have not gone away. As detection techniques for deceptive activities advance, so do evasion techniques. Thus, the fight never ends—combating spam and assessing email and domain reputation continue to be active areas of research; fake sellers, products, and reviews are still prevalent on major e-commerce platforms such as Amazon and eBay.

As with the problem of detecting email spam and fraudulent sellers, the problem of detecting fake accounts on social media platforms has been studied and tackled in different forms over the last decade. For example, only a few years after its launch in 2004, Facebook began efforts to crack down on fake accounts with a real name policy mandating that a Facebook account must match a user's real identity [11]. Accounts with names that did not follow Facebook's expectations of a "real" name—those with "unusual capitalization, repeating characters or punctuation"—were, and in many cases still are, required to submit government-issued identification to prove their authenticity [12]. Though the policy had some unintended consequences and controversies, Facebook maintained that the policy was crucial for preventing impersonations and fake accounts [13].

By around 2010, the focus on individual problematic accounts shifted, as both industry and academia dedicated their efforts to the problem of Sybil detection—identifying multiple fake accounts controlled by the same user. Around 2015, the work of Russia's Internet Research Agency to manipulate elections brought renewed interest to the problem of identifying networks of coordinated accounts. The years 2018 and 2019 brought major advancements in text generation and image manipulation technology, enabling everyone from sophisticated, malevolent actors to devious hobbyists to rapidly create plausible, automated textual content and realistic deep-learning generated images known as *deepfakes*. In turn, this has created a whole

Return to CONTENTS

new host of impersonation, manipulation, and fraudulent tactics for industry and academia to counter; Facebook, for example, has entire teams dedicated to detecting and countering fake image and video.

The problem of fake accounts is pervasive on nearly every social media platform. Note that the problem of detecting inauthenticity is not limited to users or accounts—consider the problem of fake news [14] and fake reviews [15]. And while these are significant problems, this article focuses specifically on the current state of fake account detection on social media platforms, starting first with a discussion of the nuances of the fake account problem. This is followed by an overview of current approaches for detecting fake accounts, with specific examples of recent work in academia and industry. The conclusion provides a broader discussion of the long-term challenges in this space.

Understanding the fake account detection problem

To fully understand the fake account detection problem, it is important to first learn about key terminology and concepts.

The importance of context

What defines an account as inauthentic, deceptive, or fake? The answer lies in a social media platform's terms of service and community standards. On Facebook, fake accounts are any accounts where users have misrepresented their identities (e.g., using an inaccurate name or age) [16] or constructed an entirely false identity [17]. Facebook explicitly requires accounts to reflect real people. According to their community standards, "Authenticity is the cornerstone of our community. We believe that people are more accountable for their statements and actions when they use their authentic identities" [16].

On Twitter, a fake account has a different meaning almost entirely. Accounts are not obligated to represent real people; for example, Twitter's rules and policies explicitly allow parody accounts [18]. An account on Twitter is deemed inauthentic and subject to removal if it engages in abuse against other users, impersonation, election manipulation, and certain types of account automation [19, 20, 21].

Context matters because techniques to detect fake accounts must adapt to the definition of fake on a particular platform. Consider how Twitter specifically allows parody accounts. On Twitter, it is necessary to differentiate between impersonation accounts and *parody* impersonation accounts, a distinction that may come down to subtle language elements such as tone and humor. Therefore, impersonation detection techniques that work in Facebook's environment of stringent authenticity requirements may not translate to Twitter's environment.

Additionally, context within a specific platform's environment is also important. Later sections of this article present techniques to identify suspicious accounts using social network structure, but for now, consider how the Twitter account of a celebrity differs from that of a non-celebrity. Celebrity accounts will likely be followed by many but follow few in return. Non-celebrity accounts will likely have far fewer followers than a celebrity account, but perhaps many bidirectional relationships (e.g., users follow users who follow them). This is a relatively simple example, but it illustrates how there can be networks of users that are vastly different from each other but equally plausible. The ability to root out anomalous or suspicious networks depends on an understanding of the expected structure of these two different groups.

Insider vs. outsider perspective

The prevalence of fake accounts is arguably one of the biggest problems facing social media platforms. Unsurprisingly, fake account detection is an active area of research in both industry and academia. It is important to note that when researchers external to a social media platform attempt to develop solutions to the fake account problem, they are often doing so with far less insight than those researchers internal to the company. Simply stated, researchers external to the company do not have access to the entire pool of data or insights into users that the company possesses.

