Measuring Social Spam and the Effect of Bots on Information Diffusion in Social Media

Emilio Ferrara

1 Introduction

Social media have received widespread recognition as enablers of modern society communication [14, 18, 55, 56, 58], as a tool to democratize discussion about politics [2, 10, 15, 25, 26, 61, 90] and social issues [9, 22, 23, 40, 41, 81, 84], and even as an effective system to respond to crises and emergencies [39, 57, 78, 91, 92].

The benefits of the rise to popularity of social media are hard to quantify, as they touch billions of people every day, all over the world. However, as early as 2006, concerns have been raised regarding the possibility of manipulating public opinion through social media [44]. Particularly problematic can be the fact that social media have proved effective in influencing individuals, their beliefs and behaviors [7, 17, 33, 54, 67]. These concerns have been later proved well grounded by several scientific studies, which highlighted a variety of manipulation strategies and related contexts where such forms of abuse can take place [27, 30, 32, 45, 66, 72, 73, 86].

One way to manipulate social media is by using social bots, algorithmically-controlled accounts that emulate the activity of human users but operate at much higher pace (e.g., automatically producing content or engaging in social interactions), while successfully keeping their robotic identity undisclosed [36, 46, 65, 85].

Evidence of the adoption of social media bots to attempt manipulating political communication dates back nearly a decade: during the 2010 U.S. midterm elections, social bots were employed to support some candidates and smear others, by

E. Ferrara (✉)
University of Southern California, Information Sciences Institute, Los Angeles, CA, USA
e-mail: emiliofe@usc.edu

S. Lehmann, Y.-Y. Ahn (eds.), *Complex Spreading Phenomena in Social Systems*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-77332-2_13

injecting thousands of tweets pointing to websites with fake news [71]. The research community reported another similar case around the time of the 2010 Massachusetts special election [66]. Campaigns of this type are sometimes referred to as astroturf or Twitter bombs. Unfortunately, most of the times, it has proven impossible to determine who's behind these types of operations [11, 36, 53]. Governments, organizations, and other entities with sufficient resources can obtain the technological capabilities to deploy thousands of social bots and use them to their advantage, either to support or to attack particular political figures or candidates.

Bots have been used in other contexts too, most prominently for social spamming and social phishing purposes [48, 50, 69, 74, 82, 83, 89]. A large body of scientific literature covers the challenges related to detecting social spam [38, 63, 94], spam bots [12, 59, 60, 76], fake reviews [69], etc. Differently from traditional Internet spam, distributed via email or mailing lists, social spam proliferates in online platforms, and bots have been extensively used to make its diffusion more effective. Although much work has been devoted to characterize and detect social spam campaigns or spam bots, the interplay between these two, and in particular the effect of spam bots on the diffusion of spam in social media, has not received much attention.

1.1 Contributions of This Chapter

This chapter aims at investigating both the directions of social bots influence on political discussion and spam bots influence in social spam campaigns. In particular, we will be concerned with measuring the role and effects of bots in social media information spreading dynamics. The scope and contributions of this chapter are therefore threefold:

- We will first review how social bots, and in particular Twitter bots, are created, how they operate, and what are the challenges in detecting them (see Sect. 2). The literature discussed here will be mostly aligned with a recent review paper we published on *Communications of the ACM* [36].
- We will then discuss how social bots have been used during the 2016 US Presidential Election to sway the discussion around the presidential candidates, and to frame agendas and messages attaching particular sentiments. This review (see Sect. 3.1) will be based on results we recently published [11].
- Then, we will propose novel analysis of the effects of social spam bots on the diffusion of social spam campaigns and promotional content on Twitter (see Sect. 3.2). We will investigate the differences between traditional spammers and social spam bots, provide a characterization of their most typical features, and describe their effect of the diffusion of social spam on Twitter.

2 What Social Bots Are and How They Operate

2.1 How to Create a Social Spam Bot

In the early days of online social media, over one decade ago, creating a bot was not a simple task: a skilled programmer would need to sift through various platforms' documentation to create a software capable of automatically interfacing with the platform and operate functions in a human-like manner. For example, in 2009, we spent significant amounts of efforts to create a simple bot that would navigate Facebook pages and extract basic publicly-available social network information [16]: that required the application of sophisticated Web scripting techniques [35] in conjunction with a trial-and-error approach to deal with the Web platform infrastructure. Similar efforts have been reported for other such type of early endeavors [4, 20]

These days, the landscape has completely changed: indeed, it has become increasingly simpler to deploy social bots, so that, in some cases, no coding skills are required to set up accounts that perform simple automated activities: tech blogs often post tutorials and ready-to-go tools for this purposes. Various source codes for sophisticated social media bots can be found online as well, ready to be customized and optimized by the more technically-savvy users [53].

We inspected same of the readily-available Twitter bot-making tools and this is a (non-comprehensive) list of capabilities they provide:

- Search Twitter for phrases/hashtags/keywords and automatically retweet them;
- Automatically reply to tweets that meet a certain criteria;
- Automatically follow any users that tweet something with a specific hashtag, keyword, or phrase;
- Automatically follow back any users that have followed the bot;
- Automatically follow any users that follow a specified user;
- Automatically add users tweeting about something to public lists;
- Search Google (and other engines) for articles/news according to specific criteria and post them, or link them in automatic replies to other users;
- Automatically aggregating public sentiment on certain topics of discussion;
- Buffer and post tweets automatically.

Most of these bots can run within cloud services or infrastructures like Amazon Web Services (AWS) or Heroku, making it more difficult to block them when they violate the Terms of Service of the platform where they are deployed.

Finally, a very recent trend is that of providing Bot-As-A-Service (BaaS): companies like RoboLike[1] provide "Easy-to-use Instagram/Twitter auto bots" performing certain automatic activities for a monthly price. Advanced conversational

[1]RoboLike: https://robolike.com/.

bots powered by sophisticated Artificial Intelligence are provided by companies like ChatBots.io that allow anyone to "Add a bot to services like Twitter, Hubot, Facebook, Skype, Twilio, and more".[2]

2.2 How to Detect Social Bots

The detection of social bots in online social media platform has proven a challenging task. For this reason, it has attracted a lot of attention from the computing research community. Even DARPA became interested to the point that a DARPA Challenge was organized, namely the 2016 DARPA Twitter Bot Detection [77]: over one dozen academic and industry teams participated, with University of Maryland, University of Southern California, and Indiana University topping the challenge.

For these reasons, the literature on social bot detection has become very extensive. We tried to summarize the most relevant approaches in a survey paper recently appeared on *Communications of the ACM* [36]: we refer the interested reader to that review for a deeper analysis of this problem.

In our review, we proposed a simple taxonomy to divide the social bot detection approaches proposed in literature into three classes: (1) bot detection systems based on social network information; (2) system based on crowd-sourcing and leveraging human intelligence; (3) machine learning methods based on the identification of highly-revealing features that discriminate between bots and humans. In the following, we report some examples of these three classes.

2.2.1 Graph-Based Social Bot Detection

Social bot detection has been framed as an adversarial setting [6]: an adversary may control multiple social bots to impersonate different identities and infiltrate a system. Proposed detection strategies often rely on examining the structure of a social graph, and assume that bot accounts exhibit a small number of links to legitimate users, connecting mostly to other bots. This feature is exploited to identify densely interconnected groups of bots. Yet, a wise attacker may counterfeit the connectivity of the controlled bot accounts; this strategy would make the attack invisible to these detection methods. To address this shortcoming, some systems also employ the paradigm of *innocent by association*: an account interacting with a legitimate user is considered itself legitimate. Unfortunately, the effectiveness of such detection strategies is bound by the behavioral assumption that legitimate users refuse to interact with unknown accounts. This was proven unrealistic by various experiments [13, 29, 76]. On other platforms like Twitter and Tumblr, connecting and interacting with strangers is one of the main features. In these circumstances, the

[2]Pandora bot: https://developer.pandorabots.com/.

innocent-by-association paradigm yields high false positive rates. Moreover, real-world platforms may contain many mixed groups of legitimate users who fell prey of some bots [6], and sophisticated bots may succeed in large-scale infiltration making it impossible to detect them solely from network structure information. Despite its high false-positive rate, social network information can complement other sources of information to improve prediction accuracy, as demonstrated by prior work [36].

2.2.2 Crowd-Sourcing Social Bot Detection

Some authors suggested crowd-sourcing social bot detection, assuming that it would be a simple task for humans to evaluate an account's behavior and to observe emerging patterns and anomalies associated with bots [88]. Using data from Facebook and Renren (a popular Chinese online social network), the authors tested the efficacy of human detectors, using both expert annotators and workers hired online. Although this strategy exhibited a near-zero false positive rate, it has proven unfeasible for several reasons: for existing platform with large user bases, like Facebook and Twitter, manually verify millions of suspicious accounts has a prohibitive cost; even if large social network companies could afford to hire teams of analysts for this purpose [75], such cost might not be sustainable for small social networks in their early stages; finally, exposing personal information to online workers for annotation would raise privacy issue [28].

