Contagion dynamics of extremist propaganda in social networks

Emilio Ferrara

University of Southern California, Information Sciences Institute, 4676 Admiralty Way #1001, Marina del Rey, CA (USA) 90292, USA

ABSTRACT

Recent terrorist attacks carried out on behalf of ISIS on American and European soil by lone wolf attackers or sleeper cells remind us of the importance of understanding the dynamics of radicalization mediated by social media communication channels. In this paper, we shed light on the social media activity of a group of twenty-five thousand users whose association with ISIS online radical propaganda has been manually verified. By using a computational tool known as *dynamic activity-connectivity maps*, based on network and temporal activity patterns, we investigate the dynamics of social influence within ISIS supporters. We finally quantify the effectiveness of ISIS propaganda by determining the adoption of extremist content in the general population and draw a parallel between radical propaganda and epidemics spreading, highlighting that information broadcasters and influential ISIS supporters generate highly-infectious cascades of information contagion. Our findings will help generate effective countermeasures to combat the group and other forms of online extremism.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Researchers in the computational social science community have recently demonstrated the importance of studying online social networks to understand our society [43]. New powerful technologies are usually harbinger of abuse, and online platforms are no exception: social media have been shown to be systematically abused for nefarious purposes [23]. As online social environments yield plenty of incentives and opportunities for unprecedented, even "creative" forms of misuse, single individuals as well as organizations and governments have systematically interfered with these platforms, oftentimes driven by some hidden agenda, in a variety of reported cases:

- During crises, social media have been effectively used for emergency response; but fear-mongering actions have also triggered mass hysteria and panic [23,33].
- Political conversation has been manipulated by means of orchestrated astroturf campaigns [46,52] even during election times [14,24,35].
- Anti-vaccination movements [60,62], as well as conspiracy (and other anti-science) theorists [13,21], took social media by the storm and became responsible for a major health crisis in the United States [55].
- Social media bots (non-human automated accounts) have been used to coordinate attacks to successfully manipulate the stock market, causing losses in the billions of dollars [23,25,36,64].

- Some governments and non-state actors have been active on social media to spread their propaganda. In some cases, they have allegedly "polluted" these platforms with content to sway public opinion [22,23,38], or to hinder the ability of social collectives to communicate, coordinate, and mobilize [59].

Especially related to the last point, researchers have been recently devoting more attention to issues related to online propaganda campaigns [6,53,57]. Increasing evidence provided by numerous independent studies suggests that social media played a pivotal role in the rise in popularity of the *Islamic State of Iraq and al-Sham* (viz. ISIS) [8,26,29,58,68]. Determining whether ISIS benefitted from using social media for propaganda and recruitment was central for many research endeavors [9,11,45].

Analyses by Berger and Morgan suggested that a restricted number of highly-active accounts (500–1000 users) is responsible for most of ISIS' visibility on Twitter [9]. However, Berger's subsequent work suggested that ISIS' reach (at least among English speakers) has stalled for months as of the beginning of 2016, due to more aggressive account suspension policies enacted by Twitter [10]. Other researchers tried to unveil the roots of support for ISIS, suggesting that ISIS backers discussed Arab Spring uprisings on Twitter significantly more than users who stood against ISIS [45].

These early investigations all share one common methodological limitation, namely that to collect social media data they start from keywords known to be associated to ISIS [9,10,45]. This strategy has been widely adopted in a previous research aimed at characterizing social movements [18,31,65]. However, we argue that it is not sufficient to focus on keyword-based online chatter to pinpoint to relevant actors of radical conversation.

In fact, our recent results [8] suggest that radical propaganda revolves around four independent types of messaging: *(i)* theological and religious topics; *(ii)* violence; *(iii)* sectarian discussion; and, *(iv)* influential actors and events. Here is a series of examples of possible biases introduced by the keyword-centric approach:

- Some studies [45] focused on religion-based keyword lists, but most terms typically associated to religion are not necessarily used in the context of extremism.
- Other studies [9,10] focused on influential actors or events; this can introduce biases due to the focus on popular actors rather than the overall conversation.
- Further noise can be introduced by tweets that simply link to news articles reporting on events; although these tweets may contain keywords in the predefined watchlist, they clearly do not represent extremist propaganda efforts.
- Finally, focusing on pre-determined keywords could cause incomplete data collection by missing topics of discussion that can emerge dynamically and do not adopt any of the predefined key terms.

In this work, we will leverage an alternative data collection and curation approach: we will start from a large set of Twitter users that are known to be associated to or symphatizers of ISIS. We will then collect their activity over a large time span of over one year to obtain a complete characterization of their extremist propaganda efforts.

1.1. Contributions of this work

This study aims to address the two following research questions:

RQ1: Can we define a solid methodological framework and suggest good practices for data collection, validation, and analysis to study online radicalization?.

Setting such best practices will hopefully steer the information sciences and computational social sciences research communities in the direction of producing more rigorous and reproducible work. One contribution of our work is to address this issue by focusing on manually-verified set of ISIS supporter accounts.

After describing how we collected and curated the dataset object of this study, we move forward to investigate the dynamics of online propaganda adopted by ISIS. Using computational social science tools to gauge online extremism, we aim to answer the following second research question:

RQ2: What types of network and temporal patterns of activity reflect the dynamics of social influence within ISIS supporters? And, can we quantify the adoption of extremist content in the general population?.

Our findings will help generate effective countermeasures to combat the group and other forms of online extremism.

2. Data collection and curation

Due to the limits of keyword-based data collection approaches we highlighted above, in this work we exclusively rely on data and labels obtained by using a procedure of manual data curation and expert verification. We obtained a list of Twitter accounts whose activity was labeled as supportive of the Islamic State by the online crowd-sourcing initiative called *Lucky Troll Club*. The goal of this initiative was to leverage annotators with expertise in Arabic languages to identify accounts affiliated to ISIS and report them to Twitter to request their suspension. Twitter's anti-abuse team manually verifies all suspension requests, and grants some based on evidence of violation of Twitter's Terms of Service policy against the usage of the platform for extremist purposes.

We verified that 25,538 accounts present in the *Lucky Troll Club* list have been suspended by Twitter in the period between March 17, 2015 and June 9, 2015. In this study, we focus only on this subset of twenty-five thousand accounts: we consider their activity to be unequivocally linked to the Islamic State, as determined by the two-step manual verification process described above. For each account, we have at our disposal information about the suspension date, as well as the number of followers of that user as of the suspension date. The anonymized data focus of this study can be made available to the interested researchers upon request. For contact information and further details about the dataset please visit: http://www.emilio.ferrara.name/datasets/

2.1. Twitter data collection

The next step of our study consisted in collecting data related to the activity of the 25,538 ISIS supporters on Twitter. To this purpose, we leveraged the Observatory on Social Media (OSoMe) data source set up by our collaborators at Indiana University [20], which continuously collects the Twitter data stream from the *gardenhose* API (roughly a 10% random sample of the full Twitter data stream). Using this large data stream avoids known issues derived by using the public Twitter stream API which serves only less than 1% of the overall tweets [49].

