Early Detection of Policy Violations in a Social Media Site: A Bayesian Belief Network Approach

Anna Cinzia Squicciarini, William McGill, Giuseppe Petracca, Shuo Huang
College of Information Sciences and Technology
Pennsylvania State University

***Abstract*— One of the main goals of all online social communities is to promote a stable, or perhaps, growing membership built around topics of like interest. Yet, communities are not impermeable to the potentially damaging effects resulting from those few participants that choose to behave in a manner that is counter to established norms of behavior. Typical moderators in online social communities are the ones tasked to reduce the risks associated with unhealthy user behavior by rapidly identifying and removing damaging posts and consequently taking action against the perpetrating user. Yet, the sheer volume of posts relative to the number of moderators available for review suggests a need for modern tools aimed at prioritizing posts based on the assessed risk each user poses to the community. To accomplish this, we propose a threat analysis model, referred to as TrICO (Threat requires Intent Capability and Opportunity) implemented using Bayesian Networks for early detection of damaging behavior in online social communities. To the best of our knowledge, this is the first user-centered model for policy enforcement in online sites. We apply our model to a comprehensive data set characterizing the entirety of a popular discussion forum. Our results show that the TrICO model provides accurate results relative to common alternative statistically based approaches, such as random selection.**

I. Introduction

Commenting systems on the Social Web have been growing in popularity in the past few years, from blogs and social media sites like YouTube and Flickr to major news sites like NYTimes.com. Along with it, the number of episodes of online abuse are proliferating [3], [1], [10], [17]. Abusive behavior comes in many forms ranging from minor to extremely harmful. Forms include grieving, trolling, flaming, harassment, threats trolling, multiple accounts, shared accounts, advertising, plagiarism etc. The actual definition of abusive behavior is discretionary to each site's usage policy and term of use, and usually is defined based on the site's scope, goals and targeted audience. Generally, any behavior that is destructive, negative and offensive is considered as abusive.

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 comments [19], [8]. From the academic world, most of current proposals in this space focus on classifying single deviant posts [10], [2], [4], [5], or they allow to rank user-contributed comments [13]. Both these lines of work fail to account the site's policies, and classify posts based on their nature, wording, and informativeness [6], [14]. These solutions are useful in filtering users' malicious posts, as well as in analyzing users' main interests, identifying influent commenters, etc. They are, however, inadequate to make accurate assessment on the users' compliance with respect to site policies. Further, they are often unable to track the behavior of repeated deviant users. Finally, they are typically narrowed to a specific domain, like Youtube, Digg or other popular site. As such, the ability to predict, rank or estimate the quality of comment is heavily anchored to the specific structure of the site.

In order to facilitate site policies enforcement and promote healthy and stable communities, in this paper we tackle the problem of reliably assess risk of users' damaging behavior, i.e., users' actions violating the usage policy enforced in user-contributed sites. Precisely, we are interested in the following kinds of user-committed misbehavior: posting messages that flout the sites rules (e.g. racist or adult messages), messages aimed at starting or continuing a flame (or bashing), posting malware in the form of malicious URLs, posting personal advertisements. This is in contrast to the similar problem of spam. In the context of automated spam, the spam messages are typically identified by a set of common features (e.g. use of swear words, malicious links, unwanted advertisements, repeated identical posts, etc), and can be well identified by manual or automated labeling. In our domain, the additional level of complexity is added due to the subjective nature of what is defined to be deviant.

We design and develop a user choice model able to reliably identify a malicious user from a legitimate one. At the core of our model is a risk-based warning system that alerts superusers regarding increased risk of imminent deviance (or usage policy violation). We model users choices and monitor changes in their behavior in text-based communities. Our model, referred to as the TrICO model (Threat Requires Intent Capability and Opportunity), is a probabilistic model based on the assumption that an individual must have an intention to commit an act that is undesired by others, a capability to commit this act and an opportunity to apply these capabilities in order to pose a threat [20]. There is no threat if any of these three aspects are absent. The model draws from a preliminary qualitative study

conducted with a number of real-world online site moderators, which provides guidelines and insights on the most important aspects humans use for early detection.

The TriCO model is deployed using Bayesian Networks. Bayesian networks have been used in multiple scenarios to model uncertainty [25], and have proven effective for real-time security analysis. Bayesian networks are an effective tool for assessing deviant user behavior since they are able to capture the uncertainty typical of any user-in-the system technology.

In the paper, we include also an initial architecture of our TrCo model, and extensive evaluations. The TrICO model is in fact the core component of a comprehensive tool for policy of forums and user-contribute sites, and can be used to ensure effective enforcement and help shape the policy users of the site should adhere to. For our experiments, we have applied our model to a comprehensive data set characterizing the entirety of a popular discussion forum. We compare the results of our model with relatively context-free approaches, based on moderation of posts chosen at random in a given time interval. Our experiments show that the TrICO model provides accurate results relative to common alternative statistically based approaches.

