Online Abusive Users Analytics through Visualization

Anna C Squicciarini, Jules Dupont, Ruyan Chen
College of Information Sciences and Technology, Pennsylvania State University
University Park, PA, USA
asquicciarini@ist.psu.edu, jnd5183@ist.psu.edu, ruyanc@ist.psu.edu

ABSTRACT

In this demo, we present *Abuse User Analytics (AuA)*, an analytical framework aiming to provide key information about the behavior of online social network users. AuA efficiently processes data from users' discussions, and renders information about users' activities in a easy to-understand graphical fashion with the goal of identifying deviant or abusive activities. Using animated graphics, AuA visualizes users' degree of abusiveness, measured by several key metrics, over user selected time intervals. It is therefore possible to visualize how users' activities lead to complex interaction networks, and highlight the degenerative connections among users and within certain threads.

1. PROBLEM OVERVIEW

As user-contributed sites continue to proliferate, online deviance is becoming a significant problem [1, 5, 8]. Deviance -defined as any behavior that is destructive, negative and offensive- is a practice that online users adopt for a number of selfish reasons, such as social and personal gain. For instance, many users are reported to lie in order to improve their popularity, gain access to reserved portions or entire protected sites, or engage in animus discussions to defend their personal and political beliefs [10]. Even initially innocuous users may eventually misbehave and violate terms of use of their preferred online sites. Importantly, one user's malicious behavior often influences otherwise honest users toward deceptive activities or misuse [12], eroding the overall well-being of the community.

Several real-world examples confirm a similar phenomenon of "bad" users influencing honest ones. A well-known case occurred in the popular Reddit social network. Violentacrez was a very popular Reddit moderator [12], who fiercely championed absolute free speech doctrine on Reddit. His identity was revealed by Anderson Cooper (CNN 360^o) after Violentacrez had gained great popularity, and collected a large number of fans and followers. Violentacrez was in fact founder and moderator of several very active subreddits (which are custom-made subforums on Reddit on specific areas of interest) on controversial topics (e.g. gore, sexual content) [12]. Many users followed his lead and created similar subreddits with massive amount of content even more provocative (and illegal) than the ones originally posted by Violentacrez. This anecdotal case shows that a disruptive user may become highly influential, and promote similar abusive actions (e.g. posting of obscene content) from a relatively large number of users.

To date, enforcement of usage policies in user-contributed sites is largely a manual task. Typical enforcement strategies involve careful monitoring of the shared community space by superusers. Superusers, also referred to as moderators, are often dedicated and long-running members in good standing who have been granted some authority to patrol and take action against members for deviant behavior. To assist superusers, mechanisms are often put in place to help quickly report and stop abusive behavior. For example, some automated tools exist to detect vandalism and bots, that filter malicious or inappropriate user posts and posts [9, 4]. From the academic world, current approaches mostly focus on classifying deviant or unusual posts (e.g. [2, 3]), or they allow to rank user-contributed comments [6]. Both these lines of work focus on empirical models for data classification, based on the nature, wording, and informativeness of users' comments [7]. These solutions are useful in filtering users' malicious posts, as well as in analyzing users' main interests, identifying influent posters, etc. However, they don't provide a easy-to-use way for site administrators and researchers to manage, process, and visualize users' activities. Further, they are often unable to track the behavior of repeated deviant users. Finally, they are typically customized to fit a specific domain (e.g. Youtube, Digg).

2. ABUSIVE USER ANALYTICS

2.1 Overview

In order to promote healthy and stable online communities, we present *Abuse User Analytics*, shortened as *AuA*, an analytical framework aiming to provide key information about online social network users. AuA efficiently processes data from users' discussions, and renders information about users' activities in a succinct, easy to-understand graphical fashion with the goal of identifying deviant or abusive activities. Users' activities are summarized not only using raw statistics of users' actions in a given social site, such as frequency of posts and number of friends, but also include the social interactions among users, how and if they influ-

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04. http://dx.doi.org/10.1145/2567948.2577019.

ence each other, and whether or not they contribute to the same set of threads (for threaded online communities) along with social network statistics. More importantly, using animated graphics, AuA visualizes users' degree of abusiveness, measured by several key metrics (e.g. content degeneration, language used, sentiment etc), over possibly large (user selected) time intervals.

