TrollSpot: Detecting misbehavior in commenting platforms

Tai Ching Li*, Joobin Gharibshah*, Evangelos E. Papalexakis* and Michalis Faloutsos*
* University of California - Riverside
900 University Ave, Riverside, California 92557
Email: {tli010,jghar002,epapalex,faloutsos}@cs.ucr.edu

***Abstract*—Commenting platforms, such as Disqus, have emerged as a major online communication platform with millions of users and posts. Their popularity has also attracted parasitic and malicious behaviors, such as trolling and spamming. There has been relatively little research on modeling and safeguarding these platforms. As our key contribution, we develop a systematic approach to detect malicious users on commenting platforms. Our work provides two key novelties: (a) we provide a fine-grained classification of malicious behaviors, and (b) we use a comprehensive set of 73 features that span four dimensions of information. We use 7 million comments during a 9 month period, and we show that our classification methods can distinguish between benign, and malicious roles (spammers, trollers, and fanatics) with a 0.904 AUC. Our work is a solid step towards ensuring that commenting platforms are a safe and pleasant medium for the exchange of ideas.**

I. Introduction

Any successful medium eventually attracts abusive behaviors, and commenting platforms is no exception. Over the last decade, commenting on news articles has emerged as a new form of highly social interaction. First, a small number of companies facilitate the backend management of comments for many of websites. We use the term **commenting platform** to refer to such platforms, including Disqus [1], LiveFyre [2], and IntenseDebate [3]. Second, commenting is an intense activity for many users, who spend many hours daily at it.

We list a set of definitions that we use in this paper. A **user** is defined by a platform account, which enables her to leave comments to articles on websites supported by that platform. A user may leave more than one comment for an article, which leads us to define the **engagement** of a user for that article. An engagement has a time duration and intensity in terms of number of comments. When two users comment on the same article, we say that they **collaborate** and we use the term **collaboration** to describe this activity. We use the term **collaboration intensity** to refer to the number of articles for which two users collaborate.

The key question in our work is: *Can we automatically detect malicious users in these commenting platform?* Specifically, the input to the problem is the commenting information of the users. This includes: the author of the comment, the time it was posted, and information on the article. The goal is to identify malicious behaviors and users. Detecting abusive behaviors is critical for ensuring that these platforms continue to enable the honest and safe exchange of opinions.

Commenting platforms have attracted little attention so far with only few exceptions [4] [5]. Most work on modeling and misbehavior detection focuses on Online Social Networks (OSNs), and blogs. The key related areas include: (a) detecting abusive behaviors and malware propagation in OSNs [6] [7]; (b) modeling online user behavior [8] [9] [10] [11]; and (c) analyzing text of online users [12] [13] [14]. Due to space, we defer a survey to the long version of the work.

We propose a systematic comprehensive methodology to identify malicious users on commenting platforms to enable: (a) interpretable, and (b) fine-grained classification of malicious behavior. We claim the following key novelties.

a. A behavior-based classification. We propose two classification methods, one of which introduces a two-stage classification approach. In this method, we map: (a) observable features into to behaviors, and (b) behaviors into user roles, using unsupervised and supervised learning respectively.

b. A comprehensive multidimensional feature set. We combine 73 features from four different dimensions of user interactions: (a) social interaction or user-user interaction, (b) engagement or user-article interaction, (c) temporal features, and (d) linguistic features.

c. Fine-grained malicious role identification. Our approach goes beyond a good versus bad determination to a more fined-grained classification of misbehaving roles. Here, we focus on three roles: (a) spammers, (b) trolls, and (c) fanatics, which are defined in the next section. However, it is easy to introduce more roles as long as appropriate ground truth is available.

Promising classification results. Our study is grounded on nearly 7 million comments from nearly 200K users over 9 from Discus, which is arguably the largest commenting platform. Our method identifies **misbehaving users with 0.904 AUC** and it outperforms the previously-proposed baseline method. In addition, our method provides **role classification with 80.8% overall accuracy.**

This work was supported by Bourns College of Engineering at University of California - Riverside, NSF NeTS 1518878, NSF SaTC 1314935 and DHS ST CS (DDoSD) HSHQDC-14-R-B00017 grants.

