Tweeting is Believing? Understanding Microblog Credibility Perceptions

Meredith Ringel Morris[1], Scott Counts[1], Asta Roseway[1], Aaron Hoff[1], Julia Schwarz[2]
[1]Microsoft Research, [2]Carnegie Mellon University
{merrie, counts, astar, aaronho}@microsoft.com, julenka@cs.cmu.edu

ABSTRACT
Twitter is now used to distribute substantive content such as breaking news, increasing the importance of assessing the *credibility* of tweets. As users increasingly access tweets through search, they have less information on which to base credibility judgments as compared to consuming content from direct social network connections. We present survey results regarding users' perceptions of tweet credibility. We find a disparity between features users consider relevant to credibility assessment and those currently revealed by search engines. We then conducted two experiments in which we systematically manipulated several features of tweets to assess their impact on credibility ratings. We show that users are poor judges of truthfulness based on content alone, and instead are influenced by heuristics such as user name when making credibility assessments. Based on these findings, we discuss strategies tweet authors can use to enhance their credibility with readers (and strategies astute readers should be aware of!). We propose design improvements for displaying social search results so as to better convey credibility.

Author Keywords
Twitter, Microblogging, Social Search, Credibility.

ACM Classification Keywords
H5.m. Information interfaces and presentation: *Misc*.

General Terms
Design, Experimentation.

INTRODUCTION
The popular microblogging service Twitter [twitter.com] lets users broadcast 140 character status messages known as *tweets*. Users currently assess tweets' credibility based on trust relationships with authors whose streams they elect to follow. However, consuming social media by *searching for a topic* rather than *following an author* is becoming increasingly prevalent. By June 2011, Twitter's search portal [search.twitter.com] was already servicing over 1.6 billion queries per day [35]. In addition to supporting explicit querying, Twitter also provides clickable "trending topic" terms, which launch searches for popular (and sometimes advertiser-promoted) keywords. General-purpose search engines have created separate portals specifically for searching public social streams, such as Bing Social Search [bing.com/social] and Google Real Time Search [google.com/realtime]. Google and Bing have also recently begun integrating social status updates directly into their main search results pages when appropriate [24, 33], enabling serendipitous encounters with socially-generated content.

Twitter acts not only as a social network, but as a news source [19]. Informing oneself about breaking news is a common motivation for searching public tweets [37], such as when seeking updates about local emergencies [39]. Unfortunately, social search tools amplify the audience not only of breaking news, but also of undesirable memes such as spam [36] and rumors [5]. Although some rumors, such as false reporting of celebrity deaths [5], are relatively harmless, increased reliance on social media for actionable news items (*Should I vote for candidate X? Should I donate to victims of disaster Y?*) makes credibility a nontrivial concern. Evidence of false tweets has recently been discovered in U.S. Senate campaigns [23], reporting of the Iranian election protests [8], and coverage of unfolding natural disasters such as the Chilean earthquake [22].

Factors influencing users' perceptions of the credibility of Web pages and earlier forms of social media (e.g., blogs and instant messaging) have been well-studied. Some influential factors for Web pages' credibility perception, like visual design [21], are not relevant to tweets, while others, like conflating search engine ranking with credibility [15], may apply. The relative importance of features may vary for this new medium, as well; for instance, author credentials [18] may take on heightened importance, given the social nature of tweets.

In this paper, we present an investigation of user perceptions of tweet credibility. We report the results of a survey about the features that impact users' assessments of tweet credibility. Our results indicate a discrepancy between features people rate as relevant to determining credibility and those that mainstream social search engines make available. Based on these findings, we conducted two controlled experiments to measure the impact of several tweet features (message topic, user name, and user image) on perceptions of message and author credibility. Our results indicate that tweet consumers have difficulty discerning truthfulness based on content alone, with

CSCW 2012, February 11–15, 2012, Seattle, Washington, USA.

message topic, user name, and user image all impacting judgments of tweets and authors to varying degrees regardless of the actual truthfulness of the item. Based on these findings, we discuss strategies tweet authors can use to enhance their credibility with readers, and manipulative strategies that wary social media consumers should be mindful of. We also offer ideas for redesigning tweet search result pages to better support credibility assessment.

RELATED WORK

The volume of activity on Twitter has increased at an extraordinary rate, to an average of 140 million tweets per day as of March 2011 [38] – this volume of information makes it increasingly impractical for users to monitor all messages from their network in order to identify the most relevant pieces of information. This has prompted researchers to develop novel interfaces and algorithms for filtering tweets, such as tools like Twahpic [28] or Eddi [1], which perform topic-based clustering and filtering, or algorithms for identifying Twitter authors with authority on specific topics [26]. Search-based access to tweets has become increasingly available through third-party search tools, including Google [24] and Bing [33], and researchers have begun to study how people use search to access microblog updates [7, 37]. Informing oneself about breaking news events is a common motivation for searching tweets [37], which is consistent with Twitter's increasing prominence as a news source [19].

Searching Twitter for news updates can provide users with real-time information not yet available in the mainstream media, such as when eyewitnesses' tweets provided the first information about a plane crash-landing in the Hudson River [20]. Unfortunately, the quality of news posted to Twitter is not uniform – spam [29, 23, 36], surreptitious advertising [14], false rumors [5, 8, 17, 22], and imposter accounts [25] are common occurrences. Users' ability to assess the *credibility* of tweets, therefore, has taken on increased importance. This paper presents both self-report and experimental data on the features that impact users' credibility assessments of tweets.

