Deconstructing Information Credibility on Twitter

Byungkyu Kang
Dept. of Computer Science
University of California
Santa Barbara
bkang@cs.ucsb.edu

Sujoy Sikdar
Dept. of Computer Science
Rensselaer Poly. Inst.
Troy, New York, USA
sujoy@cs.rpi.edu

Tobias Höllerer
Dept. of Computer Science
University of California
Santa Barbara
holl@cs.ucsb.edu

John O'Donovan
Dept. of Computer Science
University of California
Santa Barbara
jod@cs.ucsb.edu

Sibel Adalı
Dept. of Computer Science
Rensselaer Poly. Inst.
Troy, New York, USA
sibel@cs.rpi.edu

ABSTRACT

Many studies of information credibility in Twitter treat credibility as a singular construct. In this paper, we introduce a variety of constructs for information credibility and show that there is no one best metric to measure all of them. Next, we illustrate that any method for predicting information credibility must take into account the underlying context of the given problem domain such as the amount of uncertainty in the information and the level of familiarity the users have with each other. To this end, we introduce a study of two collections of tweets on a similar topic and time frame, and a large number of features based on user behavior, network structure and message content. Using a best-feature analysis, we show how the underlying context and the credibility construct imply significantly different models for predicting credibility. We present a detailed discussion of our findings, which includes a predictive accuracy evaluation on various credibility-based metrics, and conclude with a discussion of potential pitfalls for the design of information filtering models for microblogs.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Social Media Analytics, Credibility, Twitter

1. INTRODUCTION

Increased popularity of microblogs in recent years brings about a need for better mechanisms to extract credible or otherwise useful information from noisy and large data. Especially in times of uncertainty, people tend to use social media as an essential source of information. The main question we would like to answer in this paper is how can we effectively measure information credibility in social media, especially in Twitter? To answer this question, we take a different approach than existing literature on this topic. Instead of constructing a model to measure credibility and report on its accuracy, we deconstruct credibility altogether. First, we ask what is information credibility and what contributes to it. Then, we examine how we can construct ground truth for information credibility. Finally, we study how indicators of credibility may differ based on information context. In particular, we discuss when behavioral indicators are relevant to the study of credibility and when message level indicators are sufficient.

We are not the first researchers to study information credibility on Twitter [34, 8, 26]. Clearly, this is a topic of critical concern given the increasing prevalence of social media and social contexts in disseminating information. While most studies on credibility concentrate on predictive models for credibility, they often do not go into detail about the specific construct of credibility that they study and the limitations of their specific approach. As a result, it is not clear how generalizable the results from one study are in a different context. Instead, we introduce a meta-study of how credibility should be studied. We consider two problems: the underlying credibility construct and how the given construct may be impacted by other contextual factors.

First, we discuss the different credibility constructs we consider. Messages can be credible because of their surface appearance: they cite resources and/or are written in an authoritative manner. Messages may also be credible because of their sender: the sender may be a known authoritative source or an expert. The sender may also be a social contact who is trusted by the receiver, even though they do not have expertise in the given topic. In all cases, signals for message credibility may differ significantly. For example, a trusted source does not have to signal credibility by citing a

source. A seeming incredible message may be believed if it comes from an expert or a friend. By studying how credibility signals differ in these different contexts, we can develop more comprehensive models for measuring credibility in social media.

A recent article [16] on the use social media during Hurricane Sandy illustrates all of these different ways credibility is judged. During the hurricane, many used Twitter as a way to get and disseminate vital information. However, there were attempts to distribute misinformation too. For example a user with the Twitter id *@ComfortablySmug* (who was actually a candidate for the Congress) pushed a rumor that New York Stock Exchange floor was under three feet of water. However, this information was quickly revealed to be wrong by others on Twitter. In fact, the propagation of the rumor was stopped by other users. This did not stop the information from infiltrating other networks like Facebook where concerned people would warn their friends that they should be careful. In fact, according the 2012 Pew report on State of the Media, more and more larger group of people follow links that come from friends instead of news organizations. This means that they rely heavily on social links to assess the credibility of information.

How can these different notions of credibility be measured to obtain ground truth information? A common method is to conduct user surveys with the help of Amazon Turk in which the participants are asked to rate the credibility of the information that they read [8]. This has the advantage of obtaining large amount of credibility ratings and the assessment can be considered objective as it is not impacted by various biases of the users. However, this approach also has some limitations. First, this type of credibility assessment is often post-hoc. The raters get to see the information after it has been posted and hence may have additional information that was not accessible to the discussants at the time of message. For example, in the case of Hurricane Sandy, the raters may now know where the flooding was. But at the time, there was a great deal of uncertainty. On the flip side, the survey subjects may be less familiar with the topic than those discussing it. The raters are also unlikely to know the senders of information. Hence, their judgments are almost solely based on the textual content of the message. As a result, surveys of credibility for the most part measure whether the text is credible or whether the text "appears" credible. Studies in credibility assessment [33, 13] show that surface elements like the way message is written, whether the source has a picture in their profile can be very important in judging credibility especially if the information source is unknown to the receiver.

Credibility studies [5, 13] also reveal that people often do not consider judgments of the message content if the source of the information is known to them. As it is mentally taxing to evaluate information and information about the credibility of the source is much cheaper to evaluate, they rely on the credibility of the source. Unfortunately, most user studies cannot replicate this type of credibility assessment, which might be much more prevalent in the studied data. In fact, many discussions on credibility show the importance of fact checkers, i.e. trusted and reputable sources, in stopping the spread of misinformation. In some studies, the credibility raters are told about the social features of the message sender (like the number of friends and followers). They can use this information as an additional signal for the expertise of the user. Research shows that people can form opinions of others based on information they read about them [4] and this information can be highly relevant. However, opinions formed in this way are not equivalent to the first hand experience with the sender, which are richer, involving multiple dimensions, and are retrieved from memory much more easily.

