Abstract

This study investigated trolls' influence in online community by examining how individual members react toward trolls. Trolls are antisocial individuals provoking emotional responses and disrupt discussions. Using social identity theory and a dataset from YouTube, the study found out that individual members' centrality, discussion network's density, other members' previous response to trolls, and the community's cumulative response to trolls and negativity of troll posts are associated with individual members' likelihood of responding to trolls.

Who Will Reply to A Troll? A Network Approach to Understanding Trolls in Online Communities

Introduction

Trolls are online identities who behave antisocially to provoke emotional responses and disrupt on-topic discussions in online communities (Shin, 2008). The motivations of their behaviors can be as complex as manipulation of public opinions (Engelin & De Silva, 2016), or as simple as a desire to attract attention (Herring, Job-Sluder, Scheckler, & Barab, 2002). Previous research shows that unlike cyber-bullying, trolls do not purposefully target one person, but indiscriminately harass everyone and anyone who is provoked by their behavior (Lim, 2015). Some people are easily targeted and triggered by this uncooperative behavior while others are more resistant. By analyzing comment data from YouTube videos, we illustrate that network metrics of a discussion community provides a useful approach to understanding how community members react to trolls. Network metrics provide information not only about the community structure but also about individual members' involvement, emotional devotion, and attachment, which can form and reinforce the social identity that predict how they will react to trolls (Cover, 2012).

Theory and Hypotheses

Trolling has been defined differently in previous literature. An early work from Donath (1999) on trolling focused on deception and manipulation, which disseminate bad advice. Engelin et al. (2016) defined it as interrupting, harassing, or trying to impose opinions to others. In extreme case, trolls are considered as capable hackers (Lim, 2015). Those definitions indicate that trolling messages may contain misinformation, fallacies or a mixture of the truth and fake news. The intention is to manipulate public opinions and deception is one of the key

characteristics. However, deception is not the only feature of trolling behavior, and the intention of trolls is not always about the public.

Different perspective offered to look at the consequence of trolling behaviors. The definitions include causing disruption and triggering or exacerbating conflict for the purposes of their own amusement (Nevin, 2015); transgressions of community norms that result in anger, harm or discomfort (Bergstrom, 2011); and posts of erroneous or inflammatory information with the intention of provoking a strong reaction (Merritt, 2012). Those definitions emphasized on results of trolling behaviors that negatively affect people's internet use experience, reduce the chance of problem solving and mislead on-topic discussion, whereas this school of thought neglect the fact that trolling messages usually argue for fallacy, which is part of the reason that they can successfully provoke others.

Some other scholars believe that the definition of trolling should be a combination of those two sets (Klempka & Stimson, 2014). Herring et al. (2002) defined trolling as message from a sender who appear outwardly sincere; message designed to attract predictable responses of flame; and message waste a group's time by provoking futile argument. It can be outright swearing, personal attacks, veiled insults, sarcasm, and off-topic statement (Cheng, Bernstein, Danescu-Niculescu-Mizil, & Leskovec, 2017). A model was proposed by Hardaker (2010), which includes deception, aggression, disruption and success. This model described that a trolling message with negative emotion and false information that will lead to dysfunction of the discussion or unpleasant experience.

Existing research shows that trolling is a complex behavior with multiple intentions and different results. In this study, we will adopt the assembled perspective, since it considered both the characteristics and the outcome of the behavior. The trolling behavior is viewed as aggressive

behavior with the intention to disrupt the discussion and displease other users, and it usually manifests as providing fallacy regarding of the topic in-discussed, making snap judgment of other users or attacking people, and using profanity.

According to previous research, the victims of trolling behaviors are the ones with high self-involvement (Klempka & Stimson, 2014), which manifest as engaging in interactions, knowledge contributing and emotion involving. Data-driven research on trolls illustrated that how victims respond to trolling behaviors can be predicted with network features of online communities in computation models (Al-garadi, Varathan, & Ravana, 2016; Squicciarini, Rajtmajer, Liu, & Griffin, 2015). Moreover, social network structures maintain the performative acts of social identification online (Pearson, 2009). Thus, theories of social identity is introduced into social network analysis in this study to examine normal community members' reaction towards trolls.

Social identity theory usefully explains the mechanisms within individuals' online social engagement and their relationships with other internet users. A social identity is individual's self-concept generated from perceived membership in a relevant social group, which is dynamic, self-reflective and performative (Greenhow & Robelia, 2009), and it's a social product of given social environment and context (Zhao, Grasmuck, & Martin, 2008). Online community members can establish the social identity through the group affiliations, which requires time investment, emotional involvement, and frequent participation (Bergami & Bagozzi, 2000). The group affiliation enables an individual to interact with other group members, which may lead to various social networks. And the networks can provide an access to resources to fulfill one's needs (Zaheer & Bell, 2005). The social identity can also be created and enhanced through the knowledge of conventions and social acts of the online community (Ochs, 1993). Previous

research showed that an online community member's social identity is positively related with his or her knowledge contribution to the community (Shen, Yu, & Khalifa, 2010). The shared knowledge and social resource produce shared social characteristic that lead to identification, since the understandings of conventions in the group for doing particular social acts are also be shared (Ochs, 1993).