The credibility of information encountered online is a problem that has long perplexed educators and librarians, who have developed systems of heuristics for students to use when they encounter online content, such as assessing the accuracy, authority, objectivity, currency, and coverage of the material [18]. Researchers have studied factors influencing users' perceptions of Web page credibility [9, 10, 11, 15, 21], identifying many factors that contribute to this assessment including features of the Web page itself (e.g., visual design [21]), properties of the user (e.g., level of internet use [9]), and means of encountering the content (e.g., search engine ranking [15]). Fogg's Prominence-Interpretation theory [11] suggests that the impact that a Web page element has on perceived credibility depends on both its prominence (likelihood of being noticed) and interpretation (the meaning assigned to it). However, users are often forced to make credibility judgments about online content before having the opportunity to view a full web page – Schwarz and Morris [34] and Yamamoto and Tanaka [41] have proposed techniques for supplementing search results to support credibility assessment.

Like Web search results, tweets pose an example of a particularly challenging credibility-assessment scenario, due to their compact nature (a limit of 140 characters). The opportunity for customization of visual design, which is an important factor in the assessment of traditional Web pages' credibility, is quite limited, other than users' ability to select the avatar they use to represent themselves. While some researchers have studied the issue of credibility assessment for more social subsets of the Web, such as blogs [13, 30, 40], the issue of how users perceive the credibility of microblog updates is only just beginning to receive attention.

Schmierbach and Oeldorf-Hirsch [31] showed college students articles on the New York Times Web site as well as tweets from the official New York Times Twitter feed describing those same stories, and found that the students rated the news items less credible when reading the tweet than when viewing the website. While their finding indicates that users may have concerns about credibility when consuming tweets, it does not indicate what aspects of the tweets contribute to this impression; our research contributes findings to clarify this latter issue.

Pal et al. [27] did not examine credibility *per se*, but asked users to rate how "interesting" a tweet was and how "authoritative" its author was, manipulating whether or not they showed the author's user name. They found that authors who had more followers (and therefore presumably more recognized user names) received higher "interesting" ratings for their content when their user names were revealed. User names of organizations, rather than individuals, and those which were topically related to the tweet also received higher ratings than those which were not. User name style is one of several features we explore in this paper, though we focus on how this feature impacts credibility perceptions rather than content interestingness.

Some researchers have begun building systems to automatically or semi-automatically classify tweet credibility. Truthy [29] visually represents the diffusion of a Twitter meme; through crowdsourcing, these visualizations are inspected and flagged for potentially spam-like patterns. Castillo et al. [3] used Mechanical Turk to crowdsource judgments of tweet credibility, and used these judgments to train a machine learning system that rates the credibility of tweets on a particular topic. Our findings could enhance such automatic techniques by providing information about the features that end users rely on to make such judgments – such features could benefit automatic credibility classification by, for example, suggesting tweaks to feature weightings in machine learning approaches, or by helping determine what information to feature more (or less) prominently in crowdsourcing tasks.

SURVEY

To better understand the factors influencing users' perceptions of tweets' credibility, we conducted a survey. The following sub-sections discuss the methods used to design the survey, the question types, and the participants. We then report on our findings.

Survey Design

We started by conducting a pilot study in which we observed users thinking aloud while conducting a search on the search.twitter.com webpage. Five people, ranging in age from 17 – 49 years old, participated in the pilot. Participants had non-technical occupations (e.g., photographer, sales representative). Participants were familiar with Twitter and occasionally read tweets, but only one had an account.

Participants were given a task intended to simulate a realistic information need. Since all participants were residents of the state of Washington, they were instructed to search on Twitter's search engine for the name of a local candidate in the upcoming U.S. Senate election, in order to learn about his positions on the issues. Participants were instructed to "think aloud" while viewing the retrieved tweets. The experimenter prompted further think-aloud by asking questions about some of the tweets, such as whether participants thought certain tweets were from the candidate in question, from official news sources, etc.

The experimenter took notes on the features that participants mentioned paying attention to as they analyzed the Twitter search results. For example, participants often commented on the nature of the avatar associated with particular tweets, noting that a particular tweet seemed untrustworthy because the man in the photo "looks like a stalker," or that another seemed less official because the user's avatar was an "anime character." Participants attributed importance to user names, assuming that names that were linguistically similar to their search terms (e.g., "Dino Rossi HQ") were officially sanctioned. Repetition of similar content by multiple tweets increased participants' confidence in the veracity of a message. Some features were not explored by participants unless prompted by the experimenters, such as clicking URLs or clicking user names in order to view an author's Twitter homepage (which contains biographical information and that user's recent tweet history); when prompted to view these features, however, users noted their value – for example, one participant doubted the seemingly official nature of a particular tweet after noting the unprofessional visual design of the author's homepage.

The collection of 26 features discussed by these five pilot participants was used to design our survey. Respondents were presented with the list of features and asked to indicate whether they typically pay attention to each feature when reading tweets ("usually", "occasionally", "never"); the "never" option was split to allow the user to indicate that they never consider a feature but think they probably ought to, or that they never consider a feature and think there is no value in doing so. For each feature, respondents were also asked to assess how that feature impacts credibility on a five-point Likert scale ranging from "greatly decreases credibility" to "greatly increases credibility." The survey also asked users to indicate credibility concerns regarding various sources and topics of tweets, and gathered information about users' Twitter habits, social search habits, and demographic data. We asked only whether people pay attention to user images, but then broke out images into four types when asking about their impact on credibility.

Participants

Sampling a diverse array of Twitter users for our survey was a challenging goal, since directly purchasing advertising on Twitter was only an option for special "partner" companies as of late 2010 (the time when we distributed our survey) [32], with most advertising being conducted through tweets planted in popular accounts [14]. Advertising to our own Twitter followers was also undesirable, due to the drawbacks of snowball sampling techniques [2]. Consequently, in order to obtain a reasonably diverse sample of Twitter users for our survey, we advertised the survey in two venues: on an email list for social media users within Microsoft, and on a message board for alumni of Carnegie Mellon University. We received a total of 256 completed surveys, 101 from the corporate group and 155 from the alumni group. Reading tweets at least occasionally was a prerequisite for participation in the survey.

