Toward Formal Reasoning with Epistemic Policies about Information Quality in the Twittersphere

Brian Ulicny
VIStology, Inc
Framingham, MA, U.S.A.
bulicny@vistology.com

Mieczyslaw M. Kokar
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA, USA
m.kokar@neu.edu

Abstract – Some recent systems accurately produce high-level situational awareness by mining traffic in Twitter. Where these systems have been successful, there has been no attempt to evaluate Twitter streams for source reliability and information credibility because the situations have not been adversarial. The use of Twitter in recent political dissent in the Mideast makes the need for computationally tractable approaches to evaluating Twitter source reliability and information credibility more acute in order to produce accurate situation awareness in the face of misinformation or deliberate disinformation.

Keywords: Twitter; soft data fusion; situation awareness; information evaluation; reliability; credibility; source independence; social network analysis

1 Introduction

Twitter has become the best-known example of a global broadcast system for short "status update" messages. Such platforms have recently become associated with organizing and mobilizing political dissent and disruption [1]. In the 2011 "Arab Spring" uprisings in the Middle East in Tunisia, Egypt [2], Yemen and elsewhere Twitter and Facebook are widely believed to have played a major part in organizing and mobilizing elements of society to overthrow the governments in those countries, although some observers have stated that the contributory role of social media platforms like Twitter in similar uprisings in Iran and Moldova has been overstated [3]. As unrest continues in the Mideast, regardless of whether Twitter and similar social media platforms are essential technologies for initiating and organizing such dissent or not, it is clear that the use of technologies like Twitter cannot be ignored as a important source of situation awareness data for soft data fusion.

Twitter, on which we will focus here, is a platform by which users can sign up for a free, password-authenticated account anywhere in the world. Users can post short messages with a 140-character limit associated with their username via their computer, smartphone or SMS (text); currently, approximately 55 million tweets are sent each day [8]. Messages are time stamped. Users can address another user with an *@tag*: a username prepended with '@'. Users can annotate a message by topic with a *hashtag*: a folksonomy term prepended with '#'. Users can subscribe to the messages of other users by *following* them. Follower lists and @tag uses thus create a visible social network for Twitter users. Users can also send private messages to someone who follows them by prefixing their message with 'DM' (direct message) and the username. Users typically shorten URLs in their tweets by means of various services (e.g. bit.ly) to maximize the 140-character message length. These shortened URLs are unique to the originating message. Users can also *retweet* a message, indicating whom it came from by simply prepending the message with 'RT' and the originators username. Users can automatically associate a geolocation with their message if their phone or other device using Twitter supports this and they have turned this option on. Less than 1% of Twitter status updates are geolocated currently. Twitter messages are archived and become unsearchable after six months [4].

Twitter users can provide a short profile message, a profile picture, a profile location, and a URL to provide more background. Twitter, and other such platforms, are particularly interesting because they are public. Anyone can follow what is going on in the Twittersphere simply by 'following' users or topics (hashtags) or keywords, via applications such as TweetDeck (tweetdeck.com).

Twitter verifies some (mostly celebrity) users' identities, and indicates this status on their profile. In general, however, users are not verified, and anyone can tweet under whatever username and profile they like. Thus, it is possible to tweet under a false identity. Twitter suggests that by providing a link to one's Twitter feed on one's website, this provides a kind of user authentication.

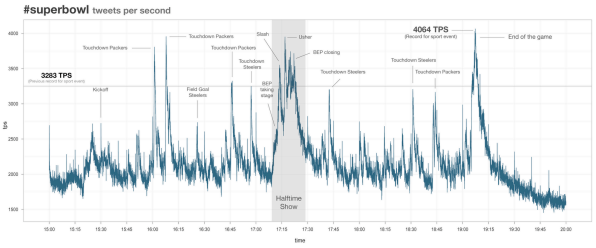
Although, we focus on Twitter here, Facebook and Google Buzz provide similar functionality. Additionally, the Ushahidi platform (ushahidi.org) combines a map – based interface with the ability to post reports by location, via cell phone texts or from Twitter or anonymously from the web, primarily in humanitarian relief situations. It has been used to monitor election fraud in Afghanistan and responses to the 2010 Haitian earthquake.

