A Model of Online Social Interactions based on Sentiment Analysis and Content Similarity

Claudiu-Cristian Musat, Boi Faltings
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
Firstname.Lastname@epfl.ch

Philippe Roussille
ÍRIT laboratory
Toulouse, France
philippe.roussille@irit.fr

ABSTRACT

In this paper we create a model of human behavior in online communities, based on the network topology and on the communication content.

The model contains eleven distinct hypotheses, which validate three intuitions. The first intuition is that the network topology alone fails to clearly distinguish between the users who contribute to the community and the troublemakers. The second intuition is that the content of the messages exchanged in an online community can separate good and insightful contributions from the rest. The third intuition is that there is a delay until the network stabilizes and until standard measures, such as betweenness centrality, can be used accurately. Taken together, these three intuitions are a solid case against indiscriminately using network measures. They also underline the importance of the communication content. We show that the sentiment within the messages, especially antagonism, can significantly alter the community perception. We create a novel sentiment analysis technique to identify antagonistic behavior.

We use real world data, taken from the Slashdot[1] discussion forum to validate our model. All the findings are accompanied by extremely significant t-test p-values.

I INTRODUCTION

Information processing has become the backbone of modern society and the analysis of communication and its role in modeling interpersonal relations has a central role. So far, it has focused on objective content expressed in numbers and facts. However, human communication is much richer and many of its facets, including the expressed sentiment, have received comparatively little attention.

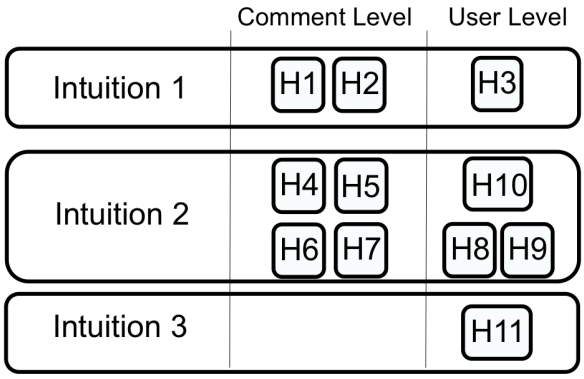

Figure 1: Behavior model hypothesis distribution

With the ever growing popularity of social networks, much research has been dedicated to understanding their inner mechanisms. Social network analysis (SNA) views social relations in terms of network theory with social actors as network nodes and social relations as links between them. The intuition is that a link between two individuals is a form of endorsement. By aggregating all the connections of a social actor one can thus obtain an indication of her value within the community.

Most of the research analyzes the implications of aggregating social links and not on what they represent. Various types of user centrality are computed with little regard to the underlying significance of how the network is formed. While undoubtedly a user's centrality is, to some extent, connected to her social importance, we believe complementary information extracted from the communication content can improve the quality of the overall picture.

Not all social interactions are positive, thus not all network links are a form of endorsement. Network measures sometimes provide good results, but in spite of this. Working with just the final value of a user's centrality is an approximation of her behavior. Potential explanations for this resilience abound. For instance, one might believe that, after an unacceptable behavior, the perpetrator is ostracized, thus ending

[1] www.slashdot.org

up with a low centrality value. This might happen although, in the short term, people have noticed that low quality users thrive on negative feedback. This led to the saying *do not feed the trolls* - noting that a short term centrality boost is what those users are after.

To the best of our knowledge, these intuitions have not been investigated. In this paper, we present a model which connects the user behavior, extracted from their communication content, with the network topology. We name this approach a user level analysis. We also perform an analysis at the communication, or comment, level and hypothesize on what affects the community response to a message. We create 11 hypothesis, pertaining to either a comment level or a user level analysis, which are grouped behind three intuitions, shown in Figure 1. We test these hypothesis using real world data from the forum of the technology news website Slashdot, where users comment on news items.

We group our findings in three intuitive classes. A first intuition is that *standard topological analysis, based on explicit reply-to links, cannot clearly distinguish valuable communication instances from potentially damaging ones*. A similar observation can be made for the users who generate the communications. We will show that user centrality cannot discriminate between malicious users, or *trolls*, and users that bring substance to the discussions. Inconsequential, average users can, however, be clearly separated from the rest using the network topology alone.

A second intuition is that *additional information, based on the content of the communication, can separate useful communications and users from the rest*. We provide two conceptually different alternatives. Within a first one, we link the relevance of an actor's communication to her social importance. When discussing a given topic (e.g. a news item), the closer the message is, semantically, to the original news item, the better its chances of receiving positive community feedback.

In our most extensive analysis we will show that the sentiment within the user created comments can discriminate between network links that represent endorsement and links that do not. This insight can be then used to improve a topological analysis of the data. We extract a special type of negative opinions, directed at the author of the parent comment, to define antagonism. Results show, with a high probabilistic significance, that antagonistic behavior leads, in the end, to low average centrality values.

A third intuition is that *centrality values have a significant delay in reaching their final values*. Within the studied dataset, certain behaviors trigger an initial centrality increase, while, in the long term, they are, on average, correlated with a centrality decrease. There is, for instance, a positive correlation between antagonistic behavior and the chance of having a temporarily increased centrality. This intuition can be linked to automatic forum moderating. If the centrality values are not instantaneously connected to the future final rating of the user, moderating based on it is impossible. Using a content based analysis leads to an instantaneous result, and can thus help in moments when centrality values alone can be misleading.

We start by reviewing the relevant previous work in Section 2. We then describe the data structures in Section 3, the experimental setting and methods used in Section 4, present the results in Section 5 and conclude our work in Section 6.

II RELATED WORK

In this paper we model the importance that society ascribes to each user and each communication instance within an online community, depending on sentiment content and topic relevance of the analyzed communications. To the best of our knowledge, no previous attempts to create such a model exist.

Traditionally, social networks are studied as part of network theory. Degree related distributions, various types of node centralities and cluster analysis [1] are just a few of the methods and metrics used to study these networks. To determine social importance, algorithms like PageRank [2] are standard practice, and the use of homophily has been proven beneficial in detecting leaders [3,4]. Using related methods on forum data, [5] detect the users most likely to have relevant expertise.