2.2.3 Feature-Based Social Bot Detection

Encoding behavioral patterns into features, in conjunction with machine learning techniques to learn the signature of human and bot behavior, may be the most popular bot detection strategy. One example of feature-based system is represented by *Bot or Not*: released in 2014, and constantly updated, this was the first Twitter bot detection tool to be made publicly available [24].[3] *Bot or Not* implements a detection algorithm relying upon highly-predictive features capturing a variety of suspicious behaviors to separate social bots from humans. The system employs off-the-shelf supervised learning algorithms trained with examples of both humans and bots behaviors. In addition to the classification results, *Bot or Not* provides a variety of interactive visualizations that yield insights on the features exploited by the system. We will later describe how we used *Bot or Not* for our studies.

Bots are continuously changing and evolving: the analysis of the highly-predictive behaviors that feature-based detection systems can detect may reveal interesting patterns and provide unique opportunities to understand how to discriminate between bots and humans. User meta-data are considered among the most predictive features and the most interpretable ones [46, 88]: we can suggest few

[3] http://truthy.indiana.edu/botornot.

rules of thumb to infer whether an account is likely a bot, by comparing its metadata with that of legitimate users. Further work, however, will be needed to detect sophisticated strategies exhibiting a mixture of humans and social bots features (sometimes referred to as *cyborgs*). Detecting these bots, or hacked accounts [93], is currently impossible for feature-based systems. Recent studies suggested that some advanced social bots may no longer aim at mimicking human behavior, but rather at misdirecting attention to irrelevant information [1]: such *smoke screening* strategies, requiring high degree of coordination among bots, can also escape feature-based detection systems.

3 Applications and Case Studies

In the following, we present two case studies. We first study the use of social bots in the context of the 2016 US Presidential Election (cf. Sect. 3.1). The results we present are based on recently published work [11]. Then, we discuss new results on the effect of bots on the diffusion of social media spam (cf. Sect. 3.2).

3.1 Case Study 1: Political Campaigns

In the introduction of this chapter, we discussed at length the widespread abuse of social media platforms. In the context of political campaigns, one could try to boost the popularity of a candidate, for example by creating the impression that there is an organic support behind that candidate; however, the apparent support can be artificially generated by means of orchestrated campaigns. This phenomenon is commonly referred to as *astroturf*, and it has long-lasting roots, starting from offline campaigns [62], and evolving, during more recent times, into various forms of Internet [52] and social media [72] campaigns. We report our study of social media astroturf in the context of the 2016 US Presidential Election next, with a special focus on the role of social bots. We discuss data collection first, then we go over the employed bot detection and sentiment analysis approaches. The case study concludes with some discussion of the insights our analysis yielded.

3.1.1 Data Collection

We manually crafted a list of hashtags and keywords related to the 2016 US Presidential Election. The list was compiled so that to contain a roughly equal number of hashtags/keywords associated with each major presidential candidate: we selected 23 terms in total, including 5 terms specifically for the Republican Party nominee Donald Trump (#donaldtrump, #trump2016, #neverhillary, #trumppence16, #trump), 4 terms for the Democratic Party nominee Hillary Clinton

(#hillaryclinton, #imwithher, #nevertrump, #hillary), and several terms relative to the four presidential debates. The full list of search terms is reported in our paper [11]. By querying the Twitter Search API at regular intervals of 10 s, continuously and without interruptions in three periods between September 16 and October 21, 2016, we collected a large dataset constituted by 20.7 million tweets posted by nearly 2.8 million distinct users. We used the Twitter Search API[4] to obtain all tweets that contain the search terms, posted during the data collection period, rather than a sample of unfiltered tweets: this avoids incurring in the issues reported in the literature related to collecting sample data from the Twitter Stream API[5] instead [68].

3.1.2 Bot Detection

Determining whether either human or a bot controls a social media account has proven a very challenging task [36, 77]. Our prior efforts produced an openly accessible solution called Bot Or Not [24], consisting of a Python API[6] and a Website.[7] As we briefly discussed earlier, Bot Or Not is a machine-learning framework that extracts and analyzes a set of over one thousand features, spanning content and network structure, temporal activity, user profile data, and sentiment analysis to produce a score that suggests the likelihood that the inspected account is indeed a social bot. Extensive analysis revealed that the two most important classes of feature to detect bots are, maybe unsurprisingly, the metadata and usage statistics associated with the user accounts.

The following indicators provide the strongest signals to separate bots from humans: (1) whether the public Twitter profile looks like the default one or it is customized (it requires some human efforts to customize the profile, therefore bots are more likely to exhibit the default profile setting); (2) absence of geographical metadata (humans often use smartphones and the Twitter iPhone/Android App, which records as digital footprint the physical location of the mobile device); (3) and activity statistics such as total number of tweets and frequency of posting (bots exhibit incessant activity and excessive amounts of tweets), proportion of retweets over original tweets (bots retweet contents much more frequently than generating new tweets), proportion of followers over followees (bots usually have less followers and more followees), account creation date (bots are more likely to have recently-created accounts), randomness of the username (bots are likely to have randomly-generated usernames). We point the reader interested in further technical details to our prior work [24, 36].

[4]Twitter Search API: https://dev.twitter.com/rest/public/search.

[5]Twitter Stream API: https://dev.twitter.com/streaming/overview.

[6]Bot or Not Python API: https://github.com/truthy/botornot-python.

[7]Bot or Not Website: https://truthy.indiana.edu/botornot/.

Bot Or Not has been trained with thousands of instances of social bots, from simple to sophisticated, and an accuracy of above 95% [24]. Typically, Bot Or Not yields likelihood scores above 50% only for accounts that look suspicious to a scrupulous analysis. We adopted the Python Bot Or Not API to systematically inspect the most active users in our dataset. The Python Bot Or Not API queries the Twitter API to extract the most 300 tweets and all the publicly available account metadata, and feed this features to an ensemble of machine learning classifiers, which produce a bot score. To label accounts as bots, we use the 50% threshold—which has proven effective in prior studies [24, 36]—an account is considered to be a bot if the bot score is above 0.5.

Since the Python Bot Or Not API incurs in the query limitations imposed by the Twitter API, it would have been impossible to test all the 2.78 million accounts. Therefore, we tested the top 50 thousand accounts ranked by activity volume. Although these top 50 thousand users account for roughly only 2% of the entire population, it is worth noting that they are responsible for producing over 12.6 million tweets, which is about 60% of the total conversation. This choice gives us sufficient statistical power to extrapolate the distribution of bots and humans for the entire population without the need to test accounts that are only marginally involved in the conversation. Out of the top 50 thousand accounts, Bot Or Not assigned a bot score greater than the established 0.5 threshold, and therefore classified as likely bots, to a total of 7183 users, responsible for 2,330,252 tweets. A total of 40,163 users (responsible for 10.3 million tweets) were labeled as humans. Bot Or Not labeled the remainder 2654 users as unknown/undecided, either because their scores does not significantly diverge from the classification threshold of 0.5, or because the accounts have been suspended/deleted. Even if all the 2654 users were bots, and Twitter suspended their accounts for violating the terms of service, this would suggest that roughly 70% of the total bot population (the remainder 7183 accounts) was still active on the platform at the time of our verification. By extrapolating for the entire population, we estimate the presence of at least 400 thousand bots, accounting for roughly 15% of the total Twitter population active in the U.S. presidential election discussion, and responsible for about 3.8 million tweets, roughly 19% of the total volume. Additional statistics are summarized in our paper [11].

3.1.3 Sentiment Analysis

To understand how bots and humans discuss about the presidential candidates we will rely upon sentiment analysis. To attach a sentiment score to the tweets in our dataset, we used SentiStrength [80]. SentiStrength is a sentiment analysis algorithm which has been specifically designed to annotate social media data. This design choice provides some desirable advantages: first, it is optimized to annotate short, informal texts, like tweets, that contain abbreviations, slang, and other non-orthodox language features; second, SentiStrength employs additional linguistic rules for negations, amplifications, booster words, emoticons, spelling corrections, etc. Applications of SentiStrength to social media data found it particularly effective

at capturing positive and negative emotions with, respectively, 60.6% and 72.8% accuracy [79]. We tested it extensively and also used it in prior studies to validate the effect of sentiment on the diffusion of information in social media [33]. The algorithm assigns to each tweet t a positive $P^+(t)$ and negative $P^-(t)$ polarity score, both ranging between 1 (neutral) and 5 (strongly positive/negative). Starting from the polarity scores, we capture the emotional dimension of each tweet t with one single measure, the sentiment score $S(t)$, defined as the difference between positive and negative polarity scores: $S(t) = P^+(t) - P^-(t)$. The above-defined score ranges between -4 and $+4$. The negative extreme indicates a strongly negative tweet, and occurs when $P^+(t) = 1$ and $P^-(t) = 5$. Vice-versa, the positive extreme identifies a strongly positive tweet labeled with $P^+(t) = 5$ and $P^-(t) = 1$. In the case $P^+(t) = P^-(t)$—positive and negative sentiment scores for a tweet t are the same—the sentiment $S(t) = 0$ of tweet t is considered as neutral (note that the neutral class represents the majority, by construction, since it contains all tweets that have equal number of positive and negative words, as well as all tweets with no sentiment-labeled terms).