To address this problem, we introduce a second notion of message credibility that considers how far a message traveled in the social network through forwarding. The intuition behind this metric is that every time the message is forwarded, it is endorsed by social links. In addition, the authoritativeness of its sender as evaluated by the receivers also plays a role in the decision to forward or not. As the forwarding happens when the message is circulating, the forwarders are acting solely on the information available to them. As a result, the length of the forward chains serve as a proxy for in-network credibility of the message at the time it was circulating. The limitation of this metric however is that credibility is not always the reason for forwarding a message: interestingness may not always imply credibility. People may forward messages because it confirms their own opinion, especially in the case of political messages. As a result, message propagation is a noisy factor. However, by considering now a single propagation but chain length, we capture a notion of in network credibility that cannot be measured easily by surveys.

Finally, we consider a weaker notion of credibility based on dyadic exchanges, i.e. a directed messages between two individiuals with a history of multiple exchanges, indicating a social link. Sometimes messages are forwarded because they come from a social trusted source, who is not an expert in the given topic. Examples of these are given in [16] where people share information that they believe to be correct to help their friends. With these three notions of credibility, we hope to obtain a more comprehensive picture of credibility and factors leading to it.

In addition to the construct of credibility, we also study the importance of topical context in judging credibility. Previous studies have considered newsworthy articles [8] or political discourse [22] to cite a few. The question we would like to answer in this paper how the topic of a tweet impacts the factor corresponding to its study. To this end, we study two different datasets corresponding to the same topic: Hurricane Sandy. The first dataset is collected during the storm using keywords "Sandy" and "Frankenstorm". It corresponds to discussion at the time of great uncertainty that is discussed in [16]. The second data set corresponds to the relief effort after the storm, collected using the keyword "OccupySandy". It is initiated by people who participated in the Occupy Wall Street movement and actually builds on the existing social capital, i.e. relationships, built during this time. The relief effort is still ongoing, but the information exchanged does not have the same urgency and criticality as in the first data set.

To study the impact of context and credibility constructs, we

construct a large set of features based on message content, the network properties of the information sources and a large set of behavioral features that capture the nature of the social relationships between people. In particular, pairwise relationship features based on network behavior are shown to be significant in distinguishing between social relations (friendship) and relations based on expertise (information sources) [3]. Furthermore, the word usage in these two relationships differ significantly. Given these two data sets corresponding different discussions on the same topic and the various annotations of credibility, we segment our data into different sets and report on the most relevant credibility indicators in each segment.

In this paper, we make the following unique contributions:

- We introduce different notions of credibility and show methods to obtain annotations that are indicative of these different notions. Through a case study, we illustrate that our methods measure different and complementary concepts.
- We introduce a case study of two related but different topic related message collections. We study different segmentations of our data. In particular, we show that the most relevant features change significantly depending on the topic context and the construct of credibility.
- We study segmentations of the data that show how the significant behavioral features change depending on the level of connectivity between the information sources.
- We conclude by discussing the implications of our study in developing robust behavioral indicators of credibility and recommendation systems.

2. CONTENT-BASED AND SOCIAL FEATURES

There are a wide range of features proposed in many different studies relying on different signals to assess credibility of messages. We consider these in two main categories: content based and social. Social features include both network based and behavioral features.

Content based features evaluate the textual content alone, whether the text contains mentions, urls, specific type of words, sentiments expressed and so on. Note that when judging credibility, the importance of the textual content cannot be disregarded. For example, information that appears plausible is much more likely to be believed. Information that is familiar to the information consumer, can be remembered quickly, and as a result may be judged more credible. These types of heuristics are often employed when judging credibility. In fact other cues such as the existence of cited sources (people or urls), whether it was retweeted or not, may be used to infer the authoritativeness of the message. The expertise of the information consumer in the given topic is an important factor in determining to which degree they will rely on the actual textual content of the message, and how much they will use these heuristic measures. The list of content based features are given in Table 2. The details of these features can be found in [15].

feature name	*description*
char	# chars
word	# words
question	# question marks
excl	exclamation marks
uppercase	# uppercases in text
pronoun	# pronouns (count by corpus)
smile	# smile emoticons
frown	# frown emoticons
url	# urls
retweeted	0: not retweeted, 1: retweeted once, 2: multiple times
sentiment_pos	positive word count based on lexicon sourced from NLTK[a]
sentiment_neg	positive word count based on lexicon sourced from NLTK
sentiment	sentiment polarity (sentiment_pos - sentiment_neg
num_hashtag	from entity metadata
num_mention	from entity metadata
tweet_type	0:none 1:RT 2:Mention 3:RT+Mention
ellipsis	counting ellipsis sign(. . .)
news	occurrence frequency of news sources
lex_diversity_tweet	proportion of unique words per tweet
lex_diversity_global	proportion of all terms in the current tweet
news_words	NLTK corpus of news article terms (sourced from Reuters), count of occurrence
chat_words	AOL messenger corpus, count of occurrence

[a]NLTK: Python Natural Language Processing Toolkit.

Table 1: The set of content-based Twitter features analyzed in our evaluation.

feature name	*description*
friend	Number of friends
follower	Number of followers
age	Number of days on Twitter
bal_soc	the ratio of follower to friends
fofe_ratio	log-log scaled ratio (followersfriends)
mutual_friends	number of mutual friendship a user has
ignoring_friends	number of single-directional relationship of a user.
listed_count	number of lists
cred_soc	deviation of user u's number of followers from the average number of followers

Table 2: The set of network-based Twitter features analyzed in our evaluation.

Social features try to assess the reliability of a person: source of the information or the person who last forwarded the information. Often these features measure either the expertise of the user or the reliability of the user based on how embedded they are in one's network. Note that these two are not necessarily the same construct. A number of features refer to the network location and properties of the user, such as number of friends and followers, number of days they have been on Twitter, etc. There are many studies that elaborate on the importance of these features. Often, they signal reputation which serves as a proxy for competence or expertise [7]. They have also been found to be relevant to certain personality traits [2]. These network features can be found in Table 2.