We obtained all tweets present in the OSoMe database that have been posted by any of the twenty-five thousand ISIS accounts prior to their suspension. We also included all retweets these tweets generated, and all tweets containing mentions to such set of ISIS supporters. The resulting dataset that we will study consists of 3,395,901 tweets. Almost 1.2 million of these tweets was generated by the ISIS accounts during the period between January 2014 and June 2015. We found that a total of 54,358 distinct other users has retweeted at least once one of the twenty-five thousand ISIS supporters. This amounts for the remainder of about 2.2 million tweets in our dataset.

Summarizing, we identified two best practices to answer **RQ1**:

- Favor starting from a manually-verified list of users involved in online propaganda, radicalization efforts, or recruitment, rather focusing on keyword-based searches. When such manually validated lists are not available, human annotations can be generated by means of services such as Amazon Mechanical Turk.
- Use large social media data streams, when available, rather than small samples that can be biased. Services like the Indiana University's OSoMe database [20] can provide data sources especially valuable for Twitter-based social media studies. Alternatively, the Twitter Search API can yield comprehensive data around specific users or topics, provided that the search is limited to short time frames.

2.2. Limitations and strategic choices of this study

The largest majority of the tweets in our dataset, over 92%, is in Arabic. This introduces a number of important challenges, from both technical and methodological perspectives. From a technical standpoint, analyzing Arabic content would require sophisticated natural language processing (NLP) techniques as well as accurate sentiment analysis tools capable of extracting emotional information. NLP toolkits with support for Arabic content are only recently starting to be developed [32], while Arabic-based sentiment analysis is still in its early developmental stage [39].

From a methodological point of view, the analysis of Arabic-language material would require interpreters with sufficient domain knowledge in extremism-related issues to yield useful and unbiased insights from the data. Access to such experts is not always available, and thus the need to develop alternative strategies of enquiry emerges.

For these reasons, in the rest of the paper our analysis will rely exclusively upon language-agnostic techniques: in particular, we will focus on statistical properties of information diffusion networks, which were proven very useful in our prior studies on criminal networks [5], as well as temporal patterns of information diffusion, which we already exploited to study the interplay between ISIS' activity online and offline [8].

Due to the simplifications introduced by our strategy of enquiry, we call for caution in the interpretation of our results. However, we believe that our approach will help identify important areas of research that warrant further development (such as Arabic-based NLP and sentiment analysis toolkits), as well as yield valuable insight about ISIS social media operations and other forms of online extremist propaganda.

3. Results

We next report our investigation to address **RQ2**. We study the activity of ISIS supporters and sympathizers on Twitter by means of the data collected as described above.

3.1. Activity and support of ISIS accounts

The first form of validation that we performed pertains the mechanisms of suspension of ISIS accounts on Twitter. As we mentioned before, we identified a set of twenty-five thousand accounts related to ISIS that have been suspended: Fig. 1 shows the timeline of suspensions of the accounts under investigation. The suspension period occurred throughout almost three months, with the first suspensions occurring on March 17, 2015 and the last occurring on June 9, 2015. After this date,
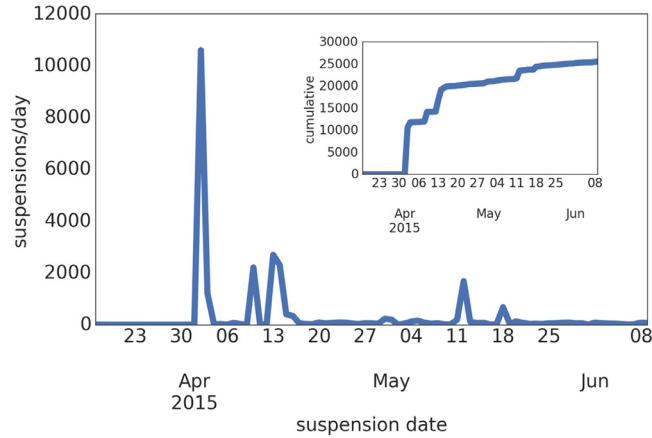
Fig. 1. Timeline of suspensions per day of the 25,538 ISIS-supporting accounts. The main figure shows the daily suspension counts, while the inset figure shows the cumulative count.
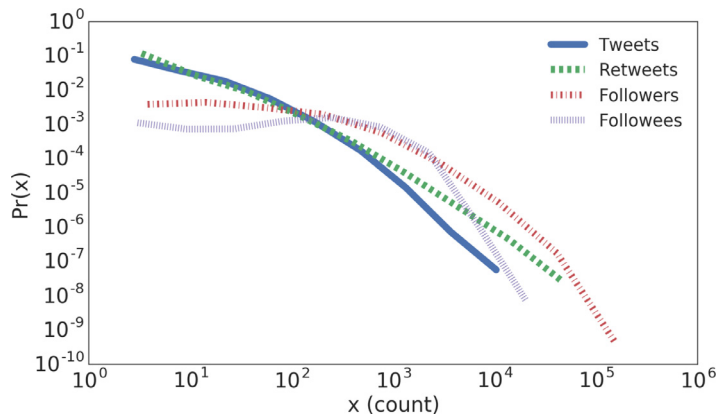

Fig. 2. Probability distributions of the number of tweets posted by ISIS supporters, generated retweets, and the number of followers and followees at the time of their suspension.

none among the twenty-five thousand accounts in our list is anymore active on Twitter. Account suspensions appear to occur in batches, some more substantial than others, with a significant spike of suspensions (over ten thousands) occurred on April 2, 2015. Our findings are consistent with *The Guardian*'s report that, between April 2015 and February 2016, Twitter's anti-abuse task force suspended more than 125,000 accounts linked to ISIS [69].

Recent literature reported contrasting evidence about the activity and popularity of ISIS supporters on Twitter [9,10]. As for today, it is unclear whether ISIS accounts obtained a significant support on Twitter, and to what extent they were active. Berger and collaborators first found that ISIS presence was very pervasive on Twitter during 2014–2015 [9], and later suggested that only a core of 500–2000 ISIS users was active after that period [10]. To shed light on this question, we calculated the distribution of the number of followers and followees (friends) of ISIS accounts at the time of their suspension, along with the number of tweets and retweets they generated. For each of these four variables, we calculated their probability distribution $Pr(x)$. These probability distributions display the probabilities of values taken by the four variables (*i.e.,* tweets, retweets, followers, followees) in our dataset: they can be thought as *normalized frequency distributions*, where all occurrences of outcomes for each distribution sum to 1. The results are shown in Fig. 2.