3.1.4 Partisanship and Supporting Activity

We next inferred the partisanship of the users in our dataset. We used the five Trump-supporting hashtags (#donaldtrump, #trump2016, #neverhillary, #trumppence16, #trump) and the four Clinton-supporting (#hillaryclinton, #imwithher, #nevertrump, #hillary) to attribute partisanships. In detail, we employed a simple heuristics based on hashtag adoption: for each user, we calculated the top ten hashtags that appear in the tweets posted by that user. If the majority of hashtags support one particular candidate, we assigned the given user to that political faction (Clinton- or Trump-supporter). This is a very strict and conservative partisanship assignment, likely less prone to misclassification that may be yield by automatic machine-learning techniques not based on manual validation, e.g., [21]. Our procedure yielded a small, high-confidence, annotated dataset constituted by 7112 Clinton supporters (590 bots and 6522 humans) and 17,202 Trump supporters (1867 bots and 15,335 humans).

3.1.5 Analytic Insight 1: Human vs. Bot Engagement

Figures 1 and 2 show the Complementary Cumulative Distribution Functions (CCDFs) of the interactions respectively replies and retweets, initiated by bot and human users. Each plot disaggregates the interactions in three categories: (1) within group (for example, bot–bot, or human–human); (2) across groups (e.g., bot–human, or human–bot); and, (3) total (i.e., bot-all and human-all). Both figures exhibit broad distributions typical of social media activity. What interestingly emerges from contrasting the two figures is that humans are engaging in replies interactions significantly more (one order of magnitude difference) with other humans than with bots (see right panel of Fig. 1). Conversely, bots fail to substantially engage humans and end up interacting via replies with other bots significantly more than

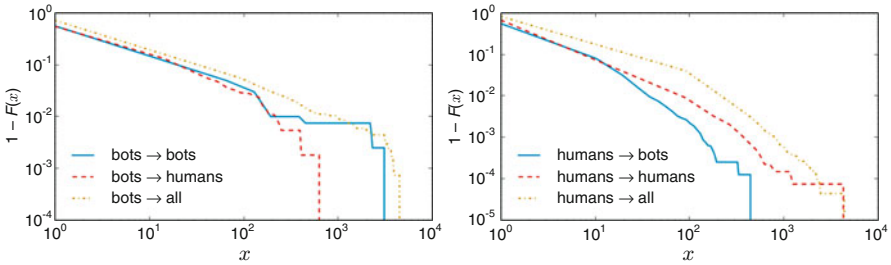

Fig. 1 Complementary cumulative distribution function (CCDF) of replies interactions generated by bots (left) and humans (right) (published in Bessi and Ferrara, 2016 [11])

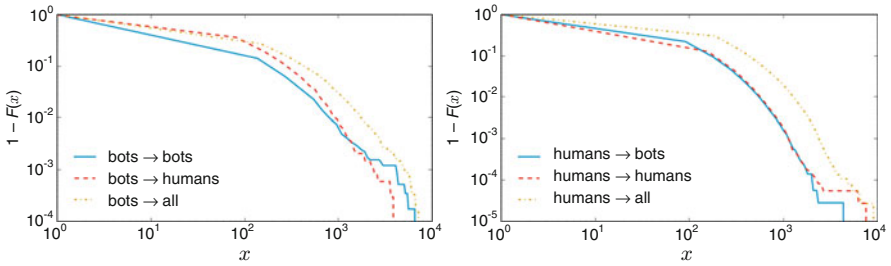

Fig. 2 Complementary cumulative distribution function (CCDF) of retweets interactions generated by bots (left) and humans (right) (published in Bessi and Ferrara [11])

with humans. Given that bots by design are intended to engage in interactions with humans, our observation goes against what we would have intuitively expected—similar paradoxes have been highlighted in our prior work [36]. One intuitive explanation to this phenomenon is that bots that are not sophisticated enough, cannot produce engaging-enough questions to foster meaningful discussions with humans. Figure 2, however, demonstrates that rebroadcasting is a much more effective channel of information spreading: there is no significant difference in the amounts of retweets that humans generate by rebroadcasting content produced by other humans or by bots. In fact, humans and bots retweet each other substantially at the same rate. This suggests that bots are being very effective at spreading information in the human population, which could have some nefarious consequences in the cases when humans fail at verifying the correctness and accuracy of such information and information sources.

3.1.6 Analytic Insight 2: Human vs. Bot Sentiment

To further understand how social media users (both bots and humans) are talking about the two presidential candidates, we explore the sentiment that the tweets convey. To this purpose, we rely upon sentiment analysis and in particular on *SentiStrength*. Figure 3 shows four panels: the top two panels illustrate the sentiment

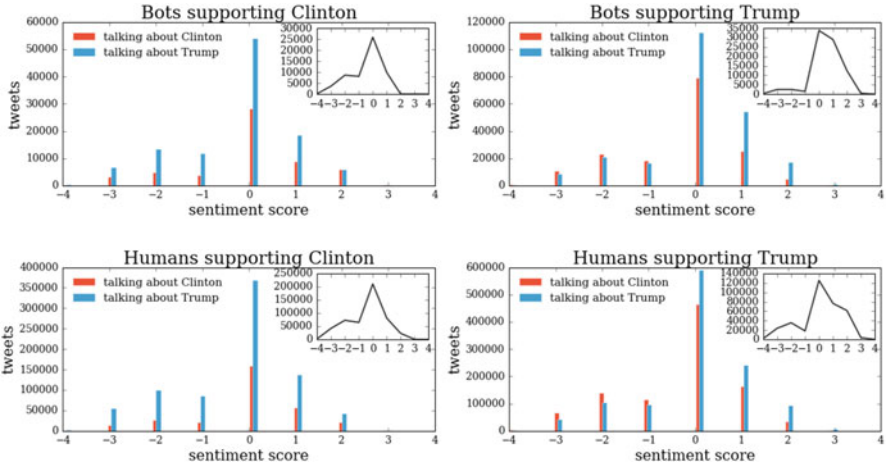

Fig. 3 Distributions of the sentiment of bots (top) and humans (bottom) supporting the two presidential candidates. The main histograms show the disaggregated volumes of tweets talking about the two candidates separately, while the insets show the absolute value of the difference between them (published in Bessi and Ferrara [11])

of the tweets produced by the bots, while the bottom two panels show the same information for tweets generated by humans. Furthermore, the two left panels show the support to Hillary Clinton (respectively by bots and humans), whereas the two right panels show the support to Donald Trump (respectively by bots and humans). The main histograms in each panel show the volume of tweets about Clinton or Trump, separately, whereas the insets show the difference between the two (this to illustrate the disproportion in support of the candidate of one's factions, as opposed to the other candidate). What appears evident from contrasting the left and right panels is that, on average, the tweets produced by Trump's supporters are significantly more positive than that of Clinton's supporters, regardless of whether the source is human or bot. If we focus on Trump's bot supporters, we note that they generate almost no negative tweets; they indeed produce the most positive set of tweets in the entire dataset—a very significant fraction of these non-negative bot-generated tweets (about 200,000 or nearly two-third of the total) are in support of Donald Trump. This generates a stream of support that is at staggering odds with respect to the overall negative tone that characterizes the 2016 presidential election campaigns. The fact that bots produce systematically more positive content in support of a candidate can bias the perception of the individuals exposed to it, suggesting that there exists an organic, grassroots support for a given candidate, while in reality it is all artificially generated. Some interesting insights emerge also from the analysis of Clinton's supporters: on average, human-generated tweets show slightly more positive sentiment toward the candidate than the bot-generated ones. Overall, a more natural distribution of tweets' sentiment emerges from the two groups of bots and human supporters, with a roughly equal number of positive and negative tweets being present in the pro-Clinton discussion. To further understand

these dynamics, we manually analyzed two hashtags, namely #NeverTrump and #NeverHillary, as emblematic examples of campaigns explicitly devoted to target the candidate of one's opposing political leaning. The hashtag #NeverTrump, used by supporters of the Democratic Candidate Hillary Clinton, accrued 105,906 positive tweets, and 118,661 negative ones, roughly an equal split; on the other hand, the hashtag #NeverHillary pushed by Trump's supporters generated significantly more negative tweets (204,418) than positive ones (171,877). The paper [11] reports various examples of tweets generated by bots, and the candidate they support. A final consideration emerges when contrasting the pro-Clinton and pro-Trump factions: the former focuses much more on their candidate, with a significant number of tweets referring to Clinton. Conversely, pro-Trump supporters (humans and bots) devote a significant number of tweets to their opponent: in fact, the majority of negative tweets generated by both humans and bots are addressing Hillary Clinton.