To the best of our knowledge, this is the first user-centered model for policy enforcement in online sites.

The rest of the paper is organized as follows. Next, in Section II, we present a classification model for deviant posts, followed by the discussion of the TrICO model. In Section III, we describe the key techniques used for probability computation. Our experimental results are reported in Section IV, whereas we discuss the related work in Section V. We conclude the paper in Section VI. The paper also includes a summary of the qualitative study conducted to ground our model, reported in the Appendix.

II. The Bayesian Belief Network Approach

In this section we discuss our methodology for posts classification, followed by the description of our uncertainty model. Our context of reference is that of open forums, operated by software such as vbulletin. Users can create an account after little or no verification at all, typically upon presenting a valid email account. We assume a forum system whereby discussions are organized in topics (or threads), and each thread has multiple posts (or comments) from different registered users. Formally, a forum is defined as $F = \{Th_1, \ldots, Th_z\}$, with $Th_k, k \leq z$ a generic thread of one or more posts. We do not consider forums where contributors can be anonymous (i.e. a permanent identifier needs to be associated to every post entry). Forums may implement user-to-user feedback, user reputation and social relationships, although these features are not mandated for the design of our approach.

A. Comment Representation and Ranking

Our approach to predicting deviant behavior, builds on the ability of classifying a post as not acceptable for the forum 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.

To all extent and purposes, therefore, ours is not a simple classification problem, since the terms of classification are subjective to the specific context being considered. Identification of "bad" posts is however *the first step toward a user-driven prediction model*, which is described in the next section. We identified few factors that help classify the overall post quality, and its level of abusiveness.

We take into account a number of features for post classification, including abusive/swear words count, community ratings on the post, the sentiment of the comments, and the content degeneration. While other metrics could be considered, such as level of informativeness, complexity, and cohesion of the comment with respect to the thread [13], these appear to be less relevant when detecting abusiveness. Short, uninformative posts may be legitimate ("Hi, how are you?" or "I think this is the right documentation for this problem"), although they may not greatly contribute to the discussion.

The following features are used for classification purposes.

- *Comment Sentiment.* This feature allows us to estimate the polarity of the overall post. We use LingPipe [16] as the tool for classifying positive and negative posts.
- *Content Degeneration.* This feature measures the extent to which a particular comment is degenerated relative to the post originating the discussion. It is measured by considering the mutual information (MI) of the comment, with respect to the category or topic of the thread. The less cohesive the comment 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.

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

  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 comments 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.
- *Frequency and Type of abusive words.* This feature considers the actual number of abusive words appearing in the post. The list of abusive words includes swear words, negative adjectives (ugly, dumb) that are not promoted by the community. Our experimental analysis uses predefined word lists. However, this list can be updated and customized by each site administrator to address any specific or topic-related comment that may not be included in general list.

- *Comment Rating.* This feature includes only a simple count of the users' feedback on the specific comment. Given the users' population U, and a user i, the overall rating for a user i in the context of a thread Th_k is

$$CR_{p_i}^{-/+} = \sum_{j \in U} PI_{ij}^{Th_k,-/+} \quad (2)$$

Here, PI_{ij} represents the the total number of indicators added by user j with respect to user i's post p_i in a given thread Th_k[1]. Indicators can be positive or negative, hence the superscript $-/+$. We will therefore have $CR_{p_i}^{+}$ and $CR_{p_i}^{-}$.

Using the aforementioned features, we use classification techniques to estimate the likelihood that the message denotes some misbehavior, and is a non conforming post (NCP) depending on out tolerance for such messages. Specifically, we work under the assumption of conditional independence, and use a Bayesian classifier to rate whether a post is representative of a misbehaving user or not.

Naïve Bayes is a classifier that uses the Bayes theorem to classify objects. Given a post p_i, with the features of above and two classes to choose from (CP, NCP), Naïve Bayes determines which class C_x is more likely under the assumption that the features $(f_1, \ldots, f_m)$ are independent. That is,

$$\begin{aligned} &argmax[P(C_x|f_1, \ldots, f_m)] = \\ &argmax\left[\frac{P(C_x) \times P(f_1|C_x) \times \ldots \times P(f_m|C_x)}{P(f_1, \ldots, f_m)}\right] \end{aligned} \quad (3)$$

Notice that it is also possible to adopt a weight-based classifier, to provide a more realistic model of the site's overall rules of usage. For example, every site typically includes in the usage agreements, statements indicating that obscene, vulgar, sexually-orientated, hateful, threatening, or otherwise violative of any laws messages are prohibited. The extent to which this is actually tolerated changes slightly from site to site. Influenced by the site audience and education level, certain sites are more tolerant with respect to use of slang words or offensive words (e.g. YouTube), whereas others strictly control the language used by the posters. Similarly, the cohesiveness of a post may be given more weight for sites that expect commenters to focus solely on their main forum's topic (e.g. technology, fashion, beauty, travels), and have low tolerance versus off-topic comments. Weighted bayesian classifiers, such as [7] and [26] can be employed for greater accuracy. In our experiments, we employ a standard Bayesian model as a baseline. As our results show, these appear to be sufficient for accurate classification.

