If it Behaves Like a Troll, it is a Troll! A Computational Mechanics Approach to Trolling and its Contagion

Background

As human interactions in digital environments increasingly have become an important part of our daily activity, the concerns about the misconducts in online platforms has increased as well. One of the misbehaviors that draws most attentions is trolling behavior. Trolling behavior is considered as an antisocial online behavior that diffuse misinformation, provoke emotional responses, or disrupt on-topic discussions (Shin, 2008). Much of the work on trolling behavior has focused on its language aspect. The premises are that if someone talks like a troll, it is a troll. We argue that trolling is a complex online assaulting behavior, and therefore much more than verbal abuse. The motivations of the behavior vary from simple attention attractions (Herring, HobSlider, Scheckler, &Barb, 2002) to well calculated manipulations of public opinions (Engelin & De Silva, 2016). In addition, the effects and consequences of the behavior are also varied across audiences and content. It is a behavior to attract responses and gain influence by making time-wasting comments on some controversial topics, using provocative language and strategies, such as referring to a person, repeatedly commenting, and using misinformation. Therefore, our premise is that if someone behaves like a troll, it is a troll.

Per definition, behavior is a dynamical process. Traditionally, Communication as a field has not paid much attention to dynamic processes that unfold in time (Pool, 2007). Our methodological contribution in this study is to employ a computational method from dynamical systems theory that quantifies behavioral dynamics with the help of the unique, minimally complex, maximally predictive model of the dynamic, the so-called predictive state model (Shalizi & Crutchfield, 2001).

Step I: Preliminary Test for Trolling Patterns

We first collected several internet users' commenting behaviors to test how dynamical systems theory can be applied to trolling context. By doing that we tried to illustrate the hidden patterns of online commenting behaviors, especially trolling behaviors, which can help us to reveal how trolling behavior can influence internet users.

Methods

We use traceable digital footprint from social media and build two kinds of models. The first one is an autoregressive model that we call the self-driven model. It quantifies the minimal size, optimally predictive behavior of a user based on its past. The so-called epsilon machine (Crutchfield & Young, 1989) assumes that a user's future behavior is only influenced by its own past behavior. Consider a social media user's behavior as a series of discrete points in a time period, at any given time instant t, a user either troll or not, which denoted by $X_t,\ X_t = 1\ or\ 0$. In the self-driven model, the probability of whether a user troll or not in the future time t is determined by whether they troll in time points before t. The derived predictive state model is essentially a unifiliar hidden Markov model of the dynamic. Its hidden states consist of "a set of histories, all of which lead to the same set of futures. It's a simple dictum: Do not distinguish histories that lead to the same predictions of the future" (Crutchfield, 2017, p. 2). Mathematically, the predictive statistic of the past for predicting the future of a conditionally stationary stochastic process is a minimal sufficient statistic for prediction. Thus, for each possible predictive distribution, we could find a class of pasts that induce this predictive distribution, and we can find a statistic that can map a past into an equivalence class for that past. In other words, the states of the unifiliar hidden Markov model represent a partition for all pasts based on the conditional futures they induce.

Our second model expands the autoregressive logic to include a second parallel time series that is considered to influence the ongoing dynamic. Following the modeling approach of computational mechanics, if the epsilon machine is finite state machine that computes the user's future behavior based on its past behavior, we now consider an input-output transducer, that computes the user's future behavior based on its past and on the influence of an external source (Sipser, 2006). Such input-output predictive state models have been called epsilon-transducer (Barnett & Crutchfield, 2015; Darmon, 2015). In our case, we refer to it as the social-driven model. It assumes that a user's future behavior is influenced by both their past behavior and the past behavior of people they interacted with. It aims at capturing contagion.