The Microsoft respondents ranged in age from 18 – 60 years old, with an average age of 32 years. 29% were female. 93.1% had a Twitter account, and all read tweets, with 91% reading them at least a few times a week, and 74% reading them at least once a day. All worked in the technology industry, albeit in a variety of job roles, including software development, marketing, legal, HR, and management.

The university alumni respondents ranged in age from 18 – 54 years old. 34% were female. 88% had a Twitter account, and all read tweets, with 91% reading them at least a few times a week, and 77% reading them at least once a day. Occupations varied. "Student" was the most common occupation (29.6%), but the majority of respondents had non-student professions such as administrative assistants, journalists, lawyers, architects, financial professionals, dentists, nurses, and customer relations specialists.

Responses to our survey from these two participant pools were similar in character, so we report their results jointly. The similarity of the two groups' results suggests the applicability of our findings beyond a single demographic. The reader should bear in mind that some demographics that consume tweets were not covered by our recruitment method, such as teenagers or adults without a college degree; such groups may have different perceptions of credibility, and studying their habits is left to future work.

Results

Here we establish that participants do encounter tweets through search and that this elicits greater concern for credibility than encountering tweets by those followed. We then describe the topic areas most pertinent to credibility concerns, and finally the extent to which participants reported using various tweet features when making credibility judgments.

Method of Encountering Tweets

In addition to reading tweets from users they followed, respondents consumed tweets by conducting searches on search.twitter.com (84%), clicking trending topics on the Twitter homepage (84%), searching for tweets using Bing's and Google's social search functionality (72%), or serendipitously encountering tweets mixed into the results of general Web searches (81%). While respondents place a great deal of trust in tweets from users they follow, tweets encountered through Twitter search ($\chi 2(2, N = 256) = 44.7$, $p < .001$) and general search engines ($\chi 2(2, N = 256) = 47.2$, $p < .001$) elicited concern (Figure 1).

Tweet Topic Type

Respondents were least concerned with credibility for celebrity news and gossip related tweets, and secondarily for movie and restaurant reviews. News, political, emergency, and consumer oriented tweets caused the greatest concern about credibility (Figure 2).

Tweet Features

Table 1 summarizes each feature's impact on credibility perceptions. Features associated with low credibility perceptions were the use of non-standard grammar and punctuation, not replacing the default account image, or using a cartoon or avatar as an account image. Following a large number of users was also associated with lower author credibility, especially when unbalanced in comparison to follower count – as one respondent articulated, "if someone is following significantly more people than they have following them, I mistrust them."

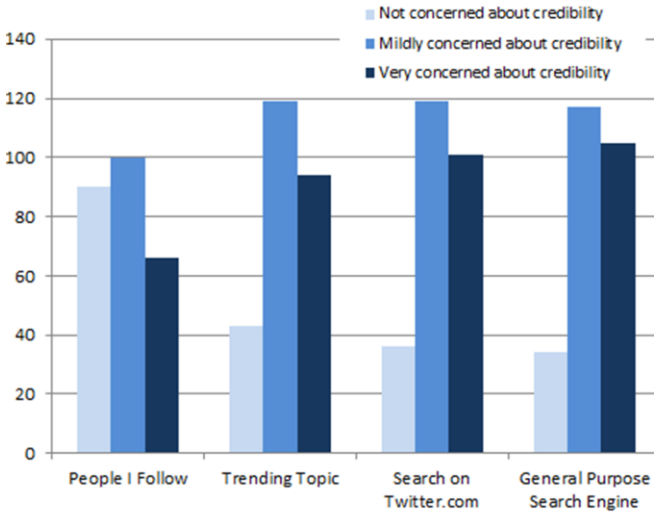

Figure 1. Histogram of respondents' credibility concern levels based on method of encountering a tweet. Tweets encountered through searching inspire greater credibility concerns than those encountered through following.

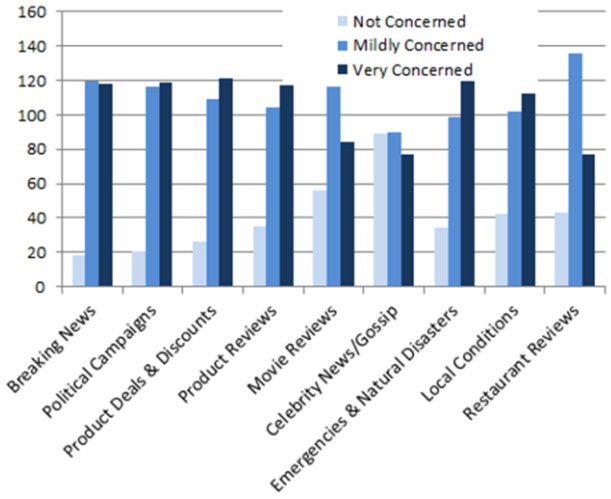

Figure 2: Histogram of credibility concerns for different types of content.

Features perceived as most enhancing a tweet's credibility generally concerned the author of the tweet. These included author *influence* (as measured by follower, retweet, and mention counts [1]), *topical expertise* (as established through a Twitter homepage bio, history of on-topic tweeting, pages outside of Twitter, or having a location relevant to the topic of the tweet), and *reputation* (whether an author is someone a user follows, has heard of, or who has an official Twitter account verification seal). Content-related features viewed as credibility-enhancing were containing a URL leading to a high-quality site, and the existence of other tweets conveying similar information.

Features respondents attend to also focused on author characteristics (whether the author is known or followed) and features immediately visible in the interface, especially properties of the tweet (e.g., is a retweet).