The third type of features are behavioral, they look at how the user behaves in the network and even more importantly how the user's followers behave towards the user. Our previous work is especially targeted towards the study of behavior both the sender and the receiver's perspective [3]. We show that two types of relationships: embedded friendship relationships and people who serve as information sources (expertise/reputation) result in different network behavior. Propagation type behavior by followers of a person combined with high number of followers, assortativity (balance of the number of friends and followers) are signals for asymmetric relationships that are based on reputation. Conversation type behavior, reciprocity of messages and propagations are a signal of friendship. These features are computed based on the statistical features of behavior between pairs of individuals without considering message contents. However, for a pair to be considered they must have at least 3 messages total and one directed message in each direction. These features are then aggregated for each user across all the friends and followers to find behavioral features for the users. Table 2 summarizes the user features that are based on behavior of the information provider and the information consumer. Note that for all features that compute the mean, we also have an equivalent feature that computes the standard deviation of the values for the same feature.

For completeness, we describe here how we compute these behavioral features. To consider conversations, we only look all directed messages between a given pair of individuals and the time of these messages. The variable τ represents the

feature name	*description*
u-url	mean # urls in tweets
u-mention	mean # hashtags in tweets
u-length	mean text length in tweets
u-balance	mean balance of number of followers
u-conv-balance	mean balance of conversations
u-tweets	total # of tweets
u-favorite	# tweets favorited
u-rt_count	total # tweets of the user that got retweeted
u-time	mean time between tweets
u-directed-ratio	# directed tweets/#broadcast tweets
u-retweet-ratio	# retweets/#tweets
u-prop-from	# users the user propagates from
u-prop-to	# users that propagate the user
u-convers-with	# users that converse with the user
u-response	mean response time to tweets
u-propagated-tweets	# tweets propagated by other users
u-propagation-energy	amount of propagation energy spent on this user by others
u-worthiness	proportion of user's tweets found worthy of propagation by others
u-propagation-reciprocation	mean propagation reciprocation
u-conv	mean # conversations
u-conv-tweets	mean # tweets per conversations (also min,max)
u-tau	mean time between messages between pairs
u-cred_rt	credibility score based on the deviation of a user's retweet rate from the normal rate
u-util_rt	utility score that considers number of retweets based on number of tweets and number of followers

Table 3: The set of behavior-based Twitter features analyzed in our evaluation.

average time between any consecutive message. We segment messages into conversations. Two messages are part of the same conversation if the time between them is less than $c\tau$ for a smoothing factor c. Given many conversations between pairs, we compute a number of features. The balance in a conversation is given by the balance of messages from A to B, and B to A. The highest value is when they are completely balanced with half of the messages in each direction. We use entropy to measure the balance (or reciprocity) of any two quantities x, y as follows. Let $p = x/x + y$, $H(p) = -p \log p - (1-p)\log(1-p)$. Conversations are undirected between a pair of users.

A propagation is directed by contrast, (A, B) means that B has propagated a message from A. For propagations, we consider both directed messages and broadcast messages. Instead of looking at whether a message was retweeted, we instead consider the timing of the messages from A and B, and use a linear time maximum matching algorithm developed in [6] between B's incoming and outgoing messages satisfying a causality constraint with respect to time. The propagation must come after the propagated message. Our previous studies have shown that there is a statistically significant correlation between our computed propagation and

Set Name	*frankenstorm +sandy*	*occupysandy*
Seed Users	2,154,735	24,463
Seed Tweets	3,801,395	60,671
Tweets in Survey	8,728	6,503
Authors in Survey	7,974	3,239

Table 4: Overview of the two topic-specific data collections mined from Twitter.

real retweet behavior [1]. Our method is a noisy indicator of propagation, but has multiple advantages. It is very difficult to assess whether a message is an individual retweet behavior as Twitter metadata only lists the original message for a retweeted message. Furthermore, information can be propagated by other means that simple retweet, such as by discussing the information content. Our previous studies have shown that the propagation behavior is different than conversation behavior and is particularly beneficial in distinguishing between different types of ties [3].

3. COLLECTION AND ANNOTATION OF TWITTER DATA

We crawled a number of topic-specific Twitter data using Twitter Streaming API starting from Oct 29th, 2012 for two weeks. We applied a few specific keywords in order to generate two topic-specific datasets:
Set1: *#frankenstorm+#sandy*, and Set2: *#occupysandy*.

The first data set was collected during Hurricane Sandy and these two terms correspond to the two main keywords used to refer to the hurricane. The second data set was collected after the Hurricane. Occupy Sandy is a coordinated relief effort to distribute resources and volunteers to help neighborhoods and people affected by Hurricane Sandy. It has been started by those who have participated in Occupy Wall Street demonstrations in 2012. Our choice of these two topics reflect two different realities about the same topic. During the hurricane, there is a great deal of uncertainty. The topic is also of great interest to a large group of people, many of whom may not know each other. After the hurricane, the relief effort likely involves a more localized group of people who are likely to know each other to some degree. We would like to compare and contrast credibility in these two different data sets.

The streaming API is not rate-limited, so is was possible to collect a large amount of tweets for our first data set (#sandy, 3,801,395 tweets). The second topic (#occupysandy) is far less popular. For this, we have collected only 60,671 tweets. More detailed statistics can be seen in Table 4. From Set1, we collected two basic samples of users. We first randomly selected 50 users from pairs of users who had exchanged 2 or more messages between themselves in each direction on the topic and identified all of their tweets (4,000 total). We then sampled the same number of tweets randomly from Set1. The users from the first set of tweets correspond to a group with some social connectivity, while the users in the second set of tweets is completely random.

In our samples, we excluded any user with more than 5K friends and 50K followers. These numbers were chosen as the mean number of friends and followers of Twitter users. We have obtained this number by crawling the user info from the 2011 NIST Twitter dataset [1]. This dataset contains 16 million tweets sampled between January 23rd and February 8th, 2011, and is meant to be a representative sample. Our reasoning for this filtering method was to remove any user who is likely to be an outlier, a power user and possibly an institution.

A similar method was followed for the second topic, which contained only 50K seed tweets. We identified users who had exchanged 2 or more messages between themselves and identified all of their tweets. We then sampled the same number of tweets randomly from Set2. For all the users in our study, we collected their list of friends and followers to be used in calculating our features.