Previous research has shown that ordinary people can be triggered into engaging in trolling behavior by discussion content and emotions (Chang, Bernstein, Danescu-Niculescu-Mizil, & Leskovec, 2017). People could get affected by prior online behaviors that both trolling in the past and participating in a discussion having previous trolling comments raise the likelihood of future trolling behavior, and such behavior can persist across those affected people to spread further. As such, trolling behavior could transmit and spread from person to person in the process of discussion. Thus, additionally to the user's behavior X_t, we introduce a new variable of social inputs Y_t, and in our case, Y_t indicates the parent comment.

$$Y_t = \begin{cases} 1, & if\ the\ parent\ comment\ is\ troll\ messages \\ 0, & otherwise \end{cases}$$

Then we have P $(X_t = x_t | X_{t-1} = x_{t-1}, Y_{t-1} = y_{t-1})$.

As in self-driven model, we derive the unique model with minimal complexity and maximal prediction power for a user's behavior. The logic of the resulting hidden Markov model is the same as self-drive model, just that we now have input and output symbols on the transitions (just in ordinary transducer from computational theory). Equivalence classes are identified over joint self and social pasts and a mapping partition from the current joint past to its equivalence class is detected.

Both self-driven and social-driven model helps to detect hidden patterns for past and future, whereby we could apply those tools from computational mechanics to answer the following questions: (1) Is the trolling behavior self-activated or situational? (2) How much is trolling behavior contagious? (3) Is there any hidden pattern of trolling behavior?

We accessed Reddit post history through Google Big Query and selected the 15 most active users from the hundred most popular posts in political subreddit on May 17, 2018, resulting in 15,558 comments with all parent comments (the comment which is the first one in a thread). Seven undergraduate students worked as human coders for categorization of troll message. By going through 4 rounds of training and discrepancy solving, the target user comments and the parent comments were labeled as "troll message" and "non-troll message", respectively, 1 and 0, with an inter-coder reliability of 83%.

Because the self-driven and socially-driven models assume that users' online behavior can be modeled as a conditionally stationary stochastic process, the distribution over futures is independent of the time index conditional on the observed past. To approximate the assumption, we consider each comment as a unique temporal point. For both kinds of predictive state models, we will use the Causal State Splitting Reconstruction (CSSR) algorithm (provided by Darmon, 2015) to infer the models from the data. Considering that the majority of the comments are not replying to the same parent post, we distinguish among predictive states if they have a similarity less than $\alpha = 0.001$, and we go two time-steps into the past to make predictions, $L_{max} = 2$.

Results

Figure 1 shows the representative topologies of the autoregressive predictive state models. Among the 15 users that we tested, 8 users only have one predictive state, 6 users have two causal states and 1 user have three states. The number of predictive states of an $\epsilon-\text{machine}$ provide a rough reflection of the complexity of the user's behavior, because each causal state is a "further refinement of the past for predictive sufficiency" (Darmon, 2015). Since our most complex model has 3 states, it is possible that self-driven trolling behavior is not a complex process. It basically switches from routines of trolling to non-trolling and *vice versa*.

Figure 1 (a) shows the $\epsilon-\text{machine}$ for one of the one-state users. For those users, troll or not troll is like flipping a fair coin that it is not a decision made based on their own past behavior. Whether they acted as a troll is random or, at least, it cannot be detected from self-driven model. There is no obvious pattern detected from their own behaviors. In Figure 1 (b) is the $\epsilon-\text{machine}$ for one of the two-state users. Their behavioral pattern distinguishes among two states that can be considered as a troll state (A) and a non-troll state (B), and they process memory for both stages: as they switch from non-troll to troll, they are more likely to remain troll, and *vice versa*. It is interesting to note that it is quite as likely to switch from trolling to non-trolling (31% and 32.7%). In addition, the unique three state user's epsilon-machine is illustrated in Figure 1 (c). There is a stage in addition to the troll (C) and the non-troll (B) state, and we call it transition state (A). With a 56.7 % chance, the user will advance to trolling from this state, and with 43.3 % chance to the non-troll state. Once the user arrives at the troll state, they will stay there with a probability of 76%. Note that the transition state (A) cannot be reached from the troll state (C) directly. It seems that the transition state serves the purpose of a threshold for becoming a troll or not. Being in a non-trolling state, the probability of staying there is quite high, 65.6 %. Being in the trolling state, the probability of staying a troll is even higher, 76%. Being in the transition state, thresholds are more evenly distributed, and the user seems to be in an uncertain and unstable situation of becoming a troll or not.