Discussion

Participants' responses indicated an awareness of the problem of tweet credibility, particularly for tweets not encountered through their following stream (Figure 1). Concern for credibility also varied across topic types (Figure 2). Participants were aware of features that differentially convey credibility, yet the features most attended to suggest that their ability to judge credibility in practice is largely limited to those features visible at-a-glance in current UIs (user picture, user name, and tweet content). Conversely, features that often are obscured in the user interface, such as the bio of a user, receive little attention despite their ability to impact credibility judgments. In the following section we experimentally examine the impact of the most salient features on credibility judgments of tweets and authors.

Feature	Credibility Impact	Attention Received
non-standard grammar/punctuation	2.71	1.46
default user image	2.87	n.a
cartoon/avatar as user image	3.22	n.a.
author is following many users	3.30	1.29
logo as user image	3.37	n.a.
contains shortened URL	3.39	1.89
customized Twitter homepage	3.41	1.22
author location near you	3.43	1.34
contains hashtags	3.48	2.05
contains a URL	3.50	1.91
author tweets frequently	3.52	1.68
contains complete URL	3.57	1.80
near top of search result list	3.58	1.66
posted recently	3.59	2.10
is a reply	3.61	2.09
author has many followers	3.65	1.56
author bio suggests topic expertise	3.66	1.60
is a retweet	3.66	2.17
username is related to topic	3.67	1.85
author location near topic	3.67	1.34
author often mentioned/retweeted	3.69	1.66
personal photo as user image	3.70	n.a.
many tweets w/ similar content	3.71	2.07
author often tweets on topic	3.74	1.96
account has verification seal	3.92	1.83
author is someone you've heard of	3.93	2.37
contains URL you clicked thru to	3.93	2.20
author is someone you follow	4.00	2.40
verified author topic expertise	4.04	1.84
is a RT from someone you trust	4.08	2.43
user image, generally	n.a.	1.75

Table 1. Mean ratings for tweet features' perceived credibility impact (5-point scale; higher = more credibility) and attention typically allotted (3-point scale; higher = more attention).

EXPERIMENTS

To better understand the impact of various tweet features on credibility perceptions, we conducted two online experiments in which we systematically altered several properties of tweets in order to measure their impact on readers' credibility assessments. We first describe our primary experiment and its results, followed by a description of a follow-up experiment we ran to explore an issue raised by our initial findings.

Method

To design a balanced experiment of a reasonable length (so as not to exhaust participants), we narrowed down the list of features examined in our earlier survey to a set of three. We selected the features to focus on in our experiment by choosing features that were self-reported as being highly influential on credibility perceptions in our survey, and that were currently and highly visible to end-users on Twitter.com and major search engines. Using these criteria, the three features we selected were *Message Topic*, *User Name*, and *User Image*.

Message Topic

Our survey participants indicated that their level of concern about credibility was affected by a message's topic; hence, we chose topic as a factor to manipulate in our experimental design. We included tweets on three different topics, each of which has been subject to false tweet phenomena: *politics*, *science*, and *entertainment*.

User Name

Survey participants indicated attributing credibility to user names, such as assuming that topically-relevant user names were associated with credible information. Additionally, survey participants expressed concerns about tweets containing non-standard grammar such as abbreviations commonly used in IM or text messaging, and other internet-age modifications to language. These factors led us to include three types of user name in our experiment: *traditional* (e.g., "Alex_Brown"), *internet* (e.g., "tenacious27"), and *topical* (e.g. "AllPolitics").

All user names were gender neutral. Traditional style user names were constructed by selecting first names from a list of popular gender-neutral baby names in the United States, and selecting last names from a list of common surnames in the United States. We verified that all user names were not actual registered Twitter account names, so that participants would not have prior assumptions about the quality of tweets from a particular author.

User Image

Survey participants indicated that a tweet author's account image (or "avatar") influenced their credibility perceptions, reporting that they attributed the most credibility to photos, followed by cartoons/icons, with the default image inspiring the least credibility. We therefore decided to include image type in our study design; we included five image types: *Male Photo*, *Female Photo*, *Topical Icon*, *Generic Icon*, and *Default*.

Male and female photos were chosen by taking photos from real users' twitter accounts, so as to achieve realistic photographic styles. Photos were obtained by using Twitter's search engine to search on stereotypically gendered topics (e.g., #nfl for men and #twilight for women). Photos from popular or celebrity accounts were not used. To avoid race- or age-based confounds in our results, we selected only headshot photos of Caucasians who appeared to be young adults (in their twenties or thirties).

Generic icon photos were also chosen from twitter accounts, to enhance realism of style. Topical icons were constructed using PowerPoint clip art, since many topical icons we encountered when searching twitter were associated with organizational accounts that may have been

prominent enough to be familiar to study participants, and we did not want familiarity to be a confound. The "default" image used was the Twitter egg image that appears if a user does not upload his/her own image.

Tweet Content

We authored original tweets for the purposes of the experiment, in each of the three chosen topic areas. All tweets were in English, and were written with standard grammar, spelling, and punctuation. All tweets described a topically-relevant current event, followed by a URL. URLs were constructed to look like they were from the URL-shortening service bit.ly [bit.ly.com], a popular service that is frequently used to compress URLs so that they fit within the 140-character limit of tweets.

To validate that participants' credibility judgments were influenced by the features we manipulated (topic, user name, and/or user image) rather than the actual truth value of the tweet itself, we designed the tweets such that half described true news events and half described events that had never taken place, but were plausible. We pilot-tested these tweets on ten members of our organization to verify that they could not determine which were true or false.

Tweets were rendered using a .css stylesheet copied from Twitter.com and saved as image files, so that they looked as realistic as possible, but without clickable links (so that participants could not verify the credibility of a tweet by clicking on the fake bit.ly URL included with each). Figure 3 shows sample tweet stimuli created for our experiment.