Social identity theory states that individual's behavior within a group is largely influenced by perceived social identity, which based on intergroup relationships. In online community, behavior such as updating, commenting, responding and tagging are performative acts that motivated by identification, in which perceived relationships are significant elements that connections constitute a social milieu that contextualize one's identity (Zhao et al., 2008). Reciprocally, as group affiliation and knowledge contribution build social identity, social identity motivates members to share knowledge, spend time in the community and interact with other community members, which maintain a positive self-defining relationship with other members and reinforce the attachment (Shen et al., 2010), and also facilitate the emergence of collective behaviors (Ackland & O'Neil, 2011).

Social identity is highly related to individual's social network, which is also a result of individual's conversations and contributions in community. By joining online groups, becoming fans of some products online and publishing in forums, people get involved within the boundary of an online community (Bergami & Bagozzi, 2000), which lead to cognitive centrality, in-group affection and in-group ties (Cameron, 2004). Cognitively, group affiliation is developed and enhanced, and structurally, a social network is built from behaviors engaging in online communities. Previous research has shown that identification with an online community is related to social tie strength and homophily (Brown, Broderick, & Lee, 2007). The strength of

ties is a combination of the amount of time spent together, the emotional intensity, the intimacy and the reciprocal services (Granovetter, 1973). Strong ties and network clusters with a lot of redundant ties provide social reinforcement for collective behaviors. In addition, the number of ties is also important for identification and performance. For an online community, the more local bridges it has, the more cohesive and more capable of acting in concert. For an individual in online community, it also requires contacting multiple sources of "infection" before convinced to adopt a behavior (Centola, 2010). Social identity is maintained by those online behaviors within the network structure (Pearson, 2009). The process of getting to know each other and communicating enable social influence of a community (Postmes, Spears, & Lea, 1998), while interpersonal ties with other group members and the information exchanged through those ties will benefit individuals socially and psychologically, that reinforce the group affiliation and enhance social identity (Millen & Patterson, 2003; Cameron, 2004). As Wang & Fesenmaier (2004) pointed out, interactions among members in a community are essential to community prosperity that it reflect members' commitment to it. In other words, people with more salient social identity are more likely to better involved in the cohesiveness of the network.

Thus, individuals at a more central position in network probably have a higher social identity attached to the community. Those individuals may tend to seek positive distinctiveness of an ingroup, the behavior they choose may reflect their perceived stability and legitimacy of their social identity. When facing trolls, those individuals may be more active or even aggressive, and they may tend to reply to and argue with trolls to defend their community and norms of the community. Here, we hypothesize that:

H1: Individuals with higher centrality are more likely to respond to trolls.

Furthermore, when most of community members hold high social identity, the community may be very active with a lot of interactions between members. Those high social identified members are willing to share information without caring too much about their own personal gaining, but rather care more about the interest of the group. By sharing information, they may construct a lot social ties among themselves which constitute a dense network. As the community members share the same high social identity, the community is likely be perceived as a concrete entity, which leads to more and stronger ties in the network (Sohn, 2009). Therefore, it's likely that community members have a high social identity attached to the community if it's a dense network. When facing trolls, the members tend to protect the shared interest of the group, acting defensive towards trolls.

H2: Individuals in a denser network are more likely to respond to trolls.

According to Tsai and Men (2013), identification with a community is both the cause and the indicator of members' engagement within the community. Members who engaging in active interactions are more likely to have resources and information that can be shared within the group. The engagement of members created the community dynamics that may further increase the social identification with the community. Therefore, those members with higher social identity attached to the community are likely to have higher level of activity level. When facing trolls, those members tend to engage trolls into discussions.

H3: individuals who have higher activity level are more likely to respond to trolls.

The process of identification with a community not only cognitively categorizes individuals into the community, but also decide how and what individuals learn from the interactions within the community (Peteraf & Shanley, 1997). According to social learning theory, behaviors can be acquired by direct experience as well as observations and imitations

(Bandura & Walters, 1977). Groups and communities provide the environment in which exposure to definitions, imitation of models and social reinforcement take place (Akers, Krohn, Lanza-Kaduce, & Radosevich, 1979). With external information provided by the environment of communities, individuals make decision about different performances of behavior by interaction and by observation, and learning can occur. Social reinforcement plays a role in learning as well (Bandura & Walters, 1977), it can motivate the learning process. Since people tend to imitate the one who is similar to themselves (Bandura & Ross, 1961), members of the group who shared the same identity serve as the reinforcement as well as the initiation when their behaviors are observed and learned. In an online community, individuals who identify themselves with the community can learn normative behaviors by direct instruction and observation. Direct instruction can be announced regulations or other community members' advice, and observation can be done by spend time in the community and interact with members, which also promote the social identity with the community.