B. Model Formulation

The TrICO is rational-actor model. That is, considering the context of a specific forum, we assume the user will make his choice to post based on a deliberate comparison of perceived benefits with perceived costs. The model probabilistically associates particular users with a certain risk or non-compliance-level by studying his behavior over time. Moreover, the model probabilistically associates the situational environment with prototypical attractiveness levels. The connection between data and user actions is established via a Bayesian Network.

[1]For simplicity, we here assume these could be added, although other forms of aggregation methods could be used.

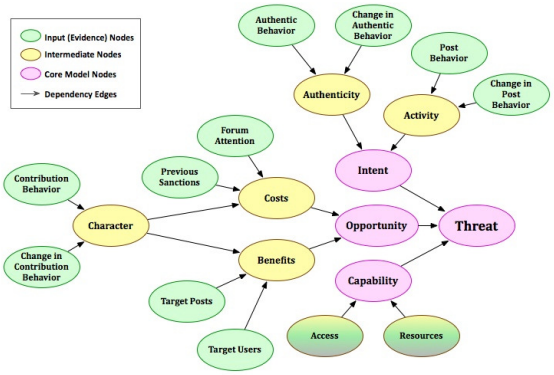


Fig. 1. The TriCO network

A Bayesian network (BN) is a graphical representation of cause-and-effect relationships within a problem domain. More formally, a Bayesian network is a Directed Acyclic Graph (DAG) in which: the nodes represent variables of interest (propositions); the directed links represent the causal influence among the judgments. The strength of an influence is captured by conditional probability tables (CPT).

The Bayesian Network implementing the TrICO is shown in Figure 1. The nodes were selected according to the insights obtained by our qualitative study, which revealed the important elements used by human moderators for detection (see Appendix). This network consists of 22 nodes (i.e. variables). 13 of the BN variables (nodes) are derived from data collected from online forums and discussion boards (see Section IV). These are the independent variables. The dependent variable is threat, that is, whether the user is a threat or not. There are various intermediate variables (colored as yellow nodes in Figure 1) at the interface between data states and the choice model, focused on: user attitude, site attractiveness as perceived by the user, and prior sanctions.

The core network that defines our anticipation model is based on three core dimensions: opportunity, intent and capability (nodes are pink). *Capability* is essentially the ability of the user to post on the site. It is for the most part a given, in that we focus on forums that are open to the public, or sites that, if access is controlled, are relatively easy to access. The *Intent* variable constitutes our main input as indicator of deviance. A user's trace sheds insight into their personality and behavior. A number of metrics can be used to infer online personality. Here, we consider the user's posts and their quality (conforming versus non-conforming), the user's overall contribution to the discussion as perceived by peers, and the overall changes in the users' attitude over a certain time period. Given the loose nature of the notion of identity in online communities, and the lack of verification mechanisms against the declared identities, we purposely exclude to consider any

identity-related information. Finally, *Opportunity* accounts for the perceived benefits or costs associated with the usage of the site. The benefits account for the possible audience of the poster's message, whereas the costs represent the possible sanctions associated with publishing non-conforming posts.

1) Independent Variables: These are metrics that we derive directly from observing users' comments and their activity throughout their lifespan as community users. We observe:

- *Post Behavior*. This measures a user's activity in the community by the number of posts the user has submitted to the community. This metric can be restricted to an interval of interest, that is a given timeframe or limited to topics of certain categories.

$$PB_i^{Th_k} = \sum_{\forall p \in Th_k, t \leq t_w} 1 \tag{4}$$

where $P \subset Th_k$ is the set of all posts in a particular thread Th_k by a user i, each post has occurred some time ago t, and t_w is the upper bound of the time span of interest. The overall number of posts can be computed to include any posts in the forum by summing all the non-zero post-sets P across all threads across the forum.

- *Authentic Behavior and Changes.* This variable measures the proportion of posts that conform to the thread or topic with respect to the total number of posts at a given time point. We define authentic behavior as the ratio of the number of conforming posts (CP) to total posts (TP) made by user i on thread Th_k at a given time point t (considering the posts since the previous time interval), used to calculate TP. CPs are defined based on the outcome of the post classification discussed in previous sections.

$$AB_i^{Th_k}(t) = \frac{CP_i^{Th_k}(t)}{TP_i^{Th_k}(t)} \tag{5}$$

Authentic Behavior may vary depending on the topic. The aggregate authentic behavior of user i across all threads, $AAB_i(t) = AB_i^*(t)$, where the $*$ denotes sum across all states of a superscript.