II. Data collection and definitions

Data Collection. We collected data from Disqus through its Application Programming Interface (API). Using the API, we collect data from four popular websites: (a) CNBC News, (b) ABC News, (c) Bloomberg Views and (d) Breaking News - a Disqus channel. The first three are well-known news websites and the last one is the most popular channel on Disqus. A channel is similar to a news-feed, whose articles are selected by the users that participate in that channel. We collect all

comments posted at articles published on these 4 sources in between Nov 1st 2015 and July 31st 2016. The dataset consists of: (a) 286,275 articles, (b) 6,994,693 comments and (c) 201,112 unique users, (d) 1,705,667 engagements.

Roles of misbehaving users and ground truth. We identify and attempt to detect three different roles of misbehaving users: trolls, spammers and fanatics. These are inherently difficult to define, so we resort to human feedback. We define three malicious roles below and show here the definitions that we gave our evaluators, as we explain in Section IV: **(a) Trolls:** Users who make inflammatory or inappropriate comments for the sole purpose of upsetting other users and provoking a response; **(b) Spammers:** Users who repeatedly make similar comments in the same or multiple articles; and **(c) Fanatics:** Users who exhibits an extreme and uncritical enthusiasm in religion or politics.

III. Features and User Behavior

In this section, We study the behavior of users along four dimensions: **(a)** engagement behavior (user-article interaction), **(b)** social behavior (user-user interaction), **(c)** temporal behavior, and **(d)** linguistic properties. The goal is to identify meaningful features that can help us detect misbehavior. In Table I, we outline the features that we use in Section IV.

A. Engagement behavior

We quantify the engagement (user - article interaction) with 7 different features. Six of them are derived from two major properties: The **engagement duration** is the time interval between the first comment and the last comment user makes on the article. If the user leaves only one comment, we consider this as zero length interval. The **engagement intensity** is the total number of comments user makes on the article.

Engagement duration: 90% last for less than 10 hours, but some can be as long as half an year. In Figure 1(a), we plot the Cumulative Distribution Function (CDF) of the duration (top x-axis) for all 1.7M engagements. We find that 90% of engagements last less than 10 hours and have less than 7 comments. Interestingly, we find 106 engagements which last for more than half year!

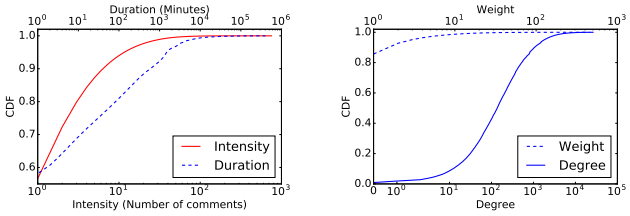
Engagement intensity: 90% have less than 7 comments, but 0.06% have more than 100 comments. In Figure 1(a), we plot the CDF of the intensity (bottom x-axis) for all engagements. We find that 90% of them have less than 7 comments, while 1,151 (0.06%) of them have >100 comments.

B. Social Behavior

We propose 17 features to model the social interaction of the users, which we define as commenting at the same articles.

Single-article collaboration Threshold: θ comments. We say that two users collaborate in one article, if they each post at least θ comments on that article. For $\theta = 1$, the graph become very dense, and the analysis is both slow and less informative. In the remaining of this work, we use a threshold $\theta = 2$.

User-user collaboration intensity and threshold: λ articles. The collaboration intensity is the number of articles



(a) (b)

Figure 1. (a) **Engagement behavior:** distribution of intensity and duration of engagements. (b) **Social behavior:** the CDF of the user degrees (bottom x-axis) and the edge weight distributions (top x-axis)

that two users collaborate for a given threshold θ. To study collaborations at difference levels of intensity, we introduce the **collaboration intensity threshold** λ, which we use below.