However traditional approaches do not always work.The authors of [6] show that in degree is an extremely poor indicator of popularity on Twitter. We also show in Section 5.1 that measures like the user centrality sometimes fail to clearly distinguish the best users from the worst. We believe that additional information, available in text format, can be used in such cases. The relation between textual data and network theory is bidirectional. While the network structure has been shown to improve text mining [7], text has also been used to reveal the social structure [8]. [9] provide an influence measurement

mechanism between communication content and social network structure. Within the field of text mining, our interest is mainly sentiment analysis.

Most work in sentiment analysis has focused on texts in isolation [10]. When analyzing the relation between social network analysis and opinion mining, so far the network has been viewed as an opinion rich environment [11], where social context can be used to improve the performance of opinion mining techniques [12]. Moreover, social networks permit the study of sentiment propagation and opinion cascade formation [13]. The link between network structure and the opinions flowing through it has been little investigated.

Opinions are a distinct class of textual content that can express homophily. A direct usage of sentiment is to assign a polarity to social links. The authors of [14] and [15] study online social networks and allow edges to have both positive and negative weights. The authors of [16] also try to distinguish positive from negative endorsements in blog analysis. they however conclude that too few links are available for a successful analysis. We, however, analyze a much denser interaction space - the forum of a well known alternative news website and thus do not face this difficulty.

Another approach is to create links between social actors, based on the opinions between their messages. The authors of [17] use a sentiment dictionary to extract the overall polarity of replies in a forum. They then analyze the average polarity of the replies to a certain comment. While this can be considered a way to quantify the community opinion about the said communication instance, the sentiment extraction method is prone to significant errors. An overall negative opinion of a comment can be directed at the subject discussed or outside factors, rather than the previous comment. To distinguish between the various types of opinions, we employed a much bigger dataset than the one used in [17]. We were thus able to focus on opinions targeted towards other users, rather than general ones.

While [17] uses opinions to quantify the community response to a comment, others use sentiment data to determine the value of users, to establish leadership status or assess their credibility [18, 19]. They use dictionary-based sentiment mining methods coupled with standard ranking algorithms [2] to obtain opinionated rankings. The network-based and opinion-based analyses are however separated and the results do not show a connection between the involved opinions and the network structure.

Aside from the topology analysis and sentiment content, we also study the content relevance of a communication message, with regard to a given topic. Content similarity is a well researched topic and multiple measures have been proposed. [20] use Latent Dirichlet Allocation to extract the discussed topics and they evaluate the context of a blog post with regard to these topics and metadata. Topic models can be replaced with ontologies in computing text similarity [21]. The authors of [22] also exploit the idea of topic relevance. They prove that a random surfer is more likely to jump to a page that is relevant to the page she is currently viewing. They thus use the text content similarity to create implicit links between the pages and ameliorate forum search results.

For the analyzed data source - the Slashdot news site - [14] analyzed the user social network topology based on the Friend/Foe option while [23] considered the comment reply structure as the relevant link information source. [23] conclude that relations within the Slashdot message board are more generic than explicit friend links.

III DATA STRUCTURE

The news site *Slashdot* is a collaborative platform sharing news items that can be commented on. It features a comment rating system based on human feedback. This can be viewed as a gold standard of the community perception of the ongoing interactions and a readily available labeling system. We gathered $934'612$ comments written by $35'045$ users on $4'922$ news stories between February 1, 2011 and February 1, 2012. The corpus is freely available for download[2].

1 COMMENT LEVEL STRUCTURES

The comments are organized into discussion threads, as the one in Figure 2. Each new thread is a tree $T = \{V, \rho\}$ that consists of a short story or news item $s \in S$, published by a member $u \in U$, which is followed by multiple comments $c \in \mathcal{C}$. Let $\mathcal{P}(\mathcal{C})$ be the power set of $\mathcal{C}$. Similarly, let the power sets of all users, stories and text elements be $\mathcal{P}(U)$, $\mathcal{P}(S)$ and $\mathcal{P}(\mathcal{C} \cup S)$ respectively. Each comment has an associated creation time, in POSIX format, $t(c)$, $t : \mathcal{C} \to \mathbb{N}$.

[2] http://lia.epfl.ch/data/Slashdot_2011_2012.zip

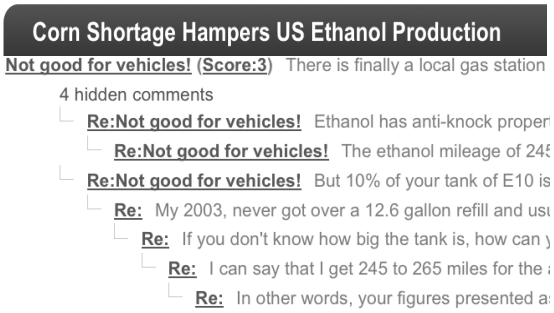

Figure 2: A Slashdot story and sample comment chain

Let $\rho \subset \mathcal{C} \times \{\mathcal{C} \cup S\}$, be the reply-to relation between a comment and another comment or news item and $\rho^* \subset \mathcal{C} \times \{\mathcal{C} \cup S\}$ its transitive closure. The number of replies a comment $c \in V_i$ receives is thus $n(c) = card\{c' \in V_i | (c', c) \in \rho^*\}$ We define a reply chain from c' to c, similar to the definition employed in [24], as $\nu : \mathcal{C} \times \{\mathcal{C} \cup S\} \to \mathcal{P}\{\mathcal{C}\}$,

$$\nu(c', c) = \{c'' \in \mathcal{C} | (c', c'') \in \rho^* || c' = c'', (c'', c) \in \rho^*\} \quad (1)$$

The length of the reply chain is $n(c', c) = |\nu(c', c)|$.

Moreover, for a single comment $c \in \mathcal{C}_i$, we define its **depth** as the length of the longest chain $\nu(c', c)$: Its depth is also the height of the largest subtree with the root in c.

$$\delta(c) = max_{c' \in V_i}(n(c', c)) \quad (2)$$

For any comment subset $\mathcal{C}' \subseteq \mathcal{C}$ we define the average number of replies as $\bar{n}(\mathcal{C}') = avg_{c \in \mathcal{C}'}(n(c))$ and average comment depth $\bar{\delta}(\mathcal{C}') = avg_{c \in \mathcal{C}'}(\delta(c))$.

