



**HAL**  
open science

## Setting the Record Straighter on Shadow Banning

Erwan Le Merrer, Benoît Morgan, Gilles Trédan

► **To cite this version:**

Erwan Le Merrer, Benoît Morgan, Gilles Trédan. Setting the Record Straighter on Shadow Banning. INFOCOM 2021 - IEEE International Conference on Computer Communications, IEEE, May 2021, Virtual, Canada. pp.1-10, 10.1109/INFOCOM42981.2021.9488792 . hal-03234771

**HAL Id: hal-03234771**

**<https://inria.hal.science/hal-03234771>**

Submitted on 25 May 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Setting the Record Straighter on Shadow Banning

Erwan Le Merrer,  
Univ Rennes, Inria, CNRS, Irisa  
erwan.le-merrer@inria.fr

Benoît Morgan,  
IRIT/ENSHEEIT  
benoit.morgan@irit.fr

Gilles Trédan,  
LAAS/CNRS  
gtredan@laas.fr

**Abstract**—*Shadow banning* consists for an online social network in limiting the visibility of some of its users, without them being aware of it. Twitter declares that it does not use such a practice, sometimes arguing about the occurrence of “bugs” to justify restrictions on some users. This paper is the first to address the plausibility of shadow banning on a major online platform, by adopting both a statistical and a graph topological approach.

We first conduct an extensive data collection and analysis campaign, gathering occurrences of visibility limitations on user profiles (we crawl more than 2.5 millions of them). In such a black-box observation setup, we highlight the salient user profile features that may explain a banning practice (using machine learning predictors). We then pose two hypotheses for the phenomenon: *i*) limitations are bugs, as claimed by Twitter, and *ii*) shadow banning propagates as an epidemic on user-interaction ego-graphs. We show that hypothesis *i*) is statistically unlikely with regards to the data we collected. We then show some interesting correlation with hypothesis *ii*), suggesting that the interaction topology is a good indicator of the presence of groups of shadow banned users on the service.

## I. INTRODUCTION

Online Social Networks (OSNs) like Twitter, Facebook, Instagram or YouTube control the visibility of the content uploaded by their users. They have the capacity to promote or demote specific contents, and have great responsibilities (*e.g.*, to moderate hate speech, prevent automation for influence gain [1] or to defend copyright ownership). OSNs often position themselves as free speech defenders.

While OSNs need to implement policies that satisfy such requirements, precise policies are rarely publicly displayed. Therefore, debates on their behavior with respect to some content they host is generally fueled by three sources: *i*) OSN’s official statements, *ii*) anecdotal evidence from users publicizing their observations (*e.g.*, particular requests such as “Clinton vs Trump” [2]), and *iii*) whistle-blowing from internal sources [3] or internal information leaks. Investigation journalism sometimes discusses the problem in a broader context with disparate methods [4].

While debates about perceived freedom of speech are inevitable, we believe it is important to explore techniques to shed light on OSNs content regulation practices. More precisely, means to observe<sup>1</sup>, assess and quantify the effects of content moderation is important for the debate on information regulation in the public sphere. However, as content is produced and consumed distributedly, and as its moderation

happens on the OSN side, collecting information about potential issues is difficult. In this paper, we explore scientific approaches to shed light on Twitter’s alleged *shadow banning* practice. Focusing on this OSN is crucial because of its central use as a public communication medium, and because potential shadow banning practices were recently commented.

*Shadow banning and moderation techniques.* shadow banning (SB or banning for short, also known as *stealth banning* [6]) is an online moderation technique used to ostracise undesired user behaviors. In modern OSNs, shadow banning would refer to a wide range of techniques that artificially limit the visibility of targeted users or user posts (see *e.g.*, ref. [7] for a position on shadow banning in Instagram).

Some people claim what they publish is discriminated by a moderation algorithm [6]. However, while platforms publicly acknowledge the use of automatic moderation, they deny the use of shadow banning. In particular, in a dedicated blog post entitled “*Setting the record straight on shadow banning*” [8], Twitter acknowledged some problems in July 2018, but presented them as patched issues or bugs.

*Observation in a black-box setup.* From a user-standpoint, observing a remote decision-making algorithm (*e.g.*, recommending people to follow, recommending topics, or searching and sorting users accounts), gaining some information about the OSN moderation practices imposes a *black-box interaction setup* (see *e.g.*, refs [9]–[11] for related research works). In such a setup, the difficulty is to be bound to observe solely input/output relations such as actions and consequences in the OSN, and to build a relevant case from them.

We follow a statistical approach in order to **address the question of the plausibility of shadow banning in Twitter**. Such an approach was also recently embraced by Jiang & al to collect the context of YouTube videos, in order to assess if the political leaning of a content plays a role in the moderation decision for its associated comments [9]. The question is addressed statistically, to validate or reject the hypothesis of bias by YouTube.

*Contributions.* We rely on three known techniques [12] to detect Twitter users or tweets with diminished visibility, that we implement in a full fledged scalable and automated crawler. We pursue a statistical and topological perspective on shadow banning, by comparing the plausibility of two hypotheses. More precisely, we make the following contributions:

- We quantify the phenomenon of shadow banning on Twitter, through an extensive data collection and analysis campaign. We collect the public profiles and interactions

<sup>1</sup>We operate a shadow banning test website: <https://whosban.eu.org> [5].

of millions of Twitter users, as well as their shadow banning status.

- We identify salient profile features that contribute to the probability to be banned, using machine learning explainable predictors.
- We test the hypothesis of a random bug  $H_0$ : *shadow banned users are uniformly spread among Twitter users*, which corresponds to Twitter’s bug claim. We show this hypothesis to be statistically unlikely.
- We propose another hypothesis  $H_1$ : the topological hypothesis. It models shadow banning as an epidemic process among interacting users. It leverages their interaction topologies, in order to capture the observed effect of groups of shadow banned users. We show this hypothesis to better match our collected observations.

The remaining of this paper is organized as follows. In Section II, we define how to test for shadow banning, and detail the collection campaign we conducted to allow focusing on the shadow banning question in Twitter. In Section III, we report statistics and analyze the presence of banned profiles. In Section IV, we have a look at which features may predict a shadow ban status on a user. In Section V we introduce and study our two core work hypotheses. We review Related Work and conclude in Sections VI and VII. Finally, we issue a data and code availability statement in Section VIII.

## II. A DATA COLLECTION CAMPAIGN FOR TWITTER

Studying shadow banning on Twitter requires two fundamental ingredients: first, means to detect whether a specific user profile is banned. Second, we need to select populations on which to apply such user-level detection. Each population should be large enough and representative, so that conclusions drawn can be meaningful.