By monitoring Twitter, in principle we can discover what users are talking about and interested in from moment to moment. Although individual tweets may not

978-0-9824438-2-8/11/$26.00 ©2011 IEEE

provide much insight, aggregated Tweets can convey a strong signal about the situation they reflect. For example, Figure 1, from the Twitter blog[1], graphs tweets per second over time for the hashtag #superbowl, as updated during the 2011 NFL Superbowl game. The spikes in the graph of tweets per second correlate strongly with important moments in the game, such as one team scoring. Other spikes correlate with moments in the game's half-time show, particularly the surprise appearance of one performer. Armed only with these tweets, it is likely that one could recreate an accurate account of what happened in the game and when, by looking for commonalities in the messages at the times corresponding to spikes.

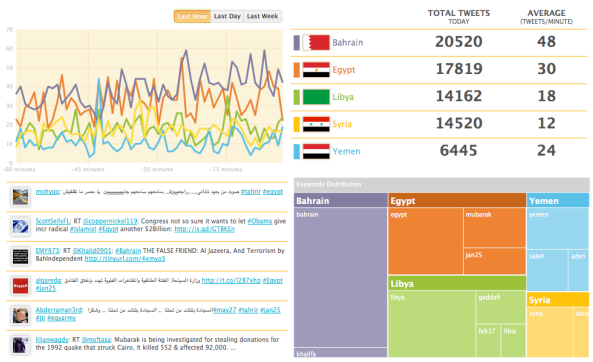
Figure 1 NFL Superbowl 2011 #Superbowl Tweets per second (from Twitter blog)



Similarly, Culotta has shown [14] that influenza outbreaks can be geospatially tracked in near-real time quite effectively just by looking for simple keywords in tweets, and mapping their geolocation or profile location. Culotta validated his output by comparing Twitter results with weekly epidemiological reports from the Center for Disease Control.

What the Super Bowl and flu situations have in common is that there is little reason for a Twitter user to publish disinformation in these scenarios. The situation is reflected by similarities among a large number of tweets. Thus, Al Jazeera's Twitter Monitor (Figure 2) provides statistics like Tweets per Minute and most common keyword by country to track the political situation.

Figure 2 Al Jazeera Twitter Dashboard (http://blogs.aljazeera.net/twitter-dashboard)



[1] http://blog.twitter.com/2011/02/superbowl.html

In this paper, our focus will be not primarily on producing situation awareness from a multitude tweets but in formally evaluating tweets for their information quality along several dimensions that are relevant to adversarial, or partially adversarial, situations. That is, while current approaches to situation awareness via Twitter treat every tweet at face value, because of the adversarial nature of the recent struggles in which Twitter plays a large part, it is prudent to treat tweets differentially in terms of their reliability, credibility, and other epistemic properties before constructing situation awareness from them.

2 Information Evaluation

NATO STANAG (Standard Agreement) 2022 "Intelligence Reports" states that [6] where possible, "an evaluation of each separate item of information included in an intelligence report, and not merely the report as a whole" should be made. It presents an alpha-numeric rating of "confidence" in a piece of information which combines an assessment of the information source's reliability and a numeric assessment of the credibility of a piece of information "when examined in the light of existing knowledge".[2]

Source **Reliability** is designated by a letter A to F signifying various degrees of confidence as follows:

A: Completely reliable. It refers to a tried and trusted source which can be depended upon with confidence.

B: Usually reliable. It refers to a source which has been successfully used in the past but for which there is still some element of doubt in particular cases.

C: Fairly reliable. It refers to a source which has occasionally been used in the past and upon which some degree of confidence can be based.

D: Not usually reliable. It refers to a source which has been used in the past but has proved more often than not unreliable.

E: Unreliable. It refers to a source which has been used in the past and has proved unworthy of any confidence.