Through the process of social learning, individuals gain the knowledge of normative behaviors and social desirable behaviors of the community. By interaction, group members construct the identification with the group, gain information about group norms and values. By observing others' behavior towards ingroup and outgroup, members learn what behavior will benefit the group, elevate self-esteem and change ingroup status. It's possible when an individual observes a series of behaviors and consequences regarding trolls, they learn from some behaviors and imitate them. If some community members gaining compliments because of arguing with trolls, others may mimic them and learn this behavior. And if arguing with trolls be adopted by a lot of members, people may be more willingly to follow the example because the behavior will be considered socially desired or even normative. Thus:

H4a: The more previous responses to a troll by other community members, the more likely an individual will respond to the troll.

H4b: The more previous response to trolls in a community, the more likely an individual in the community will respond to trolls.

Social interaction is a key element in learning (Hogg, 2016). Imitating behaviors from interacted members are more likely to occur, because interactions promote trust and anticipation reinforcement for behaviors. For an individual, if someone they know in the community gaining compliments because of arguing with trolls, the individual is more likely to mimic them and learn this behavior. And if a lot of friends in community argue with trolls, the individual may be more willingly to carry the same behavior.

H5a: An individual is more likely to respond to a troll if other community members who they have interacted with responded to the troll.

H5b: An individual is more likely to respond to trolls in a community if other community members who they have interacted with responded to trolls.

Social networks of online discussion have active boundary maintenance, that is informed by group norms held by everyone under the discussion (Kelly, Fisher, & Smith, 2006). Social norms are informal understanding that govern the attitude and behavior and characterize a social group and differentiate it from other social groups (Hogg & Reid, 2006). It's a shared thought that can be directly talked about and can be indirect guide of the communication within group. Social norms and normative behaviors is a way to generate positive distinctiveness that people are motivated to behave consistently with the shared understanding to get better ingroup identity (Christensen, Rothgerber, Wood, & Matz, 2004). Social norms and values that connoted for attributes and relationships within the community enhance the social identity. In certain situation

or environment, social norms serve as guidance for socially appropriate behavior. Situational factors and individual difference decide the use of social norms. For example, the salience of an outgroup may bring threatening feelings to ingroup members that are more likely to have group-based behaviors. In addition to situational factors, individual difference in social identity will control how individual take norms as appropriate behavior guide. Thus, high identified individuals may feel good about themselves if the social relationships (both ingroup relationships and outgroup relationships) congruent the group norms (Wood, Christensen, Hebl, & Rothgerber, 1997). People who identifying themselves with a group are more likely to conform the group norms.

By observing behavior of others and how much behavior is performed, group members learn how much the group approve the behavior and how it can be changed in different situations and environments. In an online community, community members learn the appropriate attitudes and behaviors towards trolls. If in the community, a lot of members reply to trolls and argue with them, this defensive behavior may be considered as normative behaviors toward trolls, other members may follow.

H6a: The more community members response to a troll, the more likely an individual in the community would respond to the troll.

H6b: The more community members response to trolls, the more likely an individual in the community would respond to trolls.

Prior research in psychology and organizational communication has showed that people respond differentially to positive and negative stimuli, negative events tend to elicit stronger and quicker emotional, behavioral, and cognitive responses than neutral or positive events (Rozin & Royzman, 2001). “There is a general bias, based on predispositions and experience, to give

greater weight to negative entities." (Rozin & Royzman, 2001, p. 296) And this bias is called "negativity bias". In general, negative entities are stronger the equivalent positive entities. People tend to think the negative events are more severe than positive events, which actually are equivalent in severity. The negativity of negative events grows more rapidly with approach to them in space or time. The negative stimuli will produce larger psychological effects than positive stimuli. The combination of negative and positive entities yields a more negative evaluation. In addition, negative entities are more varied with wide response, for instance, in linguistic, there are more vocabulary used to describe the qualities of negative events (Rozin & Royzman, 2001).

Recent research has shown that negative sentiment posts induce more feedback than positive sentiment (Stieglitz & Dang-Xuan, 2013). In online community, troll posts are negative, it's possible to generate more response than other posts. Moreover, those more negative troll posts are considered more salience that community members may find it more important to reply to them. Thus, it's reasonable to hypothesize:

H7: Troll post with higher negativity are more likely to provoke responses.

Methods

In online political discussion community, strong and irreconcilable comments from trolls receive more attention and response from people who are more at the core of the network that they try to engage trolls rather than ignore them (Kelly et al., 2006). On YouTube, communities were formed with the videos. People who are concerned or interested in a particular topic will watch related videos and engage in the discussions. They identify themselves as fans of YouTubers, of some characters from the videos or supporters of certain ideas. Behaviors of collective identity can be observed from YouTube video communities (Halpen & Gibbs, 2013).

