Bug ou ban ? Une Perspective Topologique sur le Shadow Banning

Erwan Le Merrer, Benoît Morgan, Gilles Trédan

Bug ou ban ? Une Perspective Topologique sur le Shadow Banning

Erwan Le Merrer[1] et Benoit Morgan[2] et Gilles Trédan[3]

[1] *Univ Rennes, Inria, CNRS, Irisa. erwan.le-merrer@inria.fr* [2] *IRIT/ENSHEEIT benoit.morgan@irit.fr* [3] *LAAS/CNRS gtredan@laas.fr*

Le *shadow banning* (SB) consiste pour un réseau social à limiter la visibilité de certains utilisateurs, sans que ceux ci ne s'en rendent compte. Pour cet article, nous prélevons 200 utilisateurs sur Twitter, ainsi que leur égo-graphe de discussion. Nous montrons tout d'abord que statistiquement les deux populations de ces utilisateurs (députés ou utilisateurs pris aléatoirement) sont affectées différement par le SB. À l'aide d'un modèle de propagation épidémique sur la topologie des égo-graphes, nous montrons ensuite une corrélation avec les cas observés de SB. Ceci met à mal l'hypothèse de bugs aléatoires et met au jour un possible aspect topologique (*i.e.*, relationnel) du problème.

Mots-clefs : Graphes de terrain, shadow banning, interactions en boite-noire, statistique.

Shadow banning (abbreviated SB hereafter, and also known as *stealth banning*) is an online moderation technique used to ostracise undesired users. In modern platforms such as Twitter, Facebook or Instagram, SB would refer to a wide range of techniques that artificially limit the visibility of targeted users or users posts. However, while platforms publicly acknowledge the use of automatic moderation, they deny the use of such practices. Recent press coverage widely relayed multiple occurrences of this debate. Twitter stated "To be clear, our behavioral ranking doesn't make judgments based on political views or the substance of Tweets" [3]. Alternatively, some problems of that sort were presented as bugs and declared as patched [2]. For an external observer of a decision-making algorithm, such as a user, SB is by definition difficult to assess. It therefore can easily be justified as a bug, or as a moderation strategy. On the other hand, because of the polemic nature of the subject, discussions should be based on solid evidence.

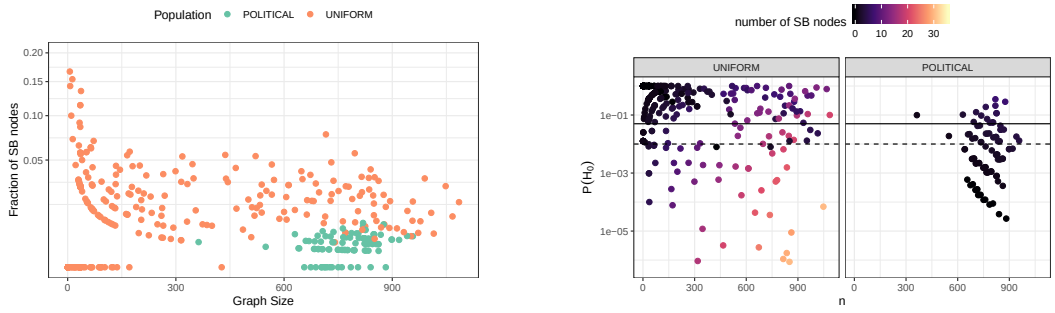
To address the question of the plausibility of SB in Twitter, we exploit a technique allowing us to detect users or tweets with diminished visibility (which we hereafter define as our reference observable of SB). We pursue a topological perspective on SB, by statistically opposing two hypotheses. H_0 : ***SB users are uniformly spread among Twitter users***, and H_1 : ***relation graph topologies in Twitter explain the SB effect***.