Let us discuss support and activity separately. Concerning ISIS support on Twitter, we notice that the distribution of followers of ISIS accounts exhibits the long tail typical of Twitter [41] and other social networks (*cf.* yellow dash-dotted line in Fig. 2). This skewed distribution has mean $\mu = 516$ ($\sigma = 1,727$), median $Q_2 = 130$, and lower and upper quartiles $Q_1 = 37$ and $Q_3 = 401$. This means that the majority of accounts has a limited number of followers (for example, one quarter of the users has less than 37 followers), yet a significant number of ISIS supporters managed to obtain a large number of followers before getting suspended (in fact, the upper quartile of users has more than 400 followers). The presence of this broad distribution of followership suggests that influence and radicalization operations of Islamic State supporters on Twitter were successful for at least several thousand of their accounts. This is in line with Berger's early results discussed above [9]. Another interesting insight is yielded by the distribution of followees (*cf.* green dotted line in Fig. 2): differently from the distribution of followers, this distribution shows an unexpected upward trend in the regime between 100 and 1000
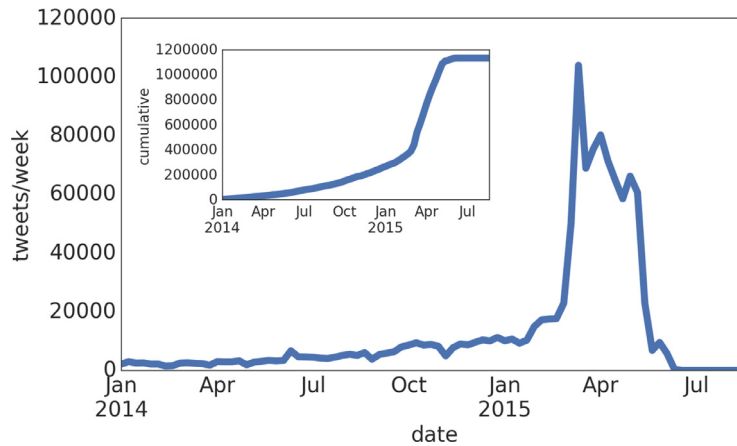
Fig. 3. Timeline of total number of tweets generated by ISIS supporters. The main figure shows the number of tweets per week; the inset figure shows the cumulative count.

followees. This characteristic behavior has been associated to forms of social network manipulation, for example attempts to create rings or cliques in which multiple accounts under the control of a same entity all follow each other to reciprocally increase their visibility and followership[52]. We thus suggest that ISIS accounts enacted strategies to artificially enhance their visibility by strengthening one another social networks.

A significant portion of ISIS accounts was very active on Twitter: both distributions of tweets (*cf.* solid blue line) and retweets (*cf.* red dashed line) shown in Fig. 2 exhibit the typical power-law shape common to social networks with heterogeneous activity patterns. This implies that a significant fraction of users posted and retweeted large amounts of tweets. For example, at least 1% of the ISIS users posted at least 30 tweets during the observation period; similar figures hold for retweeting. More importantly, there appears to be a strong core constituted by a few dozen accounts who posted and retweeted hundreds of tweets in the same period. This set of very active ISIS supporters we found is compatible with what reported by Berger's first study[9]. Furthermore, there seems to be a handful of accounts with thousands of tweets and retweets, suggesting the likely presence of some social media bot used to enhance the volume of content generated by ISIS and its spreding on Twitter[25].

To investigate ISIS accounts' activity further, we extrapolated the time series of the total volume of tweets and retweets posted every week by the ISIS users under investigation: the result is shown in Fig. 3. It's worth noting that our observation window spans 1.5 years and starts on January 2014 when the firsts among the twenty-five thousand ISIS supporters became active on Twitter: Although the activity volume slowly builds up over time, it is only in early March 2015 (15 months into our observation period) that the volume of tweets per week drastically spikes.

In the early regime between January 2014 and March 2015 the volume of tweets associated to ISIS accounts spans 1000 to 10,000 per week. This increases nearly tenfold after March 2015, with a spike of over 100,000 tweets per week, and an average of over 60,000 tweets per week in the period between March and May 2015. This period concurs with the period of strongest Twitter suspensions shown in Fig. 1, suggesting a timely reaction of Twitter to fight the activity of ISIS users on the platform. Indeed, the volume of tweets per week produced by these accounts drops in early June, and goes to zero, as expected, in the late period of observation when fewer and fewer of the ISIS accounts under investigation are left unchecked on Twitter. Cumulatively, the ISIS users under investigation produced almost 1.2 million tweets during the observation period (*cf.* inset of Fig. 3).

Our findings pinpoint to the power of the crowd-sourcing volunteer initiative that set to bring up to Twitter's attention these accounts. However, there is no evidence to quantify how many (if any) of the ISIS supporters not recorded by the *Lucky Troll Club* operation were independently suspended by Twitter anti-abuse team.

3.2. Dynamic activity-connectivity maps

We further address **RQ1** by offering a powerful tool to information scientists and computational social scientists: the *dynamic activity-connectivity (DAC) maps*. DAC maps allow to study influence and authority dynamics of online extremism in social networks with temporal activity patterns.

Fig. 4 shows an example of *dynamic activity-connectivity map*. We developed DAC maps as dynamic variants of the map proposed by Gonzalez-Bailon and collaborators—see Fig. 4 in *Broadcasters and Hidden Influentials in Online Protest Diffusion*[30]. The key intuition behind this tool is to allow investigate what effect the progression of activity levels of an user has of their connectivity evolution (and viceversa).
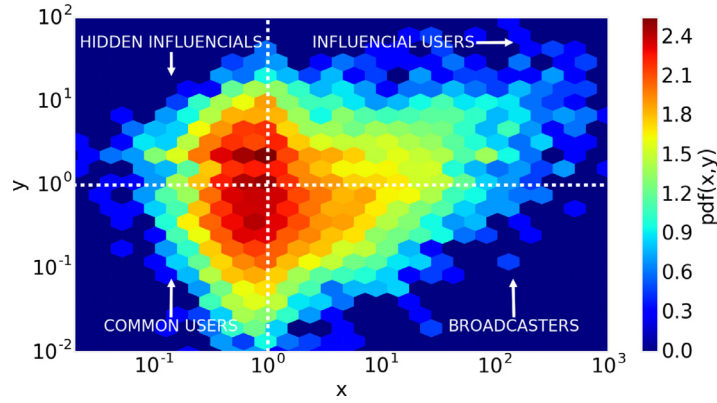
Fig. 4. The *dynamic activity-connectivity map* shows the Joint Probability Density Function *pdf(x, y)* of the connectivity growth *x* and the activity rate *y* as defined in Eq. 1.