Study Design

Running a fully within subjects study design [message topic x user name x user image x truth value] would have required each participant to rate 90 (3 x 3 x 5 x 2) tweets, which we felt would be exhausting for participants. Instead, we made user image a between-subjects factor, resulting in 18 (3 x 3 x 2) tweets to be rated by each participant.

Hence, the set of 18 tweets consisted of six in each of the topic areas (politics, science, and entertainment). Within a topic area, there were two tweets of each user name style (traditional, internet, and topical), one of which was true and one of which was false. A given user saw these 18 tweets combined with one of the five user image types (with a different user image randomly combined with each tweet, except in the "default" image condition, where the the same image was used each time). No user ever saw the same user name, user image, or tweet more than one time; seeing only one tweet from a given author is similar to a Web search scenario (as opposed to a following scenario).

Participants completed the study online, in their Web browser. They were randomly assigned to one of the five user image conditions. The 18 tweets were shown to them one at a time, in a random order. Instructions reminded the participants that the links within tweets would not be clickable (ensured through the use of pre-rendering the tweets as images rather than using live HTML), and were

Figure 3. Sample tweets constructed for our experiment. The five image types are (a) default, (b) topical icon, (c) generic icon, (d), male photo, and (e) female photo. Tweet (a) employs a *traditional* user name, (b) and (c) depict *internet* style names, and (d) and (e) feature *topical* names.

instructed not to leave the current Web page or perform supplementary Web searches. Underneath each tweet, participants saw two 7-point Likert scales (from strongly disagree to strongly agree), asking them to rate whether "this tweet contains credible information" and whether "this author is credible." After rating all 18 tweets, participants took a survey that collected basic demographic information, as well as information about their use of Twitter.

Participants

Our online experiment was conducted during a one-week period in February 2011. It was advertised via email to 1,000 adult U.S. residents who had signed up through our organization's user study recruitment website. The recruitment email indicated that to be eligible for the study, participants had to know what tweets were and read them occasionally (familiarity with tweets was also re-verified through self-report on the post-study questionnaire). Entry into a drawing for an e-commerce gift certificate was offered as incentive for completing the study. 266 participants completed the study.

Random assignment to each of the five user image conditions resulted in approximately equal distribution of participants, with 54 participants seeing the *default* image, 52 seeing *generic icons*, 54 seeing *topical icons*, 51 seeing *female photos*, and 55 seeing *male photos*.

Participation was approximately gender-balanced, at 54% male and 46% female. 28% of participants were in the 18 – 24 age range, 28% were 25 – 34, 26% were 35 – 44, and the remaining 18% were 45 or older. Participants had a wide variety of occupations, including medical professionals, clergy, homemakers, students, web designers, personal fitness trainers, financial professionals, and sales.

49% of participants reported reading tweets at least once a day, with an additional 25% reading tweets at least a few times a week; the remaining 26% read tweets less

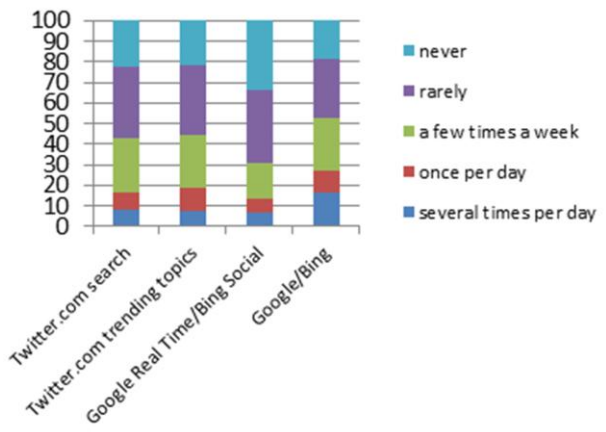

Figure 4. This figure shows the percent of respondents encountering tweets through different search mechanisms at different frequencies. About half of our experiment's participants reported encountering tweets through search (rather than following) at least a few times a week.

frequently. Participants read tweets both from people they followed (75% had Twitter accounts) and tweets they discovered using various search services. See Figure 4.

Participants who had Twitter accounts reported following a mean of 322 people (median = 30), and having a mean of 268 followers (median = 20). 44% of those with accounts reported authoring tweets at least a few times a week.

Results

Since participants' ratings of *tweet credibility* and *author credibility* were approximately normally distributed, ANOVA analyses were performed to test the impact of our experimental manipulations on credibility ratings, with follow-up pairwise t-tests when appropriate. Bonferroni corrections were used to mitigate the effect of multiple comparisons.

Tweet Credibility vs. Author Credibility vs. Truth

Participants provided two ratings for each tweet, one quantifying the credibility of the information conveyed by the tweet and one quantifying the credibility of the tweet's author, since our earlier survey results indicated that users' credibility perceptions are influenced by features of both the content and the author. We hypothesized that these concepts may not be independent of each other, since perceptions of an author may influence perceptions of their message (and vice-versa); indeed, our results indicate that participants' tweet credibility and author credibility ratings were highly correlated, with a Pearson correlation coefficient of $r = .85$ ($p < .001$). Average ratings for both tweet credibility (3.79) and author credibility (3.27) were slightly below the neutral point on our 7-point scale.

Averaging credibility scores on a per-tweet basis and comparing to the actual truth value for each tweet, we found only a moderate correlation between truth and tweet credibility rating ($r = .39$, p = .11) and between truth and author credibility rating ($r = .29$, p = .25). Neither of these correlations were statistically significant. This suggests that participants were generally unaware of the true truth value of the messages, and that their credibility judgments were mostly influenced by factors other than truthfulness.