### A. Means to Assess Shadow Banning in Twitter

In the context of Twitter, the notion of *shadow banning* can describe a handful of situations where the visibility of a shadow banned user or his posts is reduced as compared to normal visibility. The first website to provide users with the ability to check whether they are individually shadow banned is *shadowban.eu* [12]. Interestingly, its authors provided code on GitHub, as well as explanations of techniques to assert banning facts. We leveraged and incorporated these techniques to develop our crawler. Here are the types of bans we consider in the paper:

- *Suggestion Ban*: Users targeted by the suggestion ban are never suggested, as another user performs searches or mentions them in some content. This limits the possibility for users to accidentally reach a banned user profile.
- *Search Ban*: Users are never shown in search results, even if their exact user name is searched for.
- *Ghost Ban*: If a targeted user made a tweet  $t$  as a new thread, a retweet or a reply to someone else’s tweet  $t'$ , it is not shown (but is replaced by the mention “This tweet is unavailable”). No button allows to see it.

We declare a user to be banned if at least one of these bans holds. We later report their precise relative occurrence in our analysis.

It is important to highlight two properties of this detection approach. First, it does not produce false-positives (normal users accidentally appearing as banned): detected users have actual diminished visibility (at least at the performed crawl time). Moreover, our detector also produces a proof allowing a human direct confirmation of detected case. Second, these types of bans might only constitute a subset of Twitters’ banning strategy: there might be more methods to diminish the visibility of a user and for which we do not know any practical detector. As a consequence, the data collection results might underestimate shadow banning on Twitter, but not overestimate it.

### B. A Data Collection Campaign

We built a scalable crawler to retrieve user profiles, and test the types of bans we described. As all Twitter’s users obviously cannot be tested for banning (Twitter in Q1 2019 reported 330 millions of monthly users<sup>2</sup>), we resorted to the sampling of ego-graphs around selected users, which is a common practice for studying OSNs (see *e.g.*, [13], [14]).

In order to analyze if banning is concerning evenly different types of users, we selected *four* types of user *populations*. We now describe these populations, and how we extracted them from Twitter:

*a) A random population*: To uniformly select a RANDOM population of users, we exploit a property of the Twitter API that associates to each user a user ID randomly drawn in a finite subset of  $\mathbb{N}$ . To cope with the success of that social network, this user ID space has been resized from 32-bit to 64-bit in late 2015. Current user IDs seem to be still randomly drawn from this huge 64-bit space which is for now still sparse : 330 millions over 18 billion billion, leaving us a probability less than  $1.8 \times 10^{-11}$  to pick an actual account at random. Due to obvious time limitations, we decided to use the first user ID space to draw random accounts, created before late 2015. Therefore, our RANDOM population contains pre-2015 users sampled by drawing uniformly at random user IDs in the range  $[1, 2^{32} - 1]$ .

*b) Identified bots*: For collecting a population of BOTS [15], we leveraged the website <https://botsentinel.com>, that has the purpose of identifying and listing bots operating in Twitter. Bots are classified into categories accessible using a web interface. We have chosen to use the so called “Trollbot” category, because of their “perceived likelihood” of being shadow banned. We have instrumented the HTTP REST API endpoint used by the web interface in order to extract 1,500 account screen names.

*c) Celebrities*: To build a population of very visible user accounts, we denote FAMOUS, we leveraged the website <https://majesticmonitor.com/free-tools/social-explorer>. This application offers a hierarchical ranking, by topic, of the 10 most

<sup>2</sup><https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

famous Twitter accounts as follows. Topics hierarchy is a tree where its root gathers the 10 most famous Twitter accounts from all topics. Root siblings are in turn gathering the 10 most famous Twitter accounts from their own topics and subtopics, and so on. That is, the user is able to display the top ten Twitter accounts by category, by browsing in the tree based hierarchy.

Once again we have scripted browser interactions in order to extract the 1,500 most famous accounts, using a depth limited breadth-first search over the HTTP REST API.

d) *Political representatives in France*: We build a population we coin `DEPUTEES`, gathering the full list of elected deputies who have a Twitter account in France [16] (577 as of December 2019). We target this population because of its specific exposure to the media.

### C. Sampling ego-graphs in the Twitter Interaction Graph.

Rather than simply crawling individual profiles in each of these populations, we rely on "snowball" sampling from these profiles to capture ego-graphs topologies around them.

More precisely, we consider the *Twitter interaction graph* as the graph  $G_{Twitter} = (V, E)$  constituted by  $V$  the set of all Twitter user accounts, and  $E$  a set of directed edges established as follows:  $(u, v) \in E \Leftrightarrow$  user  $v$  replied to  $u$ , or retweeted one of  $u$ 's messages<sup>3</sup>. As crawling the full Twitter interaction graph is out of the question, we sample ego-graphs from that graph as follows (see e.g., ref. [13], [14] for other works extraction ego-graphs in the interaction graph).

We call each of the user profile in the four populations a *landmark*, around which the ego-graph will be recursively sampled in the interaction graph. More precisely, from each of these landmarks  $l$ , we conduct a depth-limited breadth-first search: we parse the 33 first tweets of  $l$  returned among its 1,000 most recent tweets, and list the set of users  $V_{out}(l)$  with whom  $l$  interacted. We then repeat that procedure for each  $i \in V_{out}(l)$ , to discover the two-hop neighbors of landmark  $l$ ,  $V_{out}^2(l)$ . Finally, we also keep the neighbors of those rank-two nodes. The resulting ego-graph for landmark  $l$ , is noted  $G_l$  and is the sub-graph of  $G_{Twitter}$  induced by some of its close neighboring profiles  $V_l = \bigcup_{i=1,2,3} V_{out}^i(l)$ .

Note that using this process, although we chose the initial landmarks, we do not control the population in the ego-graphs: any user interacting with a landmark (or its ego-graph neighbors) will also appear in its ego-graph.

Our crawling campaign took place in April 2020. We run the set of shadow banning tests for each visited profile in each ego-graph, on all the tweets posted since 2019 and still available. Users that did not post any tweet since 2019 are considered inactive and ignored for further analysis.

### D. Ego-Graph Collection Results.

We targeted around 1,000 graphs per category (except for the `DEPUTEES` population that is bounded below by nature). We consider a graph to be suitable if it contains at least two

<sup>3</sup>Note that this graph differs from the explicit Twitter graph in which edges capture the "follower" relationship, examined for instance in [17].