In a DAC map, for a given user *u*, x_u and y_u are defined as

$$x_u = \frac{1 + \delta f_u}{1 + \delta F_u} \qquad \text{and} \qquad y_u = \frac{1 + m_u}{1 + M_u}. \tag{1}$$

We use the notation f_u and F_u to identify the number of followers and friends, respectively, of a user *u*. The variation 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 suspension (this varies from user to user). Finally, m_u is the number of mentions user *u* received by others, and M_u is the number of mentions user *u* made to others, during *u*'s activity period. All values are added to the unit to avoid zero-divisions and allow for logarithmic scaling (i.e., in those cases where the variation is zero). The third dimension, 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 two dimensions defined over the *dynamic activity-connectivity map* are interpreted as follows: the x-axis represents the growth of connectivity formation, and the y-axis conveys the rate of messaging activity. In general, we would expect that in a dynamic activity-connectivity map, the bulk of the joint probability density mass would be observed in the neighborhood of (1, 1), that hosts the majority of accounts for which the variation of the two dimensions is comparable.

DAC maps are ideal tools to addres our second research question (**RQ2**): they capture at the same time network and temporal patterns of activity, and they can help understand how connectivity variations affects social influence (and viceversa), which are dynamics at the core of our investigation.

Let us discuss the two dimensions of Fig. 4, namely *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 10^3. 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 [25,48], to indiscriminately follow others in order to seek for reciprocal followerships. Our dynamic 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 [30] 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; the 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 mentions versus how frequently s/he mentions others. Users with values of $y > 1$ are those who receive systematically more mentions with respect to how frequently they mention others. 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 divide in influential ($x > 1$) and hidden influential ($x < 1$) users. Conversely, users with values of $y < 1$ are those who generate increasingly more mentions over time, either because they reply to many tweets, or because they address directly other users. This group generally represents 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 dynamic activity-connectivity maps has been provided, we can proceed with interpreting Fig. 4: the bottom-left quadrant reports the most common users, those with both activity and connectivity growth lesser than 1.
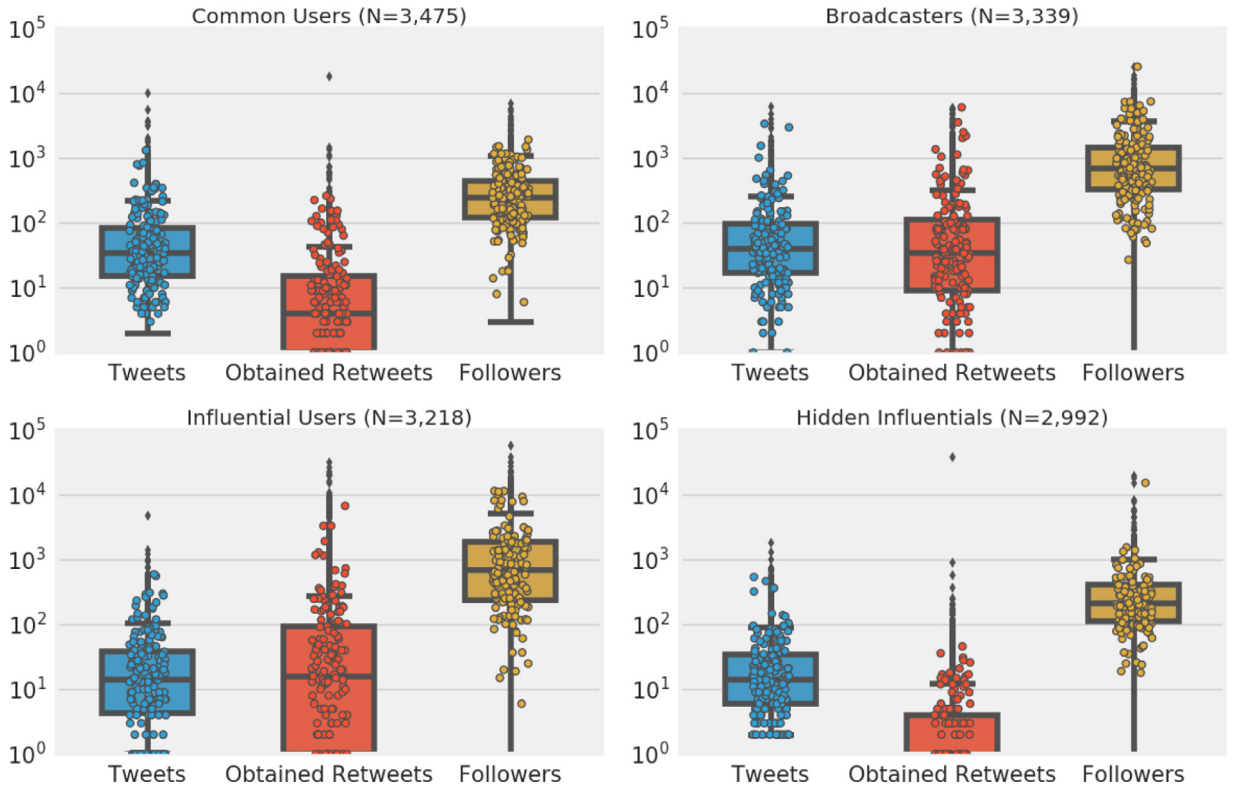
Fig. 5. Distributions of *(i)* number of posted tweets, *(ii)* number of obtained retweets, and *(iii)* number of followers, for common users, broadcasters, influencial, and hidden influencial users, respectively.

Conversely, the upper-right quadrant reports users with the higher connectivity growth and activity rates. These are influential ISIS supporters who are very active in the discussion. We note how the connectivity growth dimension spans three orders of magnitude in the positive domain, while the activity rate dimension only spans two orders of magnitude. This means that some users' followerships grows tens of times faster than the rate at which they follow others; conversely, the rate of receiving mentions is only up to tens of times higher than that of mentioning of others. In the next section, we will devote special attention to these four different classes of users to determine what types of differences emerge in the ISIS social network.

3.3. Dynamical classes of ISIS supporter behaviors

Prior research illuminated on the dynamical aspects of activity and connectivity in social media [30,44]. Next, we focus on the four classes of user behaviors highlighted by the dynamic activity-connectivity map. We first select, out of the twenty-five thousand ISIS supporters in our dataset, only the subset of those who have mentioned and have been mentioned at least once during the observation period. This will allow to focus on active accounts and correctly capture their activity rate. This filter reduces the number of users under investigation to $N = 13,024$ ISIS supporters, nearly half of the entire ISIS population initially collected. We further divide these users in the four dynamical classes defined above. The classification is obtained by simply adopting the rules defining the four quadrants of Fig. 4, which yields $N = 3,475$ common users ($x < 1$, $y < 1$), $N = 3,339$ information broadcasters ($x > 1$, $y < 1$), $N = 3,218$ influentials ($x > 1$, $y > 1$)), and $N = 2,992$ hidden influential users ($x < 1$, $y > 1$).