1 Problem modeling and data collection

Shadow Banning on Twitter. In the context of Twitter, the term of *shadow banning* can describe a handful of situations where the visibility of a SB user or her posts is reduced compared to normal visibility. The social networks provide numerous examples of individuals claiming they are SB, sometimes exhibiting a screenshot as a proof to back their claim. Yet, few information is available regarding how to actually test for SB. One of the first page to provide users with the ability to check whether they are SB is *shadowban.eu* †. Moreover, its authors provided explanations along the code ; we based our approaches on these. Starting from this code, we developed our own tests for the following potential SB methods we witnessed on Twitter. *i/ Suggestion Ban :* Users targeted by the suggestion ban are never suggested, as a user perform searches or mentions another user in some content. This limits the possibility for users to accidentally reach a profile. *ii/ Search Ban :* Users are never shown in searches, even if their exact user name is looked for. *iii/ 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.

Note that we declare a user to be SB if at least one of these ban actions holds.

†. `https://shadowban.eu/`



(a) (b)

FIGURE 1: (a) Fraction of SB nodes found in each ego-graph, as a function of graph size. Colors encode the belonging to the POLITICAL or UNIFORM population. (b) The p-value of the H_0 (bug) hypothesis for each landmark. Dashed and continuous lines respectively represent the 1 and 5% significance levels.

Sampling ego-graphs for data collection. Having the code to assess whether an individual profile is being SB or not, we can now describe our data collection campaign. As all Twitter's users obviously cannot be tested for SB (Twitter in Q1 2019 reported 330 million monthly users ‡), we resorted to sampling *ego-graphs*. 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. (Note that this graph might differ from the explicit Twitter graph in which edges capture the "follower/friend" relationship). We start by selecting *landmarks* from two populations of users : deputies at the French parliament (noted POLITICAL), and users selected at random in Twitter (noted UNIFORM). From each of these landmarks l, we conduct a depth-limited Breadth-First-Search : we parse the 33 most recent tweets returned by Twitter, and list the set of users $V_{out}(l)$ with which l interacted. We then repeat that procedure for each $i \in V_{out}(l)$, to discover the two-hop neighbors of landmark $l, V^2_{out}(l)$, and then again for the 3-hop neighborhood of l, $V^3_{out}(l)$. 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 nodes $V_l = \Cup_{i=1,2,3} V^i_{out}(l)$.

More precisely regarding the two populations : *1)* in the UNIFORM case, we exploit a property of the Twitter API that associates to each user a user ID randomly drawn in a finite subset of $\mathbb{N}$. We uniformly sample user IDs in the range $[1, 2^{32}-1]$ (that was the identification pattern up to late 2015), in order to constitute a set of landmarks. *2)* For the POLITICAL population, we select landmarks that use Twitter in a political context. To achieve this, we select as landmark candidates all of the 577 French MEPs that have an official Twitter account [1]. These two different landmark sets $L = \{L_{Unif}, L_{Pol}\}$ constitute the seeds from which we will recursively sample the Twitter interaction graph, one landmark at a time. We run our set of SB tests for each visited profile on all the tweets still available (1000 last tweets at most). We kept 100 randomly selected ego-graphs for each population, ensuring that they consist of at least 2 nodes each. We make all graphs undirected.

We report the following statistics for the two populations. For POLITICAL, we find 389 SB users among 76,497 in the 100 ego-graphs (μ_p =0.50%). Graphs are of average size 764.97, have a clustering coefficient on average of 0.243, and an average node degree of 11.9. For UNIFORM, we find 520 SB users among 28,958 in the 100 ego-graphs (μ_u =1.79%). Average graph size is 289.58, clustering of 0.27, and node degree of 4.19.

2 Analysis and Empirical Results

2.1 Hypothesis H_0 : the plausibility of bugs

We recall hypothesis H_0 : *SB nodes are uniformly distributed among Twitter users.* In this hypothesis, each node is SB with probability $\mu = \frac{\mu_p + \mu_u}{2}$. This hypothesis embodies the *bug* explanation : bugs should be random. As H_0 completely ignores topological dimension of collected data, what remains is a ball and bins sampling effect. We can easily assess the probability of observing our data under H_0.