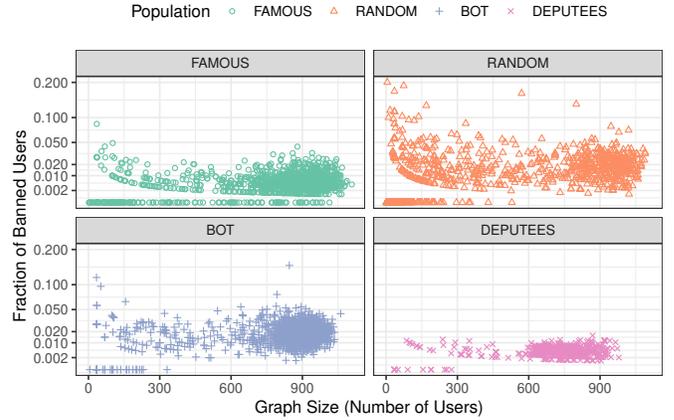


Fig. 1: Fraction of shadow banned nodes ( $y$ -axis) as a function of ego-graph sizes ( $x$ -axis), for the four populations.

nodes (*i.e.*, the landmark and one neighbor at the very least). As shown in the first column of Table II, the number of graphs to extract in order to filter out those with only one node vary greatly depending on the target population. Indeed, while the `DEPUTEES` population is very dense (filtering graphs with one node removes only 20 graphs over 512). The `RANDOM` one is crawling intensive: we needed to gather 13,991 graphs to be able to keep 947 filtered ones (*i.e.*, we filtered out the 93% of users sampled at random that never interacted with someone since 2019).

We report no throttling nor crawl limitation from Twitter during this data collection campaign. We performed a distributed crawling from 86 machines, resulting in a rate of around 100 profiles crawled per second. The total amount of crawled and tested user profiles adds up to above 2.5 millions.

## III. CRAWL STATISTICS: TRACES OF SHADOW BANNING

### A. Shadow Banning Prevalence in Populations.

We first plot in Figure 1 the fraction of shadow banned users present in each ego-graph. We observe that all populations are concerned by the phenomenon of shadow banning, across all the spectrum of ego-graph sizes. The detailed statistics are reported in Table I. The percentage of banned users in populations ranges from 0.50% for `DEPUTEES`, to 2.34% for `RANDOM` (and 0.74% for `FAMOUS`, 1.97% for `BOTS`). We note that the raw number of shadow banned profiles for `FAMOUS` looks relatively high (23,358), but this is due to the high density of ego-graphs, that are making our crawl to retrieve more than three times more profiles (1,179,949, see Table II) than for the `RANDOM` population for instance. Very noticeably, the `RANDOM` population is thus touched close to five times more by the shadow banning phenomenon than the `DEPUTEES` population. This already questions supposedly even spread of shadow banning in Twitter, due to a bug for instance.

Second, we observe significantly different statistics such as the average degrees of nodes knowing that a node is itself

	#SB nodes	% of SB nodes/graph (avg)	Degree of nodes SB—not SB (avg)	Fraction of SB neighbors: node is SB—not SB
FAMOUS	6,805	0.74	4.97 — 8.69	0.1044 — 0.0051
RANDOM	9,967	2.34	6.94 — 9.59	0.1694 — 0.0211
BOTS	23,358	1.97	11.40 — 15.04	0.0443 — 0.0184
DEPUTEES	1,746	0.50	22.18 — 14.40	0.0195 — 0.0104

TABLE I: Data collection campaign statistics: shadow banning (denoted SB) information on individual profiles, their neighbors, and their relative degree.

banned or not. In particular, the fraction of neighbors of a banned node that are also banned is much higher than for nodes that are not banned. This remark is consistent across the four populations. This constitutes a first indication of the existence of “groups of shadow banned users”, captured by the topology of ego-graphs.

### B. Co-occurrence of Types of Bans: a Graduated Response?

We presented general statistics about shadow banned users in four different populations. As we explained that a shadow ban status can come from three types of bans (and at least one was sufficient so that we declare a user as shadow banned), we now have a look at each type of ban in the dataset, in order to question a possible Twitter shadow banning policy as a reaction to user misbehaving.

Figure 2 reports a grand total of 41,071 *typeahead* banned profiles, 23,219 *search* bans, but only 3,681 *ghost* bans.

Shadow banning techniques described in Section II imply different consequences on user profiles. Their impact on visibility can be ordered by increasing severity. As a matter of fact, *typeahead* ban is less impacting than *search* ban, which in turn is less impacting than *ghost* ban. Indeed, the first two sanctions leverage access to the profile while the last one the publications themselves.

If we consider shadow banning as a punitive response against unwanted behavior as in penal law enforcement, one could say that:

- 1) on the one hand a punitive reaction could be graduated according to the severity of one misbehaving;
- 2) while on the other hand, recidivists could be disciplined several times, with increasing severity.

Observations depicted in Figure 2 seem to fit quite well with these two points. Indeed, users solely *typeahead* banned are moderately to very little *search* banned (53%) or *ghost* (9%) banned, while *ghost* banned users are almost every time *search* banned (100%) or *typeahead* (97%) banned. Let us remind that our data set is one snapshot at a given point in time. Although nicely fitting to the collected data and the severity order, being able to observe the evolution of sanctions per user in time would also have been of a great interest to strongly conclude on the second point.

### C. Graph Topology Related Statistics.

We now present general statistics on the ego-graphs.

a) *Graph sizes*: Due to our ego-graph sampling strategy, the size of the graphs we extract is upper bounded to  $1 + 33 + 33^2 = 1123$  nodes (corresponding to a depth two crawl around the landmark, with 33 neighbors at maximum per

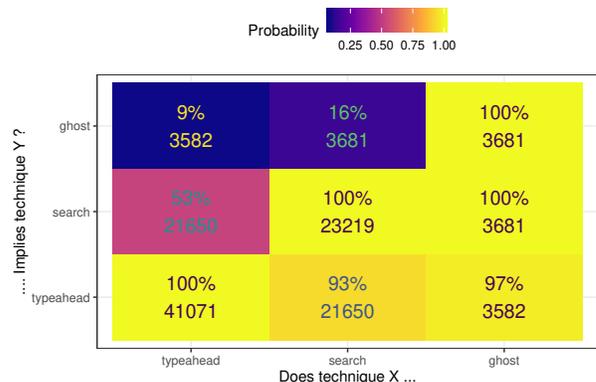


Fig. 2: The interaction of the three different types of bans we measured in our dataset. For instance, 97% of users that are ghost banned are **also** typeahead banned, while the reverse is true in only 9% of the counted cases.

node). We observe in Figure 3 the sizes of the collected graphs, per population, as a probability density function. BOTS and DEPUTEES exhibit a single mode, while FAMOUS and most notably RANDOM have two modes, consisting in a fraction of graphs with small sizes and another one of close to maximal sizes (centered at around 1,000 nodes).