2 USER LEVEL STRUCTURES

Each comment is written by a single user. We define the belong-to relation $\mu : \mathcal{C} \to U; c \mapsto \mu(c) = u \in U$ and denote the set of comments authored by a user with $\mathcal{C}(u) = \{c \in \mathcal{C} | \mu(c) = u\}$. For any user $u_i \in U$ let t_i be the moment when the first comment from that user is written.

Based on the relations between individual comments within news threads, we can define relations between individual users. The link weight between two users u_1 and u_2 is the number of comments written by u_1 as replies to comments posted by u_2. $\omega : U \times U \to \mathbb{N}; (u_1, u_2) \mapsto \omega(u_1, u_2) = card\{(c_i, c_j) \in \rho | \mu(c_i) = u_1, \mu(c_j) = u_2\}$

In the graph of user replies $G = (U, E_u)$, let σ_{st} be the number of shortest paths between users s and t, and $\sigma_{st}(u)$ the number of shortest paths between users s and t that contain user u. Vertices that have a high probability to be on shortest paths between other nodes have a high betweenness centrality.The betweenness centrality of u is defined as

$$C_B(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}} \quad (3)$$

The centrality of a user at a given point in time τ, ${C_B}^\tau(u)$, is obtained by constructing the graph G only based on comments having $t(c) < \tau$. For a user subset $U' \subset U$ the average betweenness centrality is: $\bar{C_B}(U') = avg_{u \in U'}(C_B(u))$

3 COMMENT AND USER RATINGS

Each comment $c \in \mathcal{C}$ is assigned a discrete rating $r : \mathcal{C} \to R = \{-1..5\};$. A -1 rated comment is considered unnecessary and unpleasant, while a comment rated 5 is considered meaningful. Comment ratings represent the *human generated* ground truth. The rating are given by the users themselves as +1 or -1 increments or decrements, and the final rating is an aggregated value of these contributions. We define the comment subset that share a given rating i as $\mathcal{C}_i \subset \mathcal{C}$.

A user's average rating is computed as the mean of the rating values of the comments she wrote: $r : U \to R_u = [-1, 5]$, $u \mapsto r(u) = avg_{c \in \mathcal{C}(u)} r(c)$. We use $U_{i,j} \subset U$ to denote the users having an average rating between i and j, $\forall u \in U_{i,j}, i \leq r(u) \leq j$. The average rating for the given user set is $r(U_{i,j}) = avg_{u \in U_{i,j}} r(u)$. Moreover, we define the maximum centrality of a user set with an average rating between i and j as $C_{B_{max}}(U_{i,j}) = max_{u \in U_{i,j}} C_B(u)$.

An important noise source is that some posts are not rated by users. Instead, they are assigned an automatic rating based on previous contributions. Since these ratings do not represent direct human feedback, we removed from the analysis all comments with a rating equal to the mode of all comments belonging to their author, rated in the 0-2 range.

IV PROPOSED EXPERIMENTS

We create a model of the Slashdot user behavior using topological and content measures. We formulate 11 hypothesis, shown in Figure 1, grouped around

the three aforementioned intuitions. **Intuition** 1 stated that *standard topological analysis, based on explicit reply-to links, cannot clearly distinguish valuable communication instances from damaging ones.* **Intuition 2** is that *the content of the communication, can be used to separate useful communications and users from the rest.* **Intuition 3** is that *centrality values have a significant delay in reaching their final values.* We separate the hypothesis into comment level and user level ones and we use the human generated ratings to validate them.

1 TOPOLOGICAL ANALYSIS

We aim to separate useful comments and users from the rest, using solely the topology of the reply-to generated network. In the **comment level analysis**, we used two measures of relevance. We investigate the connection between a rating value, i, and the average number of replies, $\overline{n}(\mathcal{C}_i)$, and average depth, $\overline{\delta}(\mathcal{C}_i)$, of comments rated i, We summarize the above connections in the following hypothesis:

Hypothesis 1. The engagement level and controversy created by comments that are rated as average is minimal, while it is higher for comments rated high and low.

Hypothesis 2. Informative comments, rated the highest, receive the most responses, followed by controversial ones.

In the **user level analysis**, we aim to separate users that routinely add substance to the discussion from the rest. Hypothesis 3 summarizes the correlation between the maximum betweenness centrality of users that have the rating between two values i and j $C_{B_{max}}(U_{i,j})$ and the average rating value for that user set $r(U_{i,j})$:

Hypothesis 3. High rated users are more likely to have a high centrality than the others. Low rated users are more likely to have a high centrality than average ones.

We also study whether having a higher centrality implies receiving a better rating from the community. Given all centrality values of users in U, we sort and then split them in N intervals, each containing the same number of values, with the lowest ones in interval 1. For each interval $i \in \{1..N\}$, let $C_B^{min}(i)$ and $C_B^{max}(i)$ be its start and end values. Also, let the set of users whose centrality falls between these values be $U_{CB}(i) = \{u \in U | C_B^{min}(i) \leq C_B(u) \leq C_B^{max}(i)$. We compute the average rating, $r(U_{CB}(i))$, and the variance of the ratings, $\sigma^2(i)$, for each interval. From these values, we determine the **predictive power** of a user's centrality.

2 CONTENT RELEVANCE ANALYSIS

We then inquire whether users assign better ratings to comments perceived as being *on topic*. In the Slashdot forum, we use the content similarity between comments and the original story to distinguish relevant from irrelevant comments.

We use standard preprocessing methods, such as stemming, and we determine the content similarity between comments and the original news item using the *Jaccard* distance. This measure takes the word sets W_A and W_B from two texts A and B, and outputs a percentage of similarity:

$$d_{Jaccard}(A, B) = \frac{|W_A \cap W_B|}{|W_A \cup W_B|} \tag{4}$$

We define the **topic relevance** of a comment subset $\mathcal{C}' \subset \mathcal{C}$ as its average Jaccard distance to the discussed stories: $\forall c \in \mathcal{C}'$, $\forall s_i \in S$, $c \in V_i$, $\overline{d_{Jaccard}}(\mathcal{C}') = avg_{s_i \in S}(\sum_{c \in \mathcal{C}' \cap V_i} d_{Jaccard}(c, s_i))$ We compute the topic relevance of each subset $\mathcal{C}_i$ that contains the comments rated i, $\overline{d_{Jaccard}}(\mathcal{C}_i)$, and test the following hypothesis:

Hypothesis 4. High rated comments are more likely to have a high topic relevance than low rated ones.