Moving to the predictive state model that incorporates social inputs, we notice that all users can be well-described by epsilon-transducer with 2 or 3 states, which is analogous to the epsilon-machine architecture. The knowledge of the recent past of both their own and their social input (parent comment) behaviors provides sufficient information for predicting their future behavior. Figure 2 (a) highlights a two-state user's epsilon-transducer. When the user's own previous behavior is trolling behavior, the user is more likely to switch to troll state, and if the user is replying to a troll parent comment, it will reinforce the switch to the troll state and the user will stay in that state. Figure 2 (b) is the epsilon-transducer for a user with three states. The user exhibits both self and social memory in the sense captured by the model. There are two routes from non-troll state (A) to troll state (B), one is directly from non-troll state to troll state, and the other is through transition state (C). Unlike transition state in three-state epsilon-machine, the user can switch from the transition state to both troll or non-troll behavior.

Furthermore, comparing the epsilon-machine and the epsilon-transducer of a sample user, we see that the epsilon-transducer captured additional information (see Figure 1 (a) and Figure 2 (b)). Although it seems that the user's behavior is haphazard based on their own previous behavior, the social-driven model reveals that there is a complex hidden pattern guiding the behavior. Both self and social inputs predict the future behaviors that trolling behavior can be self-motivated, social- motivated, and self-social-together-motivated. Users can be influenced by parent comments, and a previous troll may lead to more trolls in the platforms.

Step II: Data Collection & Trolling Classification

While insights from the two models for all history comments from the 15 users was insightful, the step II consists in collecting the behavioral patterns from a much larger dataset. The goal is the creation of a behavioral taxonomy of trolling behavior, quantified by computational methods from dynamical systems theory.

Data Collection & Cleaning

Based on the hundred most popular posts in political subreddit on May 17, 2018, all users participated in those posts were considered recent active Reddit users. History posts and comments from 22,583 active users detected were collected from Google BigQuery. More than 18,000,000 comments were collected with all parent comments. Some criteria were used to further filter the dataset: 1) any comments with deleted parent comment and were disqualified because it would be difficult to interpret social influence from deleted parent comment; 2) all non-English comments were disqualified as well; 3) all comments and parent comments with only pictures or emojis were disqualified as they could not be classified.

Total about 17,000,000 comments still remain in the dataset. We further proceed into preprocessing the dataset based on the requirement for obtaining vectors representations for words. We first deleted all URLs, ip addresses and emojis. And then abbreviations were replaced by proper words, and some spelling correction were done. Some special symbols were replaced by words, for instance, “&” was replaced by “and”, and “@” was replaced by “at”. In addition, we noticed some features can be good indications for emotions such as all-capital-letter words and elongated words, so we keep all those features.

The original 15 users’ history comments were taken as training dataset, in addition to human coding, sentiment and big five emotions were extracted by using IBM Watson Tone Analyzer for building machine learning classification model.

Trolling Classification

We built a model to predict if a comment is a trolling comment or not. We first converted words to vector by using GloVe, Global Vectors for Word Representations (Pennington, Socher, & Manning, 2014), which is performed basing on aggregated global word-word cooccurrence. Then the vectors were used to build the vocabulary and encode the sentences in comments by applying InferSent method (Conneau, Kiela, Schwenk, Barrault, & Borders, 2017).