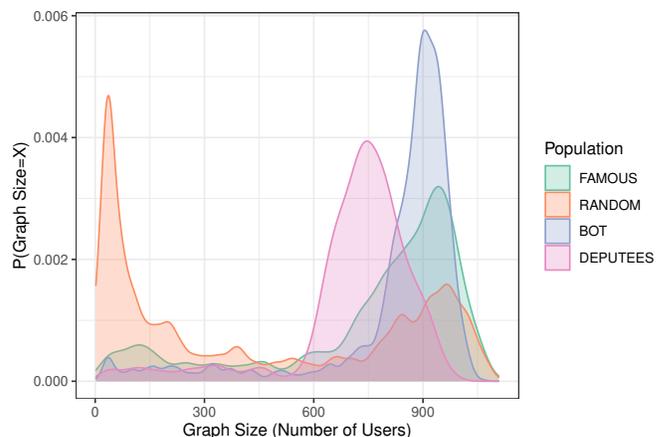


Fig. 3: A probability density function for the size of collected ego-graphs, for each of the four populations.

b) *General graph statistics*: General statistics are reported in Table II; note that we made all graphs undirected in our analysis, allowing for some topological computations and analysis such as  $k$ -cores. Because of our ego-graph sampling strategy, the average degree of nodes is expected in  $[2, 34]$ .

Population	#graphs — (unfiltered #)	Total #nodes	degree (undirected, avg)	clustering (avg)	2-core size (avg)
FAMOUS	1211 — 1400	908,131	5.92	0.2074	575.64
RANDOM	947 — 13,991	424,489	5.45	0.1965	357.21
BOTS	1436 — 1505	1,179,949	11.62	0.1754	751.64
DEPUTEES	492 — 512	348,640	12.21	0.1694	658.99

TABLE II: Statistics for the crawling campaign: topological information for extracted ego-graphs.

Indeed, if the rank-2 nodes have only neighbors outside the set of nodes already crawled, we have  $33 + (1+33) \times 33 + 1 \times 33^2 = 2244$  edges at maximum; this leads to  $2244/1123 \approx 2$  as an average graph degree. At the other extreme, since we capture only 33 interactions at maximum, maximum degree is 34. We observe in II that those averages lie close to the lower bound for the FAMOUS and RANDOM populations (with 5.92 and 5.45 respectively), and a significantly above for the BOTS and DEPUTEES populations (11.62 and 12.21). We note the average clustering coefficients to be relatively even (0.2074 for the largest one in the FAMOUS case). These statistics indicate various levels of interactions in the four different populations; we shall explore the effect of these interactions on the banning process.

#### IV. USER-FEATURES CORRELATING WITH BANNING

We are now questioning if some features in the collected individual user profiles are good predictors of a potential shadow ban status.

We leverage machine learning classifiers: the idea being that if one can predict with some reasonable accuracy if a profile is shadow banned by only looking at its features, then these features are encoding a part of the cause of a profile being banned. We choose three machine learning models that are explainable [18] by construction, that is to say that the model allows for precisely pinpoint the influence of features on the classification accuracy. Here is the considered setup.

*a) Prediction setup:* In order to train a predictor for shadow banning, we first need a labeled dataset. A first difficulty is however the unbalanced nature of the classification task at hand: over 97% of our dataset are negative instances (representing users that are not banned). Thus, a trivial classifier predicting "not banned" for any input would have a 97% accuracy, without bringing any information on relevant features. To circumvent this, we first balance the dataset.

We retain a total of 9,626 profiles of the RANDOM population, all having all of the features we leverage (we note that a large set of profiles have unset or missing features). The shadow banned and non shadow banned profiles are in an even quantity in the resulting dataset; these two sets constitute our labeled dataset. For each profile in these two sets, we use as features the data extracted from each user Twitter webpage that is either of a Boolean or integer format. In total, we exploit a set of 18 features that are listed on Figure 4 and analyzed hereafter. The naming of some of those data fields is very explicit (such as `followers_count` for instance), while some others are not (e.g., `possibly_sensitive_editable`).

We use the Scikit learn [19] library, and experiment with three explainable classifier models: a random forest algorithm

Classifier	Banned Status Prediction Accuracy
Random Forest (RF)	0.806
AdaBoost (AB)	0.766
DecisionTree (DT)	0.748

TABLE III: Accuracies of three explainable classifiers predicting the shadow banned status of 1,925 test users, based on their crawled profiles.

(RF), the AdaBoost algorithm (AB), and a decision tree (DT). The RF is the result of a grid search on the best combination of the following parameters: 'number of estimators'  $\in [50, 150, 250]$ , 'max features'  $\in [sqrt, 0.25, 0.5, 0.75, 1.0]$  and 'min samples split'  $\in [2, 4, 6]$ , leading to a optimum setup of respectively 150, *sqrt* and 2. The AB is the result of a grid search on the best combination of the following parameters: 'number of estimators'  $\in [50, 150, 250, 500]$ , 'learning rate'  $\in [0.1, 1, 2]$ , leading to a optimum setup of respectively 500 and 1. Finally, the DT is the result of a grid search on the best combination of the following parameters: 'max features'  $\in [auto, sqrt, log2]$ , 'min samples split'  $\in [2 \dots 15]$ , and 'min samples leaf'  $\in [1 \dots 11]$ , leading to the selection of respectively  $log2$ , 13 and 11.

The training is set to 80% of the dataset, leaving 20% of profiles as a test set.

*b) Predictor accuracies:* Accuracies of the three models are reported in Table III. An accuracy of 80.6% is observed for the RF model (76.6% for AB and 74.8% for DT), which clearly shows that there is some information in the features we collected that correlate with the shadow banned status of tested profiles. We believe this raw result to be encouraging for later research on even higher accuracies, for allowing for instance services like *shadowban.eu* [12] or *whosban.eu.org* [5] to rely on direct inference on public profile data, rather than on the interaction with the Twitter services to test the ban types described in Section II-A.

*c) Most salient features:* We now look at the features that are influencing the classifications of the RF model, as it has the best accuracy and is explainable. The relative contribution of each individual feature to the RF model decision is represented in Figure 4.

We first note that there is no single feature that can help differentiate between banned and non banned users: the decision might be a complex combination of several of them.

There are two features with above each 12% of influence on the result: `media_count` and `friends_count`; above 10% are also two more: `statuses_count` and `favorite_count`. Together those 4 features determine nearly half of the decision (45.8%). These features relate to a sort of acceptance from general users of the user under

scrutiny. This acceptance could lead to a good indicator on the probability to be shadow banned.

We note that the second classifier in accuracy, AB, ranks four features over 10% of influence. By decreasing order: `media_count`, `listed_count`, `normal_followers_count` and `statuses_count`. While the first feature is in a rank agreement with RF, the others are not, possibly indicating some redundancy of information in these features.

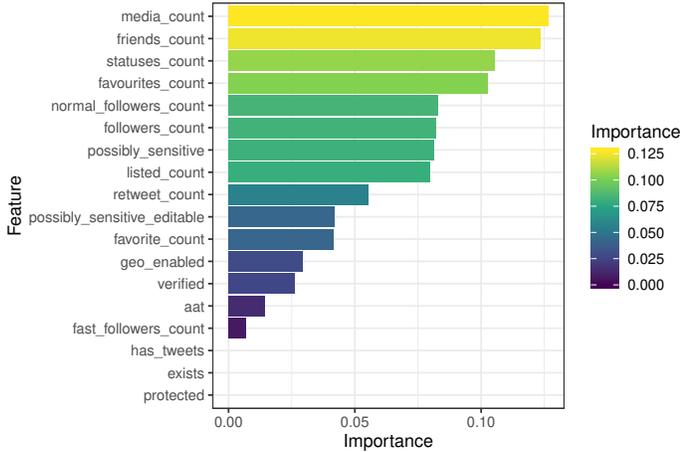


Fig. 4: The features sorted by order of importance in the random forest model (RF) prediction of shadow banned users.