For each of these users, we generated the distribution of *(i)* the total number of tweets they posted, *(ii)* the cumulative number of times they have been retweeted, and *(iii)* the maximum number of followers they gathered. Fig. 5 shows the boxplots corresponding to the four dynamical classes. Significant differences emerge: common users produce an amount of tweets very similar to that of broadcasters, but they accrue nearly one order of magnitude less retweets than the latter. Information broadcasters also appear to generate the largest followerships, on par with influential users; influentials, however, post significantly less tweets, while accumulating similar amounts of retweets than broadcasters, suggesting that our map successfully captures a notion of social influence intended as a proxy for attention generated to one's posts. The class of hidden influentials shows comparable activity to influential users, but significantly less influence, accruing about one order of magnitude less tweets and significantly smaller followerships than influentials.
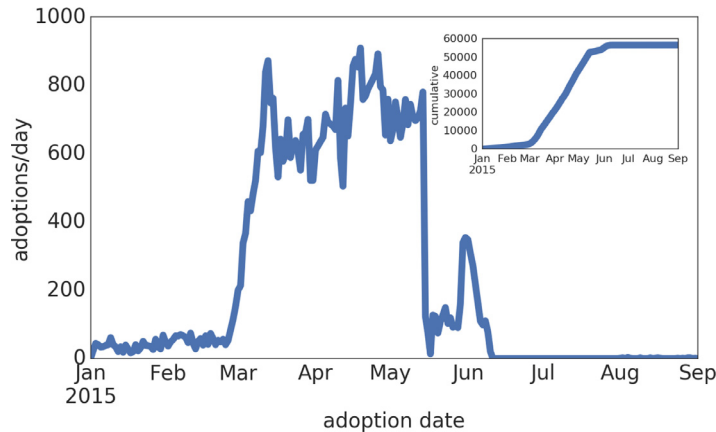
Fig. 6. Timeline of the number of individuals per day who adopted ISIS contents.

Our analysis suggests that different classes of ISIS supporters' behaviors emerge. In the future, we will study what are the characteristics of different classes that produce the most effective propaganda and make the most influential users, analyzing content and language, political and religious beliefs, motives and attitudes of the ISIS social network.

3.4. Adoption of ISIS propaganda

So far, our analysis focused on characterizing some dimensions of the behavior of ISIS supporters on Twitter. Next, we investigate whether the content they generated has been adopted by other users who have become exposed to it. This will help address the last part of our second research question (**RQ2**), namely whether we can quantify the adoption of extremist content in the general population. Our notion of adoption is very simple: when a user who does not appear set of the twenty-five thousand ISIS accounts retweets for the first time any tweet generated by one such ISIS supporter, we count this as a content adoption. Although some recent work suggested that retweeting radical propaganda can be considered as an early sign of radicalization [9,45], we call for caution about interpreting the results in such a way: this definition greatly simplifies the notion and complexity of such types of online adoption processes [15,16,27,28,40]. However, we do believe that investigating the spread of radical content in the user population has a value to determine the extent and effectiveness of ISIS propaganda operations.

Fig. 6 shows the time series of the number of ISIS content adoptions per day during the first half of 2015. Prior to that, no significant amount of adoptions could be observed, partly due to the low activity rate of the accounts under investigation. In the three months between March and June 2015, we notice a significant uptake in the number of adoptions, peaking at nearly one thousand adoptions per day. In that period, at least 10,000 tweets per day (70,000-100,000 tweets/week) were generated by ISIS accounts (*cf.* Fig 3). This suggests that a very significant fraction of tweets, about 5–10%, was actually retweeted on average at least once by other users. During this period, a total of 54,358 distinct other users has retweeted at least once one of the twenty-five thousand ISIS supporters.

If we simplify propaganda diffusion as an infectious disease, we can draw a parallel with epidemics. The *basic reproduction number* R_0 of an infection is the number of cases generated on average by an infected individual over the course of its infectious period, in an otherwise uninfected population. Given that 25,538 ISIS supporters generated 54,358 distinct *infected users*, we can derive an $R_0 = 2.13$ for the ISIS propaganda "infection". In other words, an ISIS supporter before being suspended on average "infected" 2.13 other users. For comparison, health epidemics like Ebola, [7,42] SARS, [67] HIV/AIDS, [34,50] and certain strains of influenza [47] all have similar values of $2 < R_0 < 3$.

3.5. Contagion dynamics of ISIS propaganda

We conclude our analysis by studying the cascades of content adoptions generated by the four classes of users defined above. This investigation is twofold: first, we would like to determine whether the mechanisms of receiving retweets (what we defined as content adoption) and being mentioned by out-of-sample users exposed to extremist content follow the same or different dynamics. Second, we will compute the distributions of scores for the basic reproduction number R_0 relative to both receiving retweets and mentions for the four classes of users. The goal is to reveal whether any significant difference emerge between groups of users in their content spreading efficacy, and ultimately to understand which groups of users generated the most effective information contagions.

Fig. 7 shows the distributions of retweets (top) and mentions (bottom) received by the users in the four classes defined by means of the dynamic activity-connectivity map. Although overall all distributions are broad, as expected given the heterogeneous nature of information diffusion, the dynamics of obtaining retweets and mentions are significantly different for the
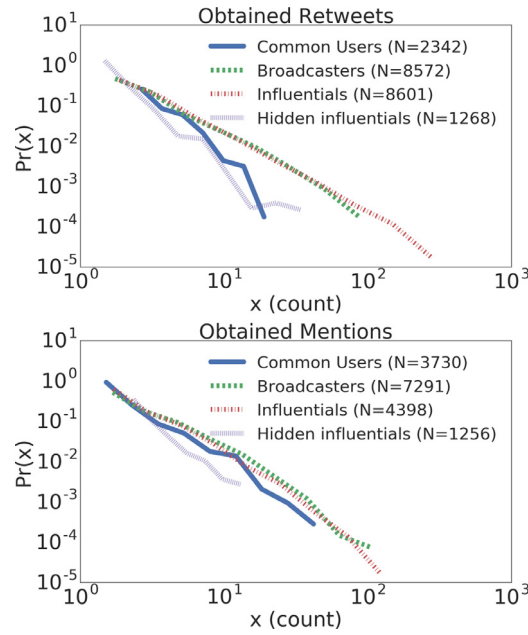
Fig. 7. Probability distributions of retweets (top) and mentions (bottom) obtained by ISIS users as divided in the four dynamical classes, namely common users, broadcasters, influential and hidden influential users. Each group reports the size of the population under investigation (different in the two scenarios).

four groups: in particular, for what concerns receiving mentions, no appreciable difference emerges among the four classes of users, which suggests that ISIS supporters are being mentioned in a similar fashion regardless of the class they belong to. However, receiving retweets shows a different mechanism: influential and broadcaster users receive generate significantly larger retweet cascades much more frequently than common users and hidden influentials.