AuA users can choose to combine the dimensions used for abusiveness analysis in various ways, by assigning weights through a user-friendly graphical user interface. In addition, by providing a network representation of users' activities and of their relation with other abusive users in the network, AuA facilitates discovery of possibly hidden users' pattern of interactions. For instance, it is possible to quickly view the threads wherein users appear to be more aggressive, how abusive users are connected, and even the "type" of abusive behavior mostly displayed. By analyzing the topic central to the most abusive threads, one may easily identify the "heated" topics leading to disagreement and, more generally, abusive behavior.

AuA is partly inspired by our previous work [11] wherein we designed a user choice model able to reliably identify a malicious user from a legitimate one. TriCO is a risk-based warning system that alerts superusers regarding increased risk of imminent deviance. We modeled users' choices and monitor changes in their behavior in text-based communities, and deployed it by means of a Bayesian Network model. In AuA, the focus is on the analysis, aggregation and visualization of users' activities and posts, using some of the features used for the TriCO model.

2.2 Information Extraction Model

Our approach to analyzing deviant behavior builds on the ability of determining whether users' posts are acceptable for the online site in question. What constitutes an acceptable post is governed explicitly by site policies and, implicitly, by the character of the community. The particular users, how they communicate and what they communicate about, combined with the oversight and policy enforcement maintained by the moderating status, define this character. Henceforth, ours is not a simple good versus bad post classification problem, since the terms of classification are subjective to the context being considered. The analysis of the quality of the posts is however the first step toward identifying persistent threats in online sites.

We identified few factors that help classify the overall post quality, including abusive/swear words count, community ratings on the post, the sentiment of the posts, and the content degeneration.

- *Post Sentiment.* This feature allows us to estimate the polarity of the overall post. We consider this variable as an indicator of the overall post nature, along with the type of language used by the author of the post. It is computed using natural language processing methods, as discussed in the next section.
- *Jargon.* The jargon is here defined as the presence of abusive or inappropriate words in the post, denoting offenses, curses, or other inappropriate wording. The list of abusive words includes swear words, negative adjectives (ugly, dumb) that are not promoted by the community. This list can be customized by each site administrator to address any specific or topic-related post that may not be included in a general list.
- *Content Degeneration.* This feature measures the extent to which a particular post is degenerated relative to the post originating the discussion. It is measured by considering the mutual information (MI) of the post, with respect to the category or topic of the thread. The less cohesive the post is with respect to the whole thread, the more degenerated or out of context it is likely to be. Precisely, we simply measure it in terms of its cohesiveness. Let TT be the overall thread topic or category, and p a post of user i. Assume each post has a set of words in it.

$$deg(p_i; TT) = \sum_{w \in p_i} MI(w, TT) \quad (1)$$

MI measures the amount of information each term w relates with the thread topic (TT). $MI(w, TT)$ is calculated as $p(w|TT)p(t)log\frac{p(w|TT)}{p(w)}$. Here $p(w|TT)$ is the probability that the term w appears in other posts of TT. $p(w)$ is the fraction of posts with term w (it is corrected to ensure that $p(w) \neq 0$). $p(TT)$ is the fraction of posts on the specified topic.

- *Post Rating.* This feature measures users' feedback on the specific Post, and can be used to measure the perceived popularity of a user's contribution.

Each metric is then normalized, and an average is computed to determine the overall user's score per metric.

The above list is by no means exhaustive. Other metrics, such as level of informativeness or complexity of a post, could be also considered, although they appear to be less relevant when detecting abusiveness.