F: Reliability cannot be judged. It refers to a source which has not been used in the past

The **Credibility** of a piece of information is rated numerically from 1 to 6 as follows:

1: If it can be stated with certainty that the reported information originates from another source than the already existing information on the same subject, then it is classified as "confirmed by other sources".[3]

[2] The same matrix is presented in Appendix B "Source and Information Reliability Matrix" of FM-2-22.3 "Human Intelligence Collector Operations" (2006) without citing STANAG 2022, and in Sections 4-24 and 4-25 of FM 2-22.9 "Open Source Intelligence" (2006). JC3IEDM [7] includes a reporting-data-reliability-code rubric that is nearly identical, with some quantitative guidance ("not usually reliable" means less than 70% accurate over time.)

[3] JC3IEDM's reporting-data-accuracy codes are nearly identical to these except that the top three categories refer to confirmation by 3, 2 or 1 independent sources, respectively. JC3IEDM also contains an additional, unrelated reporting-data-credibility-code (reported as fact, reported as plausible, reported as uncertain, indeterminate); it is not clear how it relates to the others.

2: If the independence of the source of any item of information cannot be guaranteed, but if, from the quantity and quality of previous reports, its likelihood is nevertheless regarded as sufficiently established, then the information should be classified as ``probably true''.
3: If, despite there being insufficient confirmation to establish any higher degree of likelihood, a freshly reported item of information does not conflict with the previously reported behaviour pattern of the target, the item may be classified as ``possibly true''.
4: An item of information, which tends to conflict with the previously reported or established behaviour pattern of an intelligence target should be classified as ``doubtful'' and given a rating of 4.
5: An item of information that positively contradicts previously reported information or conflicts with the established behaviour pattern of an intelligence target in a marked degree should be classified as ``improbable'' and given a rating of 5.
6: An item of information the truth of which cannot be judged.

As such, the credibility metric involves notions of source independence, (in)consistency with other reports, and the quality and quantity of previous reports. Confidence is the combination of the two values.

2.1 Current Approaches to Reliability

The STANAG 2022 standard for evaluating reliability is based on past accuracy: a source is considered reliable to the extent that its past statements have been true. Trust is a correlate of reliability: it is rational for someone to trust a source or system to the extent that it is reliable. (In human behavior, trust undoubtedly has many irrational components as well.)

It is not clear how source reliability is tracked and monitored by human analysts in practice today, but it is clear that with the multitude of Twitter users posting messages, it is impossible to individually vet each one. As of November, 2010, there were 175 million registered users on Twitter [8], and even though perhaps less than 25% of these were active users (in that they followed at least 10 users, were followed by at least 10 users and had tweeted at least 10 times [9]) it would still be practically impossible to vet the reliability of the 44 million users that met those criteria. Twitter currently adds 370,000 new users per day [8]. Moreover, as Barracuda Labs reports, in 2009 Twitter shut down 12% of new user accounts for violating their policies [9]. So, while Twitter does police itself to some extent, a potentially large number of Twitter users may be unreliable on any given day.

Further, it is known that "persona management software" has been developed and deployed that allows users to create and manage "cyber presences that are technically, culturally and geographacilly [sic] consistent" "replete with background, history, supporting details" [25][26]. Clearly, such fictitious users are not reliable.

In the contemporary operating environment, an analyst is exposed to many novel sources of information across PMESII-PT categories and has very little ability to verify their reliability directly [12]. The STANAG 2022 standard requires that novel information sources be given an unknown reliability rating (F), but that seems unreasonable. The STANAG 2022 rubric treats all novel information sources as equally suspicious, when in fact most users are comfortable with indirect estimates of unknown data reliability.

In contemporary text-based information retrieval models, an information quality metric is computed for all documents in addition to the relevance metric, matching a document to the specific information need expressed by the query. This is done independently of assessing their reliability directly. That is, contemporary search engines consider two factors when they return a document in response to a query: a representation of what the document is about, usually based on the frequency distribution of terms in a document and across other documents; and a representation of how good the document is, based on an analysis of network properties. Google, that is, does not fact-check the content of a site to evaluate its information; it uses network properties that it believes are highly correlated with information quality or reliability as a correlate of reliability; these rankings can change as user hyperlinking behavior changes.