The change of authentic behavior variable measures the change in the proportion of authentic posts with respect to the total number of posts between two consecutive time intervals.

$$CAB_i^{Th_k} = \frac{AB_i^{Th_k}(t) - AB_i^{Th_k}(t_{prev})}{AB_i^{Th_k}(t_{prev})} \tag{6}$$

- *Change of Post Behavior.* This variable measures the change in the rate at which a user posts content over a fixed time interval.

$$CPB_i^{Th_k}(t) = \frac{PB_i^{Th_k}(t) - PB_i^{Th_k}(t_{prev})}{PB_i^{Th_k}(t_{prev})} \tag{7}$$

- *Target Users.* The number of users' accessing the network. Depending upon the forum configuration, this variable may defined as the friends of the user, or the overall population. Intuitively, users with many friends (or followers) may be more popular and trusted than poorly active users. *Total Post* is similar, in that it accounts for the number of the posts in the forum, as a metric of the popularity of the site.
- *Hazard Behavior and changes.* This variable measures the degree of attention paid by members of the community to the actions taken by the user in question over the course of a fixed time. The hazard behavior with respect to the credibility relationship between users i and j is:

$$HB_{ij}^{Th_k}(t) = \frac{PI_{ij}^{Th_k,+}}{PI_{ij}^{Th_k,+} + PI_{ij}^{Th_k,-}} \tag{8}$$

where$PI_{ij}^{Th_k,+}$ is the total number of good indicators noted by user j with respect to user i post in a given thread Th_k.

To capture the change in the degree of attention paid by members of the community to the actions taken by the user in question over the course of two consecutive time intervals, we compute also the *change in hazard behavior*.

$$CHB_{ij}^{Th_k}(t) = \frac{HB_{ij}^{Th_k}(t) - HB_{ij}^{Th_k}(t_{prev})}{HB_{ij}^{Th_k}(t_{prev})} \tag{9}$$

with $HB_{ij}^{Th_k}(t_{prev}) \neq \emptyset$

- *Contribution Behavior (community rating)* This feature measures the overall contribution of the user. The contribution behavior with respect to user i for thread Th_k at time t is:

$$CB_i^{Th_k}(t) = \frac{CP_i^{Th_k}(t) - NCP_i^{Th_k}(t)}{TP_i^{Th_k}(t)} \tag{10}$$

Here $NCP_i^{Th_k}(t)$ denotes the number of known non-conforming posts for the user i in thread Th_k. If $CTP_i^{Th_k} > 1$, the user is overall a positive contributor to the thread.

A meaningful measure of the overall contribution should span several threads. Hence, we extend the above equation to include any post where the user is actually active. The aggregate contribution behavior (ACB) is computed as follows.

$$ACB_i(t) = \frac{\sum_{\forall P_{Th_k} \in F, t \leq t_w} CTP_i^{Th_k}(t)}{\sum_{P_{Th_k} \in F} TP_i^{Th_k}(t)} \tag{11}$$

where P_{Th_k} is the set of posts of a user for a thread Th_k, and F is the total set of posts in the forum. t_w is again the boundary of the time span of interest.

- *Previous Sanctions.* This feature helps understand whether the user has been sanctioned before, and what was the severity of previous punishments. It is measured as a simple count or total user points, if the site adopts a point-based system.
- *Site Attention.* We capture possible subjective and not quantifiable indicators, that can play an important role in understanding the user's role and his history. Precisely, this feature accounts for possible subjective attitude of

Variable	States	Intervals
Authentic Behavior	0: Authentic	$AB_i^{Th_k}(t) \geq AB_{auth}^{Th_k}$
	1: Ambiguous	$AB_{auth}^{Th_k} \leq AB_i^{Th_k}(t) \leq AB_{amb}^{Th_k}$
	2: Unauthentic	$AB_i^{Th_k}(t) > AB_{amb}^{Th_k}$
Hazardous Behavior	0: Hazardous	$HB_i^{Th_*}(t) \geq HB_{haz}^{Th_*}$
	1: Average	$HB_{nhaz}^{k} \leq HB_i^{Th_*}(t) < HB_{haz}^{Th_k}$
	2: Non-hazardous	$HB_i^{Th_*}(t) < HB_{nhaz}^{Th_k}$
Previous Sanctions	0: No sanctions	PS<Thr
	1: Warnings	$Thr < PS < Banning$
	2: Banning	$PS > Banning$
Post Behavior	0:Below avg	$PB_i^{Th_k}(t) < \overline{PB}(t)$-$\epsilon$
	1: average	$\overline{PB}(t)$-$\epsilon < PB_i^{k}(t) < \overline{PB}(t) + \epsilon$

TABLE I

EXAMPLES OF STATES ASSIGNMENT OF SOME OF THE BN'S VARIABLES. THRESHOLDS ARE SET EMPIRICALLY.