We thus conclude that despite there is no obvious public group of features that permit perfect inference on the shadow ban status of users, a prediction of 80% indicates the presence of relevant information in our crawl. (We make this crawl public for further research.)

## V. TWO HYPOTHESIS: BUGS AND TOPOLOGICAL EFFECT

The previous sections exploited the collected data at the individual user level. Interestingly, it also revealed that at the global scale, different populations are differently impacted by shadow banning. In other words, banning does not appear as homogeneous, but rather concentrated in some regions of the interaction graph. We next seek to confirm this intuition.

### A. Hypothesis $H_0$ : the Plausibility of Bugs

We recall hypothesis  $H_0$ : *shadow banned nodes are uniformly distributed among Twitter users*. In this hypothesis, each user is banned with a uniform probability  $\mu$ , the only parameter of this model  $H_0(\mu) : \mathbb{P}(x \in SB) = \mu$ .

To avoid sampling biases, we now focus on the 400,000+ profiles RANDOM population in the remaining of this paper (as the three other populations pertain to targeted sampling of specific populations).

Fitting  $H_0$  is trivial:  $H_0(\mu)$  is most likely given our observations when  $\mu$  is set to be the fraction of observed banned nodes in RANDOM (aka sample mean) that we write  $\hat{\mu} = 0.0234$  (see Table II). When the context is clear, we write  $H_0$  as a shorthand notation for  $H_0(\hat{\mu})$ . This hypothesis embodies the

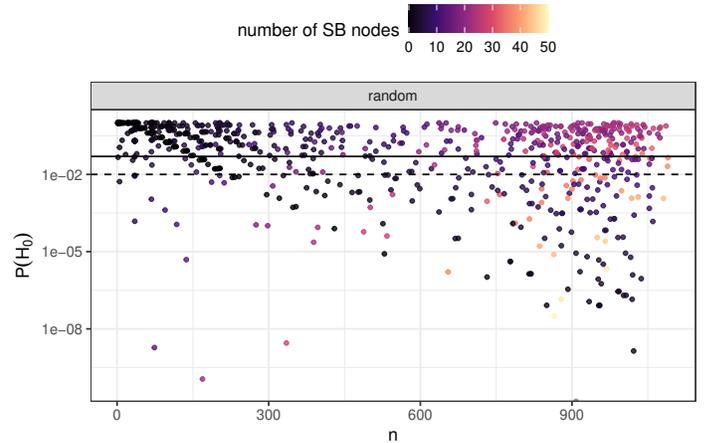


Fig. 5: The  $p$ -value of the  $H_0$  (bug) hypothesis for each landmark. Dashed and continuous lines represent the 1% and 5% significance levels, respectively.

*bug* explanation: bugs (software faults) are often considered to randomly affect users [20].

As  $H_0$  completely ignores the topological dimension of collected ego-graphs, what remains is a balls and bins sampling process: we can assess the probability of observing the amount of shadow banned nodes in each graph we collected, under  $H_0$ . In a nutshell in this hypothesis, Twitter is a big bin containing a fraction  $\hat{\mu}$  of banned balls, the rest being non-banned balls. In this  $H_0$  perspective, every time we sample a landmark  $l$  and its ego-graph  $G_l$ , we draw  $|G_l|$  balls from the bin and count how many banned balls we have drawn. In other words, every ego-graph  $G_l$  we sample is considered as  $|G_l|$  realizations of Bernoulli process of probability  $\hat{\mu}$ .

We borrow a general statistical significance testing approach, as *e.g.*, used also in ref. [7]. Given  $\hat{\mu}$ , estimating the probability to observe  $|SB|$  successes (in other words, it's associated  $p$ -value) is a process known as Binomial test.

Figure 5 displays the corresponding  $p$ -value of  $H_0$ , with regards to the size of each ego-graph and the number of shadow banned nodes it contains. In other words, it represents the probability that a Bernoulli trial with a success probability of  $\hat{\mu}$  leads to the number of banned user observed in each ego-graph. Remember that the lower the  $p$ -value, the higher the plausible rejection of the hypothesis under scrutiny. We observe an important amount of graphs that are significantly below significance levels of 1% and 5%: those are unlikely events to be observed under  $H_0$ . Moreover the number of banned nodes in each ego-graph (represented by point color) hints two types of unlikely ego-graphs: large graphs with too few banned nodes (black dots) and graphs with too many banned (clearer dots): there exist important groups of banned nodes in some of the crawled ego-graphs.

The  $y$ -axis in Figure 5 is cut under the probability  $1e-09$ , for readability. We omitted 14 samples that are even below this probability. Table IV represents the top-5 most unlikely ego-graphs we observed, around user profiles collected in our

	Ego-graph size	Ratio of SB nodes	Probability under $H_0$
Artem*	703	0.454	1.26e-315
Vlman*	605	0.443	6.42e-262
santi*	937	0.331	1.67e-255
Brows*	796	0.241	3.03e-130
ZchBr*	763	0.227	9.97e-113

TABLE IV: Top-5 most unlikely collected ego-graphs under hypothesis  $H_0$ , in the RANDOM population (along with their precise probability of observation).

crawl. User account names are truncated for privacy reasons. We observe that the ego-graph (of size 703) of user Artem\* contains 45.4% of shadow banned nodes; the likelihood of such a realization under the  $H_0$  model is 1.26e-315. In other words, observational data does not support hypothesis  $H_0$ . This conclusion calls for alternative models, such as  $H_1$ , which we now introduce.

### B. Hypothesis $H_1$ : Interaction Topology Reflects Banning

We have concluded that our shadow banning observations do not support the hypothesis of a random bug. Indeed, instead of revealing isolated cases evenly scattered in different landmarks, reveals that banned users are more concentrated around some landmarks and rarely around some others. As the ego-graphs observed around landmarks are in specific regions of the Twitter interaction graph, one can suspect a relation between the topology of the interaction graph, and the prevalence of banned users.

To investigate this, we propose an alternative hypothesis,  $H_1$ , that seeks to measure how local (with respect to the interaction topology) is the banning phenomenon. We first fit this probability, and then inspect its likelihood w.r.t.  $H_0$ .

a) *A simple Susceptible/Infected epidemic model:* We propose to adapt a simple Susceptible/Infected (SI) epidemic model [21]. Epidemic models, aside their obvious relevance in infectiology, are widely used to describe different topologically related phenomena in social networks, such as information cascades or rumor spreading [22]. While shadow banning is arguably a different phenomenon than the fact of being contaminated by a rumor, we believe the SI model to be the simplest way to capture the intuition that some groups of interacting users are differently touched by shadow banning.