Data were collected from top 3 political channels with comment section enabled (TYT, CNN and BBC news) and top 3 comedy channels (CollegeHumor, Annoying Orange, and PowerfulJRE). 23 video communities were selected as they all had more than 1000 comments within a month after videos were posted, and the videos lasted between 4 to 10 minutes to provide adequate information for potential discussions.

Trolling messages were identified from all comments about the videos, using vulgarity list, second person pronoun and the sentiment (Al-garadi et al., 2016; Squicciarini et al., 2015). For vulgarity word list, we collected several online bad word lists which contain offensive words and cursing words used by native English speakers in social media (Wang, Chen, Thirunarayan, & Sheth, 2014). A combination of word lists from banbuilder.com, bannedwordlist.com, noswearing.com and urbanoalvarez.es was used for trolling detection. According to previous research, second person pronoun is highly related to online anti-social behaviors (Squicciarini et al, 2015), so we consider it as a criterion to identifying trolls. Trolling messages are usually negative messages with high emotions such as anger and anxiety (Al-garadi et al, 2016; Squicciarini et al, 2015), thus, sentiment is the third criteria for troll detection. Based on the three criteria, trolling comments were identified as negative posts using second person pronouns and words from the vulgarity list.

To test the hypotheses, where social identity was formed before trolls came the communities, pre-troll community networks were captured. We first identified the first 5 trolling posts in the communities and construct networks of pre-troll communities for each troll. Since deleted comments were not included in our dataset, the entire interaction involving deleted comments were deleted from the networks. The pre-troll network of pervious trolls in a community were considered a part of network of later trolls, in this way, all pre-troll dynamics

were captured for each troll. In total, 874 unique individuals were identified with 3692 directed links (comments from one individual to another individual) from 115 pre-troll networks in 23 video communities. Within communities, the maximum population is 344 and the minimum is 11, with a mean of 160.52, and a standard deviation of 93.12. The interactions with one of the 23 communities with the first 5 trolls is illustrated in Figure 1.

Individual level variables, troll level variables and community level variables were identified from the networks. Degree centrality is the measurement to indicate how many people an individual integrated with. Especially, the measurement of degree has two dimensions. in-degree is the number of incoming links (M=0.98, SD=2.77), and out-degree is the number of outgoing links (M=0.98, SD=0.86). Individual's activity level in community is also an important indicator for engagement (M=1.59, SD=1.42). The distribution of activity level is right-skewed with a large number of 0 value. Thus, we use logarithmic transformation $x' = \log(x + 1)$ for activity level (M=1.25, SD=0.52). To test bandwagon effect, the number of responses to the troll before an individual engaging in the community (M=0.35, SD=1.02) and the number of responses to all trolls in the community before the individual interacting in the community (M=2.12, SD=2.69) is introduced. Similar to the measurement of activity level, both the number of response to troll before joining in community and the number of response in community before the individual join in the community are right-skewed with high frequency of 0 value. Logarithmic transformation was applied to both measurements (See Table 1). Interactions between individuals who respond to trolls is calculated and included into the models. For each individual, interaction with who respond to trolls is defined as they either reply to the responding individual or receive comments from the responding individual. interactions were calculated for each troll (M=0.001, SD=0.02) and each community (M=0.025, SD=0.21).

On troll level, variable of the total number of responses to the troll is included (M=0.61, sd=1.60), as well as the total number of individuals in troll level networks (network size, M=47.93, sd=25.24), and the density of the networks (M=0.43, sd=0.19). The cumulative number of responses to the troll is found right-skewed with 0 value, so logarithmic transformation is applied (M=0.37, SD=0.79) In social network, density is correlated with network size (r=-0.73, $p< .001$). A larger network will be sparser than a smaller network. To be able to independently test the effect of density, we used a simple regression for density and network size to get the residuals as the measure of network density. In addition, the sentiment scores for troll posts like negativity score (M=0.82, SD=0.16) is calculated. The score is based on machine learning tools for natural language processing (Kaur & Chopra, 2016; Hans, & Mnkandla, 2016). The negativity score is the probability of negativity in the text. It's ranged from 0 to 1. A higher score means more negative and higher emotion energy.

On community level, the total number of responses to troll in the community (M=3.38, SD=3.18) was introduced into the analysis. Again, a logarithmic transformation is applied to the cumulative number of response in the community to address the right-skewed distribution with 0 value. All variables of the three levels were used to predict whether an individual will respond to trolls (with response rate of 0.02) and how often they respond to trolls (M=0.03, sd=0.28). The response rate is low, but we believe it capture the feature of normal online communities.