This may seem obvious, but it can result in key differences in detection techniques. For example, in 2016, Facebook, Twitter, Netflix, Airbnb, and many other Silicon Valley-based technology companies participated in a conference called Spam Fighting @ Scale [22]. During the event, the major tech companies discussed their techniques to detect inauthentic accounts and

other activities that violated their terms of service. A key method employed across many different companies was the comparison of features associated with the network identity and connectivity of an account to the projected user identity. For example, did the geolocation of the user's IP address match the address listed on the user's profile? Was the user's IP located in a bad neighborhood of IP addresses (i.e., IP address blocks usually associated with malicious activity)? Was the user coming out of Tor nodes or making use of a virtual private network (VPN)? These are all significant red flags for detecting potentially fraudulent activity. However, these flags may not be visible to a researcher studying a platform using public-facing data alone.

The discrepancy in data access among researchers implies that it might be difficult to compare all fake account detection methods equally. Internally developed analytics have a key data advantage. Does this also suggest that external researchers provide no added value to the fake account detection problem? The ability of universities, journalists, security companies, and many others to identify accounts associated with Russia's Internet Research Agency following the 2016 US presidential election would imply otherwise. Regardless, feature sets used in research design and methods should be carefully considered when discussing the recommendations of academic researchers versus those of industry professionals.

Macro versus micro perspective

Another concept to introduce is the macro versus micro perspective for fake account detection. The macro perspective refers to detection techniques that look at the comprehensive identity associated with an account, with the goal of identifying the entire account as fraudulent or inauthentic. The micro perspective refers to detection techniques to identify fraud or inauthenticity in the components or attributes associated with an account. Inauthenticity at the attribute level may not necessarily indicate that the entire account is fake; the user may simply be lying about pieces of their identity. Inauthenticity and a lack of consistency across multiple features may point to an entirely fabricated account.

As an example, consider an account on Facebook. Attributes of a typical account may include a profile picture, a basic biography, and a collection of text posts. Analyzing a profile picture for manipulation is an example of the micro perspective—looking at an aspect of an account and attempting to determine its authenticity. Analyzing cohesiveness across an account's purported age, gender, and cultural background is the macro perspective—looking at multiple aspects of the account and attempting to spot discrepancies that may point to a fake identity.

The line between the macro and micro perspective may be blurry at times. For example, studying the language of the account's text posts may reveal differences in authorship and personality in the posts; this may indicate that multiple people are managing an account, which in turn indicates inauthenticity at both the micro level (i.e., the text posts) and macro level (i.e., the entire account).

Solutions to the fake account problem

With the nuances of the problem now in mind, what do solutions look like? This section highlights current approaches from a macro perspective.

Define normal

At a high level, most technical solutions for fake account detection involve determining "normalcy" for a given social media platform (or community of users on the platform) and identifying accounts that deviate from this norm. So what does "normal" mean? The answer is highly variable and subjective.

To define normal, first consider the social media platform in question and the attributes that compose an account on the platform. For most online social media platforms, accounts can be interpreted as collections of attributes that fall under the following major categories: the infrastructure and network connectivity of the account, the user profile associated with the account, and the user activity on the account. Under the category of infrastructure and network connectivity, attributes may include the device(s), IP address(es), and user agent string(s) associated with an account. Under the category of user profile, attributes may include the user's name, age, and gender on a site like Facebook or LinkedIn; on other sites, such as Twitter, Tumblr, or Reddit, attributes may be limited to a username and account creation date. Finally, under the category of user activity, attributes

Return to CONTENTS

may include friends and other forms of social connectivity and posts, likes, and other forms of engagement on the platform. The appearance and behavior of any or all attributes under these categories help establish a baseline for normalcy. Probability distributions, graphs, summary statistics, or even categorical values, depending on the detection technique, formalize concepts of normalcy.

Uncover the abnormal

Once the baseline for normal or expected account appearance or behavior has been established, there are a variety of techniques that can be used to identify accounts that deviate from this norm. This section introduces three frequently used strategies: graph analysis, temporal analysis, and machine learning.

Graph analysis

Graph analysis is a common fraud detection technique in which users or events are represented as vertices of a graph, and relationships or transactions are represented as edges of the graph. Fraudulent users or activities can then be identified by looking for anomalous structural patterns or subgraphs within the graph. Graph analysis has proven to be a highly effective technique for detecting fake users, fake reviews, fake financial transactions, and a variety of other abusive behaviors on online platforms.