Participants' prior experience using Twitter did not impact their ability to distinguish true from false tweets. Individual users' correlation coefficients for credibility ratings and truth value did not correlate significantly with users' Twitter account duration, frequency of reading or sending tweets, or follower/following counts. However, participants with more Twitter experience (those who reported authoring tweets at least a few times a week), gave higher tweet credibility ratings ($t(264) = 2.45$, $p = .01$) and author credibility ratings ($t(264) = 2.01$, $p < .05$) than those who had less experience, confirming prior findings that those with more experience with a given technology view it as a more credible information source [9]. We found no other interaction of demographics (age, gender, or Twitter experience) with any of our experimental manipulations.

Message Topic

Message topic influenced perceptions of tweet credibility ($F(2, 264) = 5.72$, $p = .003$), with science tweets receiving a higher mean tweet credibility rating (3.90) than those about either politics (3.74, $p = .001$) or entertainment (3.74, $p = .01$). Message topic had no statistically significant impact on perceptions of author credibility ($F(2,264) = 2.52$, $p = .08$).

User Name

Figure 5 shows how user name type impacted credibility ratings. User name type influenced perceptions of tweet credibility ($F(2, 264) = 23.36$, $p < .001$), with topical user names receiving a higher mean tweet credibility rating (3.99) than either traditional (3.69, $p < .001$) or internet (3.70, $p < .001$) name styles.

User name had an even more pronounced impact on perceptions of author credibility ($F(2, 264) = 62.64$, $p < .001$), with all pairwise differences significant at the $p < .001$ level. Topical user names inspired the highest mean

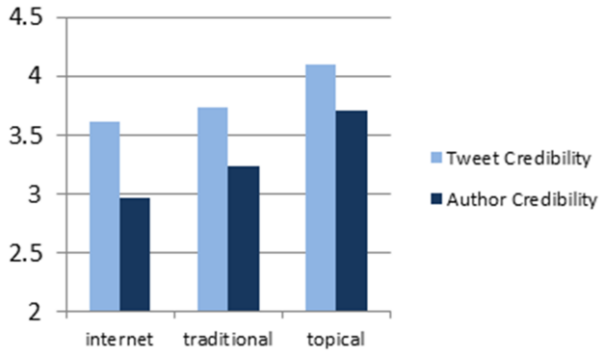

Figure 5. Users perceived topical user names as enhancing tweet credibility. Authors with topical names were considered more credible than those with traditional user names, who were in turn considered more credible than those with internet name styles.

credibility ratings (3.62), followed by traditional names (3.21), with internet style names inspiring the lowest credibility in the associated authors (2.97).

User Image

We found no significant impact of user image type on either tweet credibility ratings ($F(4, 261) = .36$, $p = .84$) or author credibility ratings ($F(4, 261) = 1.17$, $p = .33$).

Because our survey participants reported paying attention and assigning credibility value to user images, we were surprised to see little impact of image type on credibility perceptions in our experiment. We hypothesized that this lack of effect may have been due to the between-subjects nature of the user image feature (a study design selected so as to reduce the total number of tweets users would need to evaluate, thereby reducing fatigue). Specifically, seeing only a single image type may have made users ignore that factor (e.g., if everyone has the "default" image, it is not a noteworthy feature). Consequently, to further understand the impact of user image on credibility perceptions, we designed a follow-up experiment.

Follow-Up Experiment

The second study was similar to the first, except that each participant experienced two different image types (having each participant experience all five image types would have been impractical due to the large number of ratings that would be required from each user to achieve a balanced design). In order to enable us to study the relationships between each of our five user image types, participants were randomly assigned to one of ten study conditions, corresponding to the ten different possible pairings of our user image types.

We used the same 18 tweets and user names as in the first experiment. Recall that there were six tweets in each of the three topics (politics, science, and entertainment). Within each topic, there were two tweets for each user name style (traditional, internet, and topical). Each pair of tweets with a particular topic + user name style consisted of one true and one false tweet. Even though truth value had little impact in our original experiment, as a precaution we randomized the assignment of user image type with respect to truth value on a per-tweet and per-user basis. As before, all tweets were presented in a randomized order.

The follow-up experiment was conducted online during a one-week period in March 2011. It was advertised via email to a set of 1,000 adult U.S. residents that did not overlap with the set invited to participate in the original experiment. 296 people completed the study. 111 (37.5%) were women, 185 (62.5%) men. 77% of participants were aged between 25 and 54. Occupations ranged widely, including educators, retail workers, doctors, and writers. Participants followed 172 people on average (median = 44) and were followed by an average 263 people (median = 30).

Results

Use of the default Twitter icon significantly lowers ratings of content ($t=2.41$, $p = .02$) and marginally lowers ratings of authors ($t=1.93$, $p = .06$; Figure 6) in comparison to the other four image types. No other image types showed significant differences in comparison to one another. Examination of the relationship between content type and image type revealed that the default image pales most in comparison to topical, male, and female images. For instance, in the entertainment category, for ratings of authors, topical ($t=2.93$, $p < .01$), male ($t=2.21$, $p = .03$), and female ($t=3.81$, $p < .001$) images all generated significantly higher ratings than did the default Twitter icon.