We define the undirected weighted **collaboration graph** $G_\lambda = \langle V_\lambda, E_\lambda \rangle$ of collaboration intensity λ where:

1) V_λ is a set nodes v, representing users.
2) E_λ is the set of edges, where edge e_{ij} between nodes v_i and v_j exists, if and only if the collaboration intensity of the users exceeds the **threshold of λ articles**. The edge weight $w(e_{ij})$ is set to the collaboration intensity.

The collaboration graph (for $\theta = 2$ and $\lambda = 0$) has 95,527 users and roughly 21 millions edges, an average degree of 440.7 and a median degree of 137. Note that we do not include users with zero degree in this graph.

In Figure 1(b), we plot the CDF of the user degrees (bottom x-axis) and the edge weight distributions (top x-axis). We see that 90% of the users have degree lower than 1,054, the max degree goes to 26,318: this user collaborates with more than 27% of users in the graph! The figure also shows that 90% of the edges have weights less than 2. Although it is not easy to gauge from the plot, we find 12,374 edges with weight over 64, 1,779 edges with weight over 128 and 65 edges with weight over 256. In other words, there are 65 pairs of users who have collaborated on more than 256 articles.

Capturing the collaboration groups: triangles and cliques. We quantify the local connectivity of the users using the number of maximal cliques (size bigger than 3) and triangles in which a user participates. Both these metrics capture how densely connected the neighbors of that user are. In G_{128}, we find that 50% of users have less than 27 triangles on the neighborhood, 90% of have less than 376, but there are 1% with more than 868 triangles. We also find that 90% of users have less than 51 maximal cliques and there are 8 users, who participate in more than 186 cliques which is more than 50% of all maximal cliques in the graph. These highly collaborative users are suspicious and this encouraged us to consider both these features for misbehavior detection.

C. Temporal behavior

We quantify the temporal behavior of users with 25 features, but we only highlight some key observations here.

Most users exhibit persistent behavior: daily and weekly. Figure 2(a) shows the number of comments in each hour-of-day and day-of-week as a heat-map for all users, but

Table I
THE OVERVIEW OF THE 73 FEATURES WE USE PER DIMENSION

Dimension	Features	Count
Engagement	number of engagements, engagement duration*, engagement intensity*	7
Social	degrees, number of maximal cliques, number of triangles (in different level of collaboration intensity)	17
Temporal	number of comments made in 24-hour slots, highly-active hour	25
Linguistic	number of words*, number of sentences*, percentage of capital letters*, readability metrics, number of URLs	24

* We use several statistical versions (mean, maximum and minimum) of the feature per engagements, user or comment.

individual users exhibit similar daily and weekly behavior. The plot shows that Disqus users post most of their comments during the common work-hours during the week, a pattern also observed in other OSNs, like Facebook [15] and Twitter [16].

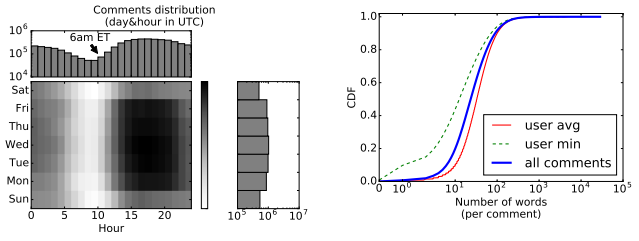
Highly-active hours is less than 4 hours for 96.7% of the users. We define **highly-active hours** to be the minimum number of hours, during which the user makes more than 50% of their total comments during a day on average. We find that 96.7% of users have highly-active hours less than four hours. Interestingly, we find 368 users who have more than 8 highly-active hours, a significantly wider spread. Upon inspection, many of these users have non-trivial activity over 14 hours in a day. This wide range of behaviors suggests that highly-active hours can be a useful metric for our classification.

D. Linguistic properties

We identify and study 24 linguistic features from the text of the posts, but we have space for only two metrics here.

Only 1.9% of comments contain URLs. We naturally consider the existence of URLs in a comment as a feature in our classification: their presence can indicate ad-oriented spamming. We find that only 1.9% of all comments contain one or more URLs. We also find that only 1.7% of the users have ever posted more than 3 comments containing URLs. In fact, our interaction with the data suggests that often users will use URLs as references in support of their opinions.