Many graph analysis-based detection techniques rely on the assumption that there are specific graph structures associated with genuine communities of users [23] and that these organic connections and structures are hard to fake. For example, in [24], the authors observe that fake accounts on both social media platforms such as Twitter and Facebook and e-commerce or review platforms such as Amazon or Tripadvisor end up with many edges, which result in large and dense regions in an adjacency matrix representation of the graph. It is also assumed and often observed that fake accounts will generally have many connections to other fake accounts and few connections to authentic accounts, making it possible to identify densely connected networks of fake accounts, especially if there are known authentic users in the graph [25]. However, here is a prime example of where context and understandings of normalcy matter. As observed by the authors of [25], on certain platforms, such as Twitter or Tumblr, it is expected that users interact with strangers, meaning that connections between a known authentic account and an account of unknown authenticity does not necessarily prove anything about the status of the unknown account.

Temporal analysis

Temporal analysis techniques involve identifying anomalous patterns of activity associated with an account's behavior over time. Temporal analysis is a highly successful technique to identify automated activity (i.e., bots). For example, in [26], the authors developed a bot detection technique for Twitter on the premise that humans are indifferent to the specific second or minute in which they Tweet, meaning that an "organic" sequence of Tweet times should appear to be randomly sampled from a uniform distribution. An automated account, however, will likely result in timing distributions that are either too uniform or not uniform enough.

Temporal approaches can incorporate insights into typical activities on a platform; for example, there is an entire body of literature to draw from to understand usage of hashtags and retweets on Twitter [27]. Research has found that the activities in which authentic accounts engage are very different from those in which inauthentic accounts engage; real users spend more time interacting with accounts that are part of their social network, while fake accounts spend more time attempting to build their social network [28, 29]. For example, on Facebook, a fake account will spend more time "friending" other users than chatting with existing friends.

Machine learning

Machine-learning approaches involve translating attributes associated with an account into features. These features are then used for clustering groups of similar accounts together or differentiating between categories of accounts (e.g., fake or real). In a study done at LinkedIn, researchers used supervised machine learning to classify clusters of accounts as either malicious or legitimate [30]. Features were derived from attributes associated with user-generated profile information, such as name, email address, and company or university. Features included distributions, frequencies, and patterns found in user-generated profile

text. Logistic regression, support vector machine, and random forest models were trained on LinkedIn data that was grouped by account registration IP address and registration date. The study proved highly successful and the random forest model, which achieved area under the curve (AUC) values as high as 0.98, was moved to LinkedIn's production environment. By 2015, when the study was published, the model had already been used to identify 250,000 fake accounts.

Machine-learning approaches for fake account detection have been widely explored in both industry and academia since about 2010. It is worth noting that the primary focus of this work is not on making significant advances in the machine-learning algorithms themselves. Rather it is on identifying novel attributes and transforming them into features or refining existing models to lower false positive and false negative rates. Use of logistic regression models, support vector machines, and random forests, as done in the LinkedIn study, is quite common.

Additional approaches to fake account detection

While the previous section focused on technical approaches to fake account detection, incorporating both technical and nontechnical approaches is an important strategy for combating fake accounts.

Phish the phishers

Some researchers take a honeypot approach to detecting fake accounts. For example, in [31], the researchers created Twitter bots that posted nonsensical messages. Any account that followed these bot accounts was determined to be a fake account, since any authentic, non-automated account would likely not follow or engage with these garbage accounts.

Leverage user reporting and manual review

In both research and practice, effective fake account detection often involves coupling human review with automated techniques. In the literature, relying on user reports of fake, abusive, or suspicious accounts is sometimes referred to as *crowdsourcing* bot detection [25]. In many fake account detection studies, manual review is a final step in the detection pipeline; automated techniques narrow potentially millions of fake accounts down to thousands or even hundreds for a human to review [32].

Human review is important because humans may be able to detect subtle differences between authentic and inauthentic accounts that feature sets do not capture or models fail to discover. Additionally, there is rarely ground truth data about which accounts are actually fake. Understanding why automated methods flag an account as fake (or fail to detect an account as fake) can also help researchers refine both their data sets and tools. For example, in the LinkedIn study referenced earlier, clusters of users were assigned a probability indicating how likely that cluster was to contain fake accounts. Depending on the probability, suspected fake clusters were either automatically suspended or passed to a human for manual review. Manually reviewed and labeled accounts then became training data in later model iterations [30].