The simplest SI model is a one step contamination process: each node is initially infected with probability  $p_0$ ; then, initially infected nodes can contaminate each of their neighbors with probability  $\beta$ . Therefore, this contamination process  $SI$  has two parameters:  $\beta$ , that captures the locality of the phenomenon, and  $p_0$  that allows to initially and uniformly spread the shadow ban status.

Let  $SI(p_0, \beta)$  be our contamination process. First, observe that  $SI(p_0 = \mu, \beta = 0) = H_0(\mu)$ : neutralizing contamination yields the random uniform spread of banned nodes described in  $H_0$ . As  $\beta$  increases, local contaminations occur around each initially infected user, and the overall number of banned users increases.

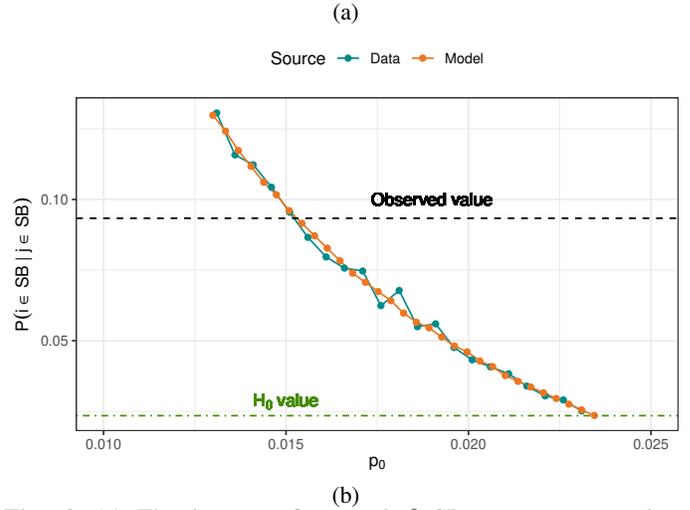
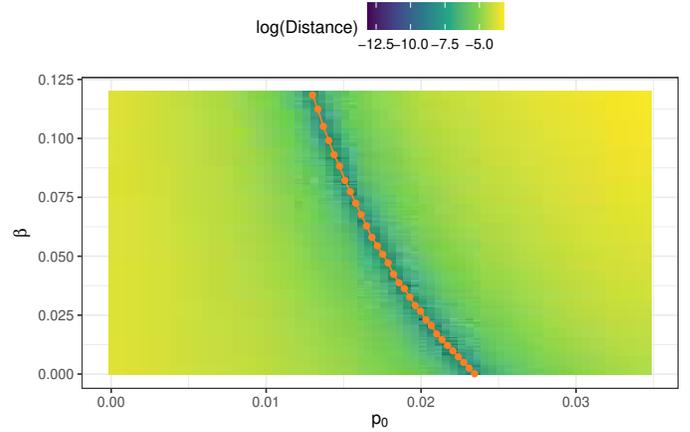


Fig. 6: (a) The impact of  $p_0$  and  $\beta$  SI parameters on the distance of simulated shadow banned user w.r.t. to the actual shadow banned users in ego-graphs. The line in green corresponds to a simple analytical model we propose. (b) Probability of neighboring contamination as a function of  $p_0$  for the  $H_1(\beta)$  model family.

b) *Fitting  $H_1$  to the observations:* We seek a couple  $(p_0, \beta)$  of parameters for  $SI$  that are the most likely given by our observations. A first observation is that such likely parameter couple should reproduce the global fraction of observed banned users  $\hat{\mu}$ . As observed above,  $(\hat{\mu}, 0)$  is one such couple, but by balancing differently initial infection and contaminations, it is possible to generate an infinity of such couples. Let  $H_1(\beta) = SI(p_0, \beta)$  such that  $\mathbb{P}(S|H_1(\beta)) = \hat{\mu}$ .

Let  $S$  be the random variable associated to the event "the user is banned". An estimation of the relation between  $\mu, \beta$  and  $p_0$  can be sketched as follows:  $\mathbb{P}(S|H_1(\beta)) \approx \mathbb{P}(\text{infected initially}) \oplus \mathbb{P}(\text{contaminated}) = p_0 + (1 - p_0)p_1$ . Where  $p_1$  is approximated as the probability of having some infected neighbors in a regular random graph of degree  $k$  and being contaminated by at least one of these:

$$p_1 = \sum_{v=1}^k \binom{k}{v} p_0^v (1 - p_0)^{k-v} (1 - (1 - \beta)^v).$$

In other words, this estimation neutralizes topological artifacts like clustering or degree heterogeneity to sketch a rough relation between  $p_0$  and  $\beta$  for a fixed  $\mu$ .

Figure 6a represents the quantity  $|\mathbb{P}(SB|SI(p_0, \beta)) - \mu|$  for varying  $p_0$  and  $\beta$ . More precisely, each point  $(p_0, \beta)$  on the figure corresponds to a SI model. We simulate this corresponding SI model on each crawled ego-graph, and count the simulated number of banned users. The color of the point corresponds to the difference between the simulated number of banned users and the (real) observed number of banned users, averaged over all ego-graphs. In other words that is the difference between fractions of shadow banned nodes observed on the ego-graphs and simulated using the SI model. A distance of 0 thus indicates that the SI simulation over the ego-graphs leads to the same amount of banned nodes that the one counted in the dataset. A darker color indicates a smaller distance.

We observe a smooth ridge, linking  $\beta$  and  $p_0$ . Note that, as expected, a  $\beta = 0$  leads to  $p_0$  being equal to the measured  $\hat{\mu}$ , that is the initial probability to be infected, without any contamination from neighbors.

The orange line represents the values derived from our analytical approximation. It follows closely the lowest experimental values that shape a valley, indicating that our model captures the process very well. As a consequence, the lowest spots of the valley and the orange line both define here a family of hypotheses  $H_1(\beta)$  in which all members approximate the total number of banned nodes as closely as the uniform infection  $H_0 = H_1(\beta = 0)$ . A natural follow-up question is "What would be a good value for  $\beta$ ?"

*c) Probability of a banned neighbor given a banned status:* Recall that  $\beta$  is the contamination probability, which is a local property. To estimate a good value, one can look at the probability that a banned node has a banned neighbor:  $\mathbb{P}(j \in SB|i \in SB \wedge (i, j) \in E)$ . While in  $H_0$  this probability is  $\hat{\mu}$  (as both events  $i \in SB$  and  $j \in SB$  are independent), in  $H_1(\beta > 0)$  the contamination drastically increases this probability. It can be roughly estimated as  $\mathbb{P}(j \in SB|i \in SB \wedge H_1(\beta)) \approx p_0 + (1 - p_0)\beta$  by again neglecting clustering in ego-graphs (and chances that two nodes contaminated by the same node are neighbors).