Results

Mixed effect hierarchy logistic regressions were completed to determine the relationship between response, a dummy variable indicating whether an individual respond to the troll, and three level variables (in-degree, out-degree, activity level, network size, density residuals, negativity, cumulative response to troll, response to troll before individual, cumulative response

in community, response in community before individual, interactions with individuals who replied to the troll and interactions with individuals who replied to trolls in the communities.). The models include nominal variables for each troll and for each community as random effects. The results are presented in Model 1 and 2 of Table 2. Moreover, mixed effect hierarchy Poisson Regressions were completed to test the relationship between weight, a count variable indicating how many times an individual reply to troll, and independent variables (Table 3). For Poisson Regression, only relied individual will be considered into model and the weight excluding the first response was the dependent variable in the models. The reason for doing so is that a large proportion of the individuals respond at most once, the logistic regression and the Poisson Regression will be highly correlated. The result is presented in Model 3 and 4 in models.

The first hypothesis was tested to investigate whether an individual's centrality is associated with the probability of response to trolls. Controlling for network size, in-degree was negatively related to response (b=-0.61, odds ratio=0.54, $p<0.01$) and out-degree was positively related to response (b=0.52, odds ratio=1.68, $p<0.01$), both dimensions of degree centrality showed a significance in predicting possibility of response, supporting H1. However, the Poisson Regression showed only a negative relation between out-degree and weight (b=-1.59, z=-2.94, $p<.01$). H2 was tested to examine whether a network's density is influential to the likelihood of response. Controlling for network size, density is positively associated with response (b=0.19, odds ratio=1.21, $p<0.05$), supporting H2. We found no significant evidence indicating the relationship between weight and network density (z=0.88, p=0.38). H3 tested whether an individual's activity level is associated with response. We found no significant relation between individual's activity level and the likelihood of response (z=1.43, p= .15), H3 is not supported,

whereas activity level significantly predicts how many times an individual reply to trolls (b=2.14, z=8.32, p<.001).

H4 tested how people's behavior is affected by previous members in troll networks and communities. H4a investigated whether previous response from others to the troll is associated with the probability of response to the troll. H4b examined whether prior responses to all trolls from other community members is associated with the likelihood of response. Controlling for network size, previous response in troll network is non-significant associated with response (z=-1.89, p=0.059), so H4a is not supported. And previous response in community is positively associated with response (b=0.34, odds ratio=1.09, p<.01), supporting H4b. Furthermore, the Poisson Regression showed no significant association between the number of responses and the number of previous response to a certain troll or the number of previous response in community. H5 tested whether the probability of response will be affect if some members an individual interacted before reply to trolls. H5a examined in a network, whether an individual's likelihood of response will be influenced if people who interacted with a individual has reply to the troll. H5b examined in the community, how is the likelihood of response be influenced by replied people who has interacted with the individual before. We found no significant evidence supporting H5a (z=-0.04, p=0.97) and H5b (z=1.54, p=0.12).

H6 tested how people's behavior is affected by the whole network and community. H6a tested whether cumulative response to a certain troll in the community is associated with the probability of response to trolls. Moreover, H6b tested whether cumulative response to all trolls in the community is associated with the response. Both H6a and H6b are supported with cumulative response in troll network (b=1.93, odds ratio=6.87, p<.001) and cumulative response in community (b=1.02, odds ratio=2.78, p<.001) positively related to response. While in Model 4

only cumulative responses in the community showed a positive relation with the number of response (b=2.67, z=2.08, p<.05). H7 tested how negativity is associated with response. Negativity is positively related to the likelihood of response (b=2.88, odds ratio=17.62, p<.001).

Discussion

Trolling behavior is a kind of aggressive behavior with the intention to disrupt the discussion and irritate other users, by providing misinformation regarding of the topic in-discussed, making misleading judgment, attacking others, and using profanity (Donath, 1999; Bergstrom, 2011; Cheng et al., 2017; Merritt, 2012; Herring et al., 2002). Trolling behavior may lead to severe consequences like dysfunction of communities and retreating from communities. By analyzing trolls in the discussion communities on YouTube, we focus on how ordinary community members' reactions towards trolls. We examined the likelihood of being provoked as a function of individual attributes such as individuals' network centrality, community attributes, including communities' density and collective behavior; and the trolling posts' negativity. Degree centrality showed a significant association with response to trolls as well as the frequency of response. However, activity level is not significantly associated with response, but only positively associated with the frequency of response. Network density is positively associated with response, as same as previous response in community, cumulative response in troll network and in community. While, previous response in troll network, interaction with replied members in troll network and in community are not significantly associated with response. Negativity of the troll post is also positively associated with response.

The first implication from these findings is that individuals who are more central in a dense community are more likely to be provoked by trolling behaviors. In previous research, emphasis is more on the network position of trolls in the network (cite), while as we

hypothesized the community's structure and individual's position in the network is also important for predicting the consequences of trolls. The relationships between the likelihood of responding to trolls and individual's network characteristics are explained by social identity approach. In general, people who has higher centrality are more likely to have higher social identity and high social status in the community, and their behaviors are governed by the sense of ingroup distinctiveness. They are more willingly to defend their group norms and normative behaviors. Respectively, community members who have higher out-degree centrality are people with higher social identity that are more active to share their information and ideas about the community. It's more likely for them to argue with trolls to try to persuade trolls with group norms. On the other hand, people with higher in-degree centrality are those receive more messages from other community members. By receiving information, they learn the group norms and values. They are more likely to be the new coming members or members with low social identity, who are less likely to argue with trolls.