Take legal action

In March of 2019, Facebook and Instagram filed a lawsuit against the People's Republic of China for "promoting the sale of fake accounts, likes and followers" [33]. By going after the industry of curated Facebook accounts and reputation, Facebook made an attempt to stem the flow of fake accounts at the "creation source" to prevent individuals and organizations (in particular, those with less resources than a nation-state actor) from simply buying accounts in order to become active players in the fake account space.

Challenge suspicious accounts

In addition to tackling the fake account industry, online platforms incorporate many checkpoints that attempt to make fake account creation as challenging as possible. CAPTCHAs and phone verification are all fraud and abuse countermeasures that most people encounter at some point, even though social network platforms try to limit the number of accounts they challenge in order the keep the user experience as frictionless as possible [30]. Facebook and other major platforms have also used the practice of quarantining users, in which suspicious accounts are sectioned off to a part of the platform where they are not interacting with the real network and then are monitored [22].

Return to CONTENTS

Fraud detection research at NSA

At NSA's Laboratory for Telecommunication Sciences, fake account detection is a key research focus area. Research is done at both the macro and micro level to assess the authenticity of cyber identities, often referred to as *digital personas*. Though a digital persona may encompass more than a social media account, the techniques and perspectives discussed in this article for social media fake account detection are still widely applicable. Personas are interpreted in terms of the three high-level categories discussed previously: the infrastructure and network connectivity of the persona, any account profiles or biographical details associated with the persona, and any online activities conducted by the persona.

Defining normalcy is central to this research. In practice, defining normalcy is a challenging problem since normal must be understood at both the micro level (e.g., the attributes that fall under each of the three categories) and the macro level (e.g., the persona as a whole). To define normalcy, open-source data is used to develop models, represented as probability distributions, which provide insight into the expected values of persona attributes. For example, market trend data may be used to construct a probability distribution representing web browser usage. A persona's use of a web browser other than Firefox, Chrome, or Microsoft Edge may then be used as a red flag.

Of course, the web browser example is oversimplified. Looking at one attribute in isolation is unlikely to provide much insight; models are much more useful if they provide insight into the cohesiveness and plausibility of attributes with respect to other attributes. For example, models summarizing web browser usage by geographic region could be used to identify a persona accessing the web in one region of the world with a browser that is generally only found in another region—this is a much more significant red flag. So, to better understand the relationship between attributes, models are also constructed to represent the expected values of persona attributes given values of other persona attributes. Bayesian probability is at the core of this approach—what is the probability of value X for attribute A given known value Y for attribute B? Suspicious or inauthentic personas are uncovered by looking for co-occurrences of attribute values that rarely, or never, exist in the data.

To solidify this research approach with an example, consider again Facebook's real name policy and the specific language stipulating that real names must not contain "unusual capitalization, repeating characters or punctuation" [12]. This policy has been highly controversial because there are many cases where genuine names do not meet these requirements because they do not fit what is inherently a biased interpretation of normalcy. Bias in data sets and definitions of normalcy result in false positives in practice. For example, Native American names are frequently flagged as inauthentic, resulting in wrongly suspended accounts [34]. Facebook has reportedly introduced a process allowing users to claim an "ethnic minority" or other exception if their name does not meet the real name policy. Though the approach seems well-intended, it does not solve the real issue here: Facebook—and likely many other social media companies—have limited insights into just how diverse normal can actually be. This is also why comprehensive analysis of an account at the macro level is crucial. "Does this name appear real given what I know about the user's ethnicity, cultural background, and various other demographics?" can be a much more useful question than asking "does this name appear real?" without any context.

Conclusion

So what is the state of fake account detection? If we look at reporting from the major social media platforms, we may be inclined to think detection methods are relatively successful. Facebook estimates that roughly 5% of its monthly active users are fake and reports a decline in fake account takedowns since the first quarter of 2019 thanks to their ability to detect fraud early at the account registration step [1]. However, if the last couple of decades of online fraud research can tell us anything, the reality is probably less comforting—low-grade fake accounts may be easy to detect, but sophisticated attackers have likely just become more sophisticated at dodging authenticity checks. These accounts are perhaps the ones we should worry the most about, as time and resources likely went into their curation.