3.2 Case Study 2: Social Spam Campaigns

In the second part of this chapter, we study social spam campaigns. The widespread use of social media makes them an ideal target as a vector to diffuse spam campaigns. Indeed, spam has evolved, moving away from traditional vectors like emails and mailinglists [43], due to the increasing effectiveness of email spam filters, and migrating to social platforms like social media [19, 38, 94] and digital marketplaces [51, 64, 70], etc. In the former scenario, the use of bots has been documented to generate artificial promotional campaigns, to advertise dubious products (whose sale is sometimes illicit), etc. In the latter, bots are exploited to generate and diffuse fake product reviews. Next, we study social media spam, focusing on the effects of social bots in the diffusion of spam campaigns on Twitter. We first discuss social spam data collection, then introduce a tool named *dynamical activity-connectivity map* we recently proposed to study the mechanisms of influence in social media. We conclude studying spam campaigns' sentiment and its interplay with bots' efficacy.

3.2.1 Data Collection

Similarly to the political discussion scenario, we manually crafted a list of hashtags and keywords to collect our data. We focused on the tobacco-related discussion, and in particular electronic cigarettes. We identified this case study by noticing how spam seems to be a pervasive presence in this topic of discussion on Twitter [5]. The list included over one hundred terms covering nicotine-related products (e.g., *tobacco*, *cigar*, *cigarettes*, etc.), electronic cigarettes (multiple variants like *ecig*, *e-cig*, *ecigs*, *e-cigs*, *e-cigarette*, *ecigarette*, etc.), vaping products (e.g., *vape*, *ehookah*, *ejuices*, *eliquids*, etc.), popular vaping brands (e.g., *green smoke*, *eversmoke*,

etc.), health-related terms (e.g., *second-hand smoke*, *second-hand vape*), health campaigns terms (e.g., *still blowing smoke*, *not blowing smoke*, *tobacco free kids*, etc.), and more. We queried the Search API at regular intervals from January 1 to September 30, 2015 and collected a large dataset constituted by over 9 million unique tweets.

3.2.2 Spam Detection

Detecting social spam has proven a challenging and tedious task. The lack of a rigorous definition of what spam is makes detection a complex problem. Although various detection techniques have been proposed in the machine learning literature, they carry some limitations: they are either outdated, being trained and tested on early (2008–2010) Twitter spam data [12, 59, 60, 76], or overly-specific to detect certain types of campaigns [37, 38, 63, 94]. The first limitation becomes a problem due to the fact that bots evolve, becoming increasingly sophisticated thus rendering detection less effective if training data is not current; the latter issue hinders the applicability of detection systems to a broader range of problem domains.

For the reasons above, to detect spam campaigns in our data and separate legitimate tobacco-related discussion from social spam, we implemented a novel strategy. We first performed traditional data cleaning operations on the texts of the tweets in our dataset, namely removing stop-words and punctuation, then tokenizing and stemming the terms. Afterwards, we elaborated the following iterative three-stages detection procedure:

1. We generated a list of keywords appearing in the tweets, ranked by frequency.
2. Then, two independent human annotators manually identified and labeled keywords associated to spam campaigns appearing in the list of the top 250 most common keywords (to provide contextual information, the annotators had access to the full text of some example tweets where such keywords occur).
3. Finally, all tweets containing spam-associated keywords are moved into a separate repository that we will call *spam dataset*; the iterative process then restarts. It is worth noting that, at each next iteration of the algorithm, the ranked list of keywords changes because the spam keywords identified at stage 2 are removed.

The process ended when the list of top 250 most common keywords did not contain any spam-associated term. This yielded a manually-curated list of 87 spam keywords,[8] that appear in the *spam dataset* accounting for 3.06M unique tweets posted by over 850 thousand distinct users. Of these users, about 74K posted more than one tweet. We will focus our attention, for the rest of our analysis, on these 74K active spammers.

[8]The combination of the top 250 non-spam keywords, plus the 87 spam keywords, accounts for over 90% of all tweets in the original dataset.

The top ten most recurring spam keywords, in order of frequency, are: *win*, *dvd*, *movies*, *giveaway*, *deals*, *horror*, *bluray*, *ebay*, *gameofthrones*, *movie*. Manual inspection of the 87 keywords suggests that three main types of social media spam campaigns occur in this scenario:

- Tobacco-related product promotions (sales, coupons, discount codes, etc.);
- Tobacco-unrelated product promotions (sales, coupons, discount codes, etc.), in particular related to entertainment products (dvd, music, books, etc.);
- Topic-hijacking campaigns, i.e., spam that includes tobacco-related keywords to attract the attention of users to tweets related to completely different topics, including movies and TV shows (keywords like *gameofthrones*, *fiftyshades*, *hungergames*, *celebs*, *ageofultron*, *insurgent*, and many others), and offline news events (e.g., *charlestonshooting*, *ericgarner*).

The phenomenon of Twitter hashtag hijacking has been documented extensively [19, 42, 47, 49]. In the following analysis, we do not make a specific distinction between different types of spam campaigns. However, in the future, we will try to determine whether campaign types, as well as different scopes and intents lead to different social spam dynamics.

3.2.3 Descriptive Data Statistics

Our initial exploratory analysis aims at highlighting the temporal dynamics of social spam production. Figure 4 shows the timeline of the volume of spam tweets per day in our dataset. Overall, we can note a mild upward trend over the course of

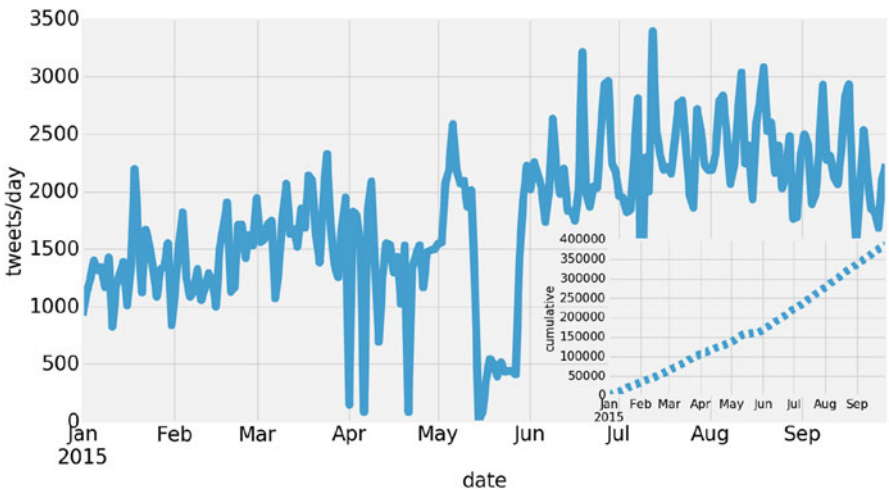

Fig. 4 Timeline of the volume of spam tweets per day during the observation period. The inset shows the cumulative count. A few drops visible in April and May are associated with Twitter data collection service outages

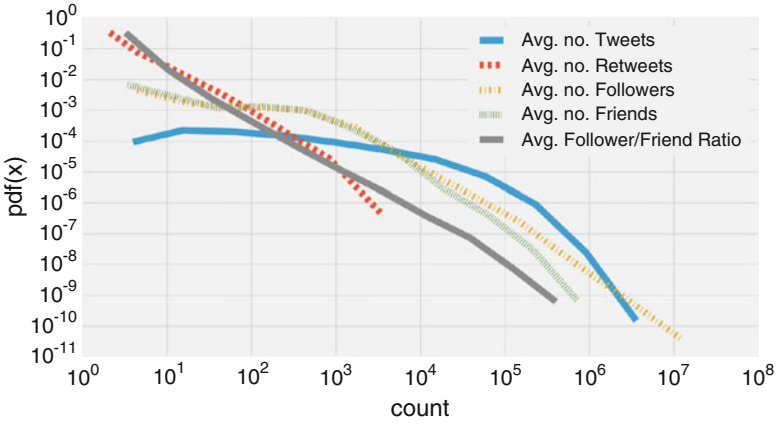

Fig. 5 Distributions of the average number of tweets, retweets, followers, friends, and follower vs friend ratio of the users in our spam dataset

the 9 months of observation. By the end of the year, the volume of tweets per day is roughly twice that of the beginning. This growth suggests the effectiveness of social spam in the tobacco-related context: if ineffective, the cost associated with running social spam campaigns would outweigh their benefits and therefore we would observe declining trends.