(Neg.) Contrib. Behavior	UnAuth. Behavior	Below Avg Post Behavior	P(Intent)	P($\neg$ intent)
T	T	T	1	0
T	F	T	0.2	0.8
T	T	F	0.4	0.6
T	F	F	0.2*0.4=0.08	0.92
F	T	T	0.3	0.7
F	F	T	0.3*0.2=0.06	0.94
F	T	F	0.3*0.4=0.12	0.88
F	F	F	0.3*0.2*0.4=0.024	0.976

TABLE II

EXAMPLE OF NOISY-AND TABLE.

the moderator(s) with respect to the user. For example, personal relationship, or history that affects the superuser overall opinion. This is a qualitative metric, and we assume it takes one among three possible values, [0,1,2], ranging from no-sympathy (0) to close friendship (2).

We note that some systems do not support some of the features needed to implement all such metrics. For example, some sites do not support explicit user-to-user feedback. The equation of the Hazard metric will need to be adapted accordingly. In this case, an indirect indicator of the users' reactions to a post X can be measured by considering the number of posts subsequent to that of the original post X, and their sentiment. If negative sentiments are reported after a single post, then this may be considered as a negative indicator.

In case of a negative explicit feedback from one user to another, it is important to carefully define criteria and policies for interpretation. The indications given on SNs or social communities often indicate only whether or not user j likes or dislikes a post made by user i. This is not necessarily an indication of whether or not the post constitutes misbehavior. We apply the following rule: a $PI_{ji}^{-/+}$ is correct if and only if one of the following conditions is satisfied:

- The post p is rated positively and it is neither a spam post nor a violator post
- The post p is rated negatively, and it is in fact a NCP.

2) Intermediate variables: There are two sets of intermediate variables: inner and outer variables. The three inner intermediate variables within this network are as follows:

◇ *Intent (3 parents)* This variable describes a particular user in terms of prototypical behaviors. The specifics of an individual influence his intentions, how he/she perceived benefits, perceived costs, and how he/she balances utility across the two. Intent has three parent nodes: Relative Contribution behavior, Relative Authentic Behavior, Relative Post Behavior.

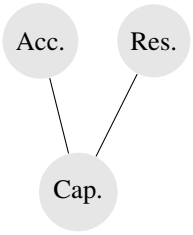
◇ *Capability (2 parents)* Capability has two cause variables: *Access* and *Resources*. Access is controlled by black-list look up such as [21], the idea being that if a email appears to have been used by previous offender access could be denied. Resources accounts for the resources of the user, i.e. his ability to contribute to the forum.

◇ *Opportunity (2 parents)* is a variable roughly describing the site attractiveness. It is further described by two additional variables: *cost* and *benefit*. *Cost* is representative of the risk incurred by the attacker of being caught: it is defined in terms of *site attention*, and *previous sanctions*. *Benefit* quantifies the popularity of the site: what would be the impact of the post? how many users would read it?

III. CONDITIONAL PROBABILITY TABLES ASSIGNMENT

In order to develop the network, the next step is to map each of the external variables listed above into discrete state-variables, and compute the conditional probabilities of the effect variables (i.e. intermediate nodes), down to the core dependent variable (i.e. *Threat*). We report in Table I, examples of some of the independent variables discretized. The complete mapping is omitted for lack of space. The conditional probabilities, described by conditional probability tables (CPT), represent the probability distribution of the nodes' possible states conditioned on the parents' states. The state of node without parents, i.e. root nodes or independent variables, is calculated through analysis of the available data or through expert's feedback and evaluations, based on the behavior of the monitored user. For example, for post behavior, one can derive the average number of posts per user and set the corresponding state values accordingly. In particular, in the experiments discussed in the following section, we derive our root nodes' state values either taking advantage of the experts, i.e. superusers who have direct experience with malicious users and their typical behavior, or by analyzing available data. For example, post behavior frequency can be obtained by considering the average number of posts per user. Similarly, the threshold to set Hazardous behavior can be set by considering the average number of warnings and sanctions received by known offenders.

To determine CPTs, we further employ some of the most widely used models for supporting such elicitation, that is the *leaky noisy-OR* and the *noisy-AND* model. The *leaky noisy-OR* model [18] assumes that a number of binary causes (i.e. variables $X_1, \ldots, X_k$) can produce an effect (i.e. verify a certain variable Y). Their interaction is expressed by a logic OR gate. The noisy-OR gate encodes the assumption that $X_1, \ldots, X_k$ fail independently, from which the number of



Acc.	Res.	P(Cap)	P(¬Cap)
T	T	1	0
T	F	0.7	0.3
F	T	0.5	0.5
F	F	0.35=0.7*0.5	0.65

Fig. 2. Example of Noisy-Or for the Capability Variable

probability parameters to be assessed per CPT is reduced to linear on the number of parent variables. Further, the leaky noisy-OR model introduces the possibility that the effect variable Y can be true even if all its causes are false. The hypothesis is that the predecessor nodes (e.g. access, resources) are independent on one another. We adopt the noisy-OR for some of the intermediate nodes, such as the *Capability* variable (see Figure 2). In our case, we assume that even if the user may not have been given access, given the weak forms of authentication used in online communities, he may have the capability to enter the site and add comments to it. Conversely, one user may be able to have access but limited resources, and yet have the capability to post, for example through proxy sites (i.e. twitter feed). The noisy-OR gate allows us to efficiently compute the complete CPT for the capability node. For example, let $P(capability|\neg access) = 0.5$, whereas $P(capability|\neg resources) = 0.7$. The leaky parameter is computed equal to 0.35. Here, the intuition is that a user may overcome access restrictions and lack of resources and still be able to post comments.