3.1 Annotating Twitter Data

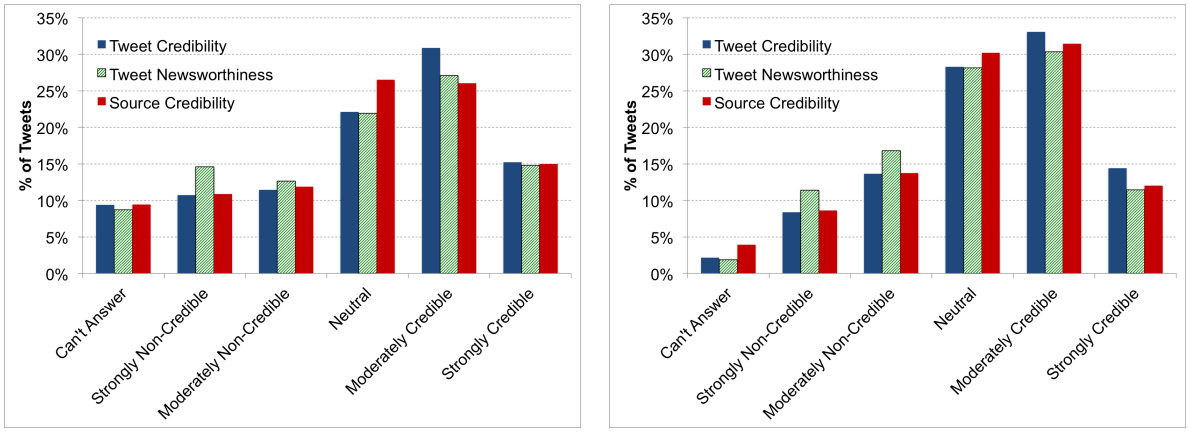
To analyze the tweets in terms of credibility, we sought three different types of annotations related to information credibility. These annotations capture three aspects of information credibility: (a) the message is credible, (b) the source is an expert in the given topic, (c) the source is a social tie who is reliable.

(a) Survey Questions

Figure 2: Screen shot from the Amazon Turk tweet assessment survey.

CRED: Credibility Survey. Human-provided assessments on groups of tweets were sourced from an online evaluation. Tweet data was presented to Amazon Mechanical Turk users who were paid a nominal amount for their participation.

[1]http://trec.nist.gov/data/tweets/



(a) #Sandy #Frankenstorm data

(b) #OccupySandy data

Figure 1: Rating Distribution from our Credibility Assessment Survey.

Figure **??** shows three screen shots of the web-based survey. Participants were presented with instructions, followed by a pre-survey questionnaire and a set of simple filtering questions to test for bots and other noise such as rapid tab-click behavior. Each participant's ability to rate was also tested using this set of pre-test questions. Those who did not answer the set reasonably were discarded, although this was unknown to them at the time of the study. (Figure 2) shows the actual survey questions. Each tweet is presented in a context similar to the standard Twitter web page, with profile name, profile icon and number of retweets displayed next to it (Figure 2). For each tweet, participants were asked to provide three Likert-scale (1-5) assessments, for message credibility, message newsworthiness and source credibility. In total 381 participants took part. Participants also had an option to select "can't" answer.

Overall participants were 35.9% female and 64.1% male, varying in age between 20s - 60s. Some post-hoc information was collected from participants after the study. For example, participants were generally familiar with the Twitter domain (3.7637/5) rating on average. In total, 381 individual credibility assessments were collected for our two topics. Participants were encouraged to provide feedback comments. From analysis of the comments it was evident that the presence of provenance features such as URLs provided a sense of credibility. However, not all users agreed with this sentiment, and one user commented that "social network activity should have no bearing on credibility". Overall, the study took less than 36 hours to reach the target number of participants using Amazon Turk, which was much faster than previous similar studies performed by the authors in the past.

The existence of images in the survey may significantly impact the evaluation of credibility as faces are often used to identify whether a source is trustworthy or not [37, 38, 36, 9]. In fact, facial evaluation is often much faster than the evaluation of text due to the dedicated processing of this signal in the brain. We expect that the source credibility judgments is highly impacted by this signal as there are only few other signals relating to the source.

CHAIN: Propagation Chains. As mentioned in the introduction, surveys are particularly useful for evaluating the text content of messages. But, there are a number limitations. The survey subjects are unlikely to be familiar with the survey topic and are more likely to use heuristics to evaluate the credibility of the message. Furthermore, as it is very unlikely that the survey subjects are familiar with the senders of the information, source credibility information will not be based on prior information regarding the source. Hence, the use of the survey is limited in measuring the credibility of the message as a function of the expertise and reliability of the source.

To overcome this problem, we compute a secondary measure of credibility based on the fact that the message was propagated in the network. This means that others in the network has endorsed the message in some way. A benefit of this method is that while the survey is a post-hoc analysis, propagation looks at how credible the message was at the time it was traveling in the network. Information that was uncertain at the creation time may be known by the time the survey is conducted. The propagation information also incorporates how credible the source of the message was. The longer a message has traveled in the network, the more credible we consider it to be. Also, messages that are part of a long chain are likely to originate from users with higher credibility and reliability. We compute a chain based on our features as follows: if B propagated a message from A, and C propagated the message from B, then we consider this a chain of length 2 for the message from A.

To compute propagations, we used the methodology described in Section 2 instead of using the retweet marker in the tweets. Note that retweets only identify the original tweet, hence are not useful for tracing how the tweet has traveled. Instead, we use our maximum matching method to determine possible propagations. Given these propagations, we computed chain length for all the messages. In total, we had 35,292 messages that were propagated, with an average chain length of 1.6108. We classified the chains into two contexts: `long-chain` (corresponding to higher credibility), having two or more propagations, and `short-chain` (corresponding to lower credibility) having one or none prop-

Context	*Description*
CRED	Message credible or not
NEWS	Message newsworthy or not
SRC_CRED	Source credible or not
CHAIN	Propagation chain long or short
DYAD	Message dyadic or not
Topic	Set1 or Set2
User_Type	Random or with Social Links (both by default)
Features	In Topic Behavior (default) or General Behavior

Table 5: Multiple segments used for evaluation

agations.