3 ANTAGONISM AND SOCIAL ENGAGEMENT

We found that hostile behavior impacts the dynamics of social interaction, both at a comment and a user level. We define the antagonism within a comment as a direct negative reference to the writer of the parent comment or story.

We perform a dictionary based subjectivity analysis. For instance, "*nice*" has a mildly positive polarity and "*idiot*" carries a strongly negative one. We attached polarity values $p(w)$ to the comment terms $w \in W_c$. The values are obtained from OpinionFinder [25] and online curse and foul language collections.

A **direct opinion** is an instance of a polarized word, or modifier, that refers to a second person pronoun

or possessive determiner. A relation is defined as the presence of a syntactic dependency chain [26] between the two words. For example, the sentence *"Your opinion is not right."*, has a polarity of -1, because *right* has a dependency chain to *your*, has a polarity of $+1$, and is negated. For all valid constructs, let $p(w)$ be the polarity of the modifier and $neg(w) \in \{0,1\}$ states whether it is negated. A comment is **antagonistic** if contains more negative than positive direct opinions. We define the antagonism property of a comment as $p_\alpha : \mathcal{C} \to \{0,1\}$

$$c \mapsto p_\alpha(c) = I_{(\sum_{w \in W_c} p(w) \times neg(w)) < 0} \qquad (5)$$

Antagonism is also a property of reply to links: $\alpha : \mathcal{C} \times \{\mathcal{C} \cup S\} \to \{0.1\}$, $(c, c') \mapsto \alpha(c, c') = p_\alpha(c)$. In this relation, c is the **source** of antagonism and c' the **destination**.

Let $\mathcal{C}_{\alpha s} \subset \mathcal{C}$ be the set of all antagonistic comment sources. The average depth of these comments is thus $\overline{\delta}(\mathcal{C}_{\alpha s})$. For the antagonistic source comments that have a rating i, the average depth is $\overline{\delta}(\mathcal{C}_{\alpha s i})$. We use similar notations for the malicious comment destinations: $\mathcal{C}_{\alpha d} \subset \mathcal{C}$, with an average depth $\overline{\delta}(\mathcal{C}_{\alpha d})$ and $\overline{\delta}(\mathcal{C}_{\alpha d i})$ for a rating i.

We incorporate the effect of hostile interaction into the Slashdot user behavior model. We believe that antagonism has a significant effect on the connectivity of its sources and destinations. In this respect we test the following hypothesis:

Hypothesis 5. A **high rated** comment is **less** likely to create controversy when it is the **destination** of antagonism.

Hypothesis 6. A **low rated** comment is **more** likely to create controversy when it is the **destination** of antagonism.

Hypothesis 7. A **high rated** comment is **more** likely than a low rated one to create controversy when it is the **source** of antagonism.

A user is defined as an antagonism source $u \in U_{\alpha s} \subset U$ if he has written an antagonistic comment: $\alpha_{su} : U \to \{0,1\}$;

$$u \mapsto \alpha_{su}(u) = max(p_\alpha(c) | \forall c \in \mathcal{C}_u(u)) \qquad (6)$$

Let $U_{\alpha s i,j} \subset U_{\alpha s}$ contain all users having $i \leq r(u) \leq j$, with $\overline{C_B}(U_{\alpha s i,j})$ their mean centrality. Similarly, we denote the set of all users who are the destination of antagonism with $U_{\alpha d} \subset U$ and $U_{\alpha d i,j} \subset U_{\alpha d}$ contains those with the rating bounded by i and j. We correlate the value of their centrality with their participation in antagonistic exchanges:

Hypothesis 8. Users that have an antagonistic behavior are less central to the community than the ones who do not.

Hypothesis 9. The centrality of users who are the destination of antagonism is lower than the centrality of users who are sources of antagonism.

Hypothesis 10. Low ranking users that are also antagonistic become pariah, with the smallest possible centrality.

But, earlier, Hypothesis 6 stated that low ranking comments receive greater community feedback if they are the antagonized. Hypothesis 11 explains the short term effects of antagonism on the centrality of those involved:

Hypothesis 11. The centrality of antagonistic users increases shortly after the negative interaction.

Taken together, Hypotheses 8 and 11 show the evolution of community interest in a user involved in negative interactions. While at first his centrality increases (possibly as a result of angry direct responses), the community then avoids further interactions, which leads to a final low centrality.

We define short term using a *user dependent* definition of time. For each user $u_i \in U_{\alpha d} \cup U_{\alpha s}$, we study the centrality variation after their negative interactions. Let $t_{0,i}$ be the moment the first negative interaction involving u_i occurs. Also, let $t_0 = min_i(t_{0,i})$ be the moment of the first negative interaction that involves any user. We mark the end of the experiment with t_{100}, the same for all users.

We divide each time interval $(t_{0,i}, t_{100})$ into 100 equal parts and denote the start of the k'th interval with $t_{k,i}$. We compute the centrality of u_i, at the moments $t_{k,i}$, $C_B^{t_{k,i}}(u_i)$ with a focus on low k values, $k \in D_k = \{1, 2, 5, 10\}$, and compare it with the starting centrality for that user, $C_B^{t_{0,i}}(u_i)$. We thus inquire whether the difference, expressed relative to the starting value, is positive $\delta_{C_B} : U \times D_k \to \mathbb{R}$:

$$\delta^k_{C_B}(u_i) = \frac{C_B^{t_{k,i}}(u_i) - C_B^{t_{0,i}}(u_i)}{C_B^{t_{0,i}}(u_i)} \times 100 \qquad (7)$$

We define the centrality variation of a user set given time increment k as $\delta_{C_B} : \mathcal{P}(U) \times D_k \to \mathbb{R}$, $\delta^k_{C_B}(U') = avg(\delta^k_{C_B}(u_i)), \forall u_i \in U'$, with $\delta_{C_B}(U')$ the average over all considered k values. For different values of

k, we inquire whether $\delta^k_{C_B}(U_a)$ are positive. A further question is if the aggregated centrality variation is larger for antagonistic users, $\delta_{C_B}(U_a)$, than for non antagonistic ones, $\delta_{C_B}(U_{n\alpha})$.