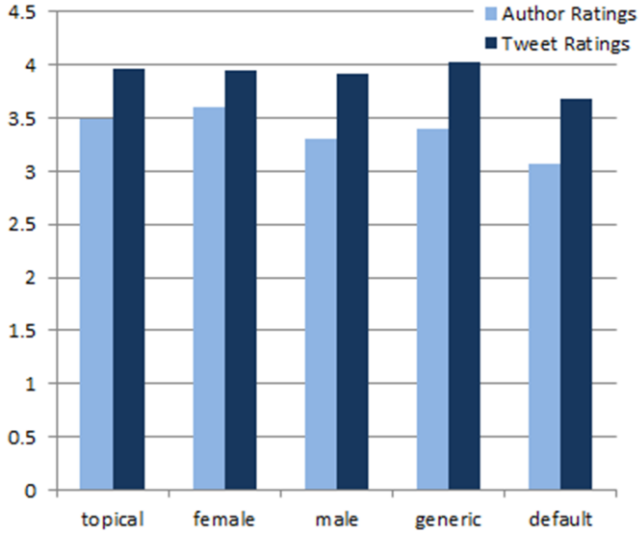

Figure 6. Histogram of respondents' credibility concern levels based on image type. The default Twitter image lowers credibility.

DISCUSSION

Our survey showed that users are concerned about the credibility of content when that content does not come from people the user follows (Figure 1). In contexts like search, users are thus forced to make credibility judgments based on available information, typically features of the immediate user interface (Table 1). Our survey results indicated features currently underutilized, such as the author bio and number of mentions received, that could help users judge tweet credibility.

It is sensible that traditional microblog interfaces hide some of these interface features because they aren't necessary when only consuming content from known authors. Without these established relationships, errors in determining credibility may be commonplace. Participants were poor at determining whether a tweet was true or false, regardless of experience with Twitter. In fact, those higher in previous Twitter usage rated both content and authors as more credible. This mirrors findings with internet use generally [9], and may be due to a difficulty in switching from the heavily practiced task of reading content from authors a person follows to the relatively novel task of reading content from unknown authors. Even topical expertise may not support reliable content validity assessments. We did find that for politics, those higher in

self-reported expertise (by a median split) gave higher credibility ratings to the true political tweets ($t=3.67$, $p<.001$) and their authors ($t=2.00$, $p=.05$), yet these effects disappear for the science topic and for entertainment where those *low* in expertise actually gave slightly (though non-significantly) higher ratings to the true content.

In the absence of the ability to distinguish truthfulness from the content alone, people must use other cues. Given that Twitter users only spend 3 seconds reading any given tweet [6], users may be more likely to make systematic errors in judgment due to minimal "processing" time. Indeed, participants rated tweets about science significantly more credible than tweets on politics or entertainment, presumably because science is a more serious topic area than entertainment. Other types of systematic errors, such as gender stereotyping based on user image, did not appear to play a role. Although our survey respondents reported finding non-photographic user images less credible, our experiment found that in practice image choice (other than the detrimental default image) had little effect on credibility judgments. It is possible that image types we did not study (such as culturally diverse photographs) might create a larger effect.

The user name of the author showed a large effect, biasing judgment of both content and authors. Cha et al. [4] discuss the role of topically consistent content production in the accumulation of followers. We see a similar phenomenon reflected here in users incorporating the degree of topical similarity in an author's user name and tweets as another heuristic for determining credibility.

Implications

What are the implications of these difficulties in judging credibility and how can they be mitigated? Our experimental findings suggest that for individual users, in order to increase credibility in the eyes of readers, they should start by avoiding use of the default twitter icon. For user names, those who plan to tweet exclusively on a specific topic (an advisable strategy for building a large follower base [4]), should adopt a topically-aligned user name as those generated high levels of credibility. If the user does not want a topical username, she should choose a traditional user name rather than one that employs "internet" styled spelling.

Other advice for individual tweet authors stems from our survey findings. For instance, use of non-standard grammar damaged credibility more than any other factor in our survey. Thus, if credibility is a goal, users are encouraged to use standard grammar and spelling despite the space challenges of the short microblog format, though we note that in some user communities non-standard grammar may increase credibility. Maintaining a topical focus also increases credibility, as does geographic closeness between the author and tweet topic, so users tweeting on geographically-specific events should enable location-stamping on their mobile devices and/or update their bio to accurately identify location, which is often not done [16].

Tweet consumers should keep in mind that many of these metrics can be faked to varying extents. Selecting a topical username is trivial for a spam account. Manufacturing a high follower to following ratio or a high number of retweets is more difficult but not impossible. User interface changes that highlight harder to fake factors, such as showing any available relationship between a user's network and the content in question, should help. The Twitter website, for instance, highlights those in a user's network that have retweeted a selected item. Search interfaces could do something similar if the user were willing to provide her Twitter credentials. Generally speaking, consumers may also maintain awareness of subtle biases that affect judgment, such as science-oriented content being perceived as more credible.

In terms of interface design, we highlight the issue that users are dependent on what is prominent in the user interface when making credibility judgments [11]. To promote easier credibility assessment, we recommend that search engines for microblog updates make several UI changes. Firstly, author credentials should be accessible at a glance, since these add value and users rarely take the time to click through to them. Ideally this will include metrics that convey consistency (number of tweets on topic) and legitimization by other users (number of mentions or retweets), as well as details from the author's Twitter page (bio, location, follower/following counts). Second, for content assessment, metrics on number of retweets or number of times a link has been shared, along with who is retweeting and sharing, will provide consumers with context for assessing credibility. In our pilot and survey, seeing clusters of tweets that conveyed similar messages was reassuring to users; displaying such similar clusters runs counter to the current tendency for search engines to strive for high recall by showing a diverse array of retrieved items rather than many similar ones – exploring how to resolve this tension is an interesting area for future work.

CONCLUSION

Social media are increasingly being incorporated into general search engine results. Google and Bing both feature Twitter and Facebook in "social search" results. Twitter's own search was used by a significant percentage of our survey respondents. While a potentially valuable source for news and information, this transition removes a critical element of social media: that users are friends or followers of the content author. The result is that users must judge the credibility of content authored by people they do not know.