Google's PageRank algorithm [10] and variants to it have been highly successful in presenting users with reliable information without direct fact-checking. The PageRank algorithm calculates a document's quality recursively, weighing inlinks from high-quality documents (those that are themselves pointed to by high quality documents) more highly. The PageRank algorithm is recursive and typically computed for only a small number of iterations for which it is assumed to converge, because it would be too computationally expensive to extend the computation to the entire Web graph. Hyperlinks are assumed to be made by disinterested parties, not for the sake of PageRank itself. Google ferrets out "Link-farming" designed to inflate PageRank.

Many other highly successful information evaluation technologies have evolved that all rely, to one degree or another, on network analysis properties: centrality, overlap, distance and so on. These networked-based metrics, like PageRank, are clearly applicable to many online open-source intelligence sources, such as news sites and blogs, to provide an estimate of reliability, even when they have not been encountered previously.

Blogs, for example, have been an important venue for political mobilization and recruitment. Technorati, a blog search engine, uses the relatively simple metric of in-link centrality, the number of links from other blogs over the last six months, as their blog quality metric, rather than PageRank. The present authors have shown that a metric combining both Technorati authority and reader engagement, as measured by blog comment counts, as well as accountability-enhancing profile features, outperforms both PageRank and Technorati Authority alone in ranking social-political blogs, in this case in

Malaysia, by their authoritativeness or influence [11].

Vark (Vark.com), recently acquired by Google, is a social question-answering application that attempts to automatically identify the person in a user's social network (gleaned from their Facebook, Twitter, IM (instant messenger) contacts and the like) that is most likely to be able to answer the question, i.e. the most reliable source for the user's question with respect to their social network. This user-respondent quality metric is computed over a feature vector that includes both social network proximity and overlap metrics as well as metrics of topic overlap (vocabulary and stated interests) and demographic overlap. The Vark service manages connecting the asker and respondent and handling their interaction. In other social question-answering services, like Yahoo! Answers[4], unfamiliar users can be assessed by means of statistics compiled for the number of times a user's answers have been voted the best answer, and the number of questions answered overall. Social search metrics such as those incorporated by Vark and Yahoo! would be applicable to estimating reliability among teammates or coalition partner information sources, such as non-governmental organizations (NGOs) and the like, whose information is likely to be important in full spectrum counterinsurgency environments. Such metrics are also applicable to estimating the reliability of unfriendly or potentially hostile sources with respect to their social networks.

All of these metrics depend on identifying central figures in a network. A highly central figure has accumulated more authority, and is therefore more likely to be reliable than a marginal figure in a social network, at least with respect to information that relevant to its participants. If someone were unreliable, they wouldn't gain followers or citations. We propose, then, that network-theoretic centrality metrics used in civilian information retrieval applications, should be investigated for systematically estimating source reliability where direct assessment is impractical or unfeasible, such as in the Twitter network.

2.2 Current Approaches to Credibility

STANAG 2022's credibility rubric ranks a piece of information's credibility on the basis of (i) assertion of the same information, by (ii) an independent source (iii) consistent with other reports. STANAG 2022's highest credibility ranking goes to information that is independently confirmed and not contradicted, by other reports. The lowest credibility ranking goes to those reports that contradict previous information.

Many information portals on the Web address the credibility of the information they provide by either limiting the information they provide to highly regarded sources (e.g. Wolfram Alpha [19]) or by "crowdsourcing" the policing of the accuracy of the information by letting anyone revise the information until a consensus is reached (e.g. Wikipedia). Neither approach is applicable to Twitter since Twitter tries not to censor tweets, nor is there a common version of every assertion that can be edited, as in Wikipedia and other wikis.