We focus on the variation of the centrality, relative to the initial value, rather than in absolute terms. Let user u_A have a very high centrality and user u_B a near zero one, $C_B(u_B) = \epsilon \ll C_B(u_A)$. A centrality variation of ϵ for u_B is very significant, while for u_A it is hardly noticeable. It is thus important to consider the variation in the context of the user's past, rather than focusing on a narrow time window.

The computational cost of determining the centralities of users at various moments is high. We perform the analysis on a subset of antagonistic users, $U_\alpha \subset (U_{\alpha d} \cup U_{\alpha s})$, and non antagonistic users $U_{n\alpha} \subset U \backslash \{U_{\alpha d} \cup U_{\alpha s}\}$. We focus on users who had previous activity before the negative comment in $t_{0,i}$ and whose centrality at that moment is not zero.

V RESULTS

1 TOPOLOGICAL ANALYSIS

1.1 COMMENT LEVEL ANALYSIS

We show that the social engagement a comment produces can be linked to the importance it is assigned by the community. Figure 3 shows the variation of the average number of replies, $\overline{n}(\mathcal{C}_i)$, for comments rated $i \in R$.

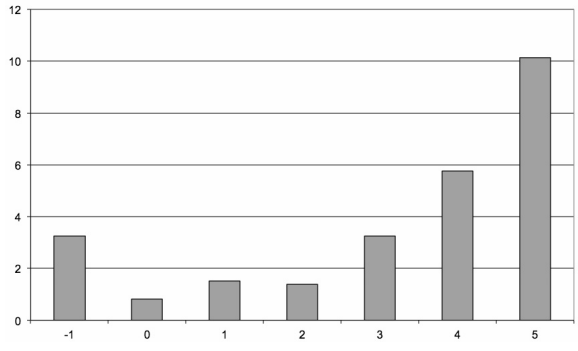

Figure 3: Comment reply number distributed by its rating

The dependence is not linear. A minimum response is reached for average rating values $\overline{n}(\mathcal{C}_{0..2})$. Higher response values are obtained for extremely negative $\overline{n}(\mathcal{C}_{-1})$ and positive $\overline{n}(\mathcal{C}_{3-5})$ rating values. This validates the social engagement side of Hypothesis 1.

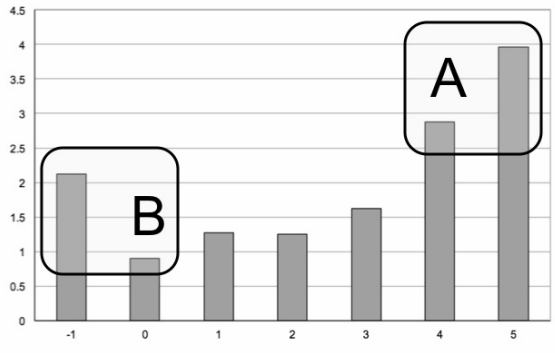

Figure 4: Depth of a comment distributed by its rating

Figure 4 depicts the comment set **depth**, for the possible rating values, $\overline{\delta}(\mathcal{C}_i)$. The connection is again not linear and we observe that engagement rises for extreme rating values. This validates the controversy part of Hypothesis 1.

Moreover, both measures show that overall engagement and also controversy reach a maximum for the informative comments (rated 5), highly regarded by the community.

$$\overline{n}(\mathcal{C}_5) \gg \overline{n}(\mathcal{C}_{-1}) \gg \overline{n}(\mathcal{C}_{0..2}) \tag{8}$$

$$\overline{\delta}(\mathcal{C}_5) \gg \overline{\delta}(\mathcal{C}_{-1}) \gg \overline{\delta}(\mathcal{C}_{0..2}) \tag{9}$$

These two observations validate Hypothesis 2. However, as $\overline{n}(\mathcal{C}_{-1}) \approx \overline{n}(\mathcal{C}_3)$ and $\overline{\delta}(\mathcal{C}_{-1}) > \overline{\delta}(\mathcal{C}_3)$, we cannot distinguish useless comments from ones rated mildly positive.

We performed a Welch test and computed the two-tailed p-values to establish the statistical significance. Let the sets containing number of replies, constructed based on the comment rating k, be ${D^n}_k = \{n(c), c \in \mathcal{C}_k\}$. We found that the p-value when comparing ${D^n}_5$ to ${D^n}_{-1}$ is 5.44×10^{-5}; 1×10^{-8} when comparing ${D^n}_5$ to ${D^n}_{0,1,2}$; 5.36×10^{-3} for ${D^n}_{0,1,2}$ Vs. ${D^n}_{-1}$. We thus reject the null hypothesis and prove that the differences between the means are significant.

We computed a similar test for the depth distributions ${D^\delta}_k = \{\delta(c), c \in \mathcal{C}_k\}$. We found that the p-value when comparing ${D^\delta}_5$ to ${D^\delta}_{-1}$ is 5.73×10^{-4}; 1.8×10^{-7} for ${D^\delta}_5$ Vs. ${D^\delta}_{0,1,2}$ and 3.16×10^{-3} for ${D^\delta}_{0,1,2}$ Vs. ${D^\delta}_{-1}$

1.2 USER LEVEL ANALYSIS

The above conclusion also holds at the user level. In Figure 5, individual user centralities are plotted on a rating scale $C_B(u) = f(r(u))$. The horizontal axis plots all the 1987 different user average rating values, with an equal number of values in the rating ranges $(-1, k)$, (k, l) and $(l, 5)$., with $k = 2.08$ and $l = 2.41$. $U_{-1,k}$ thus represents users that are rated lowest, $U_{k,l}$ average users and $U_{l,5}$ top rated users.