After assessing that social spam was "alive and well" during our analysis period, we moved forward to provide a statistical characterization of the actors therein involved: the Twitter spammers. Figure 5 shows the distribution of the average number of posted tweets, obtained retweets, number of followers and friends, and follower vs. friend ratio, for the set of users in our spam dataset. The averages are calculated across the 9-month observation period. A few observations are in order. Firstly, although all distributions exhibit the heavy tails typical of social networks [3, 8], some are significantly different from others. For example, the distribution of posted tweets is somewhat unexpected; if compared with the distribution of obtained retweets, which exhibits the typical power-law like behavior (i.e., a truncated straight line in the log-log plot of Fig. 5), the distribution of posted tweets appears anomalous. In particular, it appears that there is roughly the same probability of observing accounts with a number of posted tweets that spans from a few to over ten thousands: this is represented by the nearly-flat slope of the blue solid curve in the regime $10 \leq x < 10^4$. After that point, the probability decreases very rapidly. This unusual behavior is commonly linked to the activity of social bots. Their activity, however, does not catch up with the lack of influence they are typically characterized by, and therefore the amount of average retweets that most of these accounts receive is orders of magnitude lesser than the amount of tweets they post. Concluding, both the friends and follower distribution exhibit uncommon shapes, suggesting the presence of two different regimes, one for $10 \leq x < 10^3$ and one for $x \geq 10^3$. The slope in the former regime is nearly flat, whereas in the

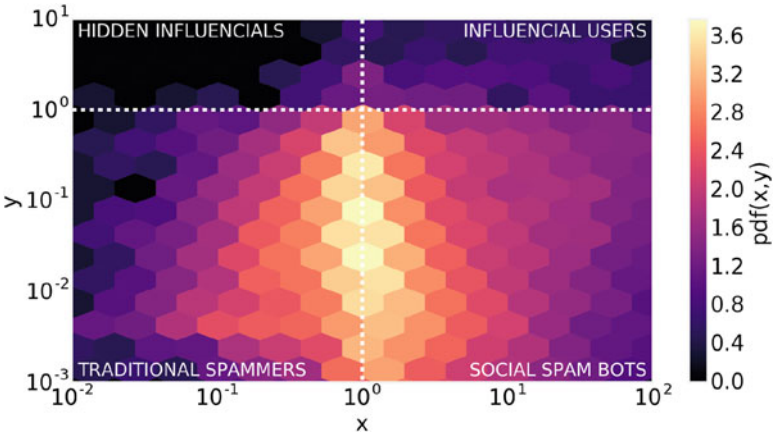

Fig. 6 Dynamical activity-connectivity map of the users in our dataset. The x axis represents the proportional variation of followers/friends for each user over the accounted time period. The y axis represents the proportional variation of received/posted tweets of each user over the time period

latter both distributions decay with more typical heavy tails suggesting the presence of accounts with a very large number of friends and followers, another interesting behavior associated with two types of users: influential individuals, or social bots. Next, we study in detail the relation between activity and connectivity patterns.

3.2.4 Dynamical Activity-Connectivity Maps

The analysis above was static: taking the average values of the five features above made the results oblivious of the temporal dynamics of activity and connectivity as they unfold over the observation time. We now plan to investigate what effect the progression of activity levels of a user has on their connectivity evolution (and viceversa). In Fig. 6 we provide a *Dynamical Activity-Connectivity map*: we recently introduced this type of maps [31, 84] as dynamic variants of the map proposed by Gonzalez-Bailon and collaborators—see Figure 4 in the paper titled *Broadcasters and Hidden Influentials in Online Protest Diffusion* [41].

Figure 6 shows the probability density of users in the two-dimensional space where the x-axis represents the growth of network connectivity, and the y-axis conveys the messaging activity rate. For a given user u, x_u and y_u are here defined as

$$x_u = \frac{1 + \delta f_u}{1 + \delta F_u} \qquad \text{and} \qquad y_u = \frac{1 + \delta r t_u}{1 + \delta t_u}.$$

We use the notations f_u and F_u to identify the number of followers and friends, respectively, of a user u. The variations of followers and friends of user u over a period of time t are thus defined as $\delta f_u = \frac{f_u^{\max} - f_u^{\min}}{t}$ and $\delta F_u = \frac{F_u^{\max} - F_u^{\min}}{t}$; the length of time t is defined as the number of days of u's activity, measured from

registration to last observed activity (this varies from user to user). Finally, the variations of received retweets, and posted tweets, are defined as $\delta rt_u = \frac{rt_u^{\max} - rt_u^{\min}}{t}$ and $\delta t_u = \frac{t_u^{\max} - t_u^{\min}}{t}$, respectively, where rt_u and t_u are the number of obtained retweets and posted tweets by user u during the period of activity t.

All values are added to the unit to avoid zero-divisions and to allow for logarithmic scaling (i.e., in those cases where the variation is zero). The "heat" (the color intensity) in the map represents the joint probability density $pdf(x, y)$ for users with given values of x and y. The plot also introduce a bin normalization to account for the logarithmic binning.

The *Dynamical Activity-Connectivity map* we conceived is interpreted as follows: the bulk of the joint probability density mass should be observed in the neighborhood of $(1, 1)$, as the majority of accounts would usually exhibit a comparable variation along the two dimensions. That would be in line with what all previous social media studies where this type of map was employed reported [31, 41, 84]. However, the results Fig. 6 shows are unprecedented: we hypothesize that this is due to the spam dynamics characterizing this dataset. Let us discuss the two dimensions of *connectivity growth* and *activity rate* separately.

The *connectivity growth* is captured by the x axis and, in our case, ranges roughly between 10^{-2} and 10^2. Users for which $x > 1$ (i.e., 10^0) are those with a followership that grows much faster than the rate at which these users are following others. In other words, they are acquiring social network popularity (followers) at a fast-paced rate. Note that, if a user is acquiring many followers quickly, but s/he is also following many users at a similar rate, the value of x will be near 1. This is a good property of our measure because it is common strategy on social media platforms, especially among bots [11, 36], to indiscriminately follow others in order to seek for reciprocal followerships. Our Dynamical Activity-Connectivity map will discriminate users with fast-growing followerships, who will appear in the right-hand side of the map, from those who adopt that type of reciprocity-seeking strategy. The former group can be associated with highly popular users with a fast-paced followership growth. According to Gonzalez-Bailon and collaborators [41] this category is composed by two groups: *influential users* and *information broadcasters*, depending on their activity rates. Values of $x < 1$ indicate users who follow others at a rate higher than that they are being followed; they fall in the left-hand side of the map. According to Gonzalez-Bailon and collaborators, these are mostly the *common users*, although the so-called *hidden influentials* also sit in this *low-connectivity* regime.

As for what concerns the y axis, it measures the *activity rate*, i.e., the rate at which a user receives retweets versus how frequently s/he tweets. Users with values of $y > 1$ are those who receive systematically more retweets with respect to how frequently they tweet. This group of users can be referred to as *influentials*, i.e., those who are referred to significantly more frequently than others in the conversation; they fall in the upper region of the map, and according to Gonzalez-Bailon et al., depending on their connectivity growth can be divided in influential ($x > 1$) and hidden influential ($x < 1$) users. Conversely, users with values of $y < 1$ are those who post exceedingly more tweets than the retweets they receive. This group

would generally represent the common-user behavior ($x < 1$), although information broadcasters ($x > 1$) also exhibit the same *low-activity* rate. These users fall in the lower region of the map.

Now that a reading of dynamical activity-connectivity maps has been provided, we can proceed with interpreting Fig. 6: the bottom-left quadrant reports the most common users, those with both activity and connectivity growth lesser than 1. In our case, we identify these accounts as traditional spammers. Manual validation of some of these accounts revealed that they employ simple automatic posting strategies, thus they generate a very large number of tweets, but they never attract other users' attention and thus they are rarely retweeted. We identified over 27K such accounts.

Conversely, the upper-right quadrant reports users with the higher connectivity growth and activity rates. These are influential accounts: they systematically attract other users' attention by receiving lots of retweets compared with how often they tweet, and their followerships grow at a very fast pace. Influential users are quite rare in this context, and in fact we identified only 438 users according to our method. Manual inspection of all these users revealed that our technique correctly detects influential users which are not bots: accounts in this category include official accounts of movies and TV shows (e.g., *Avengers*, *CaptainAmerica*, *Divergent*, *GameOfThrones*, etc.), and various official accounts of tobacco-related sellers.

Lastly, social spam bots sit in the bottom-right quadrant. Differently from traditional spammers, their connectivity growth is much more similar to that of influential accounts. Their followership increases at a pace higher than their following others. They still produce disproportionately more tweets than the retweets they receive, but their embeddedness in the social network looks somewhat effective. Further analysis reveals that many of these spam bots tend to reciprocate followership to external users (accounts not present in the spam dataset) but also tend to follow each other; this coordinated behavior gives the appearance of network influence. We identified over 46K social spammers, the majority class by far in our spam dataset. Finally, we detected only 47 hidden influentials, too few to warrant further analysis.

Figure 7 provides a different view on the five features characterizing the users in the three classes. As opposed to spammers, influential users receive significantly more attention (retweets), significantly more followers than friends (thus a much higher followers/friends ratio), and on average post one order of magnitude fewer tweets than bots. Concluding, the only significant difference between traditional spammers and social spam bots is their social network: social bots exhibit more followers than friends on average; the vice versa is true for traditional spam bots.