3. ARCHITECTURE

In Figure 1 we report the initial architecture and interaction flow of the monitoring tool hosting the AuA.

AuA is deployed using C++ for the back-end, whereas Qt (at qt-project.org) is used for visualization. The social network data is stored and managed in MySQL database. The back-end includes few core modules, related to (1) database connection and information retrieval, (2) users' posts aggregation and labeling, (3) computation of users' behavioral metrics, and (4) processing for data rendering.

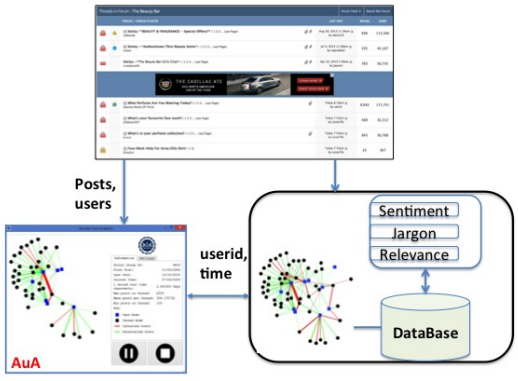

Figure 1: AuA Architecture

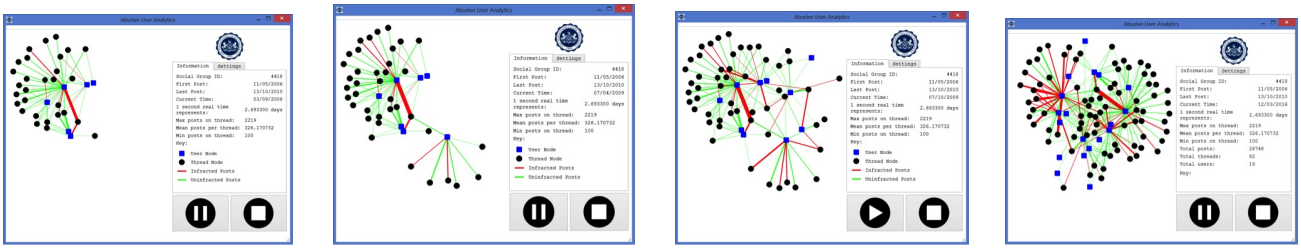

Figure 2: AuA user interface, and progressive visualization (a total of 26746 posts, 82 threads, and 19 users are visualized)

3.1 Back-end

We here discuss some implementation details of some of the basic dimensions computed for social network analysis.

Sentiment is used to determine the attitude of the user with respect to some topic or the overall contextual polarity of a thread or a subnetwork. In order to make this determination, a naïve Bayesian classifier is used to categorize each post into one of three broad labels: positive, negative, or spam. The positive category represents normal, everyday communication, the negative category represents aggressive or otherwise abusive communication, and the spam category represents posts unrelated to the content of the forum as a whole (e.g., ad-bot posts). The classifier is implemented using the bag-of-words model, in which a forum post is decomposed into its constituent words and those words become features in the classifier. Additionally, the classifier considers features such as capitalization and the frequency of certain punctuation marks to produce a more refined categorization of posts. The training set for the classifier consists of five hundred posts from each of the three classes. In the current implementation, posts for the positive class were randomly selected from the forum database and manually confirmed as positive. Posts for the negative and spam categories came from a table maintained by moderators of the forum, who had manually flagged posts as infracted for spam or for abusive content.

Upon execution, for any given user, her posts are first consolidated, and cleaned up of possible quotations and other irrelevant external references. Posts are consolidated according to different criteria, depending on the type of analysis to be completed. To show users' behavior across threads, a user's posts are grouped by threads and a sentiment value computed for each. Differently, if the emphasis is on the user interaction with peers, the sentiment is computed over the posts directed toward others in the same social group or subnetwork. Next, the sentiment score is computed.