In information retrieval, text-based question-answering systems have used sameness of text in search snippets to identify credible answers to factual questions in a textual corpus. The AskMSR system [16], for example, identified the most frequent phrases proximate to query terms in highly ranked search result snippets as the answer to a "factoid" question, such as "What is the capital of Sweden". The intuition here is that if a phrase appears in the context of question terms in search results snippets for many URLs, then it is likely that this phrase is the correct answer to the question. Or, at least, this is a way to identify the consensus answer to a question. Leveraging data redundancy in raw Web documents, rather than relying on curated reports, helps the system to provide more accurate answers. Such systems are less useful if the correct answer can change quickly with time.

In [13], the authors provide a sophisticated method for estimating the proportion of texts expressing the same sentiment in a corpus (e.g. Twitter updates expressing the same attitude about the State of the Union) without training individual classifiers for each type. This has been incorporated into the Crimson Hexagon social media analytics service[5]. Crimson Hexagon identifies sameness of attitude across messages rather than sameness of propositional content.

Typically, contemporary search engines do not evaluate source independence in ranking results. If two documents are from different domains, they are taken to be independent. A search for a phrase in Google News may return multiple URLs that all quote or derive from the same source [18]. News content is often syndicated across many different publications by wire services and the like.

Aside from curated sites, search engines make no attempt to evaluate the consistency of the information returned, as opposed to evaluating the information source itself via some centrality metric. While social question-answering systems incorporate metrics for source quality, we are not aware of social search systems that attempt to validate a respondent's answer by calculating its consistency with a body of prior knowledge. One exception (although not really a social search system, per se) is the winning team from MIT at DARPA's Network Challenge, in which ad hoc teams, recruited and interacting via social media, competed to identify the location of ten balloons placed across the continental US. Teams were competing for money, and substantial disinformation from other teams was encountered. The MIT team evaluated the proximity of a balloon reporter's IP address to the reported location of a balloon, among other factors, in evaluating a report's credibility [20].

[4] http://answers.yahoo.com

[5] http://www.crimsonhexagon.com

In conclusion, it is clear that innovative metrics are required for evaluating Twitter feeds according to the STANAG 2022 rubric.

3 Applying STANAG 2022 to Twitter

In order to reason about the STANAG 2022 rubric as applied to Twitter, we represent Twitter data as an RDF graph, using the Twitter-to-RDF conversion service called "Shredded Tweet" provided by Mark Borkum.[6] Shredded Tweet converts Twitter search results into RDF/XML (Resource Description Framework), using a variety of namespaces and properties from well-known ontologies, including the Dublin Core Metadata Initiative[7], the SIOC (Semantically-Interlinked Online Communities) Core Ontology[8], and the FOAF (Friend of a Friend) vocabulary specification.[9] A simple tweet produces at least 24 RDF triples: five with the user as subject, and ten with the tweet as the subject. We reason with rules over the resulting RDF graph using BaseVISor, a semantic web inference engine, to annotate reports with STANAG 2022 metrics.

3.1 Source Reliability in Twitter

As we have said, it is impractical to vet the reliability of individual Twitter users directly by evaluating the ratio of accurate to inaccurate reports that they produce. However, in other contexts, it has become common to use network centrality metrics as a proxy for source quality. Since Twitter is a network structure, this is an attractive option here as well. Clearly, source quality and reliability must be correlated, at least for sources that make factual assertions (as opposed to non-factual jokes or opinions).

Our choices for centrality measures include simple indegree centrality (the number of followers a user has) or some variant of eigenvector centrality (PageRank): the number of high quality followers that a user has, where quality is determine recursively by the number of high quality followers those users have. Indegree centrality can easily be inflated via fake users [27]. Since a Twitter account can be obtained at the cost of a valid email address, it is relatively easy to automatically create an account with many followers. Therefore, simple indegree centrality (follower counts) is not a good proxy for reliability in adversarial situations.