Second, the activity level is associated with the probability of response but positively associated with the frequency of response once the individual respond to trolls. One potential explanation is that a high activity level cannot guarantee a high social identity or high ingroup status. It's possible that people who has high engagement are in low ingroup status with low social identity that engaging in the community is a individual mobility strategy to elevate status in the group. They may not have solid understanding of the group norms and values; thus, they don't care about trolls. However, when an active individual has already respond to trolls, it is more likely that this individual has higher social identity attached to the community. The more active they are, the more likely for them to maintain the positive distinctiveness of ingroup. So, it's probably for them to argue with trolls.

Third, network density of an community before troll comes is also associated with members' response to trolls. Members of dense network tend to respond to trolls. In a dense network, members are more connected with each other and the whole network is considered as an concrete entity with shared social identity and social norms. Members of dense network may be more likely to resist the attack of trolls, because they have beliefs in their group values.

Fourth, other members' previous behaviors are also associated with the likelihood of response to trolls. When there are a lot of records of responding to trolls in the community history, individuals tend to respond to trolls as well, whereas the record of responding to a certain single troll doesn't have the same effect. One possible explanation is that community members tend to consider trolls as an outgroup entity without differentiating each, so the cumulative records in the community may stimulate individuals to imitate the behavior, while the previous response to a certain single troll may only be considered as part of the response that is not as valuable. The more previous response in a community, the more likely for members to perceive responding to trolls as normative behaviors that members are more willingly to respond to trolls. Furthermore, unlike our expectation, interactions between members didn't show significant association with response to trolls. One possible reason for that is that the mean density of the networks is small, and there are not so many ties to test the effect of interactions.

Fifth, responding to trolls can also be considered as collective behavior that driven by social norms. cumulative responses to trolls in the community and cumulative responses to a certain troll are both positively associated with the likelihood of response to trolls. social norms govern people's behavior in different situation and environment, when facing trolls, more response to trolls in the community and more response to a troll means the responding to trolls is socially approved behavior in the community that consisting norms. Community members tend

to follow those normative behaviors when they are in the same situation. Similarly, since replying to trolls is considered normative, when an individual has already responded to trolls, they tend to engage themselves into conversations with trolls as following normative attitude towards trolls.

Finally, post sentiment is also associated with the likelihood of responding. The more negative a troll post is, the more likely the post will receive response. More negative post will be viewed as more severe that community members may feel more urgent to correct. In addition, more negative post has more emotion energy to trigger others, which lead to more community members get emotional involved and respond to trolls.

In conclusion, this study makes important contribution by connecting social identity approach with social network in communities to understand how normal community members interact with trolls. We also provided empirical evidences for the individual level, the community level and the post level influential factors on likelihood of responding to trolls and frequency of responding to trolls. The statistical analysis provided significant evidences for individuals' centrality, network density, previous response, cumulative response and negativity.

Limitation and future research

The first limitation is that, the captured communities in our dataset are all political orientated even some communities are from comedy channels. The content helped generated more trolls but may only able to capture some characteristics of the online communities. Future research can be conducted for different contents and platforms. Second, the dataset is cross-sectional data. Although in our analysis, each network is constructed by the time of troll appearing, we didn't capture the dynamics of community formation, which may have important influence on the environment of community. Third, since there is no well-developed definition of

trolls, the categorization of trolls may not fit other communities. Further research can be done for the definition of trolling behaviors, which may bring better understanding of thus online antisocial behavior. In addition, the natural language processing approach used here entails some limitations. To achieve a high accuracy, natural language processing required certain amount of words, while some comments analyzed here contain fewer words, which may result in less accurate negativity scores.

The study is also limited by the characteristics of the dataset. In general, all the networks have relatively low density with little variation. In addition, the captured networks are not highly active networks with low total response rate to trolls. Future research could collect discussion networks with more variation.

Given the findings of our research, future studies can be done for some insignificant variables like interactions with replied members. Also, the dynamic of the community can be taken into consideration, for instance, the history of community members, the lifetime of communities. More detailed sentiments, like discrete emotions can be studied to understand the emotional disclosure and interactions with trolls in online community.

Reference

Ackland, R., & O'neil, M. (2011). Online collective identity: The case of the environmental movement. Social Networks, 33(3), 177-190.

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433-443.

Akers, R. L., Krohn, M. D., Lanza-Kaduce, L., & Radosevich, M. (1979). Social learning and deviant behavior: A specific test of a general theory. American sociological review, 636-655.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. The Journal of Abnormal and Social Psychology, 63(3), 575.

Bandura, A., & Walters, R. H. (1977). Social learning theory (Vol. 1). Englewood Cliffs, NJ: Prentice-hall.