Determining the state of the art in the fake account detection space is also challenging because there are few, if any, public data sets that researchers can use to test, validate, and assess their methods. A quick scan of published studies over the past five years reveals

that most work is conducted on different data sets, even if the same platforms (e.g., Facebook, Twitter, and LinkedIn) are the focus of the work. There are indications that this may be changing though. For example, in September of 2019, Facebook announced the Deepfake Detection Challenge, which provided researchers with deepfake image and video data that were "freely available for the community to use…[with] few restrictions on usage" [35]. Not only did this challenge provide researchers with real data sets to use, but it also made it possible to compare competing solutions for deepfake detection. Moreover, it provided one of the first opportunities to establish benchmarks in the deepfake detection community. The challenge ended in March 2020, with 2,114 participants submitting 35,109 models for deepfake detection using the training corpus of 115,000 videos provided by Facebook. Participant models were tested against a "black box data set with challenging real world examples" [36]. The winning team's model had an accuracy of 65.18%, which Facebook touts as the "new shared baseline" in the artificial intelligence community [36]. The shared data set and new baseline represent a significant step for fake account detection research.

Of course, public data sets and methods have the danger of becoming stale: as defenders learn what techniques to employ to detect fake accounts, attackers can learn how to improve their methods for creating fake accounts. Regardless, one thing is certain—as the world continues to feel the ripple effects of elections manipulated by fake accounts and as social media companies and international organizations work to counter the potentially deadly COVID-19 conspiracies populated by fake accounts—this is a problem in need of critical attention and not going away any time soon.

References

[1] Facebook. "Community standards enforcement report: Fake accounts." Available at: https://transparency.facebook.com/community-standards-enforcement#fake-accounts.

[2] Twitter. "Transparency report: Platform manipulation." Available at: https://transparency.twitter.com/en/platform-manipulation.html.

[3] Bessi A, Ferrara E. "Social bots distort the 2016 US presidential election online discussion." *First Monday.* 2016;21(11).

[4] Bradshaw SHoward PN. "Challenging truth and trust: A global inventory of organized social media manipulation." *The Computational Propaganda Project.* Oxford Internet Institute, University of Oxford; 2018.

[5] Merriam-Webster. "Words we're watching: 'Infodemic.'" Available at: https://www.merriam-webster.com/words-at-play/words-were-watching-infodemic-meaning.

[6] Ferrara, E. "What types of COVID-19 conspiracies are populated by Twitter bots?" *First Monday.* 2020;25(6). doi: 10.5210/fm.v25i6.10633.

[7] "Coronavirus 'fake news' Twitter accounts shut down." *BBC News.* 2020 Mar 10. Available at: https://www.bbc.com/news/uk-england-hampshire-51805311.

[8] Facebook. "July 2020 coordinated inauthentic behavior report." 2020 Jul. Available at: https://about.fb.com/wp-content/uploads/2020/08/July-2020-CIB-Report.pdf.

[9] Zhang H, Duan H, Liu W, Wu J. "IPGroupRep: A novel reputation based system for anti-spam." In: *2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing;* 2009 Jul 7–9; Brisbasne, QLD, Australia. doi: 10.1109/UIC-ATC.2009.15.

[10] Jøsang A, Ismail R, Boyd C. "A survey of trust and reputation systems for online service provision." *Decision Support Systems.* 2007;43(2):618–644.

[11] Ortutay B. "Real users caught in Facebook fake-name purge." *SFGATE.* 2009 May 25. Available at: https://www.sfgate.com/business/article/Real-users-caught-in-Facebook-fake-name-purge-3231397.php.

[12] Facebook. "Help center: What names are allowed on Facebook?" Available at: https://www.facebook.com/help/112146705538576.

[13] Facebook Safety. 2015 June 1. Available at: https://www.facebook.com/fbsafety/posts/861043117266861.

[14] Lazer DMJ, Baum M, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, et al. "The science of fake news." *Science.* 2018;359(6380):1094–1096.

[15] Heydari A, Tavakoli MA, Salim N, Hedari Z. "Detection of review spam: A survey." *Expert Systems with Applications.* 2015;42(7):3634–3642.

[16] Facebook. Community Standards. 17. Misrepresentation. Available at: https://www.facebook.com/communitystandards/misrepresentation.