Despite the noisy nature, we found propagations to be useful behavioral indicators in our previous study [3] and that they correlated with real retweet behavior [1]. Note that some of our behavioral features based on individual propagations, not the chain length at network level.

DYAD: Dyadic Tweets. We have also considered a much weaker notion of credibility. A directed message exchanged between two people who have a social connection is likely to be credible to the receiver. To assess a social tie, we identify users who exchanged 2 or more directed messages between them, at least one message in each direction. We then annotate the directed messages between such users as dyadic. The remaining message are considered not-social.

Note that this third notion of information credibility comes from the construct of trustworthiness. A friend is trustworthy because we expect that a friend will not tell a lie. However, there is no implication of competence in this definition. A friend may not be an expert in a given topic. We are trying to model expertise with the chains discussed above. However, a message that is not credible at face value may be accepted if it comes from a friend under some circumstances. If we believe the friend to be capable of correctly processing the information (even if she is not an expert at the network level), we might believe information from her.

4. RESULTS

The main subject of our study is to understand the best indicators of credibility in different contexts and for different constructs. The distinct segments we have compiled in our study are given in Table 5. The first three segments are based on the annotations in the user survey, and the other two segments are obtained by processing the messages as described in Section 3.1. In all cases, a message is either in one segment or another (1 or 0). All these segments are computed for the tweets for which we also have survey annotations. The first five segments correspond to a type of ground truth. We will analyze them separately for each topic, and check whether there is a difference in terms of topic in each type of ground truth.

As mentioned earlier, we have two distinct data sets, with 8K tweets in Set1 and 6.7K in Set2. As mentioned earlier, half the tweets in each set come from a set of users with at least 2 or more exchanges and one in each direction, and the other half are random tweets. By default, we consider both segments in our study. But, we will also look at these two segments separately.

We computed all the features in our study based on the tweets for a specific topic and the friend/follower information of the seed users. This corresponds to topic specific behavior. This is the default features we will report here. We will also consider when non-topic specific behavior could be useful separately.

To analyze these different segments, we have run a best feature analysis to identify the most important features for annotations in a specific segmentation. In our tests, we first compute the features described in Section 2 for each tweet. Some features correspond to the message content and some features are based on the author of the message. The behavioral features are computed based on all the tweets in a specific topic, not just the ones sampled for the survey. For each tweet, we also have an annotation of 0 or 1 based on a specific segment, which we treat as the ground truth that we are trying to predict. We use a heuristic based forward subset selection based regression (FSS) to find a linear combination of the features that best predict the annotations in a given segment [3]. FSS first finds the best single feature that approximates the given ground truth annotation. Then, it adds the next feature that minimizes the leave-one-out cross validation (LOO-CV) error until no improvements can be made to the LOO-CV error. This process typically produces a very sparse set of features and prevents overfitting. We rank the features in terms of their magnitude.

4.1 Correlation

Table 6 shows the correlations between different ground truth segments for the two data sets. We can see that in both cases CRED, NEWS and SRC_CRED are highly correlated. However, this correlation is a bit weaker in set2. This might be due to the fact that in the relief effort, most tweets are about seeking and forwarding resources and are not news items. Also CHAIN and DYAD are correlated with each other as both are social dimensions, but the correlation is weaker compared to the one between CRED and NEWS. Furthermore, in set2, the correlation between CHAIN and DYAD is lower, again, most of the relief activity is about getting the word out and hence chains are a crucial component of this type of activity.

Similarly, in both sets, CRED, NEWS and SRC_CRED are not correlated with the social segments, i.e. CHAIN and DYAD, whereas CHAIN and DYAD are correlated with each other. This clearly shows that annotations CRED, NEWS, SRC_CRED and CHAIN, DYAD measure different things. In particular, we will concentrate on CRED and CHAIN next.

4.2 Credible v/s Non-Credible Tweets

Next, we examine the most relevant features for credibility (CRED) using FSS analysis. We list the top 12 features by magnitude and their weights in Table 7. Our analysis returns 22 features for set1 and 11 features for set2. First,

Correlation	*Set1*	*Set2*
CRED & NEWS	0.60	0.42
CRED & SRC_CRED	0.56	0.42
CRED & CHAIN	-0.01	-0
CRED & DYAD	-0.06	-0.03
CHAIN & DYAD	0.38	0.24
CHAIN & SRC_CRED	-0.02	+0
DYAD & SRC_CRED	-0.05	-0

Table 6: Correlation between different contexts

Set1		*Set2*	
24	news_words	19	char
-23	word	-9.8	word
-13	mutual_friends	8.5	news_words
7.8	ignoring_friends	-4.6	num_mention
-6.3	status_count	-4.5	u-mention
5.9	url	3.3	u-prop-to
5.2	pronoun	3.3	u-cred_rt
4.5	age	-3.1	u-propagated-
4.3	sentiment_pos		tweets (std)
4.1	listed_count	2.2	u-time (std)
3.6	u-mention (std)	-1.8	u-response (std)
3.6	u-url (std)	-1.7	u-response

Table 7: Best features for measuring credibility (std refers to the standard deviation of the values for that feature)

we notice that as expected, most of the top features have to do with message content. This validates our earlier assessment that surveys mostly capture the message content. In particular, for the more newsworthy topic of set 1, almost all features are related to message content. However, in Set 2, we see many prominent behavioral features as well. In particular, those related to propagations and response time are significant indicators of credibility. Users who are propagated from and quickly respond to messages of others are active users with more credible message content. As set 2 contains a closer community of users, there is significantly more behavioral data, which contributes more to credibility than the general user case. This is a good example of how the reliability of different features depend significantly on the underlying context.

4.3 Long and Short Chains

Next, we examine the top features for predicting the chain length. Table 8 lists all the predictors for CHAIN which are mostly behavioral features. As CHAIN and CRED are not correlated, the top predictors for both annotations are quite different. One interesting thing we note is that in set 1, the chain length is almost completely predicted by behavioral features while in set 2, content based features play a role. A possible explanation for this is as follows. In smaller social groups, there are some established behavior norms that may become significant in understanding the type of content. However, in the more random user group of set 1, no such norm exists. This is an interesting insight, that a social behavior can also be predicted by normative textual content.