One question that thus rises concerns whether the actual contagion effectiveness varies between classes of users. In other words, what are the classes of users that are generating the most effective propaganda campaigns, in terms of adoption (*i.e.,* retweets) and engagement (*i.e.,* mentions)? To this purpose, for each ISIS user who obtained at least one retweet (resp., mention), we calculated the fraction $\frac{RT}{T}$ of received retweets RT (resp., mentions) over the total number of his/her tweets T that have been retweeted at least once (resp., mentioned) by an out-of-sample user (*i.e.,* a user not labeled as ISIS by our list). This measures the rate of diffusion of tweets in the otherwise uninfected population. We note that this is a simplification of the more traditional notion of information contagion where we would consider all the tweets generated and more complex diffusion mechanisms accounting e.g., for exposures, due to the limitation of the platform under study (namely, we do not have any information about information exposure on Twitter). With some abuse of notation, we thus consider the fraction $\frac{RT}{T}$ to convey the meaning of the basic reproduction number R_0 typical of epidemiology.

Finally, ISIS accounts are divided in their four classes, according to the usual classification, and the frequency of the basic reproduction scores R_0 is shown in Fig. 8, separately for the two dynamics of receiving retweets (top) and mentions (bottom). While no significant difference emerges, in either scenario, among the four different classes of users for the lowest scores (i.e., $2 < R_0 < 2.5$ for retweets, and $1 < R_0 < 1.2$ for mentions), strong class differences emerge for users whose content are more contagious ($R_0 > 2.5$ for retweets, and $R_0 > 1.2$ for mentions): concerning retweets, influential users, followed by information broadcasters, are receiving systematically more attention than users in other classes; the class differences for mentions are less pronounced.

4. Related work

One of the first computational frameworks, proposed by Bermingham *et al.* [12] in 2009, combined social network analysis with sentiment detection tools to study the agenda of a radical YouTube group: the authors examined the topics discussed within the group and their polarity, to model individuals' behavior and spot signs of extremism and intolerance, seemingly more prominent among female users. The detection of extremist content (on the Web) was also the focus of a 2010 work by Qi *et al.* [51]. The authors applied hierarchical clustering to extremist Web pages to divide them into different pre-imposed categories (religious, anti immigration, etc.).

Scanlon and Gerber proposed the first method to detect cyber-recruitment efforts in 2014 [56]. They exploited data retrieved from the Dark Web Portal Project [17], a repository of posts compiled from 28 different online fora on extremist religious discussions (e.g., Jihadist) translated from Arabic to English. After annotating a sample of posts as recruitment ef-
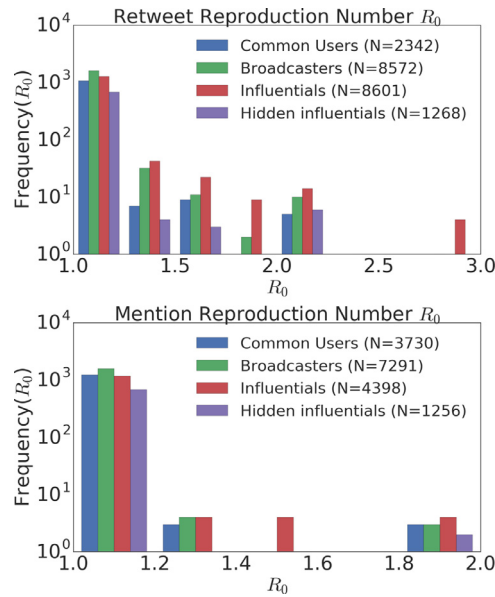
Fig. 8. Frequency of basic reproduction scores R_0 relative to obtained retweets (top) and mentions (bottom) calculated for the four dynamical classes of ISIS users.

forts or not, the authors use Bayesian criteria and a set of textual features to classify the rest of the corpus, obtaining good accuracy, and highlighted the most predictive terms.

Along the same trend, Agarwal and Sureka proposed different machine learning strategies [1,3,4,61] aimed at detecting radicalization efforts, cyber recruitment, hate promotion, and extremist support in a variety of online platforms, including YouTube, Twitter and Tumblr. Their frameworks leverage features of contents and metadata, and combinations of crawling and unsupervised clustering methods, to study the online activity of Jihadist groups on the platforms mentioned above.

A few studies explored unconventional data sources: one interesting example is the work by Vergani and Bliuc [66] that uses sentiment analysis (Linguistic Inquiry and Word Count [63]) to investigate how language evolved across the first 11 Issues of Dabiq, the flagship ISIS propaganda magazine. Their analysis offers some insights about ISIS radicalization motives, emotions and concerns. For example, the authors found that ISIS has become increasingly concerned with females, reflecting their need to attract women to create their utopia society, not revolving around warriors but around families. ISIS also seems to have increased the use of internet jargon, possibly to connect with the identities of young individuals online.

Concluding, two very recent articles [37,54] explore the activity of ISIS on social media. The former [54] focuses on Twitter and aims at detecting users who exhibit signals of behavioral change in line with radicalization: the authors suggest that out of 154K users only about 700 show significant signs of possible radicalization, and that may be due to social homophily rather than the mere exposure to propaganda content. The latter study [37] explores a set of 196 pro-ISIS aggregates operating on VKontakte (the most popular Russian online social network) and involving about 100K users, to study the dynamics of survival of such groups online: the authors suggest that the development of large and potentially influential pro-ISIS groups can be hindered by targeting and shutting down smaller ones. For additional pointers we refer the interested reader to two recent literature reviews on this topic [2,19].

5. Conclusions

Since the appearance of the Islamic State (viz. ISIS), a consensus has emerged on the relationship between extremism and social media, namely that ISIS' success as a terrorist organization is due at least in part to its savvy use of social media. It is widely believed that ISIS has managed to increase its roster to tens of thousands of members by broadcasting its savage attacks over social media platforms such as Twitter, which helps radicalize and ultimately recruit fighters from around the world.