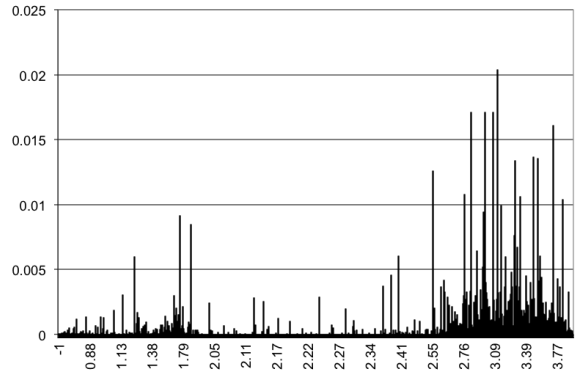

Figure 5: User centrality distributed by average user rating

We notice that positively rated users have the highest maximum centrality, followed by the ones in the lowest third,which validates Hypothesis 3.

$$C_{B_{max}}(U_{l,5}) > 2 \times C_{B_{max}}(U_{-1,k}) > 1.5 \times C_{B_{max}}(U_{k,l}) \quad (10)$$

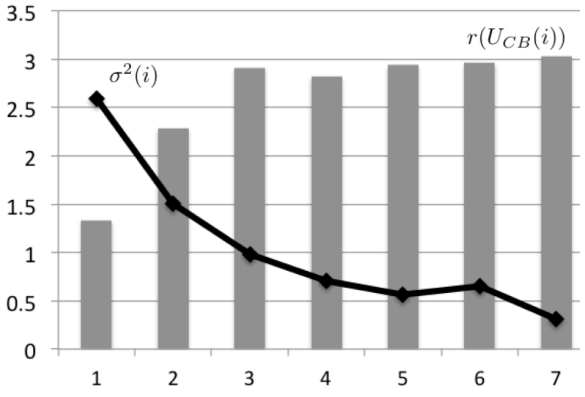

Figure 6: Average rating and rating variance for each user centrality interval

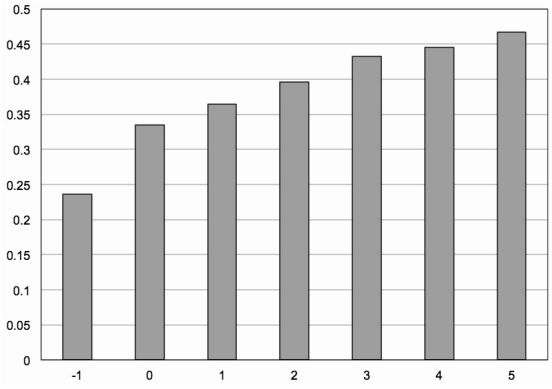

Figure 7: Content similarity between comments and stories

We then answer whether centrality can predict a user's rating. Figure 6 shows the average rating $r(U_{CB}(i))$ and rating variance $\sigma^2(i)$ for each centrality interval $i \in \{1..7\}$, as detailed in Section 4.1. We observe that, for any interval starting from the third, the average rating for users having a centrality value in that range is close to 3. Thus, in more than 70% of cases, a higher centrality value does not imply, on average, a better community perception. It is noteworthy that users with the lowest centrality values, in the first two intervals, have an average rating smaller than the rest. This lower value is, however, accompanied by high variance values, showed by the line in Figure 6. The high variance signals that a low centrality value is not useful for *individual* rating prediction. A given centrality maps into a large selection of different user ratings (low and high).

2 COMMUNICATION CONTENT RELEVANCE

The topology of the network can differentiate the best and worst users and comments from the average ones. To separate relevant users from malicious ones, we employ content relevance. Figure 7 shows the content relevance of comments sharing different ratings $\overline{d_{Jaccard}}(\mathcal{C}_i), i \in R$. We observe its steady increase as the ratings vary from a minimum of -1 to a maximum of 5.

$$\overline{d_{Jaccard}}(\mathcal{C}_i) > \overline{d_{Jaccard}}(\mathcal{C}_j), \forall i, j \in R, i > j \quad (11)$$

Poor and valuable comments can thus be separated using their textual similarity to their discussed news item. This validates Hypothesis 4 and shows that content relevance can be a valuable complement to traditional topological analysis.

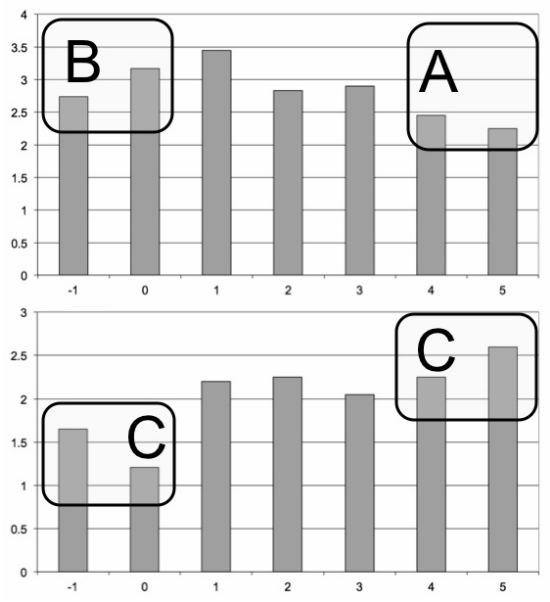

Figure 8: Response depth distributed by comment rating - upper chart - antagonism destinations; lower one - sources

3 ANTAGONISM AND SOCIAL ENGAGEMENT

3.1 COMMENT LEVEL ANALYSIS

A comment's reply depth is a measure of the controversy it causes. The bottom chart in Figure 8 shows the average depth for comments that are antagonism sources, and have a given rating i, $\overline{\delta}(\mathcal{C}_{\alpha s i})$, and the upper one for antagonism destination comments, $\overline{\delta}(\mathcal{C}_{\alpha d i})$. We compare with the respective average depth for all comments, $\overline{\delta}(\mathcal{C}_i)$, shown in Figure 4.

We find that for highly rated comments $i \in \{4, 5\}$ the controversy they create decreases if they are antagonized:

$$\overline{\delta}(\mathcal{C}_{\alpha d 4,5}) < \overline{\delta}(\mathcal{C}_{4,5}) \quad (12)$$

This validates Hypothesis 5 and shows that overt criticism is destructive for pertinent arguments. In a constructive argument, politeness is thus the key. The fact that the two pairs of means are different is backed by two p-values that compare $\{\delta(c), c \in \mathcal{C}_i\}$ and $\{\delta(c'), c' \in \mathcal{C}_{\alpha d i}\}$ for $i \in \{4, 5\}$: 1×10^{-8} for $i = 4$ and 1×10^{-5} for $i = 5$.