Set1		*Set2*	
1.9	u-balance	7.8	num_mention
1.7	num_mention	-6.8	news
-1.7	news	2.3	u-propagation-energy
1.5	u-directed-	2.2	u-retweet-ratio
	ratio	1.6	u-balance
-1.1	u-conv	1.5	u-propagation-
0.8	u-retweet-		reciprocation
	ratio	1.3	favourites_count
0.8	u-	-1.1	u-conv-tweets-max (std)
	propagation-	-0.8	frown
	reciprocation	-0.8	num_hashtag
0.6	u-prop-from	-0.8	ellipsis
-0.5	u-tweets (std)	-0.8	u-tau
		0.6	question

Table 8: Best features for measuring chain length (std refers to the standard deviation of the values for that feature)

We have also observed that the tweets in the long chains in Set2 were in general more informative than the tweets in the long chains of Set1. In Set2, users are more organized and are actively trying to disseminate information.

4.4 Impact of User Type

To understand whether the user type makes a difference in our model, we compare credible tweets from socially connected users (annotation 1) to credible tweets from random users (annotation 0). Using all our features, the FSS analysis for Set1 returns no features. This means that these two segments are statistically indistiguishable from each other. In Set2, we get 33 features, almost all of them are behavioral and are mostly based on propagations. Note that these features are based solely on the behavior of the users in the topic. This reinforces our initial view that Set2 contains users with a great deal more social behavior, which results in these features being more reliable.

To see whether more detailed information about the users would change the results drastically, we collect additional information for the users. We collect their last 3,200 tweets, we then find the tweets of all their friends and followers. This data gives us features on the general behavior of the users, not just for the specific topic. In particular, we look at how features predicting Set1 differ. First of all, we find a great deal of similarity between the results. Message features are still the most relevant with few behavioral features. However, the type of behavioral features change. Instead of more general features such as number of urls and mentions in that given topic, the reciprocation of propagations and number of messages per conversation become more relevant. In essence, both types of features are likely to be useful. When the more detailed data is not available, topic level data provides fairly useful information.

4.5 Other ground truth segments

We also look at the features that best predict the remaining ground truth segments. Dyadic tweets are best predicted by behavioral features, especially those corresponding to conversations and having metnions. They happen to be less correlated (or sometimes negative correlated) with propagations. A large number of features are predictive of this more noisy indicator. However, the difference in the features show that CHAIN is a different indicator than DYAD in general.

NEWS is best predicted in Set1 by news_words, (-) word, url, listed_count, (-) status_count, (-) ignoring_friends, u-length (std), sentiment_pos, etc. In Set2, (-) chat_words and (-) num_mention, dialogue_act_type and uppercase play a role. In both data sets, what constitutes news is quite different.

SRC_CRED is best predicted by a mix of behavioral and text features in both sets. In Set1, network level properties are more prominent, while in Set2, almost all social features are behavioral. However, these features mostly point to users who send a lot messages, but are not propagated by other users. This leads us to believe that source credibility is not likely to be a reliable measure of the users, which is understandable given the limited information about each user.

4.6 Accuracy Experiments

Up to this point we have focused on discussion and analysis of the behavior of individual credibility indicators in Twitter, and how they behave across different segments of the microblog. However, it remains important to provide a concrete example of how these features can be incorporated into a prediction model and used to filter and recommend previously unseen content. Figure **??** shows a bar graph of predictive accuracy for a J-48 (C4.5 rules) learning algorithm. The algorithm was chosen because it was the best performer in a range of preliminary tests, and for comparison with [8]. The goal of the algorithm was to predict human annotation scores (Cred, News, Src-Cred) based on a set of training examples. In this graph, the Y axis shows percentage of instances correctly classified, and the X-axis shows the different segmentations that the algorithm was trained and tested on. The three bars in each segment represent tweet credibility, tweet newsworthiness and source (author) credibility respectively. The primary result from this analysis is that our feature sets are powerful enough to predict manually labeled credibility assessments to a high accuracy, similar to that reported in [8] and [15], with the exception of the long chain segment. Here, we can attribute the drop in predictive accuracy to a lack of data points to train the algorithm (hundreds compared with thousands for the other segments). Furthermore, it is interesting that the algorithm performs best when predicting different annotations across different topics. This is an indication that the learned models are leveraging underlying features in different ways as they are trained in each different topic (#frankenstorm, #occupysandy). To investigate this further, a second accuracy-based analysis was performed, this time focusing on the portability of learned models across our two data sets.

4.7 Portability of Prediction Models

Our initial hypothesis was that a drop off in predictive accuracy would be exhibited if we applied learned models from one topic to predict credibility annotations on the other.

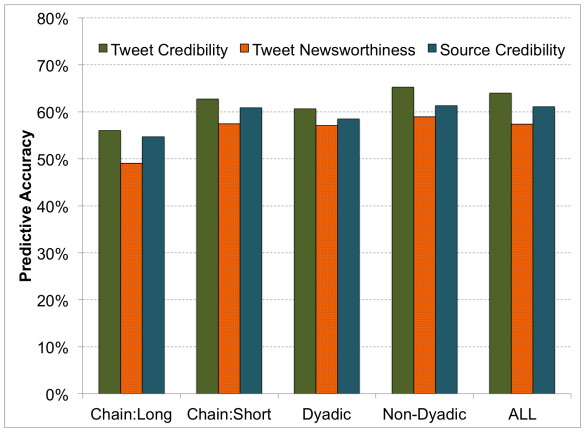


Figure 3: Results of a portability experiment in which predictive accuracy was measured for models trained on data from different topics. T1 represents the #frankenstorm data, and T2 represents the #occupysandy data

. Predictions are show for ported models, and for same-topic analysis, over Chain Length and Tweet Credibility. 10 fold cross validation was applied for all tests.

Figure 3 shows a predictive accuracy analysis for CRED and CHAIN annotations in four parts. $T1 - - > T1$ represents a model learned on #frankenstorm data and tested on an isolated portion of the same data set. $T1 - - > T2$ represents a model learned on #frankenstorm data and tested on #occupysandy data, and so on. NEED TO ADD COMMENTS ON RESULT GRAPHS HERE!!!!