Lengthy comments are more likely to be spam. In Figure 2(b), we plot the distribution of the number of words in a comment. We see that 91% of comments have less than 100 words, while roughly 20% of the comments with less than 10 words. Interestingly, we also find 14,838 (0.002%) comments with more than 500 words. We examine theses lengthy comments and find out that 47% of them are verbatim copies of at least one other comment from the same user which is an indication of spamming. This suggests that comment length is a helpful feature in detecting misbehavior.



(a) (b)

Figure 2. (a) **Temporal behavior:** Time and day of the week plot. (b) **Linguistic property:** Number of words in a comment.

IV. FEATURE-BASED MISBEHAVIOR IDENTIFICATION

We propose a method to identify misbehaving users using the features, which we outline in Table I.

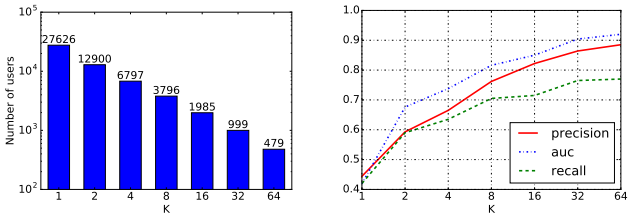
A. Establishing the ground truth. We rely on "proxy signals" and we use the community's own opinion: any user can report (a.k.a. flag) a comment as "inappropriate". We explain how we use this community feedback to construct the ground truth.

Ground-truth: Reportings per malicious comment. To increase our confidence, we set a minimum threshold of reports, ϕ, that a comment must have to labeled malicious. The rationale is that a single reporting can be created even accidentally (the authors have regrettably done this once). After analysis and deliberation omitted here, we settled on $\phi = 3$ reportings. We use the term **reported comment** to refer to a comment with more than ϕ reports.

Ground-truth: Reported comments per malicious user. In the same vain, we want to be careful in labeling a user as malicious based on the number of reported comments. We use the **reported comments threshold**, r, to control tune the definition of malicious user and consider users with zero reported comments as **benign** for the purpose of establishing the ground truth.

Building the ground truth datasets: D_r. We create a set of labeled datasets as follows. First, we distinguish reported users into groups, $R_{r(i)}$, with $r(i) = 2^i, for \quad i = 0, 1, ...$ A user is in group $R_{r(i)}$, if the number of her reported comments are greater or equal $r(i)$. It turns out that no user has more than 128 reported comments. The numbers of users in each group with threshold $r(i)$ are shown in Figure 3(a). Second, we create datasets $D_{r(i)}$ by randomly selecting 200 reported users from $R_{r(i)}$, and combining them with 200 benign users (zero reported comment).

B. The benign-malicious classification. For the classification, we use the Random Forest classifier provided by Weka, which gave the best results among many that we tried. We perform ten-fold cross validation and report the precision, recall, and **ROC curve (AUC)** of each dataset in Figure 3(b).



(a) (b)

Figure 3. (a) Number of users having k reported comments. (b) Classification results as a function of the number of reported comments that "incriminate" a user.

The plot shows that our features can identify reported users with more than 80% precision when the threshold $r > 16$.

Selecting reported comments threshold $r = 32$. We manually examined reported users in D_{16}, D_{32} and D_{64} by sampling 20 users from each group. We find that users with more than 16 reported comments exhibit a persistent misbehavior throughout their lifetime. The 40 reported users sampled from D_{32} and D_{64} are 100% labeled as misbehaving users by our independent human evaluators. Thus, we consider users with 32 or more reported comments as misbehaving users and, we use dataset D_{32} as reference below.

The accuracy of our method exhibits 90% AUC. We adapt and use a previously proposed algorithm [4] as the baseline in our study. Our method performs better than the baseline classifier (90% vs 72% **AUC**), with the caveat that both approaches are restricted to the publicly available features. We conclude that our features have good discriminatory power in classifying misbehaving and benign users.

C. The fine-grained malicious role classification. One of our key novelties is identifying misbehaving roles.

Ground truth for misbehaving roles. We resort to manual labeling to create our reference data: we examine 200 misbehaving users in D_{32} and categorize them into three different roles. Each user is labeled by three evaluators, and we use the majority rule in non-unanimous cases, to obtain: 104 trolls, 21 spammers and 75 fanatics.