The empirical value we measured in the dataset for that probability is 9.3%: A user having at least one banned neighbor has nearly 4 times more chances of being banned. This observation again weakens a scenario such as  $H_0$ .

Figure 6b represents the probability of being banned if one has a banned neighbor for the family of  $H_1(\beta)$  hypotheses obtained above. We note a very good fit of the SI model with the measurements from the dataset. As expected, as  $p_0$  decreases,  $\beta$  increases, which in turn increases neighboring contamination chances. The dashed line represents the empirical observed value  $|(SB \times SB) \cap E|/|(SB \times V) \cap E|$ .

The model closest to this experimental line is  $H_1(\hat{\beta} = 0.0955)$  corresponding to the SI model where  $p_0$  is just above 0.015. This model would explain both the global number of shadow banned nodes, and the local co-occurrences of shadow

banning in the data. In the following, we set  $H_1$  to represent our fitted values:  $H_1 := H_1(\hat{\beta}) = SI(0.015, 0.0955)$ . In this model, contaminations are  $(\hat{\beta})/0.015 = 5.4$  times more likely to occur through neighbor contamination than through initial (random) contamination. We now can evaluate the likelihood of this new hypothesis.

### C. Comparing the Likelihood of Observations in $H_0$ and $H_1$

In order to conclude on both hypotheses, and to compare the occurrence of observations in both of them, we must estimate the likelihood of  $H_1$ . Thanks to its simplicity, assessing the likelihood of  $H_0$  given our observations was simple; it is not the case for  $H_1$ , as one has to handle the exact impact of the topology on neighbor contaminations.

To circumvent this difficulty, we again resort to numerical simulations. On each of the 9,967 ego-graph topologies, we simulate 10,000  $H_1$  model infections to estimate the resulting number of contaminated nodes. More precisely, for each ego-graph  $G_l$ , let  $S_l$  be the random variable representing the number of banned nodes obtained by simulating  $H_1$  of  $G_l$ , and let  $\hat{s}_l = |\{i \in SB, \forall i \in V(G_l)\}|$  the observed number of banned users in  $G_l$ . By simulation, we experimentally sample the probability density function of  $S_l$  and retain the probability of having exactly  $\hat{s}_l$  observations:  $P(S_l = \hat{s}_l|H_1) = \mathcal{L}(H_1|G_l)$ , which is likelihood of model  $H_1$  on  $G_l$ .

Figure 7 reports these results. To compare the likelihood of  $H_1$  and  $H_0$ , we bin the likelihoods into classes of probability occurrences for both hypotheses. This allows for a fair comparison of  $H_1$  with  $H_0$ : because we resort to numerical evaluation of  $\mathcal{L}(H_1|G_l)$ , we cannot estimate by sampling the very low likelihoods (e.g.,  $L < 1e - 4$ ).

Results show that likely observations under  $H_1$  occur 2.68 times more. Conversely, unlikely observations occur 5.35 times more in  $H_0$  than in  $H_1$ . This stresses that the  $H_1$  hypothesis manages to capture a part of what is at stake in the shadow banning process in Twitter: the topology of ego-graphs, that is the interactions of users, is at play.

To conclude, we have seen that hypothesis  $H_0$  is unlikely. We propose an alternative hypothesis  $H_1$ , that captures the locality of ban observations with respect to the interaction graph. This model is substantially more likely than  $H_0$ , revealing the tight relation of shadow banning with respect to locality. In other words, bans appear as clusters in certain areas typically referred to as communities in the context of Twitter.

## VI. RELATED WORK

*a) Moderation in online social networks:* Moderation of user contributions dates back to early Internet forums such as USENET [23]: moderation was then defined with a parallel to "professional journal editors" on users contributions. This practice is now widely spread in modern platforms under many different shades: through a survey of 519 users who have experienced content moderation, West. Authors in [6] explore users' folk theories of how content moderation systems work.

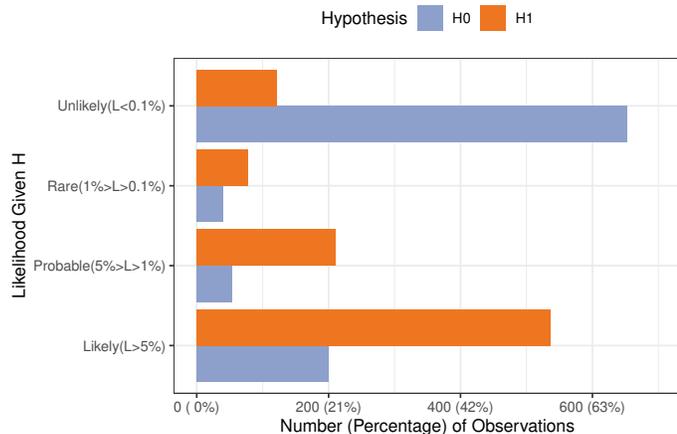


Fig. 7: Likelihood for observing the shadow ban statuses in the RANDOM population, binned by plausibility, under both the  $H_0$  and  $H_1$  hypotheses. The  $H_1$  hypothesis accounts for much more likely events than the  $H_0$  hypothesis, and much less unlikely events. This underlines that the topology of the interaction ego-graphs correlates with in the shadow banning phenomenon.

Another user-sided approach to try to infer properties of moderation in a major OSN is conducted in a recent short paper [9]; authors are asking the question of a potential political bias in the moderation of YouTube comment, and forms two null hypothesis to examine it. They leverage message content; their conclusion is that bias is supported by one hypothesis, but not by the second. Looking at moderation under another aspect, Dosono et al. [24] study how moderators shape communities on Reddit.

The question of automatic moderation is of interest for operators to sustain the mass of user-produced information now available [25] (here leveraging the semantics of usage and content).

Our work differs from studies on moderation, since Twitter denies the use of shadow banning as described in Section II, so we had to develop hypotheses to check their plausibility.

*b) Black-box observation of a remote service:* There is a growing literature interested in the means to extract or infer properties from a remotely executing algorithm, from a user standpoint. In the context of online ads, the XRay [11] approach proposes a Bayesian angle for inferring which data of a user profile, given as an input, is associated to a personalized ad to that user. Authors propose in [10] a graph theoretic approach to retrieve which *centrality* metrics are being used by peer-ranking platforms. Work in [26] shows that machine learning models can be extracted by a user from the cloud platform where they are executed, in order to leverage the leaked model to issue new predictions. Reference [27] observes online auction systems. In the domain of recommender systems, paper [28] exposes the users perspective on what they expect from recommendation.