Daniel Tunkelang's TunkRank metric [21] can be adapted as an apt measure of eigenvalue centrality for Twitter. Tunkelang's algorithm recursively produces a TunkRank score based on the expected number of people that will see a message that X tweets and the (assumed) constant probability *p* that a user will retweet a post that they have seen from someone that they follow (Equation 1). TunkRank differs from indegree centrality in that a user with many followers who are not themselves followed by anyone would receive a TunkRank of zero. Thus, TunkRank cannot be easily inflated simply by providing a Twitter user with many fake followers and is therefore immune to manipulation by persona management software.

$$Influence(X) = \sum_{Y \in Followers(X)} (1 + p * Influence(Y))/||Following(Y)||$$

Equation 1 TunkRank equation

The top Twitter users by TunkRank are listed at http://tunkrank.com/score/top. At first glance, the empirical results seem to be troubling. Although some reasonable figures are present (Barack Obama, BBC Breaking News, The White House), other figures rank high that are not known for their reliability (e.g. the satirical fake news site "The Onion", and prankster Ashton Kutcher). Although this seems to argue against the utility of the TunkRank measure, it is worth remembering that what TunkRank really measures is the pass-along value of a Twitter user to his or her followers. For most sources, not known for their wit or celebrity, the only reason to follow them is their factual reliability.

Entertainers who tweet have a different kind of value to their followers than their accurate reporting of facts. They say amusing things that their followers wish to pass along to their followers. However, since there are fewer entertainers on Twitter, and since they can be somewhat reliably identified independently, we embrace the TunkRank algorithm as an appropriate basis for assigning STANAG 2022 reliability scores.

The TunkRank API provides a percentile for each user, indicating the percent of Twitter users that have a lower TunkRank. We map these percentiles to STANAG 2022 values as in Table 1:

Table 1 Mapping TunkRank Scores to STANAG 2022 Reliability

TunkRank	**Stanag 2022 Reliability**
> 90th percentile	A: Completely Reliable
> 80th percentile	B: Usually Reliable
>50th percentile	C: Fairly Reliable
< 50th percentile	D: Not Usually Reliable
< 10th percentile	E: Unreliable
Undefined	F: Cannot Be Determined

This stands in contrast to previous work [22] in which we counted any Twitter user that was a news organization, as determined by their profile URL, as A: Completely Reliable and all other users as F: Reliability Cannot Be Determined. As in that work, the reliability of a Twitter user is annotated as a triple in the RDF graph of the relevant tweets.

In general, the accuracy of what a source reports via a tweet cannot be determined by formal reasoning. It requires external verification. Therefore, we make no attempt to modify a Twitter user's reliability based on what they say. We let other Twitter users 'vote' on their reliability via their decision to follow or not follow a

[6] http//shreddedtweet.org/
[7] http://dublincore.org/documents/2010/10/11/dcmi-terms/
[8] http://rdfs.org/sioc/spec/
[9] http://xmlns.com/foaf/spec/

Twitter user, and more importantly, to pass along what they say.

However, we can directly determine that a Twitter user is unreliable in a certain class of cases, through application of formal rules. One such type of case in when a Twitter user misrepresents the provenance of a (purportedly) retweeted message. Perhaps because the retweeting convention is something that arose after Twitter was established, nothing in Twitter prevents a user from posting a retweet and falsely attributing it as originating with another user. For example, nothing prevents a user from posting:

```
RT @whitehouse Zombie uprising in
Scranton, PA!
```

Such a tweet has the appearance of retweeting a report by the US President's staff that there is a Zombie outbreak in Scranton, PA. The user makes the assertion while attributing it (falsely) to someone else. False rumors can easily be promulgated this way, by leveraging the popularity and perceived reliability of the retweeted user to start a rumor cascade [28].