Role classification: 80.8% overall accuracy. We apply 10-fold cross validation with the same classifier and the same 73 features that we used before. Our method can effectively classify the role of misbehaving users with an overall accuracy of 80.8% as shown in Table II. For all the classes the recall is above 73% and the precision above 81% except the Spammers.

The community shows tolerance to non-provocative spammers. Intrigued by the low precision for spammers, we find that spam comments without provocative language, swear words and sarcasm often do not get reported by the community. Although unusually long comments are more likely to be spam, as we saw in Section III-D, this behaviors either escapes detection or is met with tolerance.

Our method identifies un-reported spammers. We examine the 12 false positive in the spam category: our spam label is not corroborated by the community. We actually find that at least one of these users exhibits clear spamming behavior, as she repeats the exact same comment 3 times in one article and 5 times in another. This suggests that we could catch misbehavior that avoids community detection, and consequently, our accuracy could be better than reported here, especially for spammers.

Table II
ROLE CLASSIFICATION RESULT

Role	Precision	Recall
Trolls	86.4%	73.1%
Spammers	58.6%	81%
Fanatics	86.1%	73.3%
Benign	81%	87.5%
Overall accuracy	80.8%	

V. CONCLUSION

We develop a systematic and comprehensive methodology to identify malicious users on commenting platforms with fine-grained classification of malicious behavior. The overall classification accuracy of our approach is 80.8% for the fine-grained 4-class problem. Our work is a first step towards safeguarding commenting platforms from malicious users. In the future, we aim at: (a) create a public labeled datasets to facilitate research, and (b) consider more malicious roles.

REFERENCES

[1] Disqus, "Disqus: Comment hosting service," https://disqus.com, 2015.
[2] LiveFyre, "LiveFyre: Real-time Content Marketing and Engagement," http://web.livefyre.com.
[3] Intense Debate, "Intense Debate: Imagine better comments," http://intensedebate.com.
[4] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," *arXiv preprint arXiv:1504.00680*, 2015.
[5] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *CSCW*, 2017.
[6] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Depend. Sec. Comput.*, vol. 9, no. 6, pp. 811–824, 2012.
[7] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, "A new approach to bot detection: striking the balance between precision and recall," in *ASONAM*. IEEE, 2016, pp. 533–540.
[8] A. Ferraz Costa, Y. Yamaguchi, A. Juci Machado Traina, C. Traina Jr, and C. Faloutsos, "Rsc: Mining and modeling temporal activity in social media," in *SIGKDD*. ACM, 2015, pp. 269–278.
[9] P. Devineni, D. Koutra, M. Faloutsos, and C. Faloutsos, "If walls could talk: Patterns and anomalies in facebook wallposts," in *ASONAM*. ACM, 2015, pp. 367–374.
[10] X. Gu, H. Yang, J. Tang, and J. Zhang, "Web user profiling using data redundancy," in *ASONAM*. IEEE, 2016, pp. 358–365.
[11] M. S. V. A. C. K. P. E. E. P. Joobin Gharibshah, Tai Ching Li and M. Faloutsos, "Inferip: Extracting actionable information from security discussion forums," in *ASONAM*, 2017.
[12] D. Sculley and G. M. Wachman, "Relaxed online svms for spam filtering," in *SIGIR*. ACM, 2007, pp. 415–422.
[13] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement." in *AIRWeb*, vol. 5, 2005, pp. 1–6.
[14] A. Sureka, "Mining user comment activity for detecting forum spammers in youtube," *arXiv preprint arXiv:1103.5044*, 2011.
[15] C. Warren, "When are facebook users most active? [study]," 2010. [Online]. Available: http://mashable.com/2010/10/28/facebook-activity-study/
[16] Sysomos, "Inside twitter: An in-depth look inside the twitter world," Apr. 2014. [Online]. Available: http://sysomos.com/sites/default/files/Inside-Twitter-BySysomos.pdf