3.2.5 The Interplay Between Sentiment of Spam Bots

We conclude our analysis with a high-level investigation of the interplay between spam sentiment and spam bot characteristics. We applied the same Sentiment Analysis technique, i.e., *SentiStrength*, as in the previous case study, to our spam dataset. Figure 8 shows the distribution of sentiment scores for the tweets in our

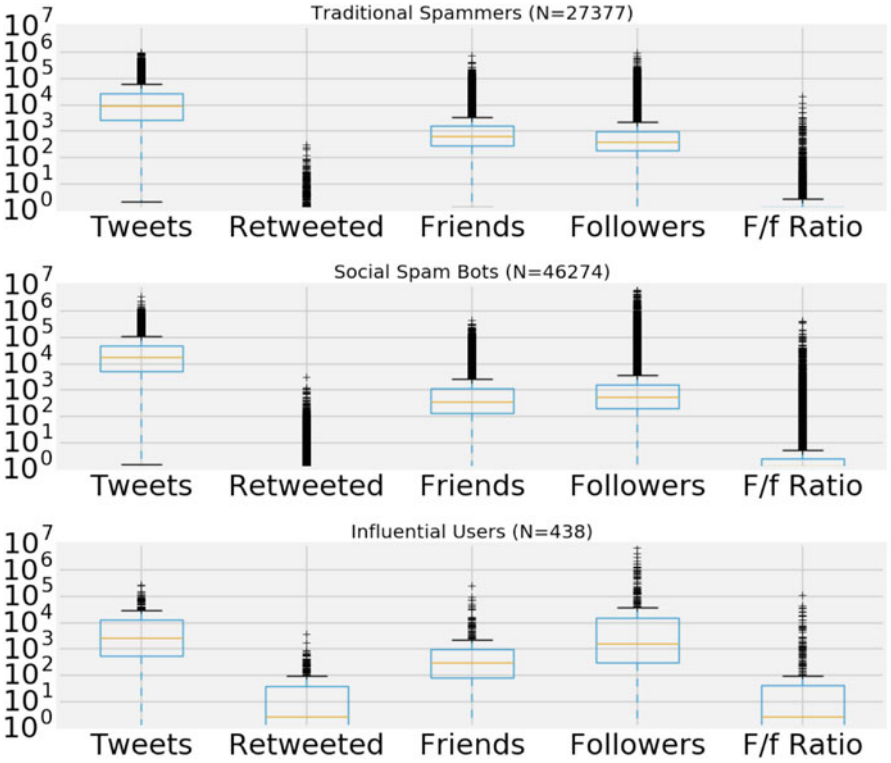

Fig. 7 Box plot of the distributions of posted tweets, obtained retweets, number of friends and followers, and follower/friend ratio for the main three classes of users in our spam dataset

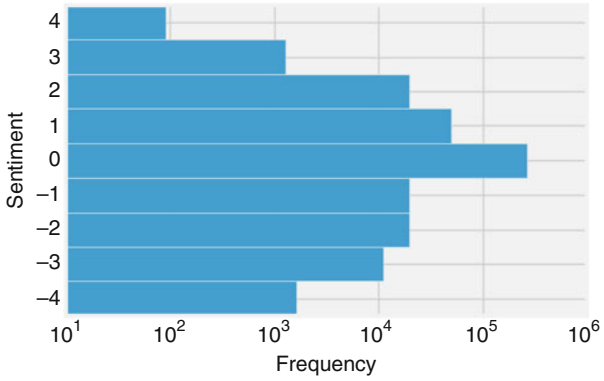

Fig. 8 Distribution of tweet sentiment scores (SentiStrength) in the spam dataset

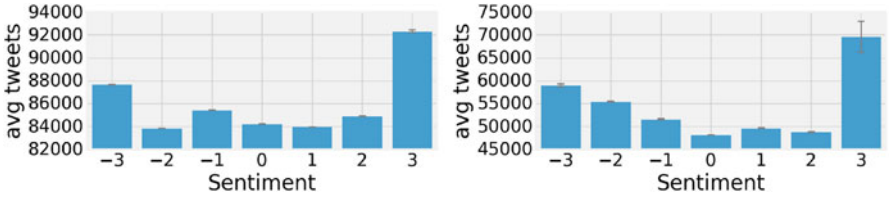

Fig. 9 Average number of tweets posted as a function of tweet's sentiment, calculated only on tweets retweeted at most once (left) and on those that have been retweeted more than once (right)

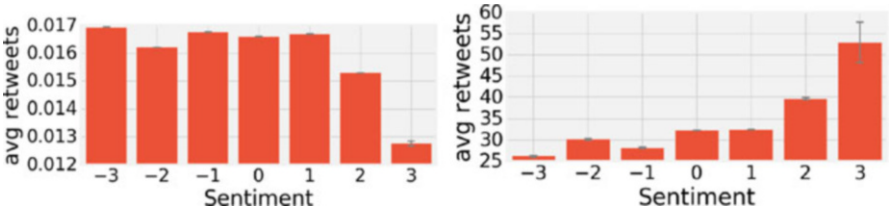

Fig. 10 Average number of obtained retweets as a function of sentiment, calculated only on tweets retweeted at most once (left) and on those that have been retweeted more than once (right)

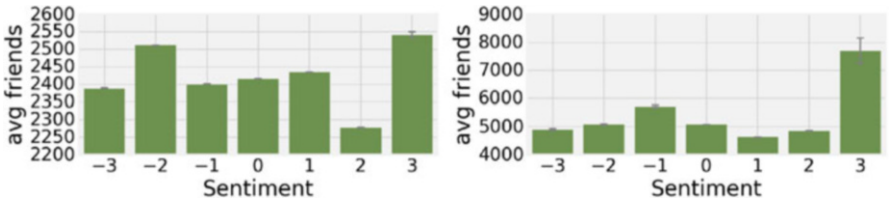

Fig. 11 Average number of user friends as a function of sentiment, calculated only on tweets retweeted at most once (left) and on those that have been retweeted more than once (right)

corpus. The distribution exhibits its typical peak around zero [34, 79]. However, in contrast with respect to previous findings on Twitter sentiment obtained using SentiStrength [34], the distribution in the spam dataset appears skewed toward negativeness. In particular, roughly one order of magnitude more strongly negative tweets ($S \leq -3$) appear than strongly positive ones ($S \geq 3$).

Worth noting, this dataset is significantly smaller and topically biased (i.e., it covers only spam) than the comprehensive Twitter dataset we previously studied [34]: we hypothesize that some correlation may exist between this atypical sentiment distribution and the role of spam bots.

To this purpose, in Figs. 9, 10 and 11 we plotted four features we used to characterize the bots (i.e., *number of posted tweets, obtained retweets, friends, and followers*). All figures report error bars (obtain hardly noticeable) that convey the standard error of the sampled average feature distributions. We will use them for diagnostic purpose, i.e., to highlight anomalies in spam dynamics with respect to

organic social media sentiment [34]. Given the exiguous number of tweets with extremely positive or negative sentiment (i.e., $S = 4$ or $S = -4$), next we will limit our analysis to values of sentiment in the range $-3 \leq S \leq 3$.

The interpretations of the bar plots in Figs. 9, 10 and 11 is the following: given a fixed value of sentiment x, then y is the average value of the selected feature for all tweets with sentiment equal to x. Plots on the left are for the subset of tweets retweeted at most once; plots on the right are for tweets retweeted more than once. The separation is carried out to address the issue of activity heterogeneity highlighted before (cf. Fig. 5) and is necessary to avoid problems like the *Simpson Paradox* [87].

For the sake of example, let us discuss the left panel of Fig. 9 that shows the distribution of the *average number of tweets* posted by users, which were retweeted at most once, as a function of sentiment.

Let us consider sentiment $S = 3$ (there are about 1300 such tweets in our dataset, cf. Fig. 8): the average number of tweets posted by the users who posted one such tweet with sentiment $S = 3$ is about 92K. This is significantly higher than for every other sentiment score, denoting the fact that users who post strongly positive tweets (e.g., promotional tweets) on average posted significantly more tweets than the others. It is also worth noting that an average value of tweets nearing the hundred of thousands clearly denotes very highly-active accounts, and likely some form of automatic posting—a common feature of spam bots.

The right panel of Fig. 9 shows how this pattern is preserved even for the set of tweets that have been retweeted more than once: moreover, the distribution takes a U-like shape, suggesting that also accounts that post negative tweets exhibit much more activity than average. This suggests that some spam campaigns may not be necessarily positive. Indeed, if one compares this result with the previous case study on the manipulation of political campaigns, some interesting similarities emerge. In other words, spam at times can aim to smear some products, e.g., those from competitors.