Bergami, M., & Bagozzi, R. P. (2000). Self-categorization, affective commitment and group self-esteem as distinct aspects of social identity in the organization. British Journal of Social Psychology, 39(4), 555-577.

Bergstrom K (2011) ''Don't Feed the Troll'': shutting down the debate about community expectations on Reddit.com, First Monday 16(8) Available at: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3498/3029, (accessed 6 September 2011).

Brown, R. (2000). Social identity theory: Past achievements, current problems and future challenges. European journal of social psychology, 30(6), 745-778.

Cameron, J. E. (2004). A three-factor model of social identity. Self and identity, 3(3), 239-262.

Centola, D. (2010). The spread of behavior in an online social network experiment. science, 329(5996), 1194-1197.

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.

Christensen, P. N., Rothgerber, H., Wood, W., & Matz, D. C. (2004). Social norms and identity relevance: A motivational approach to normative behavior. Personality and Social Psychology Bulletin, 30(10), 1295-1309.

Cover, R. (2012). Performing and undoing identity online: Social networking, identity theories and the incompatibility of online profiles and friendship regimes. Convergence, 18(2), 177-193.

Donath, J. S. (1999). Identity and deception in the virtual community. Communities in cyberspace, 1996, 29-59.

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

Granovetter, M. S. (1973). The strength of weak ties. American journal of sociology, 78(6), 1360-1380.

Greenhow, C., & Robelia, B. (2009). Informal learning and identity formation in online social networks. Learning, Media and Technology, 34(2), 119-140.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. Computers in Human Behavior, 29(3), 1159-1168.

Hans, R. T., & Mnkandla, E. (2016, November). Work in progress—Design and development of a project management intelligence (PMInt) tool. In Advances in Computing and Communication Engineering (ICACCE), 2016 International Conference on (pp. 308-313). IEEE.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions.

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing" trolling" in a feminist forum. The Information Society, 18(5), 371-384.

Hogg, M. A. (2016). Social identity theory. In Understanding peace and conflict through social identity theory (pp. 3-17). Springer, Cham.

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. Communication theory, 16(1), 7-30.

Kaur, A., & Chopra, D. (2016, September). Comparison of text mining tools. In Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2016 5th International Conference on (pp. 186-192). IEEE.

Kelly, J. W., Fisher, D., & Smith, M. (2006, May). Friends, foes, and fringe: norms and structure in political discussion networks. In Proceedings of the 2006 international conference on Digital government research (pp. 412-417). Digital Government Society of North America.

Klempka, A., & Stimson, A. (2014). Anonymous Communication on the Internet and Trolling.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. Computers in human behavior, 19(3), 335-353.

Lim, K. (2015). They do it for the lulz: Examining Trolls in the Context of YouTube, Taiwan, MA.

Merritt, E. (2012). An analysis of the discourse of Internet trolling: A case study of Reddit. com (Doctoral dissertation).

Millen, D. R., & Patterson, J. F. (2003, April). Identity disclosure and the creation of social capital. In CHI'03 extended abstracts on Human factors in computing systems (pp. 720-721). ACM.

Nevin, A. D. (2015). Cyber-Psychopathy: Examining the Relationship between Dark E-Personality and Online Misconduct (Doctoral dissertation, The University of Western Ontario).

Ochs, E. (1993). Constructing social identity: A language socialization perspective. Research on language and social interaction, 26(3), 287-306.

Pearson, E. (2009). All the World Wide Web's a stage: The performance of identity in online social networks. First Monday, 14(3).

Peteraf, M., & Shanley, M. (1997). Getting to know you: A theory of strategic group identity. Strategic Management Journal, 165-186.

Postmes, T. T., Spears, R., & Lea, M. (1999). Social identity, normative content, and" deindividuation" in computer-mediated groups.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. Personality and social psychology review, 5(4), 296-320.

Shen, K. N., Yu, A. Y., & Khalifa, M. (2010). Knowledge contribution in virtual communities: accounting for multiple dimensions of social presence through social identity. Behaviour & Information Technology, 29(4), 337-348.

Shin, J. (2008, March). Morality and Internet Behavior: A study of the Internet Troll and its relation with morality on the Internet. In Society for Information Technology & Teacher Education International Conference (Vol. 2008, No. 1, pp. 2834-2840).

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015, August). Identification and characterization of cyberbullying dynamics in an online social network. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 280-285). ACM.

Sohn, D. (2009). Disentangling the effects of social network density on electronic word-of-mouth (eWOM) intention. Journal of Computer-Mediated Communication, 14(2), 352-367.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. Journal of management information systems, 29(4), 217-248.

Tsai, W. H. S., & Men, L. R. (2013). Motivations and antecedents of consumer engagement with brand pages on social networking sites. Journal of Interactive Advertising, 13(2), 76-87.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014, February). Cursing in English on twitter. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 415-425). ACM.