‡. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Figure 1a plots the ratio of SB nodes in both populations as a function of each ego-graph size. A salient observation is the higher fraction of SB nodes in the UNIFORM population, specially for smaller size ego-graphs. POLITICAL ego-graphs contain consistently fewer SB nodes. The average fraction of SB nodes over both populations (0.509% for POLITICAL vs 1.796% for UNIFORM) also hints for an uneven SB distribution.

Figure 1b gives a sharper view on this hypothesis, by plotting the p-value of H_0, with regards to the size of each ego-graph, the number of SB nodes it contains, and for both populations. Recall that the lower the p-value, the higher the plausible rejection of the hypothesis under scrutiny. For the UNIFORM on the left-hand side, we observe an important amount of graphs that are significantly below both significance levels. Moreover the number of SB nodes in each ego-graph (represented by point color) hints two types of unlikely ego-graphs (from H_0 perspective) : graphs with too few SB (black dots) and graphs with too many SB (pink dots). For the POLITICAL on the right-hand side, we note that the majority of ego-graphs lies below the confidence levels, leading to the same conclusion.

From those two Figures, we may safely conclude that the H_0 hypothesis can be rejected [§]. This conclusion calls for an alternative model H_1.

2.2 Hypothesis H_1 : Influence of the Topology

We have seen in Figure 1a that our SB observations cannot be explained by a uniform spread. In other words, SB nodes are locally concentrated in some regions of the Twitter graph. We here explore an approach to capture this relation to the topology. To do so, we propose to adapt a simple Susceptible/Infected (SI) epidemic model [5]. This model was widely used to describe different topologically related phenomenons, like infections or rumour spreading. While SB is arguably a different phenomenon, the SI model is perhaps the simplest way to capture the locality of graph observables. This captures the intuition that some communities of users are impacted by SB actions, and where users potentially propagate SB due to their behavior.

Our SI model is a simple 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 β. In other words, β captures the locality of the phenomenon, while p_0 allows to uniformly spread the disease. Let $SI(p_0, \beta)$ be this process.

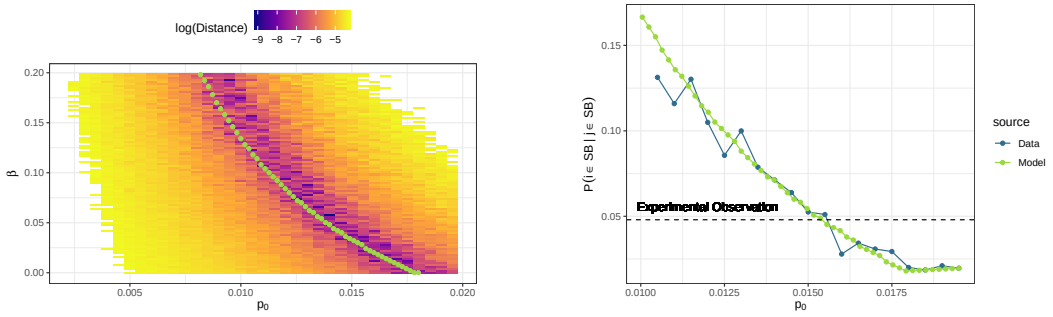
A simple analytical model under SI to explain SB. First, observe that $SI(p_0 = \mu, \beta = 0) = H_0(\mu) = H_0$ where μ is the fraction of SB nodes in the system $\mu = |SB|/n$: neutralising contamination yields the random uniform SB of nodes. As β increases, local contaminations surround each initially infected node, and p_0 must be adjusted to fit the overall number of SB nodes. Let $H_1(\beta) = SI(p_0, \beta)$ s.t. $\mathbb{P}(SB|H_1(\beta)) = \mu$.

A back-of-envelope estimation of the relation between μ, β and p_0 can be established as follows : $\mathbb{P}(SB|H_1(\beta)) \approx \mathbb{P}(\text{infected initially}) \oplus \mathbb{P}(\text{contaminated}) = p_0 + (1 - p_0)p_1$. 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 neutralises topological artifacts (clustering, degree heterogeneity) to sketch a relation between p_0 and β for a fixed μ.