The boxes labeled B in Figure 8 show that disagreeing with impertinent users *significantly increases* the feedback they receive. This result, that validates Hypothesis 6, is backed up by similarly computed highly significant p-values of 1×1^{-8} for $i = -1$ and 1×1^{-10} for $i = 0$. An interpretation of Hypothesis 6 is that we *shouldn't feed the trolls*.

$$\overline{\delta}(\mathcal{C}_{\alpha d -1,0}) > \overline{\delta}(\mathcal{C}_{-1,0}) \quad (13)$$

A third observation, that supports Hypothesis 7, is that constructive arguments create more controversy than less constructive ones. The boxes marked C in in Figure 8 show that comments rated 4 and 5 which criticize others have a higher depth than comments rated -1 and 0.

$$\overline{\delta}(\mathcal{C}_{\alpha s 4,5}) > \overline{\delta}(\mathcal{C}_{\alpha s -1,0}) \quad (14)$$

The difference is shown to be significant by the p-value of 4.81×10^{-3} obtained when comparing $\{\delta(c), c \in \mathcal{C}_{\alpha s -1} \cup \mathcal{C}_{\alpha s 0}\}$ to $\{\delta(c), c \in \mathcal{C}_{\alpha s 4} \cup \mathcal{C}_{\alpha s 5}\}$.

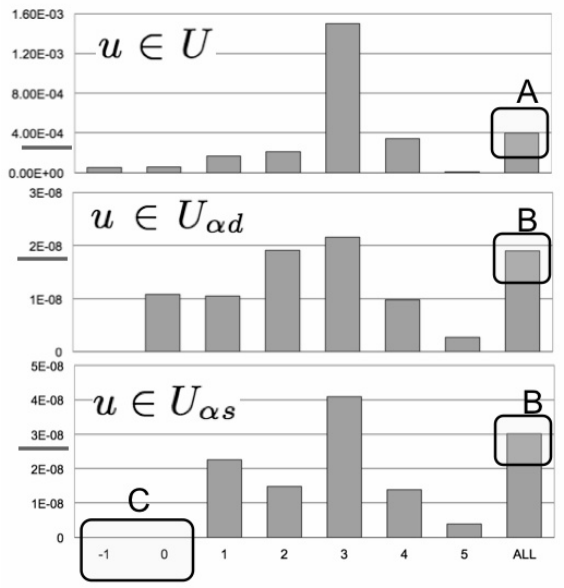

Figure 9: Centrality distributed by user rating, for all users (top), antagonism targets (middle), and sources (bottom)

3.2 USER LEVEL ANALYSIS

We prove a user's centrality depends heavily on whether he emits or is targeted by antagonism. Figure 9 presents the betweenness centrality in three different settings. The top chart portrays all users, $\overline{C_B}(U_{i,j})$, the middle one antagonistic comment destinations, $\overline{C_B}(U_{\alpha d i,j})$, and the bottom one the users that are sources of antagonism, $\overline{C_B}(U_{\alpha s i,j})$. The rightmost bars show the centrality values for all the users, regardless of their rating: top - $\overline{C_B}(U)$, middle - $\overline{C_B}(U_{\alpha d})$, bottom - $\overline{C_B}(U_{\alpha s})$.

A first observation, shown in boxes labeled B versus the box labeled A, supports Hypothesis 8. We prove that antagonistic users have a much lower centrality than the average:

$$\overline{C_B}(U) \approx \overline{C_B}(U_{\alpha d}) \times 10^4; \overline{C_B}(U) \approx \overline{C_B}(U_{\alpha s}) \times 10^4 \quad (15)$$

The differences are backed by highly significant Welch test p-values of 6.9×10^{-7} for antagonism sources and 6.54×10^{-4}) for destinations. The test compared $\{C_B(u), u \in U\}$ with $\{C_B(u), u \in U_{\alpha d}\}$ for destination users and with $\{C_B(u), u \in U_{\alpha s}\}$ for antagonism source users.

A second user level observation is that user centrality is significantly higher if you emit a negative opinion than if you are being criticized, as in Hypothesis 9. The values highlighted in the boxes labeled B on Figure 8 show that

$$\overline{C_B}(U_{\alpha s}) = \overline{C_B}(U_{\alpha d}) \times 1.59 \quad (16)$$

This finding is however less significant statistically, as the p-value obtained by comparing $\{C_B(u), u \in U_{\alpha d}\}$ with $\{C_B(u), u \in U_{\alpha s}\}$ is only 9.57×10^{-2}.

A third observation, that supports Hypothesis 10, is that giving negative explicit feedback when having a low average score leads to marginalization. The null subsequent engagement values are highlighted in box C in Figure 9.

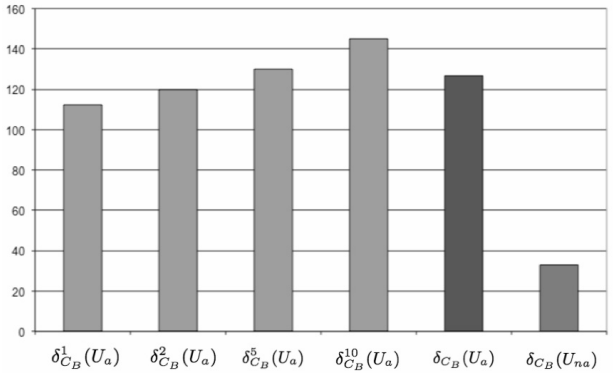

Figure 10: Short term centrality variation of antagonistic and regular users

3.3 CENTRALITY VARIATION

A final experiment regards the short term variation of antagonistic user centrality. As previously mentioned, the experiment was performed on two sets of users: antagonistic $|U_\alpha| = \{112\}$ and non antagonistic ones $|U_{n\alpha}| = \{133\}$. The first four columns of Figure 10 portray the values of $\delta^k_{C_B}(U_\alpha)$, with k varying from 1 to 10. All four values verify the claim in Hypothesis 11, that, in the short term, antagonistic users increase their centrality. We observe that, on average, they more than double their centrality for a short time period after the negative interaction.

We compare the aggregate variation value for antagonistic users, $\delta_{C_B}(U_\alpha)$ with the corresponding one for regular users, $\delta_{C_B}(U_{n\alpha})$. Results show that the short term average centrality variation for antagonistic users is almost four times larger than for regular users (126.7 versus 32.9 per cent). This result is backed by a p-value of 1×10^{-5}.

VI CONCLUSIONS

The main contribution of this work is the creation of a behavior model of people posting online comments, that was validated using real world data from Slashdot. Through this model we validated three intuitions about the usefulness of traditional network oriented methods and the proposed content based ones. We proved that, for the considered network, direct reply-to links are less informative then the content of the comments. Including this class of information based connections into the analysis of social networks improves the understanding of the analyzed societies.