Using the RDF graph constructed from tweets, we can formally check for this, however. A rule can be asserted, in a semantic web rule language such as BaseVISor rule language [23], saying that if a user retweets a tweet for which there is no corresponding original, then that user is E: Unreliable, i.e.:

False Retweet Rule: If (?a rdf:type b:MicroBlogPost) & & (?a sioc:created ?t1) & (?a sioc:has_creator ?user1) & (?a sioc:content ?c) & (?c sioc:body ?d) & (?d matches "^RT ?user2 ?text") & not((?e rdf:type b:MicroBlogPost) & (?e sioc:created ?t2) & (?t1 > ?t2) & (?e sioc:has_creator ?user2) & (?e sioc:content ?f) & (?f sioc:body ?g) & (?g matches "^?text$"))), then (?user1 has_reliability "E: Unreliable")

This rule states that if there is no way to assign variables (indicated by ?) to elements of the RDF graph that satisfy the retweet pattern, then the retweet is bogus and, therefore, the retweeter is unreliable.

Similar rules can be asserted that if a user retweets the same URL as a link from a tweet on different days with different content, none of which is reflected in the content of the URL, then that Twitter user is unreliable. Twitter itself polices users for similar violations. It is a common scam on Twitter for users to identify trending topics, via the Twitter API, and create tweets using those terms that point to unrelated URLs in order to drive traffic to those sites. Such users, too, should be downgraded.

If A tweets message M, and B retweets A's tweet, and C retweets B's tweet, then the same message may be associated with sources of increasing or decreasing reliability, depending on A, B and C's followers. Unless the user is caught faking the chain of custody for a tweet, or reusing URLs unrelated to the content of the tweet, the reliability of a user depends only on the TunkRank of that user and his followers, not the content of the message.

To illustrate, we have selected a set of 20 tweets, posted from 3:27 PM April 20, 2011 and 12:02 PM April 21, 2011, reporting the death of photojournalist Tim Hetherington[10] in Misrata (Misurata), Libya (Figure 3). The 20 Twitter users reporting this event have an average of 10,768 followers, far greater than the general Twitter average of 27 followers.[11]

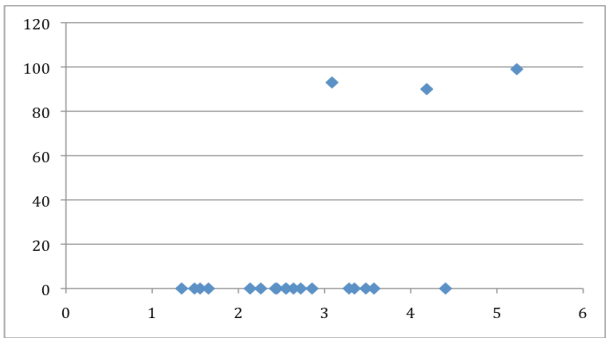


Figure 3 TunkRank (y axis) vs Followers (log scale) for Twitter Users Reporting Hetherington Death, Apr, 2011.

Based on TunkRank, however, 17 of the users have a TunkRank of 0 (corresponding to STANAG 2022 D:Unreliable), despite having an average of 2200 followers. All three users with a nonzero TunkRank have a percentile greater than 90%, corresponding to STANAG 2022 A:Completely Reliable. It is only these users that would count as reliable sources here.

The preceding suggests that the TunkRank metric would be more useful than simple indegree centrality (# of followers) for distinguishing reliable death reports from unreliable death rumors, which periodically spread across Twitter.[12] It is likely that unsubstantiated death rumors have very few users with high TunkRank who spread the rumor, even if the rumor spreaders have relatively high indegree centrality.

3.2 Source Independence on Twitter

According to the STANAG 2022 rubric, a message is credible to the extent that multiple, *independent* users assert the same thing. How can source dependence or independence be determined on Twitter?

Suppose an analyst sees Twitter status updates from two different accounts A and B each saying "The Archduke has been shot". It is premature to say that the two Twitter sources are ipso facto independent and therefore that each report independently confirms the other. Both Twitter updates might merely be retweeting what a mutual contact, C, had said previously, without the

[10] http://en.wikipedia.org/wiki/Tim_Hetherington
[11] http://themetricsystem.rjmetrics.com/2010/01/26/new-data-on-twitters-users-and-engagement
[12] http://www.digitalspy.com/celebrity/news/a317866/dwayne-the-rock-johnson-im-not-dead.html

conventional retweet attribution. On social media platforms, it is often possible to trace how information flows between users directly, but these conventions aren't always used.