A multi-layer perceptron (MLP) classifier to predict trolling comments. Previous work related to trolling identification or toxicity identification used a variety of approaches, including MLP approach (Alorainy, Burnap, Liu, &Williams, 2018). We adapted from Alorainy, et. al. (2018) and used 20 hidden layers and an activation function of relu. A 10-fold cross validation was used, for each iteration, 14,000 comments were chosen to be the testing set. The MLP classifier achieved an F1-score of 89.06%. To further explore the quality of the MLP classifier, we compared the trained labels with the original labels. Anger from big five emotions and sentiment were found highly important in the classification process. Most comments that were classified as trolling message has negative sentiment and higher score for anger. Then the trained model was applied to the larger dataset for classification. In total, about 5,000,000 comments were classified as trolling messages. All comments were rated for big five emotions and sentiment by using IBM Watson Tone Analyzer.

Step III: Identify Trolling Patterns in Large Scale Dataset

Traditional approaches to trolling detection focused on the latter (language aspect), while our approach focused on the former (behavior). As we indicated from small sample size, a comprehensive approach to the definition of trolling will eventually have to consider both aspects. In Step III, we are feeding epsilon-machine and epsilon-transducer with the larger dataset. By look at the hidden patterns of internet users' behavior, the dynamic of how emotions, attitude and behavior diffuse on internet can be studied. In addition, during the process of classification, we identified a number of bots that were suspiciously promoting certain ideas. Insights can be learned from looking at bots' behavior patterns and more can be studied about how those bots can influence people's attitude and behaviors, especially in political communication context.

At the time of paper presenting, only 1103 users have been analyzed. Majority of the epsilon-machines for analyzed users' trolling behavior is one state machine, while 31% of the analyzed users have 2 states, and only 4.99% of the users have 3 states. The distribution is similar with the small sample. Moreover, 62.01% of the users' epsilon-transducer has 2 states and 36.99% of the users' epsilon transducer has 3 states. Based on those preliminary results, a troll's behaviors can influence others, and trolling behaviors are contagion.

References

Alorainy, W., Burnap, P., Liu, H., & Williams, M. (2018). Cyber Hate Classification:'Othering'Language And Paragraph Embedding. arXiv preprint arXiv:1801.07495.

Barnett, N., & Crutchfield, J. P. (2015). Computational Mechanics of Input–Output Processes: Structured Transformations and the \epsilon -Transducer. *Journal of Statistical Physics*, 161(2), 404–451. https://doi.org/10.1007/s10955-015-1327-5

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. arXiv preprint arXiv:1702.01119.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017) *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*

Crutchfield, J. P. (2017). The Origins of Computational Mechanics: A Brief Intellectual History and Several Clarifications. *ArXiv:1710.06832* [Cond-Mat, Physics:Nlin]. Retrieved from http://arxiv.org/abs/1710.06832

Crutchfield, J. P., & Young, K. (1989). Inferring statistical complexity. *Physical Review Letters*, 63(2), 105–108. https://doi.org/10.1103/PhysRevLett.63.105

Darmon, D. (2015). Statistical methods for analyzing time series data drawn from complex social systems (Doctoral dissertation, University of Maryland, College Park).

Engelin, M., & De Silva, F. (2016). Troll Detection: A comparative study in detecting troll farms on Twitter using cluster analysis.

Herring, S., Job-Slider, K., Scheckler, R., & Barab, S. (2002). Searching for Safety Online: Managing "trolling" in a Feminist Forum. The Information Society, 18(5371-384).

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation.