Content Degeneration is computed as follows. First, from each post, tokens are extracted, so that they include only core lexical. These tokens are then queried in the Wordnet dictionary to form a list of synsets. Each synset represents a group of synonyms. If the token does not appear in the Wordnet dictionary, it is added to a list of unrecognized words. After processing the post, the very same procedure is run on the thread the post associated with (or the post to compare the original one with). Two lists of synsets and unrecognized words are produced for the post and the thread. The list is used to calculate mutual information between the post and the thread, per Equation 1. If the emphasis is on the user-to-user interaction rather than on the overall behavior of the user, content degeneration can be also measured

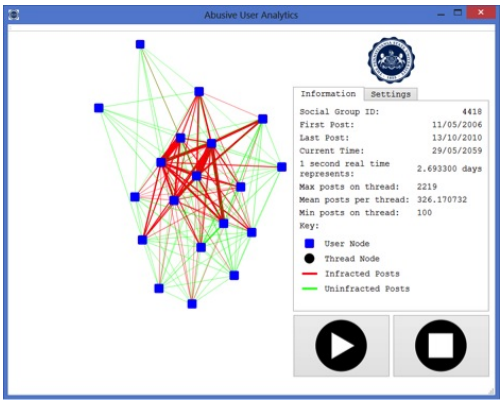

Figure 3: User-to-User interaction Network

by considering relevance with respect to the previous post or the original thread where the post is attached.

The *language* appropriateness or *jargon* is measured using known lists of prohibited words and jargon, and combined with sentiment, provides an indicator on whether the user displays an aggressive, and offensive behavior.

3.2 Front-end and Networks Visualization

AuA enables quick analysis and visualization of: (1) abusive or deviant behavior in user-selected and site-defined social groups, to assess their interaction pattern, their cohesiveness degree and uncover their social structure (2) specific users (individuals or groups) in the community. Users' connections and contributions are displayed by means of colored networks. Networks can be *static* or *dynamic*. A static visualization shows a user's or group's network at a given time point. A dynamic visualization shows animated networks, displaying data related to a manually entered time interval.

The AuA API along with examples of networks for two groups of users is reported in Figure 2 and Figure 3, respectively. The input to the API (entered in the Settings pane) is either a single group (as in Figure 2) or user's ID, or it is a range of user groups or users input via the GUI's input box. AuA users can set the API to compute degree of abusiveness based on their chosen combination of the *Sentiments, Jargon* and *Content Relevance* metrics, using sliders. The output's format is selected using a checkbox, which indicates the format of the visualization to the API. Networks are visualized using a spring layout. The API also includes an input button for users to enter the timeframe to visualize (e.g. from 1/12/2000 to 15/10/2004). Users can stop the visualization at any time, or fast forward it. The speed of visualization can be configured by the user in the settings

pane. See Figure 2 for an example, representing a group's activity over the course of two years. The network is bipartite as it shows two types of nodes: threads (black dots) and users (blue squares). This visualization facilitates an understanding of the threads mostly contributed by abusive users. Abusive contributions are colored red. Edges direction can be added if desired. As shown, statistics about the data being visualized are also given, including the first and the last input data points, maximum and minimum posts per thread the interval of speed, etc.

To visualize animated networks (as in Figure 2), we decompose the graph into a series of frames, where each frame is an exact representation of the social network at a given date and time. To meet this goal, the back-end of AuA generates a list of nodes and edges from posts. When each node and edge is generated, associated temporal data is stored alongside. For nodes, time of creation is also stored. For edges, the implementation tracks both the number of posts at a given time (so edges can be weighted) and the infraction status (so edges can be colored). When generating a graph at a given time, the implementation queries the temporal data associated with each edge and node to generate an accurate picture of the social network at that time. Once frames are generated, they are animated with Qt' s standard animation suite.