Recent attacks on American and European soil demonstrate ISIS' potential to reach, organize, and mobilize lone wolves and sleeper terrorist cells among westerners. Many of these actors are known to have consumed radical material online and many have claimed to gravitate towards Islamic State because of it.

This paper posed two research questions: the former was concerned with proposing good practices for data collection, validation, and analysis of online radicalization. The latter aimed at revealing the network and temporal activity patterns of ISIS influence on Twitter. To address these questions we analyzed the activity of a group of twenty-five thousand users associated with ISIS. These accounts have been manually identified, reported to Twitter for verification, and subsequently

suspended due to their involvement with radical propaganda. This process yielded a human-curated dataset containing over three million tweets generated during a period of one and half year, 92% of which in Arabic language.

Regarding the first research question (**RQ1**), we highlighted the challenges related to studying Arabic content, due to the limits of existing NLP and sentiment analysis toolkits. We therefore suggested the adoption of content-agnostic statistical and network techniques to dissect the users' temporal activities and connectivity patterns. By leveraging a computational tool named *dynamic activity-connectivity map*, we highlighted the dynamics of social influence within ISIS support.

For what concerns the second research question (**RQ2**), our findings suggest complex strategies carried out by these users to manipulate and influence others: four dynamical classes of ISIS supporters emerged (common sympathizers, information broadcasters, influential and hidden influential users), each with distinct activity and connectivity patterns. We concluded by quantifying the extent to which ISIS support and extremist content are adopted in the general population: by drawing a parallel between propaganda and epidemics spreading, we determined that each ISIS supporter "infected" on average 2.13 other users before Twitter suspended his/her account, highlighting that receiving retweets and mentions follow different dynamics, and that broadcasters and influential users generate much more widespread contagions.

Although we call for caution in the interpretation of these results, due to the great simplifications introduced by our framework, we believe that our findings will help design and implement effective countermeasures and responses to ISIS social media operations and other forms of online extremist propaganda.

Acknowledgments

References

[1] S. Agarwal, A. Sureka, A focused crawler for mining hate and extremism promoting videos on youtube, in: Proc. 25th ACM conference on Hypertext and social media, ACM, 2014, pp. 294–296.

[2] S. Agarwal, A. Sureka, Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats, arXiv preprint arXiv:1511.06858 (2015).

[3] S. Agarwal, A. Sureka, Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter, in: Distributed Computing and Internet Technology, 2015, pp. 431–442.

[4] S. Agarwal, A. Sureka, Spider and the flies: focused crawling on tumblr to detect hate promoting communities, arXiv preprint arXiv:1603.09164 (2016).

[5] S. Agreste, S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, Network structure and resilience of mafia syndicates, Inf. Sci. 351 (2016) 30–47.

[6] W.H. Allendorfer, S.C. Herring, ISIS *vs.* the US government: a war of online video propaganda, First Monday 20 (12) (2015).

[7] C.L. Althaus, Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa, PLOS Currents Outbreaks (2014).

[8] A. Badawy, E. Ferrara, The rise of jihadist propaganda on social networks, arXiv preprint arXiv:1702.02263 (2017).

[9] J. Berger, J. Morgan, The ISIS twitter census: defining and describing the population of ISIS supporters on twitter, Brookings Project US Relations Islamic World 3 (20) (2015).

[10] J. Berger, H. Perez, The islamic states diminishing returns on twitter, GWU Program Extremism 02–16 (2016).

[11] J. Berger, B. Strathearn, Who matters online: measuring influence, evaluating content and countering violent extremism in online social networks, Int. Centre Study Radicalisation (2013).

[12] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, A.F. Smeaton, Combining social network analysis and sentiment analysis to explore the potential for online radicalisation, in: 2009 International Conference on Advances in Social Network Analysis and Mining, IEEE, 2009, pp. 231–236.

[13] A. Bessi, M. Coletto, G.A. Davidescu, A. Scala, G. Caldarelli, W. Quattrociocchi, Science *vs* conspiracy: collective narratives in the age of misinformation, PLoS ONE 10 (2) (2015) e0118093.

[14] A. Bessi, E. Ferrara, Social bots distort the 2016 us presidential election online discussion, First Monday 21 (11) (2016).

[15] D. Centola, The spread of behavior in an online social network experiment, Science 329 (5996) (2010).

[16] D. Centola, An experimental study of homophily in the adoption of health behavior, Science 334 (6060) (2011) 1269–1272.

[17] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, G. Weimann, Uncovering the dark web: a case study of Jihad on the web, J. Am. Soc. Inf. Sci. Technol. 59 (8) (2008) 1347–1359.

[18] M.D. Conover, E. Ferrara, F. Menczer, A. Flammini, The digital evolution of occupy Wall Street, PloS One 8 (5) (2013) e64679.

[19] D. Correa, A. Sureka, Solutions to detect and analyze online radicalization: a survey, arXiv preprint arXiv:1301.4916 (2013).

[20] C.A. Davis, G.L. Ciampaglia, L.M. Aiello, K. Chung, M.D. Conover, E. Ferrara, A. Flammini, G.C. Fox, X. Gao, B. Gonçalves, et al., Osome: the Iuni observatory on social media, PeerJ Comput. Sci. 2 (2016) e87.

[21] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H.E. Stanley, W. Quattrociocchi, The spreading of misinformation online, in: Proceedings of the National Academy of Sciences, 113, 2016, pp. 554–559.

[22] Y. Dong, Z. Ding, L. Martínez, F. Herrera, Managing consensus based on leadership in opinion dynamics, Inf. Sci. 397 (2017) 187–205.

[23] E. Ferrara, Manipulation and abuse on social media, ACM SIGWEB Newslett. (Spring) (2015) 4.

[24] E. Ferrara, Disinformation and social bot operations in the run up to the 2017 French presidential election, First Monday 22 (8) (2017).

[25] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots, Commun. ACM 59 (7) (2016) 96–104.

[26] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, A. Galstyan, Predicting online extremism, content adopters, and interaction reciprocity, in: International Conference on Social Informatics, Springer, 2016, pp. 22–39.

[27] E. Ferrara, Z. Yang, Measuring emotional contagion in social media, PLoS One 10 (11) (2015).

[28] E. Ferrara, Z. Yang, Quantifying the effect of sentiment on information diffusion in social media, PeerJ Comput. Sci. 1 (2015) e26.

[29] A. Fisher, How jihadist networks maintain a persistent online presence, Persp. Terrorism 9 (3) (2015).

[30] S. González-Bailón, J. Borge-Holthoefer, Y. Moreno, Broadcasters and hidden influentials in online protest diffusion, Am. Behavioral Scientist (2013). 0002764213479371.

[31] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, Y. Moreno, The dynamics of protest recruitment through an online network, Sci. Rep. 1 (2011).