*c) Observation and statistics in Twitter:* The specific case of Twitter was consistently studied for multiple research leads, including in the INFOCOM community [13], [29]. Recently, Gilani et al. [15] study the behavioral characteristics of both populations of bots and humans in Twitter. Twitter data is often represented as graphs in order to extract relevant information, such as relationship structures [13], the “follower” mechanism [17], general dynamics [29], or influencers [30].

## VII. CONCLUSION

Allegations of shadow banning practices have been countless in the media and the population in the recent years. Yet, no objective approach ever quantified this practice. We proposed in this paper to remedy this lack, by observing at large scale shadow banning practices on a major online social network. We then presented statistical approaches leveraging the collected dataset to shed light on the phenomenon.

First, we explored public Twitter-user features to seek a relation between these features and ban statuses. Then, through two statistical modeling hypotheses, we compared the likelihood of two narratives commonly encountered around shadow ban questions. Our conclusions indicate that bans appear as a local event, impacting specific users and their close interaction partners, rather than resembling a (uniform) random event such as a bug.

As of future work, we believe one crucial notion to be analyzed is the temporal dimension of the shadow banning phenomenon: *e.g.*, how does a shadow ban status evolves among neighbors? Can the beginning of a ban be correlated in time with other observables? Is the appearance of the shadow ban statuses mostly happening in batch, or is the propagation smooth among the monitored user profiles? Another important observation to be conducted is the possible reversibility of this status: can we observe user profiles retrieving their initial visibility (*i.e.*, losing the shadow ban status they had), after they for instance interacted less with shadow banned users? Lastly, we have chosen to use both a statistical and topological approach in our study; there are probably several other interesting approaches to address shadow banning under other angles, for instance at the semantic level by analyzing the contents of the messages. We think these other interesting dimensions to be of great interest for scientists, algorithm designers and the general public.

## VIII. DATA AND CODE AVAILABILITY STATEMENT

We release an anonymized version of the dataset we gathered for this study, as well as the code for our core experiments, at the following location: <https://gitlab.enseeiht.fr/bmorgan/infocom-2021>.

## IX. ACKNOWLEDGEMENTS

We thank the *shadowban.eu* initiative, providing tests and code to individuals for spotting shadow banning practices.

## REFERENCES

- [1] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *AAAI Conference on Web and Social Media*, 2018.
- [2] "Twitter says supposed 'shadow ban' of prominent republicans is a bug," <https://www.engadget.com/2018/07/26/twitter-says-republican-shadow-ban-is-a-bug/>, accessed: 2019-12-01.
- [3] "A leaked excerpt of tiktok moderation rules shows how political content gets buried," <https://www.technologyreview.com/2019/11/25/102440/tiktok-content-moderation-politics-protest-netpolitik/>, accessed: 2020-07-01.
- [4] "The tiger mom tax: Asians are nearly twice as likely to get a higher price from princeton review," <https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review>, accessed: 2020-08-01.
- [5] "[whosban.eu.org] who is banned ?" <https://whosban.eu.org/>, accessed: 2020-01-12.
- [6] S. M. West, "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms," *New Media & Society*, vol. 20, no. 11, pp. 4366–4383, 2018.
- [7] C. Middlebrook, "The grey area: Instagram, shadowbanning, and the erasure of marginalized communities," in *SSRN*, February 2020.
- [8] "Setting the record straight on shadow banning," [https://blog.twitter.com/official/en\\_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html](https://blog.twitter.com/official/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html), accessed: 2019-12-30.
- [9] S. Jiang, R. E. Robertson, and C. Wilson, "Reasoning about political bias in content moderation," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*, February 2020.
- [10] E. Le Merrer and G. Trédan, "Uncovering influence cookbooks: Reverse engineering the topological impact in peer ranking services," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1413–1418. [Online]. Available: <https://doi.org/10.1145/2998181.2998257>
- [11] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "Xray: Enhancing the web's transparency with differential correlation," ser. USENIX Security Symposium, 2014.
- [12] "Is @username shadowbanned on twitter?" <https://shadowban.eu/>, accessed: 2019-12-01.
- [13] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni, "Ego networks in twitter: An experimental analysis," in *2013 Proceedings IEEE INFOCOM*, 2013, pp. 3459–3464.
- [14] D. Ediger, K. Jiang, J. Riedy, D. A. Bader, C. Corley, R. Farber, and W. N. Reynolds, "Massive social network analysis: Mining twitter for social good," in *2010 39th International Conference on Parallel Processing*, 2010, pp. 583–593.
- [15] Z. Gilani, R. Farahbakhsh, G. Tyson, and J. Crowcroft, "A large-scale behavioural analysis of bots and humans on twitter," *ACM Trans. Web*, vol. 13, no. 1, Feb. 2019.
- [22] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on twitter," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, ser. SNAKDD '13. New York, NY, USA: Association for Computing Machinery, 2013.
- [16] "Nos députés - observatoire citoyen de l'activité parlementaire à l'assemblée nationale," <https://www.nosdeputes.fr/>, accessed: 2019-12-30.
- [17] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network? the structure of the twitter follow graph," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14 Companion. New York, NY, USA: Association for Computing Machinery, 2014, p. 493–498.
- [18] C. Molnar, *Interpretable Machine Learning*, 2019, <https://christophm.github.io/interpretable-ml-book/>.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [20] J. B. Dugan, S. J. Bavuso, and M. A. Boyd, "Dynamic fault-tree models for fault-tolerant computer systems," *IEEE Transactions on Reliability*, vol. 41, no. 3, pp. 363–377, 1992.
- [21] W. O. Kermack and A. McKendrick, "A contribution to the mathematical theory of epidemic," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 1927.
- [23] Greening and Wexelblat, "Experiences with cooperative moderation of a usenet newsgroup," in *ACM/IEEE Workshop on Applied Computing*, 1989.
- [24] B. Dosono and B. Semaan, "Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–13.
- [25] S. A. Ríos, R. A. Silva, and F. Aguilera, "A dissimilarity measure for automate moderation in online social networks," in *Proceedings of the 4th International Workshop on Web Intelligence & Communities*, ser. WI&C '12. New York, NY, USA: Association for Computing Machinery, 2012.
- [26] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *USENIX Security Symposium*, 2016.
- [27] Z. Feng, O. Schrijvers, and E. Sodomka, "Online learning for measuring incentive compatibility in ad auctions?" in *WWW*, 2019.
- [28] S. Sinha, K. S. Rashmi, and R. Sinha, "Beyond algorithms: An hci perspective on recommender systems," 2001.
- [29] Y. Zhang, X. Ruan, H. Wang, and H. Wang, "What scale of audience a campaign can reach in what price on twitter?" in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014, pp. 1168–1176.
- [30] Z. Zengin Alp and Şule Gündüz Ögüdücü, "Identifying topical influencers on twitter based on user behavior and network topology," *Knowledge-Based Systems*, vol. 141, pp. 211 – 221, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0950705117305439>