If the input is in form of user IDs, then the API calculates the output for all the threads in which the user's ID appears, representing the user's interaction as the graph with the nodes being the other users whom the user in question interacted with (see Figure 3). If the input is a social group, the API reports the interactions among users in a given social group. Users are represented as the nodes of the graph, whereas the edge is used to connect two users who post on a same thread or reply to each other's post. The resulting network uses colored edges to display abusive interactions - determined according to the entered inputs and relative weight. Malicious users may be given (i.e. if moderators' data is stored in database, the infracted posts are known) or identified based on our abusiveness metrics and their relative weights. Edges may further be labeled, to visualize specific information on the type of connection in place (e.g. the thread ID or the keywords for the thread).

Finally, note that because a major concern when analyzing big data is the presence of large amount of "noisy" data, AuA employs several filters prior to generating network graphs. For instance, users with fewer than a certain - AuA user-defined- number of posts are eliminated. Upon generation of the social networks of interest, users with low centrality may also be discarded.

4. DEMONSTRATION

Conference participants will be invited to use the application using our pre-loaded datasets. The datasets currently at hand refer to an online gaming forum, as well as discussion threads gathered from Youtube. The gaming forum is a large threaded forum (over 3 million posts) distributed across over 1 million threads. We count a total of 756 social groups (average group size is 16, st.d. 5.4), 95% of these groups contribute to at least one group-specific discussion. The Youtube commentary dataset counts 500,000 posts, labeled by users as positive, negative or neutral. Using portions of these data sets, participants will be given the opportunity to compare various network visualizations, highlight defective users' social interactions, and their behavior over time.

Participants will also be able to visualize detailed data on chosen nodes through a zooming feature. We are deploying two types of zooming functions (1) If users are zoomed, we will display users' network information - user ID, betweenness, number of connections etc (2) If nodes are zoomed, we will display relevant topical keywords associated with a thread. With these representations, the interaction among users will be more evident, and will facilitate analysis on popular trending topics and topics that instead are more prone to trigger intervention from deviant users. For ease of understanding, we plan to show the textual data streams for similar groups on a separate machine, in parallel with the AuA analytic results and visualizations. A simple screencast is available at: http://asquicciarini.ist.psu.edu/demo.html

5. ACKNOWLEDGEMENT

We would like to thank Smitha Sundareswaran and Candice McKune, for their help on an earlier version of the project. Portion of the work from Dr. Squicciarini was supported from Army Research Office grant W911NF-13-1-0271. Ruyan Chen's work was supported by the Distributed Research Experiences for Undergraduates (DREU) program, which is funded in part by the NSF program (NSF CNS-0540631).

6. REFERENCES

[1] HaltAbuse statistics, 2012. http://www.haltabuse.org.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying video spammers in online social networks. In *Workshop on Adversarial information retrieval on the Web*, pages 45–52. 2008.

[3] N. Christin, S. Yanagihara, and K. Kamataki. Dissecting one click frauds. In *17th ACM conference on Computer and Communications Security*, pages 15–26. 2010.

[4] Fassim. Fassim: a forum spam prevention plugin. http://www.fassim.com/about/.

[5] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in Microblogging. In *23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[6] S. Kashoob, J. Caverlee, and K. Y. Kamath. Community-based ranking of the social Web. In *ACM Hypertext Conference (HT)*, pages 141–150, 2010.

[7] S. Kleanthous Loizou. Intelligent support for knowledge sharing in virtual communities. 2010.

[8] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. *Proc. of ICWSM*, 2011.

[9] S. F. SPam, 2012. http://www.stopforumspam.com.

[10] A. C. Squicciarini and C. Griffin. An informed model of personal information release in social networking sites. In *SocialCom/PASSAT*, pages 636–645, 2012.

[11] A. C. Squicciarini, W. McGill, G. Petracca, and S. Huang. Early detection of policies violations in a social media site: A Bayesian belief network approach. In *IEEE Policies for Distributed Systems and Networks*, 2012.

[12] The Daily Dot. Dot 10: The 10 most important people on Reddit, 2011.