[32] S. Green, C.D. Manning, Better arabic parsing: baselines, evaluations, and analysis, in: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 394–402.

[33] A. Gupta, H. Lamba, P. Kumaraguru, $1.00 per rt# bostonmarathon# prayforboston: analyzing fake content on twitter, in: eCrime Researchers Summit (eCRS), 2013, IEEE, 2013, pp. 1–12.

[34] T.D. Hollingsworth, R.M. Anderson, C. Fraser, HIV-1 transmission, by stage of infection, J. Infect. Dis. 198 (5) (2008) 687–693.

[35] P.N. Howard, B. Kollanyi, Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum, Available at SSRN 2798311 (2016).

[36] T. Hwang, I. Pearce, M. Nanis, Socialbots: voices from the fronts, Interactions 19 (2) (2012) 38–45.

[37] N. Johnson, M. Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, P. Manrique, D. Johnson, E. Restrepo, C. Song, et al., New online ecology of adversarial aggregates: ISIS and beyond, Science 352 (6292) (2016) 1459–1463.

[38] G. King, J. Pan, M.E. Roberts, How the chinese government fabricates social media posts for strategic distraction, not engaged argument, Am. Political Sci. Rev. (in press) 2 (2017). Copy at http://j.mp/1Txxiz1.

[39] M. Korayem, D. Crandall, M. Abdul-Mageed, Subjectivity and sentiment analysis of arabic: a survey, in: International Conference on Advanced Machine Learning Technologies and Applications, Springer, 2012, pp. 128–139.

[40] A.D. Kramer, J.E. Guillory, J.T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, Proc. Nat. Acad. Sci. 111 (24) (2014) 8788–8790.

[41] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media? in: Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 591–600.

[42] M.S. Lau, B.D. Dalziel, S. Funk, A. McClelland, A. Tiffany, S. Riley, C.J.E. Metcalf, B.T. Grenfell, Spatial and temporal dynamics of superspreading events in the 2014–2015 west Africa ebola epidemic, Proc. Nat. Acad. Sci. 114 (9) (2017) 2337–2342.

[43] D. Lazer, A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., Life in the network: the coming age of computational social science, Science (New York, NY) 323 (5915) (2009) 721.

[44] J. Lehmann, B. Gonçalves, J.J. Ramasco, C. Cattuto, Dynamical classes of collective attention in twitter, in: Proceedings of the 21st international conference on World Wide Web, 2012, pp. 251–260.

[45] W. Magdy, K. Darwish, I. Weber, #FailedRevolutions: using twitter to study the antecedents of ISIS support, First Monday 21 (2) (2016).

[46] P.T. Metaxas, E. Mustafaraj, Social media and the elections, Science 338 (6106) (2012) 472–473.

[47] C.E. Mills, J.M. Robins, M. Lipsitch, Transmissibility of 1918 pandemic influenza, Nature 432 (7019) (2004) 904–906.

[48] B. Mønsted, P. Sapieżyński, E. Ferrara, S. Lehmann, Evidence of complex contagion of information in social media: an experiment using twitter bots, arXiv preprint arXiv:1703.06027 (2017).

[49] F. Morstatter, J. Pfeffer, H. Liu, K.M. Carley, Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose, 7th Int AAAI Conf on Weblogs and Social Media, 2013.

[50] J. Prejean, R. Song, A. Hernandez, R. Ziebell, T. Green, F. Walker, L.S. Lin, Q. An, J. Mermin, A. Lansky, et al., Estimated HIV incidence in the United States, 2006–2009, PloS One 6 (8) (2011) e17502.

[51] X. Qi, K. Christensen, R. Duval, E. Fuller, A. Spahiu, Q. Wu, C.-Q. Zhang, A hierarchical algorithm for clustering extremist web pages, in: 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2010, pp. 458–463.

[52] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, F. Menczer, Detecting and tracking political abuse in social media., ICWSM 11 (2011) 297–304.

[53] S. Reardon, Terrorism: science seeks roots of terror, Nature 517 (7535) (2015) 420–421.

[54] M. Rowe, H. Saif, Mining pro-ISIS radicalisation signals from social media users, in: Proceedings of the Tenth International AAAI Conference on Web and Social Media, 2016, pp. 329–338.

[55] S. Salzberg, Forbes: anti-vaccine movement causes worst measles epidemic in 20 years, (http://www.forbes.com/sites/stevensalzberg/2015/02/01/anti-vaccine-movement-causes-worst-measles-epidemic-in-20-years/).

[56] J.R. Scanlon, M.S. Gerber, Automatic detection of cyber-recruitment by violent extremists, Secur. Inform. 3 (1) (2014) 1–10.

[57] Q. Schiermeier, Terrorism: terror prediction hits limits., Nature 517 (7535) (2015) 419.

[58] J. Stern, J.M. Berger, ISIS: The State of Terror, HarperCollins, 2015.

[59] P. Suárez-Serrato, M.E. Roberts, C. Davis, F. Menczer.

[60] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, F. Menczer, The Darpa twitter bot challenge, Computer 49 (6) (2016) 38–46.

[61] A. Sureka, S. Agarwal, Learning to classify hate and extremism promoting tweets, in: Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, IEEE, 2014. 320–320

[62] T.R. Tangherlini, V. Roychowdhury, B. Glenn, C.M. Crespi, R. Bandari, A. Wadia, M. Falahi, E. Ebrahimzadeh, R. Bastani, "Mommy Blogs" and the vaccination exemption narrative: results from a machine-learning approach for story aggregation on parenting social media sites, JMIR Public Health Surv. 2 (2) (2016).

[63] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54.

[64] O. Varol, E. Ferrara, C. Davis, F. Menczer, A. Flammini, Online human-bot interactions: detection, estimation, and characterization, in: International AAAI Conference on Web and Social Media, 2017, pp. 280–289.

[65] O. Varol, E. Ferrara, C.L. Ogan, F. Menczer, A. Flammini, Evolution of online user behavior during a social upheaval, in: Proceedings of the 2014 ACM conference on Web science, 2014, pp. 81–90.

[66] M. Vergani, A.-M. Bliuc, The evolution of the ISIS' language: a quantitative analysis of the language of the first year of Dabiq magazine, Secur. Terrorism Soc. 1 (2) (2015) 7–20.

[67] J. Wallinga, P. Teunis, Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures, Am. J. Epidemiol. 160 (6) (2004) 509–516.

[68] M. Weiss, H. Hassan, ISIS: Inside the Army of Terror, Simon and Schuster, 2015.

[69] D. Yadron, The Guardian: twitter deletes 125,000 ISIS accounts and expands anti-terror teams, (http://www.theguardian.com/technology/2016/feb/05/twitter-deletes-isis-accounts-terrorism-online).