The *leaky Noisy-AND* extends from the deterministic AND logic. With this model, the hypothesis is that the causes are dependent on the effect. A leaky parameter (or node) is also introduced, to model the uncertain situation where despite some causes are not verified true, the effect variable may still verify true. For example, assume that the user has a positive contribution behavior and authentic behavior. Hence, the intermediate variable intent would result as "false". Yet, his (bad) intentions may be only disguised by an apparently "good" behavior, or change suddenly due to an external event. Let 0.024 be the probability that a well-behaving user is a threat for the community. 0.024 represents therefore the leaky probability that a user is a threat while not showing any threatening intentions: $(p(I|\neg RCB, \neg RAB, \neg RPB)$, where I is the intent and the other variables are the relative behaviors. Notice that the computation of the leaky parameters is defined separately and independently, therefore greatly simplifying the overall elicitation of the probability distribution. An example of using Noisy-And is reported in Table II. Notice that to build the probability values, we consider one state for each variable (e.g. Authentic Behavior) and we assign true or false based on whether the state of the variable was verified true.

IV. Architecture and Experimental Evaluation

A. Architecture

The TriCO is a core component of a comprehensive tool for policy enforcement and management of user-contributed sites.

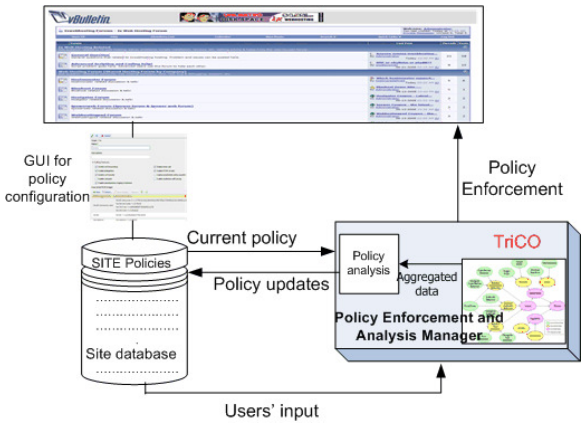


Fig. 3. Example of TrICO-powered monitoring system

By obtaining early information about users who are deviant, site moderators can quickly check to what extent site usage policies are repeatedly violated, and act accordingly against the deviant users. In addition, moderators and sites' owners can also use the outcome of the TrICO for analysis of the site policies and possible revisions. In Figure 3 we report the initial architecture and interaction flow of the monitoring tool hosting the TrICO. As shown, the tool's core component is an implementation of the TrICO model. The TrICO takes as input the users' posts and actions, and, after classifying the level of maliciousness of each post, it estimates the user's risk level. The risk is returned as a probability value, and is used for a dual purpose. First, it serves as a user monitoring tool and facilitates policy enforcement from moderators. Upon obtaining an estimate of the user's deviance, moderators can choose whether to act against him, or simply to keep a closer look at him. If a punishment is needed, they can opt to send warnings, ban the user or temporarily suspend his account. Second, the outcome of the TrICO can be used to estimate the overall decay of the site, and narrow the specific type of policies infringement and deviant behavior observed. By collecting the outcome of the TrICO for multiple users and over a certain amount of time, the TrICO-powered tool can provide feedback to site moderators on the site policies effectiveness, as well as indicate whether they need to be reconfigured. For example, if most deviant users are repeated and returning offenders, the site can consider restrictive access policies. Or, if users' feedback appears to be uninformative of users' deviance the site can enforce additional policies on moderating and tracking users' feedback, ranking and ratings.

B. Experimental Evaluation

We tested the accuracy of the TrICO model using a complete dataset from a well-known gaming forum. The site's data is maintained using Vbulletin software. The "Zelda Universe" is a popular community discussing "The Legend of Zelda" game. It includes a forum of over 61,000 users, with over 3.5 million posts (as of October 2010). Out of all the enrolled users, the forum counts roughly 21,000 users with more than

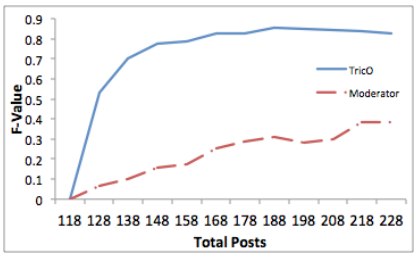


Fig. 4. F-value with increasing number of NCP

one post. 1,400 of them have infracted the site policies at least once. The reported users' age is 24, on average. We obtained the whole forum database, collecting seven years of data. The dataset includes all the textual comments, users' profile, moderator lists, and the list of the infracted users, with the punished posts and the following punishment to the user (warning, banning etc). We also have a list of the banned users. For testing purposes, we used the tables listing the infracted posts as ground truth (NCP posts).