We now focus on the UNIFORM population for our experiments. Figure 2a represents the quantity $|\mathbb{P}(SB|SI(p_0, \beta)) - \mu|$ for varying p_0 and β. It is obtained by simulating each SI model on the extracted Twitter ego-graphs. In other words that is the difference between actual fractions of SB nodes, and the ones simulated with SI. A distance of 0 thus indicates that the SI simulation lead to the same amount of SB nodes that the actual one.

The green line represents the optimal values numerically obtained using the analytical model, while the colored squares indicate the SI simulation, the darker the smaller the distance. The green line follows closely the lowest experimental values that shape a valley, indicating that the toy model captures the process rather well. The agreement decreases as β increases, probably as a consequence of the clustering neglected in the analytical model.

§. Combining p-values of independent trials is a big debate ; a harmonic mean of these p-values yields a probability of 10^{-80} to be wrong rejecting H_0.



(a) (b)

FIGURE 2: (a) The impact of p_0 and β SI parameters on the distance of simulated SB user w.r.t. to the actual SB 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 ($k = 5$).

As a consequence, the lowest spots of the valley and the green line both define here a family of hypotheses $H_1(\beta)$ in which all members approximate the total number of SB nodes as closely as the uniform contamination $H_0 = H_1(0)$. A natural follow-up question is "what would be a good value for β" ?

Recall that β is the contamination probability, which is a local property. To estimate a good value, one can look at the probability that a SB node has a SB neighbor : $\mathbb{P}(j \in SB | i \in SB \wedge (i,j) \in E)$. While in H_0 this probability is μ, 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 (and chances that two nodes contaminated by the same node are neighbors).

Figure 2b represents this probability for $H_1(\beta)$ (estimated using simulation and this model). As expected, as p_0 decreases, β increases, which in turn increases neighboring contamination chances. The dashed line represents the empirical value $|(SB \times SB) \cap E| / |(SB \times V) \cap E|$. The model closest to this experimental line is $H_1(\beta = 0.0341)$ corresponding to the SI model where $p_0 = 0.0154$. This model would explain both the global number of SB nodes, and the local co-occurences of SB in the data by using a 3.41% chances of contamination (almost twice the initial contamination probability). This experiment maps well the actual topology of ego-graphs and the number of SB nodes, thus highlighting a possible connection between these.

3 Conclusion

In summary, we have shown that H_0 is very unlikely, and defined an alternative hypothesis H_1 that would better explain local co-occurences of SB : the observed SB feature is likely not a uniform bug. While decision-making algorithms are increasingly deployed online, we believe it is important to develop techniques and frameworks to observe them from a user perspective. Some recent works for instance proposed to observe online auction systems [4], or advertisement platforms [6] from this perspective. To the best of our knowledge, this paper is the first to question the problem of shadow banning. Limitations of our work include a restricted set of ego-graphs and populations, and a current lack of understanding of how deep/far one analyst must crawl information to maximize the likelihood of a sound analysis in this setting. Future-work also includes the analysis of the temporal dimension of the shadow banning process : can we prove it is spreading among neighbors, for instance due to debates around sensitive topics ? We believe there are many more dimensions to be of general interest for algorithm designers and the public.

Références

[1] Observatoire citoyen de l'activité parlementaire à l'assemblée nationale. `https://www.nosdeputes.fr/`. Accessed : 2019-12-30.

[2] 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`. Accessed : 2019-12-30.

[3] What is a 'shadow ban,' and is twitter doing it to republican accounts ? `https://www.nytimes.com/2018/07/26/us/politics/twitter-shadowbanning.html`. Accessed : 2019-12-30.

[4] Z. Feng, O. Schrijvers, and E. Sodomka. Online learning for measuring incentive compatibility in ad auctions ? In *WWW*, 2019.

[5] 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.

[6] 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. USENIX Security Symposium, 2014.