[17] Facebook. Community Standards. 20. Inauthentic behavior. Available at: https://www.facebook.com/communitystandards/inauthentic_behavior.

[18] Twitter. Help Center. Twitter rules and policies. Impersonation policy. Available at: https://help.twitter.com/en/rules-and-policies/twitter-impersonation-policy.

[19] Twitter. Help Center. General guidelines and policies. Abusive behavior. Available at: https://help.twitter.com/en/rules-and-policies/abusive-behavior.

[20] Twitter. Help Center. General guidelines and policies. Civic integrity policy. Available at: https://help.twitter.com/en/rules-and-policies/election-integrity-policy.

[21] Twitter. Help Center. General guidelines and policies. Automation rules. Available at: https://help.twitter.com/en/rules-and-policies/twitter-automation.

[22] Spam Fighting 2016: Spam Fighting@Scale. 2016 Nov 3. Available at: https://atscaleconference.com/events/spam-fighting-2016/.

[23] Prakash A, Sridharan A, Seshadri M, Machiraju S, Faloutsos C. "EigenSpokes: Surprising patterns and scalable community chipping in large graphs." Available at: http://www.cs.cmu.edu/afs/cs.cmu.edu/user/christos/www/PUBLICATIONS/pakdd10-eigenspokes.pdf.

[24] Hooi B, Song HA, Beutel A, Shah N, Shin K, Faloutsos C. "FRAUDAR: Bounding graph fraud in the face of camouflage." In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining;* 2016 Aug 13–17; San Francisco, CA: pp. 895–904. doi: 10.1145/2939672.2939747.

[25] Ferrara E, Varol O, Davis C, Menczer F, Flammini A. In: *Communications of the ACM;* 2016 Jun. "The rise of social bots." doi: 10.1145/2818717.

[26] Zhang CM, Paxson V. "Detecting and analyzing automated activity on Twitter." In: Spring N, Riley GF, editors. *Passive and Active Measurement. PAM 2011. Lecture Notes in Computer Science, vol 6579.* Springer, Berlin, Heidelberg. pp. pp 102–111. doi: 10.1007/978-3-642-19260-9_11.

[27] Bruns A, Stieglitz S. 2012. "Quantitative approaches to comparing communication patterns on Twitter." Queensland University of Technology. Brisbane, Australia. Available at: https://eprints.qut.edu.au/55823/1/Quantitative_Approaches_to_Comparing_Communication_Patterns_on_Twitter.pdf.

[28] Wang G, Konolige T, Wilson C, Wang X, Zheng H, Zhao BY. "You are how you click: Clickstream analysis for Sybil detection." In: *SEC '13: Proceedings of the 22nd USENIX Security Symposium;* 2013 Aug 14–16; Washington, DC: pp. 241–256.

[29] Yang Z, Wilson C, Wang X, Gao T, Zhao BY, Dai Y. (2014). "Uncovering social network Sybils in the wild." *ACM Transactions on Knowledge Discovery from Data.* 2014;8(1). doi: 10.1145/2556609.

[30] Xiao C, Freeman DM, Hwa T. "Detecting clusters of fake accounts in online social networks." In: *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (AISec '15);* 2015 Oct 16; Denver, Colorado: pp. 91–101. doi: 10.1145/2808769.2808779.

[31] Lee K, Eoff BD, Caverlee J. "Seven months with the devils: A long-term study of content polluters on Twitter." In: *Fifth International AAAI Conference on Weblogs and Social Media;* 2011 Jul.

[32] Cao Q, Sirivianos M, Yang X, Pregueiro T. "Aiding the detection of fake accounts in large scale social online services." Available at: https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final42_2.pdf.

[33] Grewal P. "Sale of fake accounts, likes, and followers." 2019 Mar 1. Facebook. Available at: https://about.fb.com/news/2019/03/sale-of-fake-accounts-likes-and-followers/.

[34] Bowman J. "Facebook flags aboriginal names as not 'authentic'." CBC. 2015 Feb 25. Available at: https://www.cbc.ca/news/trending/facebook-flags-aboriginal-names-as-not-authentic-1.2970993.

[35] Schroepfer M. "Creating a data set and a challenge for deepfakes." *Facebook AI.* 2019 Sep 5. Available at: https://ai.facebook.com/blog/deepfake-detection-challenge/.

[36] Facebook AI. "Deepfake Detection Challenge Results: An open initiative to advance AI." 2020 Jun 12. Available at: https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.