5. RELATED WORK

Much of the existing literature on information credibility in Twitter focuses on the development and testing of models that *predict* credibility in some way. Approaches range from using human assessments [7, 8, 26], content-based methods [31], to model-based/machine-learned predictions [8, 15] for example. Some applications, such as Truthy [29] use visual representations to help users understand information flow and credibililty information. This paper presents a somewhat different tack, focusing primarily on understanding the behavior of underlying features that enable credibility models to function. Recent work by Microsoft in [25] describes an evaluation of credibility perception in Twitter, highlighting the increasing use of social search across the major social networks (Twitter, Facebook etc.). Given this new search paradigm, content is no longer confined to flow through existing social connections. Now, an information consumer can have an arbitrary amount of knowledge about an information producer, so it becomes important to learn about message credibility in isolation, as well as in its original network context. We believe that in order to create adaptive credibility models, capable of handling the frequent paradigm shifts that are typical of social networking applications, we must first understand the building blocks of credibility. This is the motivation behind our content-based, network-based and behavioral feature sets.

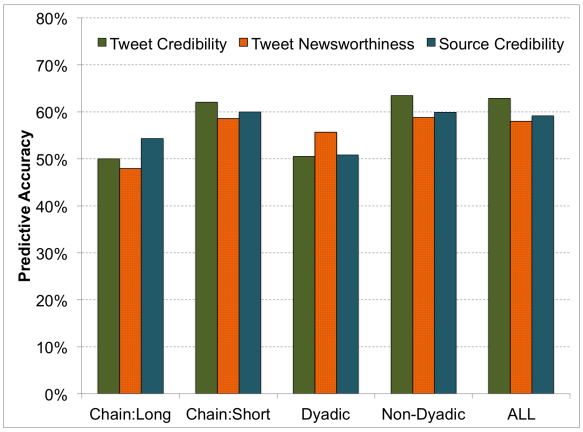


Figure 4: Results of a predictive accuracy experiment. Credibility, Newsworthiness and Author Credibility were predicted using a J-48 (C4.5 rules) learning algorithm over the #Frankenstorm data set. The five groupings are for predictive accuracy in long chains, short chains, dyadic and non-dyadic segments and all data. A 10 fold cross validation was applied in all cases.

Classification Approaches. Some approaches to modeling credibility distinguish between content and social features. [35], describe a credibility model that relies on either a content-based classifier or the social network individually to assess credibility of a message. Other approaches harness information from both sources in tandem. Canini et al. [7] present a good example of the latter, using a topic-modeling approach to infer meta-links between content and authors in the network. Like the analysis in this paper, they concentrate on topic-specific credibility, defining a ranking strategy for users based on their relevance and expertise within a target topic. Based on user evaluations they conclude that there is "a great potential for automatically identifying and ranking credible users for any given topic". Canini et al. also evaluate the effect of context variance on perceived credibility. In this paper, we provide a brief overview of a similar study performed on our data, correlating with the findings in [7] that both network structure and topical content of a tweet have a bearing on perceived credibility. Kang et al. [15] describe three computational models of credibility, derived from underlying classes of features. Their classes cover message content only, network structure (friend links etc), and a hybrid of both. The evaluation of the hybrid approach in [15] however, is focused on predictive accuracy at the level of the model, and does not question the particular attributes of the different sets of predictive features individually.

Behavioral Indicators of Credibility. The effect of trust and credibility in networks is evident across a variety of research areas, from social web search [20], semantic web [17] [39], online auctions [14] [30] [28], personality and behavior prediction [12] [1], to research of election outcomes [10, 23] and many others. Due to scale (>350 million Tweets per day)[2], network complexity, rich content and dynamic information flow, Twitter is an ideal forum for research on trust and credibility, resulting in a number of recent research efforts in the area. Moreover, the importance of credibility on Twitter has been highlighted in [32], for its effect on news distribution [16], on visibility for marketers and particularly for the potential negative effects that Twitter can have on the perceived credibility of established media such as the New York Times. The important role of credibility in preventing the spread of misinformation on Twitter has also been addressed. [8, 16] Online business applications that sell personalized statistics such as TwitterCounter [3] are becoming increasingly popular, since the dissemination of information in the microblog can have significant business impact. Twitter has been studied extensively from a media perspective as a news distribution mechanism, both for regular news and for emergency situations such as natural disasters for example [8][21][18][15]. Castillo et. al. [8] describe a recent study of information credibility, with a particular focus on news content, which they define as a statistically mined topic based on word co-occurrence from crawled "bursts" (short peaks in tweeting about specific topics). They define a complex set of features over messages, topics, propagation and users, and train a classifier that predicts at the 70-80% level for precision/recall against manually labeled credibility data. Mendoza et. al [21] also evaluate trust in news dissemination on Twitter, focusing on the Chilean earthquake of 2010. They statistically evaluate data from the emergency situation and show that rumors can be successfully detected using aggregate analysis of Tweets. Recent work by Metaxas in [22] examines credibility in politically motivated tweets during the 2010 Massachusetts senate race. They describe interesting findings about motivations for retweeting, showing that all groups studied were highly selective about what they propagated, generally only forwarding information that agreed with their own agenda.

Building Blocks of Credibility. In this section, we have described a number of different approaches to modeling credibility on Twitter from the literature. Now we must discuss the potential relevance of our contributions to the design of these models in general. Here we have proposed a set of three classes of "credibility predicting" features, and evaluated the predictive capacity of each class across a selection of different "segments" of Twitter, including long and short propagations (retweet chains), credible and non-credible assessed tweets, etc. Using the feature-distribution, correlation and rankings described in **??**, we can attempt to design a prediction model capable of adapting to the problem of varying information context, such as that described by Morris in [25]. Furthermore, we can effectively predict the impact of such transitions on our ability to detect the credibility of information in Twitter using existing models.