Figure 10 shows another interesting patterns. The left panel again captures tweets that have been retweeted at most once; the right panel captures more popular tweets and exhibits a striking difference if compared to the left one: increasingly positive sentiment yields significantly more retweets. This is known as *positivity bias*, i.e., the emergence of a strong preference for retweeting positive messages; such bias was already observed in our prior Twitter analysis [34]. Strongly positive tweets obtain on average more than twice the number of retweets than negative or neutral ones. It is worth hypothesizing that, in the spam scenario, this pattern may also conceal some form of coordinated activity, i.e., bots may retweet other bots' spam in an orchestrated fashion.

Further clues supporting this hypothesis come from Fig. 11, in particular the right panel: users associated with positive tweets that are retweeted very often all exhibit a number of friends that are nearly twice as much as others. Inspecting users who follow on average over 7K accounts revealed strong reciprocity—another very common bot characteristic highlighted multiple times above.

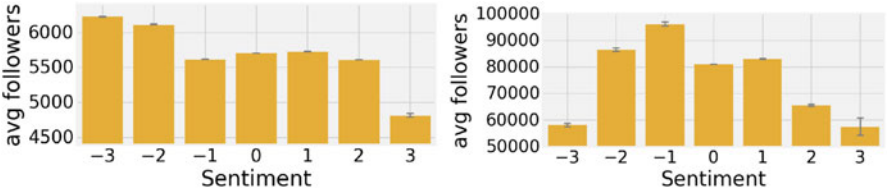

Fig. 12 Average number of user followers as a function of sentiment, calculated only on tweets retweeted at most once (left) and on those that have been retweeted more than once (right)

Looking at the complementary picture, i.e. the distribution of followers reported in Fig. 12, reinforces our hypothesis: left and right panels illustrate two very different scenarios, with the latter showing how users who post very positive or very negative tweets attracted significantly fewer followers than others: bots involved in spam campaigns do not commonly exhibit large followership (cf. Fig. 6).

Concluding, our diagnostics revealed characteristic patterns that may conceal clues to decode the strategies employed by spam bots to spread the content they produce, and try giving spam a legitimate appearance.

4 Conclusions

Social bots have become a pervasive presence in social media platforms. Applications of social bots have been documented in a variety of scenarios, including for public opinion manipulation and for social spam campaigns. The focus of this chapter was to investigate both these domains, and in particular to study the interplay between bots and information diffusion in the two scenarios.

In Sect. 2, we reviewed how social bots are created, and how they operate in social media platforms. We also briefly discussed the challenges of, and the methods to detecting them, covering techniques based on graph-centric detection, crowdsourcing, and traditional feature-based supervised learning.

Section 3.1 presented our first case study, discussing how social bots have been used during the 2016 US Presidential Election to sway the conversation around the presidential candidates. In this section we revised in detail the tools we used for social bot detection, namely *Bot Or Not*, for Sentiment Analysis, namely *SentiStrength*, and for partisanship detection.

We also summarized the results of our study on political manipulation [11], providing in particular two data-driven insights: first, we noted that social bots generate as much engagement, at least in terms of obtained retweets, than humans, suggesting the fact that humans cannot tell apart bots from other humans very easily when rebroadcasting politics-related information on Twitter. Second, we illustrated the interplay between content sentiment and social bots, highlighting a few partisanship differences (e.g., Trump bots single-handedly generated the most positive supporting content of their candidate in the entire analyzed dataset).

Finally, in Sect. 3.2 we proposed a second case study, and new results and analyses about the effects of social spam bots on the diffusion of social spam campaigns within the tobacco-related conversation on Twitter. First, we identified the presence of three types of spam campaigns: (1) relative to tobacco products; (2) relative to products unrelated to the tobacco industry, e.g., entertainment products; and, finally, (3) instances of topic hijacking, namely the use of hashtags and keywords related to the tobacco industry to attract individuals' attention on issues completely unrelated to that, e.g., social issues connected to news events in the offline world.

By means of a newly-introduced method named *Dynamical Activity-Connectivity map*, we also revealed the existence of different classes of spam accounts, including traditional spammers and social spam bots; we also discussed a statistical characterization of their most typical features. In conclusion, we provided an analysis of the interplay between sentiment and spam bots, revealing patterns that may conceal strategies of bot coordination, and the resulting effects in terms of spam diffusion.

Our findings in both case studies exemplify the potential for social media abuse: whether at stakes is the right to exercise unbiased elections and therefore democracy itself, or the exposure to illegitimate spam and propaganda, social media manipulation can have devastating societal effects. This study encourages future efforts of the research community to address the various facets of this form of abuse.

References

1. Abokhodair N, Yoo D, McDonald DW (2015) Dissecting a social botnet: growth, content, and influence in twitter. In: Proceedings of the 18th ACM conference on computer-supported cooperative work and social computing. ACM, New York
2. Adamic LA, Glance N (2005) The political blogosphere and the 2004 us election: divided they blog. In: 3rd international workshop on link discovery. ACM, New York, pp 36–43
3. Ahn Y-Y, Han S, Kwak H, Moon S, Jeong H (2007) Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th international conference on world wide web. ACM, New York, pp 835–844
4. Aiello LM, Deplano M, Schifanella R, Ruffo G (2012) People are strange when you're a stranger: impact and influence of bots on social networks
5. Allem J-P, Ferrara E (2016) The importance of debiasing social media data to better understand e-cigarette-related attitudes and behaviors. J Med Internet Res 18(8):e219
6. Alvisi L, Clement A, Epasto A, Lattanzi S, Panconesi A (2013) Sok: the evolution of sybil defense via social networks. In: 2013 IEEE symposium on security and privacy. IEEE, Piscataway, pp 382–396
7. Aral S, Walker D (2011) Creating social contagion through viral product design: a randomized trial of peer influence in networks. Manag Sci 57(9):1623–1639
8. Barabasi A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435(7039):207–211
9. Barberá P, Wang N, Bonneau R, Jost JT, Nagler J, Tucker J, González-Bailón S (2015) The critical periphery in the growth of social protests. PLoS One 10(11):e0143611
10. Bekafigo MA, McBride A (2013) Who tweets about politics? Political participation of twitter users during the 2011gubernatorial elections. Soc Sci Comp Rev 31(5)

11. Bessi A, Ferrara E (2016) Social bots distort the 2016 US presidential election online discussion. First Monday 21(11):1–14
12. Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2011) The socialbot network: when bots socialize for fame and money. In: Proceedings of the 27th annual computer security applications conference. ACM, New York, pp 93–102
13. Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2013) Design and analysis of a social botnet. Comput Netw 57(2):556–578
14. Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc 15(5):662–679
15. Carlisle JE, Patton RC (2013) Is social media changing how we understand political engagement? An analysis of facebook and the 2008 presidential election. Polit Res Q 66(4):883–895
16. Catanese SA, De Meo P, Ferrara E, Fiumara G, Provetti A (2011) Crawling facebook for social network analysis purposes. In: ACM WIMS '11: international conference on web intelligence, mining and semantics. ACM, New York, pp 52–59
17. Centola D (2011) An experimental study of homophily in the adoption of health behavior. Science 334(6060):1269–1272
18. Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in twitter: the million follower fallacy. In: Fourth international AAAI conference on weblogs and social media (ICWSM 2010). AAAI Press, Palo Alto, pp 10–17
19. Chu Z, Widjaja I, Wang H (2012) Detecting social spam campaigns on twitter. In: International conference on applied cryptography and network security. Springer, Berlin, Heidelberg, pp 455–472
20. Coburn Z, Marra G (2011) Realboy: believable twitter bots. http://ca.olin.edu/2008/realboy/
21. Conover M, Ratkiewicz J, Francisco MR, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on twitter. ICWSM 133:89–96
22. Conover MD, Davis C, Ferrara E, McKelvey K, Menczer F, Flammini A (2013) The geospatial characteristics of a social movement communication network. PLoS One 8(3):e55957
23. Conover MD, Ferrara E, Menczer F, Flammini A (2013) The digital evolution of occupy wall street. PLoS One 8(5):e64679
24. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) Botornot: a system to evaluate social bots. In: WWW '16 companion proceedings of the 25th international conference companion on world wide web. ACM, New York, pp 273–274
25. DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More tweets, more votes: social media as a quantitative indicator of political behavior. PLoS One 8(11):e79449
26. Effing R, Hillegersberg JV, Huibers T (2011) Social media and political participation: are facebook, twitter and youtube democratizing our political systems? In: International conference on electronic participation. Springer, Berlin, pp 25–35
27. El-Khalili S (2013) Social media as a government propaganda tool in post-revolutionary Egypt. First Monday 18(3)
28. Elovici Y, Fire M, Herzberg A, Shulman H (2013) Ethical considerations when employing fake identities in online social networks for research. Sci Eng Ethics 20:1–17
29. Elyashar A, Fire M, Kagan D, Elovici Y (2013) Homing socialbots: intrusion on a specific organization's employee using socialbots. In: Proceedings of the 2013 international conference on advances in social networks analysis and mining. ACM, New York, pp 1358–1365
30. Ferrara E (2015) Manipulation and abuse on social media. ACM SIGWEB Newsletter (4). ACM, New York
31. Ferrara E (2017) Contagion dynamics of extremist propaganda in social networks. Inf Sci 418:1–12
32. Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday 22(8)
33. Ferrara E, Yang Z (2015) Measuring emotional contagion in social media. PLoS One 10(11):e0142390
34. Ferrara E, Yang Z (2015) Quantifying the effect of sentiment on information diffusion in social media. Peer J Comput Sci 1:e26