In our experiments, we aimed to check if the TrICO can help the moderator identify high-risk users by associating a high risk probability value. To this end, we conducted two experiments. First, we studied how the TrICO reacts to an increased number of negative posts for a same user. Second, we tested how sensitive TrICO is to the CPT tables.

For our first experiment, we consider all the posts from a single user. We specifically selected a user with 35 infractions (or NCP), and a total of 148 posts. We initially removed his infracted posts, and started running the model with 118 posts, using 10% of the posts for training. We ran multiple rounds, and increased the number of NCPs at each round, up to a total of 110 NCPs. The first 35 of the added NCPs were infracted posts of the original user, and were added chronologically, according to when and where (i.e. what thread) they were originally posted. The remaining 75 NCP were injected by us. The injected posts were selected randomly from a large pool (about 300) of actual NCP posts moderated by the site moderators, available in the dataset. The overall F-rate (computed using classic precision and recall metrics) is reported in Figure 4. We compare the performance with that of a moderator checking 50% of the posts at each round. The moderator is ground truth, so when he extracts a post, he will either moderate it or not- and cannot be wrong. As shown, the TrICO consistently outperforms the moderator, and reaches the best performance when a large number of deviant posts are in the user's trace.

Our second test was performed using signal detection analysis [24]. We first ran the algorithm for the users with highest number of reported infractions. In this case, the selected users were 50, and the total number of posts 23,402. On average each user had 7 infractions, and the average number of posts per user was 135. We obtained a *d'* score close to 2 (1.89). We reported a larger number of false positive than false negative (26% versus 23%). The performance demonstrates the ability of our network to properly detect a large number of infractions. The relatively high error rate is due to some inherent limitations with our post classification algorithm, that does not account for some of the features considered by Zelda moderators. A typical example is a repeated post or a duplicate post. Further, some of the NCP were edited by moderators (swear words obscured or replaced in part with **** etc), therefore affecting our classification algorithm. We also hypothesized that some users had too little infractions for the TrICO. On this basis, we suspected that the accuracy could improve with a higher number of NCP posts.

We conducted an additional test to validate this hypothesis. For this test, we randomly injected in the dataset up to 30 malicious posts for each infracted user. We purposely chose infracted users to ensure that the users already had a trace of malicious or deviant activities. The malicious posts included clear offensive statements and swear words (non-censored), with other less blunt and obvious posts (i.e. posts with several slang words, not informative, out of context etc). They were, as in the first test, obtained from the actual list of moderated posts in the database. Clearly, in this case, we had to correct the way the hazardous behavior is calculated, in that no feedback from other users was obtained. We set the hazardous for the injected posts equal to the hazardous behavior for other infracted posts of the same user. Also, for our classification of posts, we did not consider mutual information. We tested 125 users, each with posts ranging from 61 to 850 posts, adding to a total of 54,321 posts tested. The outcome of the experiment was again evaluated using signal detection. We obtain a higher detection rate, with d' of 3.83. Our results show a false positive rate of 8%, and a false negative rate of only 3.5%. The results are therefore overall very encouraging.

V. Related Work

Our work is closely related with the recent body of work on social spam [22], [2], [10], [15]. Authors have focused on specific sites, video spamming and spam campaigns. Using primarily machine learning techniques, this body of work attempts to identify spam posts placed in user-contributed sites. We build on this body of work to study users misbehavior in online communities. Our approach is however substantially different, both in goal and approach. First, we are not interested in assessing the quality of a given post. Rather, we focus on users' deviance from site's policies. Therefore, in our model, even low quality posts are acceptable so long as they do not violate the enforced site usage policy. Second, we develop a rational-actor model that factors in several dimensions related to the users' displayed behavior.

Recently, authors studied the effect and implications of review spam, in recommender systems [12]. While similar to us to the extent that we both focus on identifying and analyze deviant comments from sites populated by end users, our solutions differ greatly. In our domain, classification as such is not sufficient, in that too coarse-grained. Rather, we need a more-fine grained assessment of users potential to harm the community. Classifiers have also been used by West and colleagues [23], in the context of vandalism on Wikipedia.

At the core of the West's solution is a lightweight classifier capable of identifying vandalism. The classifier exploits temporal and spatial features, extracted from revision metadata of articles. This work is representative of a growing body of work analyzing the entities on Wikipedia articles, authors, and individual edits, which primarily build on edit-level analysis and classification based on corpus. While similar in goal, we do not focus on the Wikipedia domain, but propose a generic and extensible framework for assessing behavior risk. Further, we attempt to estimate the risk of future malice rather than solely consider single user contributions.