6. CONCLUSIONS AND FUTURE WORK

This paper has described a range of different indicators of credibility, and analyzed their predictive ability in a range of different contexts within microblog data. While the results of this analysis can potentially be useful for the design of credibility-based information filtering algorithms for microblogs, the authors believe that it is also important to

[2]http://blog.twitter.com/2012/03/twitter-turns-six.html

[3]www.twittercounter.com

combine this information with existing and proven theory on information credibility from psychological, cognitive [19] and social science [11, 24] disciplines. A follow-up paper on this topic is planned, which will integrate information about context-based feature utility with established theory to produce more efficient information filtering mechanisms for microblogs.

7. REFERENCES

[1] S. Adalı, R. Escriva, M. K. Goldberg, M. Hayvanovych, M. Magdon-Ismail, B. K. Szymanski, W. A. Wallace, and G. T. Williams. Measuring behavioral trust in social networks. In *ISI*, pages 150–152, 2010.

[2] S. Adalı and J. Golbeck. Predicting personality with social behavior. In *IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM 2012)*, 2012.

[3] S. Adalı, M. Magdon-Ismail, and F. Sisenda. Actions speak as loud as words: Predicting relationships from social behavior data. In *Proceedings of the WWW Conference*, 2012.

[4] D. Ames, S. Fiske, and A. Todorov. Impression formation: A focus on others' intents. In J. Decety and J. Cacioppo, editors, *The Oxford Handbook of Social Neuroscience*, pages 419–433. Oxford University Press, 2011.

[5] B. R. Bates, S. Romina, R. Ahmed, and D. Hopson. The effect of source credibility on consumer's perceptions of the quality of health information on the internet. *Medical Informatics and the Internet in Medicine*, 31(1):45–52, 2006.

[6] J. Baumes, M. Golberg, and M. Magdon-Ismail. Efficient identification of overlapping communities. In *Proc. International Conference on Intelligence and Security Informatics (ISI)*, 2006.

[7] K. R. Canini, B. Suh, and P. L. Pirolli. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.

[8] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

[9] L. J. Chang, B. B. Doll, M. vanâĂŹt Wout, M. J. Frank, and A. G. Sanfey. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, 61(2):87–105, 2010.

[10] T. DuBois, J. Golbeck, and A. Srinivasan. Predicting trust and distrust in social networks. In *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.

[11] A. Flanagin, M. Metzger, R. Pure, and A. Markov. User-generated ratings and the evaluation of credibility and product quality in ecommerce transactions. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1 –10, jan. 2011.

[12] J. Golbeck, C. Robles, M. Edmonson, and K. Turner. Predicting personality from twitter. In *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.

[13] B. Hilligoss and S. Y. Rieh. Developing a unifying framework of credibility assessment: Construct, heuristics and interaction in context. *Information Processing and Management*, 44:1467–1484, 2008.

[14] D. Houser and J. Wooders. Reputation in auctions: Theory, and evidence from ebay. *Journal of Economics and Management Strategy*, 15(2):353–369, 2006.

[15] B. Kang, J. O'Donovan, and T. Hollerer. Modeling topic specific credibility on twitter. In *IUI*, pages 179–188, 2012.

[16] J. Keller. How truth and lies spread on twitter. (businessweek news article)., Oct. 2012.

[17] U. Kuter and J. Golbeck. Semantic web service composition in social environments. In *International Semantic Web Conference*, pages 344–358, 2009.

[18] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[19] C. Lebiere, C. Gonzalez, and M. Martin. Instance-based decision making model of repeated binary choice. *Department of Social and Decision Sciences*, page 88, 2007.

[20] K. McNally, M. P. O'Mahony, B. Smyth, M. Coyle, and P. Briggs. Towards a reputation-based model of social web search. In *Proceedings of the 15th international conference on Intelligent user interfaces*, IUI '10, pages 179–188, New York, NY, USA, 2010. ACM.

[21] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics (SOMA '10)*. ACM Press, July 2010.

[22] P. T. Metaxas and E. Mustafaraj. From obscurity to prominence in minutes: Political speech and real-time search. In *Proceedings of Web Science Conference*, 2010.

[23] P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello. How (not) to predict elections. In *SocialCom/PASSAT*, pages 165–171, 2011.

[24] M. J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *JASIS*, 58(13):2078–2091, 2007.

[25] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[26] J. O'Donovan, B. Kang, G. Meyer, T. HŽllerer, and S. Adali. Credibility in context: An analysis of feature distributions in twitter. In *In Proceedings of the IEEE International Conference on Social Computing. SocialCom/PASSAT, Amsterdam, The Netherlands*, 2012.

[27] J. O'Donovan, B. Smyth, V. Evrim, and D. McLeod. Extracting and visualizing trust relationships from online auction feedback comments. In *IJCAI*, pages 2826–2831, 2007.

[28] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 249–252, New York, NY, USA, 2011. ACM.

[29] P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *The Economics of the Internet and E-Commerce. Volume 11 of Advances in Applied Microeconomics.*, December 2002.

[30] M. Schall, J. O'Donovan, and B. Smyth. An analysis of topical proximity in the twitter social graph. In *In proceedings of the 4th International Conference on Social Informatics, SocInfo 2012, Lauzanne, Switzerland.*, 2012.

[31] M. Schmierbach and A. Oeldorf-Hirsch. A little bird told me, so i didn't believe it: Twitter, credibility, and issue perceptions. *Communication Quarterly*, 60(3):317–337, 2012.

[32] E. Sillence, P. Briggs, P. R. Harris, and L. Fishwick. How do patients evaluate and make use of online health information? *Social Science and Medicine*, 64:1853–1862, 2007.

[33] B. Suh, L. Hong, P. Pirolli, and E. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2010.

[34] Y. Suzuki. A credibility assessment for message streams on microblogs. In *3PGCIC*, pages 527–530, 2010.

[35] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall. Inferences of competence from faces predict election outcomes. *Science*, 308:1623–1626, 2005.

[36] A. Todorov and N. N. Oosterhof. Modeling social perception of faces. *IEEE Signal Processing Magazine*, 117, 2011.

[37] V. Wout and Sanfey. Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108(3):796–803, 2008.

[38] H. Zhao, W. Kallander, T. Gbedema, and F. Wu. Read what you trust: An open wiki model enhanced by social context. In *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.