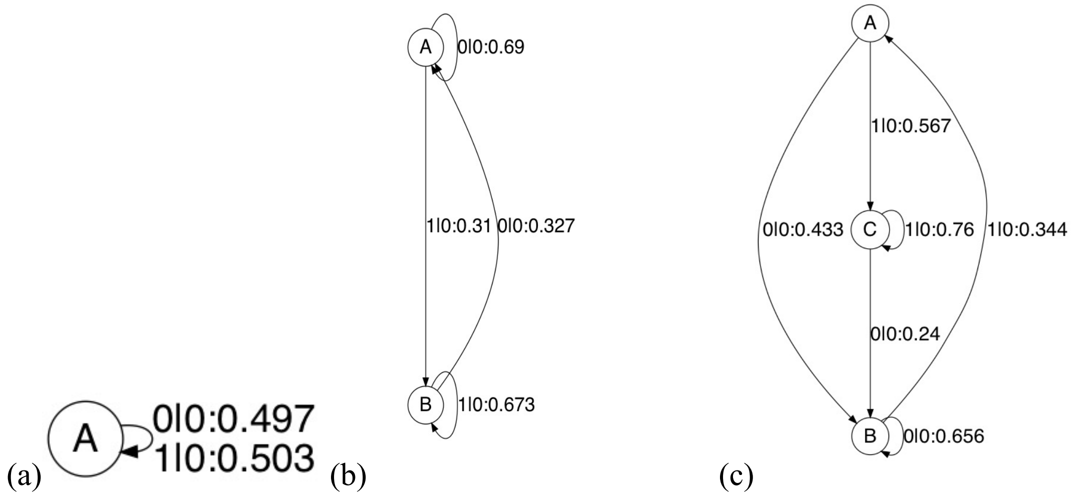
Poole, M. S. (2007). Generalization in process theories of communication. *Communication Methods and Measures*, 1(3), 181–190. https://doi.org/10.1080/19312450701434979

Shalizi, C. R., & Crutchfield, J. P. (2001). Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *Journal of Statistical Physics*, 104(3–4), 817–879. https://doi.org/10.1023/A:1010388907793

Shin, J. (2008). Morality and Internet Behavior: A study of the Internet Troll and Its relation with morality on the Internet. Social for Information Technology &Teacher Education International Conference (pp. 2834-2840). Association for the Advancement of Computing in Education (AACE).

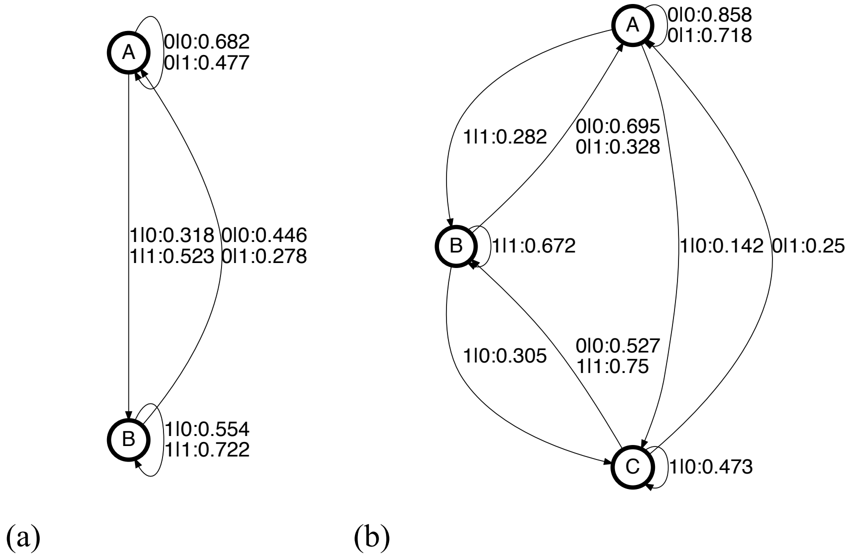
Sipser, M. (2006). *Introduction to the Theory of Computation* (2nd edition). Boston: Course Technology.

Figure 1: The ϵ − machines for the users' trolling behavior.



Notes: Note, the input is always […|0] (no social influence). (a) The one-state ϵ − machines. (b) The two-state ϵ − machines. (c) The three-state ϵ − machines.

Figure 2: The ϵ − transducers for the users' trolling behavior



Notes: (a) The two-state ϵ − transducer. (b) The three-state ϵ − transducer.