35. Ferrara E, De Meo P, Fiumara G, Baumgartner R (2014) Web data extraction, applications and techniques: a survey. Knowl-Based Syst 70:301–323
36. Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun. ACM 59(7):96–104
37. Ferrara E, Varol O, Menczer F, Flammini A (2016) Detection of promoted social media campaigns. In: 10th international AAAI conference on web and social media, pp 563–566
38. Gao H, Hu J, Wilson C, Li Z, Chen Y, Zhao BY (2010) Detecting and characterizing social spam campaigns. In: Proceedings of the 10th ACM SIGCOMM conference on internet measurement. ACM, New York, pp 35–47
39. Gao H, Barbier G, Goolsby R (2011) Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell Syst 26(3):10–14
40. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Sci Rep 1:197
41. González-Bailón S, Borge-Holthoefer J, Moreno Y (2013) Broadcasters and hidden influentials in online protest diffusion. Am Behav Sci 57:943–965. https://doi.org/10.1177/0002764213479371
42. Hadgu AT, Garimella K, Weber I (2013) Political hashtag hijacking in the us. In: Proceedings of the 22nd international conference on world wide web. ACM, New York, pp 55–56
43. Heymann P, Koutrika G, Garcia-Molina H (2007) Fighting spam on social web sites: a survey of approaches and future challenges. IEEE Internet Comput. 11(6):36–45
44. Howard PN (2006) New media campaigns and the managed citizen. Cambridge University Press, Cambridge
45. Howard PN, Kollanyi B (2016) Bots, #strongerin, and #brexit: computational propaganda during the uk-eu referendum. Available at SSRN 2798311
46. Hwang T, Pearce I, Nanis M (2012) Socialbots: voices from the fronts. Interactions 19(2):38–45
47. Jackson SJ, Welles BF (2015) Hijacking# mynypd: social media dissent and networked counterpublics. J Commun 65(6):932–952
48. Jagatic TN, Johnson NA, Jakobsson M, Menczer F (2007) Social phishing. Commun ACM 50(10):94–100
49. Jain N, Agarwal P, Pruthi J (2015) Hashjacker-detection and analysis of hashtag hijacking on twitter. Int J Comput Appl 114(19):17–20
50. Jin X, Lin C, Luo J, Han J (2011) A data mining-based spam detection system for social media networks. Proc VLDB Endowment 4(12):1458–1461
51. Jindal N, Liu B (2007) Review spam detection. In: Proceedings of the 16th international conference on world wide web. ACM, New York, pp 1189–1190
52. Klotz RJ (2007) Internet campaigning for grassroots and astroturf support. Soc Sci Comput Rev 25(1):3–12
53. Kollanyi B, Howard PN, Woolley SC (2016) Bots and automation over twitter during the first us presidential debate. Technical report, COMPROP Data Memo
54. Kramer AD, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci 111(24):8788–8790
55. Kümpel AS, Karnowski V, Keyling T (2015) News sharing in social media: a review of current research on news sharing users, content, and networks. Social Media+ Society 1(2):2056305115610141
56. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, pp 591–600
57. Latonero M, Shklovski I (2013) Emergency management, twitter, and social media evangelism. In: Using social and information technologies for disaster and crisis management. IGI Global, Hershey, pp 196–212
58. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M et al (2009) Life in the network: the coming age of computational social science. Science (New York, NY) 323(5915):721

59. Lee K, Caverlee J, Webb S (2010) The social honeypot project: protecting online communities from spammers. In: Proceedings of the 19th international conference on world wide web. ACM, New York, pp 1139–1140
60. Lee K, Caverlee J, Webb S (2010) Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 435–442
61. Lutz C, Hoffmann CP, Meckel M (2014) Beyond just politics: a systematic literature review of online participation. First Monday 19(7)
62. Lyon TP, Maxwell JW (2004) Astroturf: Interest group lobbying and corporate strategy. J Econ Manag Strateg 13(4):561–597
63. Markines B, Cattuto C, Menczer F (2009) Social spam detection. In: Proceedings of the 5th international workshop on adversarial information retrieval on the web, pp 41–48
64. Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: an empirical investigation of online review manipulation. Am Econ Rev 104(8):2421–2455
65. Messias J, Schmidt L, Oliveira R, Benevenuto F (2013) You followed my bot! transforming robots into influential users in twitter. First Monday 18(7)
66. Metaxas PT, Mustafaraj E (2012) Social media and the elections. Science 338(6106):472–473
67. Mønsted B, Sapieżyński P, Ferrara E, Lehmann S (2017) Evidence of complex contagion of information in social media: an experiment using twitter bots. PLos One 12: e0184148
68. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In: 7th international AAAI conference on weblogs and social media
69. Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on world wide web, pp 191–200
70. Pang B, Lee L et al (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135
71. Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. ICWSM 11:297–304
72. Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, Flammini A, Menczer F (2011) Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th international conference companion on world wide web. ACM, New York, pp 249–252
73. Shorey S, Howard PN (2016) Automation, algorithms, and politics| automation, big data and politics: a research review. Int J Commun 10:24
74. Song J, Lee S, Kim J (2011) Spam filtering in twitter using sender-receiver relationship. In: International workshop on recent advances in intrusion detection, pp 301–317
75. Stein T, Chen E, Mangla K (2011) Facebook immune system. In: Proceedings of the 4th workshop on social network systems, p 8. ACM, New York
76. Stringhini G, Kruegel C, Vigna G (2010) Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference, p 1–9. ACM, New York
77. Subrahmanian V, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Zhu L, Ferrara E, Flammini A, Menczer F et al (2016) The DARPA Twitter bot challenge. IEEE Comput 49(6):38–46
78. Sutton JN, Palen L, Shklovski I (2008) Backchannels on the front lines: emergency uses of social media in the 2007 Southern California wildfires. University of Colorado, Boulder
79. Thelwall M (2013) Heart and soul: sentiment strength detection in the social web with sentistrength. In: Proceedings of the CyberEmotions, pp 1–14
80. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 61(12):2544–2558
81. Theocharis Y, Lowe W, van Deth JW, García-Albacete G (2015) Using twitter to mobilize protest action: online mobilization patterns and action repertoires in the occupy wall street, indignados, and aganaktismenoi movements. Inf Commun Soc 18(2):202–220

82. Thomas K, Grier C, Song D, Paxson V (2011) Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM conference on internet measurement conference. ACM, New York, pp 243–258
83. Thomas K, McCoy D, Grier C, Kolcz A, Paxson V (2013) Trafficking fraudulent accounts: the role of the underground market in twitter spam and abuse. In: Usenix security, vol 13, pp 195–210
84. Varol O, Ferrara E, Ogan CL, Menczer F, Flammini A (2014) Evolution of online user behavior during a social upheaval. In: Proceedings 2014 ACM conference on web science, pp 81–90
85. Varol O, Ferrara E, Davis C, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: International AAAI conference on web and social media
86. Varol O, Ferrara E, Menczer F, Flammini A (2017) Early detection of promoted campaigns on social media. EPJ Data Sci 6(1):13
87. Wagner CH (1982) Simpson's paradox in real life. Am Stat 36(1):46–48
88. Wang G, Mohanlal M, Wilson C, Wang X, Metzger M, Zheng H, Zhao BY (2013) Social turing tests: crowdsourcing sybil detection. In: NDSS. The Internet Society, Reston
89. Yang C, Harkreader R, Zhang J, Shin S, Gu G (2012) Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In: Proceedings of the 21st international conference on world wide web. ACM, New York, pp 71–80
90. Yang X, Chen B-C, Maity M, Ferrara E (2016) Social politics: agenda setting and political communication on social media. In: International conference on social informatics. Springer, Berlin, pp 330–344
91. Yates D, Paquette S (2011) Emergency knowledge management and social media technologies: a case study of the 2010 haitian earthquake. Int J Inf Manag 31(1):6–13
92. Yin J, Lampert A, Cameron M, Robinson B, Power R (2012) Using social media to enhance emergency situation awareness. IEEE Intell Syst 27(6):52–59
93. Zangerle E, Specht G (2014) "Sorry, I was hacked" a classification of compromised twitter accounts. In: SAC: the 29th symposium on applied computing
94. Zhang X, Zhu S, Liang W (2012) Detecting spam and promoting campaigns in the twitter social network. In: IEEE 12th international conference on data mining (ICDM), 2012. IEEE, Piscataway, pp 1194–1199