Several tools exist to help moderators identifying bots and vandalism (e.g. [8], [19], [21]). Automated bots (e.g., Cluebot), filters (e.g., abusefilter), and editing assistants (e.g., Huggle and Twinkle) all aim to locate acts of vandalism. Such tools work via regular expressions and manually-authored rule sets. In our approach, individual comments represent the input for risk assessment. We could use one of such tools to assess the quality of an individual post and input it to our system.

Our approach also shares some similarities with work on intrusion detection systems (IDS) [11] and on free-riding in peer-to-peer systems. Approaches to intrusion detection include anomaly detection. Anomaly detection relates to our work in that it also involves the collection of data related to the behavior of users over a period of time [11]. Similarly, the TrICO analyzes individual user risk projected over time, by analyzing the collected evidence, which is however, more arbitrary and less tangible than a typical IDS. As in anomaly-based IDSs, the TrICO does not require prior knowledge of anomalies, although it represents one possible variable for assessment. Furthermore, different from the common IDSs, as shown by our experimental evaluation, we do not require an extensive training of the bayesian network to obtain accurate risk assessment values. Finally, our work parallels to the body of work on free-riding in peer-to-peer systems [9]. Peer-to-peer systems are designed to allow users to connect with others and share resources. Similar to online communities, users are free to access and contribute as much as desired, and few controls are in place. As for online communities, punishments, although applied are shown not to be truly effective. To tackle these issues, the common solution is to implement incentive-based mechanisms. Incentives are applied in certain online forums, whereby end users are given special roles and privileges as a result of their good-standing. As discussed in the appendix, this is typically not sufficient to combat deviant actions.

VI. Conclusion

We presented an innovative model for detecting deviant users in online communities. Our model is based on the threat requires intent and capability philosophy. To the best of our knowledge, this is the first support tool for detecting non-compliant users in online sites. The work presented so far is the core component of a larger system, aimed at managing policy compliance in online communities. We plan to further elaborate on the architecture proposed for such system, and in particular with respect to post classification. Further, our main limitation so far is that we are unable to account for returning offenders, who appear under a different name. We will study ways to integrate this type of detection with our model.

References

[1] HaltAbuse statistics, 2012. http://www.haltabuse.org/resources/stats/index.shtml.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying video spammers in online social networks. In *4th international workshop on Adversarial information retrieval on the web*, pages 45–52. ACM, 2008.

[3] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/$ocial: the phishing landscape through short urls. In *8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 92–101, 2011.

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

[5] G. Cormack and T. Lynam. Online supervised spam filter evaluation. *ACM Transactions on Information Systems (TOIS)*, 25(3):11, 2007.

[6] M. De Choudhury, H. Sundaram, A. John, and D. Seligmann. Contextual prediction of communication flow in social networks. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 57–65. IEEE Computer Society, 2007.

[7] L. Diao, K. Hu, Y. Lu, and C. Shi. A method to boost Nave Bayesian classifiers. In *Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference*, 2002.

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

[9] M. Feldman, C. Papadimitriou, J. Chuang, and I. Stoica. Free-riding and whitewashing in peer-to-peer systems. *Selected Areas in Communications, IEEE Journal on*, 24(5):1010–1019, 2006.

[10] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *10th annual conference on Internet measurement*, IMC '10, pages 35–47, 2010.

[11] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28, 2009.

[12] N. Jindal and B. Liu. Analyzing and detecting review spam. In *7th IEEE International Conference on Data Mining (ICDM 2007)*, pages 547–552, 2007.

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

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

[15] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[16] LingPipe. http://alias-i.com/lingpipe/.

[17] B. on Line. Issues related to cyber bullies, 2011. http://www.bullyonline.org/related/cyber.htm.

[18] A. Onisko, M. Druzdzel, and H. Wasyluk. Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 27(2):165–182, 2001.

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

[20] A. Steinberg, E. Shahbazian, G. Rogova, and W. DeWeert. An approach to threat assessment. In *Harbour Protection Through Data Fusion Technologies*, pages 95–108, 2009.

[21] StopForumSpan. http://www.stopforumspam.com/. retrieved December 15th 2011.

[22] A. Sureka. Mining user comment activity for detecting forum spammers in youtube. *Arxiv preprint arXiv:1103.5044*, 2011.

[23] A. G. West, S. Kannan, and I. Lee. Stiki: an anti-vandalism tool for wikipedia using spatio-temporal analysis of revision metadata. In *6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, pages 32:1–32:2, New York, NY, USA, 2010. ACM.

[24] T. Wickens. *Elementary signal detection theory*. Oxford University Group, 2002.

[25] P. Xie, J. Li, X. Ou, P. Liu, and R. Levy. Using bayesian networks for cyber security analysis. In *International Conference on Dependable Systems and Networks (DSN)*, pages 211–220. IEEE, 2010.

[26] R. R. Yager. An extension of the naive bayesian classifier. *Inf. Sci.*, 176:577–588, March 2006.