Basic network topology, captured as reply-to links between the Slashdot story comments, allowed the separation of regular users from high risk and high value ones. As a benchmark we used the comment ratings, which represent an aggregated value of the community and moderator explicit opinions about the said comment. The same finding also applied to comments, as average rated comments received the least responses.

We then showed that the content relevance of forum messages and antagonistic opinions within them can be used to obtain a finer grained analysis. Content relevance, expressed as the content overlap between the news story and the comments, is correlated with the comment ratings. The higher the relevance to the original story, the higher the probability that that comment will be viewed positively by the peers and moderators. Moreover, we proved that this measure gives consistent results, as comments with a high content overlap also have a high probability of having the same rating.

From an opinion mining perspective, we focused on the role of antagonism, defined as the presence of negative opinions regarding the author of the previous comment. Results showed that the response to a comment is influenced by being antagonized, which only helps low rated posts. Regarding the users, we found that being rude and having an antagonistic stance leads in the end to social marginalization. This effect is more visible for users who are targeted by negative opinions than for users who generate them. Another observation is that user centrality only decreases for antagonistic users after a long period of time, while the initial effect is reversed.

1 IMPACT

We believe the impact of our work is twofold. Firstly, the presented results are a solid case against indiscriminately using network measures, like the betweenness centrality. We hope this will increase the efforts put into finding other ways to define connections between social actors. We have shown that text based connections are a reliable alternative.

Secondly, a direct application of this work is into automatic moderating of discussion boards. While some rely on user moderating, through rating schemes, many rely solely on human moderating, which significantly increases the costs. Others, like Slashdot, use a hybrid approach. Creating an automatic moderating tool would greatly reduce the costs and improve the user experience. While for specific tasks such as spam filtering there have been some successful methods, automatically identifying malicious users has so far not been achieved. We have shown that using user centrality values for moderating is only helpful after a significant period of time. This means that alternative methods have to be employed and we proposed two complementary text based ones. We thus propose using the topic relevance of forum posts and their negative opinion content as better and instantaneous moderating tools.

References

[1] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[3] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 7–15.

[4] K. Lewis, M. Gonzalez, and J. Kaufman, "Social selection and peer influence in an online social network," *Proceedings of the National Academy of Sciences*, 2011.

[5] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 221–230. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242603

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *in ICWSM Õ10: Proceedings of international AAAI Conference on Weblogs and Social Media*, 2010.

[7] Z. Kozareva and E. Hovy, "Insights from network structure for text mining," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1616–1625. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002666

[8] J. Diesner and K. Carley, *Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis.*, ser. Causal Mapping for Information Systems and Te. Harrisburg, PA: Idea Group, 2005, ch. 4.

[9] S. Wang and P. Groth, *Measuring the dynamic bi-directional influence between content and social networks*. Springer, 2010, vol. 6496, pp. 814–829. [Online]. Available: http://www.springerlink.com/index/WV88V4666P3Q2Q61.pdf

[10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Empirical methods in natural language processing-Volume 10*. ACL, 2002, pp. 79–86.

[11] K. Jedrzejewski and M. Morzy, "Opinion mining and social networks: A promising match," in *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 599–604. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2011.123

[12] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1397–1405. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020614

[13] M. Miller, C. Sathi, D. Wiesenthal, J. Leskovec, and C. Potts, "Sentiment flow through hyperlink networks," in *ICWSM 11, Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[14] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The slashdot zoo: mining a social network with negative edges," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 741–750.

[15] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 641–650.

[16] G. Berardi, A. Esuli, F. Sebastiani, and F. Silvestri, "Blog distillation via sentiment-sensitive link analysis." in *NLDB*, ser. Lecture Notes in Computer Science, G. Bouma, A. Ittoo, E. Metais, and H. Wortmann, Eds., vol. 7337. Springer, 2012, pp. 228–233. [Online]. Available: http://dblp.uni-trier.de/db/conf/nldb/nldb2012.html#BerardiESS12

[17] A. Stavrianou, J. Velcin, and J.-H. Chauchat, "A complementary model to the social networks for mining forums," in *From Sociology to Computing in Social Networks*, 2010, pp. 59–79.

[18] K. S. Cho, J.-S. Ryu, J.-H. Jeong, Y.-H. Kim, and U.-M. Kim, "Credibility evaluation and results with leader-weight in opinion mining," *2010 International Conference on Distributed Computing and Knowledge Discovery*, pp. 5–8, 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5615503

[19] B. He, C. Macdonald, and I. Ounis, *Ranking opinionated blog posts using OpinionFinder*. ACM Press, 2008, vol. 31, pp. 727–728. [Online]. Available: http://dx.doi.org/10.1145/1390334.1390473

[20] L. Fan and B. Li, "Blog-based online social relationship extraction," in *Cognitive Informatics, 2009. ICCI'09. 8th IEEE International Conference on*. IEEE, 2009, pp. 457–463.

[21] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, and E. Milios, "Semantic similarity methods in wordnet and their application to information retrieval on the web," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM, 2005, pp. 10–16.

[22] G. Xu and W.-Y. Ma, "Building implicit links from content for forum search," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06, 2006, pp. 300–307. [Online]. Available: http://doi.acm.org/10.1145/1148170.1148224

[23] V. Gómez, A. Kaltenbrunner, and V. López, "Statistical analysis of the social network and discussion threads in slashdot," in *Proceedings of the 17th international conference on World Wide Web*. New York, USA: ACM, 2008, pp. 645–654. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367585

[24] A. Stavrianou, J. Velcin, and J.-H. Chauchat, "Definition and measures of an opinion model for mining forums," in *ASONAM*, 2009, pp. 188–193.

[25] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Opinionfinder: a system for subjectivity analysis," in *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Stroudsburg, PA, USA: ACL, 2005, pp. 34–35. [Online]. Available: http://dx.doi.org/10.3115/1225733.1225751

[26] M.-C. de Marneffe, B. MacCartney, and C. D. Manning, "Generating typed dependency parses from phrase structure trees," in *LREC*, 2006. [Online]. Available: http://nlp.stanford.edu/pubs/LREC06_dependencies.pdf