Wang, Y., & Fesenmaier, D. R. (2004). Towards understanding members' general participation in and active contribution to an online travel community. Tourism management, 25(6), 709-722.

Wood, W., Christensen, P. N., Hebl, M. R., & Rothgerber, H. (1997). Conformity to sex-typed norms, affect, and the self-concept. Journal of personality and social psychology, 73(3), 523.

Zaheer, A., & Bell, G. G. (2005). Benefiting from network position: firm capabilities, structural holes, and performance. Strategic management journal, 26(9), 809-825.

Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. Computers in human behavior, 24(5), 1816-1836.

Table 1. *Descriptive statistics (N= 3692)*

Variable	*M*	*SD*
Individual level		
Indegree	0.98	2.77
Outdegree	0.98	0.86
Activity level	1.59	1.42
Activity level(log)	1.25	0.52
Number of previous responses to the troll	0.35	1.02
Number of previous responses to the troll(log)	0.25	0.61
Number of previous response in community	2.12	2.69
Number of previous response in community(log)	1.13	1.20
interaction with response to the troll	0.001	0.02
interaction with response member in the community	0.025	0.21
Troll level		
Number of responses to the troll	0.61	1.60
Number of responses to the troll (log)	0.37	0.79
Density	0.03	0.0
Density residual (standardized)	7.09E-18	1
Negativity (%)	0.83	0.16
Community level		
Number of response in community	3.38	3.18
Number of response in community (log)	1.66	1.23
Control variables		
Number of individuals in the network	47.93	25.2
Dependent variables		
response		2.38%
weight	0.034	0.28

Table 2. *Logistic Regression Models Predicting the Likelihood of Response to Troll*s

	model 1				model 2			
variable	*B*		*SD*	*OR*	*B*		*SD*	*OR*
Outdegree	0.518	**	0.173	1.679	0.515	**	0.159	1.673
Indegree	-0.605	**	0.197	0.546	-0.648	**	0.212	0.523
Cumulative response	1.921	***	0.200	6.828				
Previous response	-0.310	.	0.169	0.733				
interaction with response members	-14.632		362.039	0.000				
Cumulative response in community					1.048	***	0.190	2.851
Previous response in community					0.332	**	0.116	1.394
Interaction with response members in community					0.514		0.334	1.671
Negativity	1.079		0.971	2.940	2.797	**	0.858	16.401
Activity level	0.451		0.320	1.570	0.428	.	0.260	1.534
Network size	-0.038	**	0.012	0.963	-0.064	***	0.011	0.938
Density residual	0.190	*	0.078	1.209	0.084		0.063	1.088
Model AIC	422.4				610.2			

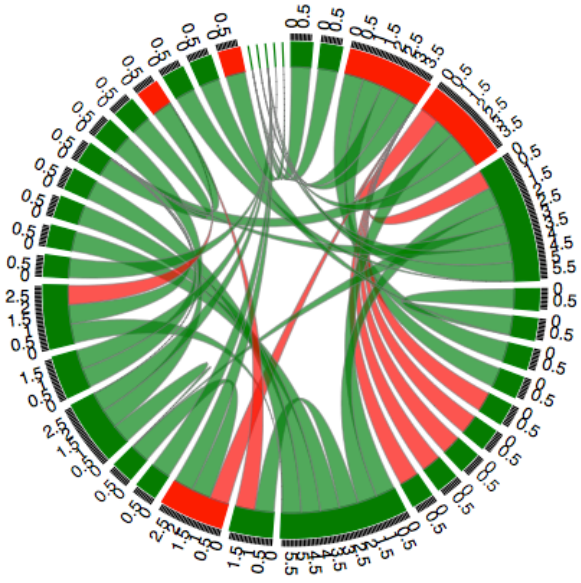
‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 3. *Poisson Regression Models Predicting the Frequency of Response to Troll*s

	model 3				model 4			
variable	*B*		*SD*	*z*	*B*		*SD*	*z*
Indegree	0.030		0.484	0.06	0.290		0.639	0.45
Outdegree	-1.599	**	0.545	-2.94	-2.406	***	0.726	-3.32
Cumulative response	-0.280		0.394	-0.71				
Previous response	0.362		0.335	1.08				
Cumulative response in community					2.665	*	1.282	2.08
Previous response in community					-0.107		0.252	-0.42
interaction with response members in community					1.651		1.342	1.23
Negativity	-0.481		1.308	-0.37	-0.744		1.136	-0.66
Activity level	2.142	***	0.258	8.32	2.569	***	0.382	6.73
Network size	0.014		0.022	0.64	0.025		0.023	1.06
Density residual	0.094		0.108	0.88	0.076		0.108	0.70
Model AIC	95.8				91.6			

‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’1

Figure 1. *Members and comments in community 1*



Note. Each bar represents an individual in community. The red bars represent trolls. The wide of the bar reflect the outdegree. The strip linking two bars represents a communication between those two individuals. The wide end of the strip is the individual who make the comment, and the narrow end is the individual whose post is commented.