In this work, we examined key elements of the information interface for their impact on credibility judgments. We showed that users had difficulty determining the truthfulness of content and that their judgments were often based on heuristics (e.g., an item has been retweeted) and biased systematically (e.g., topically-related user names

seen as more credible). In our discussion, we highlight pieces of information deemed helpful to credibility judgments that typically are buried in the interface. Many of these elements, such as an author's bio, may be minimally important when the reader knows the author but highlight critical interface changes needed to help users determine validity of content in social search contexts.

REFERENCES

1. Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S., and Chi, E. Eddi: Interactive Topic-based Browsing of Social Status Streams. *UIST 2010*.
2. Bernstein, M., Ackerman, M, Chi, E., and Miller, R. The Trouble with Social Computing Systems Research. *alt.chi 2011*.
3. Castillo, C., Mendoza, M., and Poblete, B. Information Credibility on Twitter. *WWW 2011*.
4. Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM 2010*.
5. Corcoran, M. Death by Cliff Plunge, With a Push From Twitter. *The New York Times*, July 10, 2009.
6. Counts, S. and Fisher, K. Taking it All In? Visual Attention in Microblog Consumption. *ICWSM 2011*.
7. Efron, M. Information Search and Retrieval in Microblogs. *Journal of the American Society for Information Science and Technology*, March 2011.
8. Esfandiari, G. The Twitter Devolution. *Foreign Policy*, June 7, 2010.
9. Flanagin, A.J. & Metzger, M.J. (2000). Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly*, 77(3), 515-540.
10. Fogg, B.J. Persuasive Technology: Using Computers to Change What We Think and Do. *Morgan Kaufmann*, 2002.
11. Fogg, B.J. Prominence-Interpretation Theory: Explaining How People Assess Credibility Online. *CHI 2003 Extended Abstracts*.
12. Fox, S., Zickuhr, K., & Smith, A. Twitter and Status Updating, Fall 2009. *Pew Internet &American Life Project*, Oct. 21, 2009.
13. Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., & Konig, A.C. BLEWS: Using Blogs to Provide Context for News Articles. *ICWSM 2008*.
14. Grover, R. Ad.ly: The Art of Advertising on Twitter. *Businessweek*, January 6, 2011.
15. Hargittai, E., Fullerton, F., Menchen-Trevino, E., & Thomas, D. Trust Online: Young Adults' Evaluation of Web Content. *Int'l. Journal of Communication*, 2010.
16. Hecht, B., Hong, L., Suh, B., and Chi, E. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. *CHI 2011*.
17. Kanalley, C. Facebook Shutting Down Rumor Goes Viral: Site Said to be Ending March 15, 2011. *The Huffington Post*, January 9, 2011.
18. Kapoun, J. Teaching Undergrads WEB Evaluation: A Guide for Library Instruction. *C&RL News*, 1998.
19. Kwak, H., Lee, C., Park, H., & Moon, S. What is Twitter, a Social Network or News Media? *WWW 2010*.
20. Lamont, I. Plane Lands on the Hudson, and Twitter Documents it All. *Computerworld*, Jan. 15, 2009.
21. Lazar, J., Meiselwitz, G., & Feng, J. Understanding Web Credibility: A Synthesis of the Research Literature. *Foundations and Trends in Human-Computer Interaction*, 1(2), 2007, 139-202.
22. Mendoza, M., Poblete, B., & Castillo, C. Twitter Under Crisis: Can We Trust What We RT? *KDD 2010 Social Media Analytics Workshop*.
23. Mustafaraj, E. & Metaxas, P. From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search. *Web Science 2010*.
24. Official Google Blog. An Update to Google Social Search. February 17, 2011.
25. Owens, S. How Celebrity Imposters Hurt Twitter's Credibility. *Mediashift*. February 20, 2009.
26. Pal, A. & Counts, S. Identifying Topical Authorities in Microblogs. *WSDM 2011*.
27. Pal, A. & Counts, S. What's in a @name? How Name Value Biases Judgment of Microblog Authors. *ICWSM 2011*.
28. Ramage, D., Dumais, S., and Liebling, D. Characterizing Microblogs with Topic Models. *ICWSM 2010*.
29. Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Patil, S., Flammini, A., and Menczer, F. Truthy: Mapping the Spread of Astroturf in Microblog Streams. *WWW 2011*.
30. Rubin, V.L. & Liddy, E.D. Assessing Credibility of Weblogs. *AAAI Spring Symposium*, 2006.
31. Schmierbach, M. and Oeldorf-Hirsch, A. A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions. *Association for Education in Journalism and Mass Communication*, August 4, 2010.
32. Schroeder, S. Twitter Takes a Step Toward a Turn-Key Advertising Solution. *Mashable*, December 15, 2010.
33. Schwartz, B. Bing Adds Twitter Smart Answers. *Search Engine Land*, July 1, 2009.
34. Schwarz, J. and Morris, M.R. Augmenting Web Pages and Search Results to Support Credibility Assessment. *CHI 2011*.
35. Siegler, M.G. At 1.6 Billion Queries Per Day, Twitter Finally Aims To Make Search Personally Relevant. *TechCrunch*, June 1, 2011.
36. Sullivan, D. Twitter's Real Time Spam Problem. *Search Engine Land*, June 6, 2009.
37. Teevan, J., Ramage, D., & Morris, M.R. #TwitterSearch: A Comparison of Microblog Search and Web Search. *WSDM 2011*.
38. Twitter Blog. *blog.twitter.com.* "#numbers," March 14, 2011.
39. Vieweg, S., Hughes, A., Starbird, K., and Palen, L. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. *CHI 2010*.
40. Weerkamp, W. and de Rijke, M. Credibility Improves Topical Blog Post Retreival. *ACL 2008*.
41. Yamamoto, Y. & Tanaka, K. Enhancing Credibility Judgment of Web Search Results. *CHI 2011*.