Automatically identifying source independence in Twitter is a challenge, but some automated reasoning techniques can be applied here as well as network metrics.

Clearly, a retweet by one user of another's tweet is a case of source dependence. This can easily be captured in an RDF matching rule, so the content of the retweet does not count as independently verifying the source tweet.

Less obviously, a Twitter user who posts something from a media source is an independent source of the information in that link, only if it does not come from a third party. Posting a link to a third party media report is just like retweeting a report from another user, except that Twitter doesn't have a convention for indicating this use. For example, the following tweet cites a third party BBC report.

@WillofArabia : `BBC News - British journalist Tim Hetherington dies in Libya - http://bbc.in/ejB40c`

This is not much different from the same user retweeting another user's (firsthand) report, it simply passes along a report from someone else. For our purposes, we take the source of this report to be the cited media source, rather than the Twitter user. We use cited URLs from media sources (which often have identifiable URL shortening services, like bbc.in, here) as indicators of third party origination.

Shortened URLs are unique to the originator, so if any two messages contain the same shortened URL, they are not independent. However, URL shortening is many-to-one, and there may be several shortened URLs all pointing to the same dereferenced URL. Thus, if two tweets cite the same dereferenced URL, their sources are also not independent. Both users have cited the same third-party URL. A URL is third-party if it doesn't originate from the same domain as the URL cited in the user's profile. For example, if a user has a blog at http://example.org and tweets a citation to a blog at that domain, that is not a third party citation. However, if the same user tweets a citation to a BBC article at http://bbc.in/abc that clearly is a third-party citation.

In our sample data set, for example, five of the twenty tweets cite the same dereferenced third-party URL from the *New York Times*. This same URL is referenced by four distinct shortened URLs. Thus, none of these sources are independent; the source for each is the *NY Times*.

Finally, if two tweets have exactly the same textual content, even if they are not linked by retweeting or shared URL, the sources are likely not to be independent, particularly for longer messages. To be safe, we take all string-identical message bodies as indicating a common source between two users.

These are content-based indicators of source (in)dependence, but since Twitter is a social network, there are obviously network based metrics of independence as well. In a network of sources, independent confirmation requires independence of sources. Almost all users of Twitter fail to qualify as independent if independence requires that no path exists from one source to another through the Twitter social network graph. In fact, the average path length between any two users on Twitter has been determined empirically to be only 4.12 links [17]. Since a relatively short path exists between any two users on average, we take source independence to mean that A and B have a shortest path between them of at least 4 (~4.12) hops, or the average distance between any two randomly selected users on Twitter.

3.3 Twitter Credibility

A message is 1:Independently Confirmed according to the STANAG 2022 rubric if another message from an independent source says the same thing. At first glance, this would seem to mean that if a message from A and a message from B are string identical, then the messages are independently confirmed, as long as A and B are independent. However, we have said that Twitter string identify is prima facie evidence of source *dependence*: two users who tweet the exact same message are likely to have a common source. Therefore, we need to identify how to identify messages that mean the same thing but say it in a different way.

We use the Rouge-S metric, developed for automatically computing the similarity of document summaries [24], as our measure of tweet similarity. Rouge-S computes the number of shared "skip bigrams" between a source message and a target message. A skip bigram is a pair of words, in left to right order, where the first word is to the left of the second word in the message. Thus, the set of skip bigrams for a message consists of: the first and second words, the first and third words, … the first and last words, the second and third word, the second and fourth word, … and finally, the penultimate and last word. Two identical strings have a ROUGE-S score of 1. Two strings that consist of the same, unrepeated words in reverse order, have a ROUGE-S score of 0. The bigram order constraint thus preserves an element of sentence structure.

Thus, since message (A) has 6 skip bigrams in common with message (B), which has a total of 21 skip bigrams (6 + 5 + 4 + 3 + 2 + 1), then message A is 28.6% similar to message B. On the other hand, message B is 6/10 = 60% similar to message A. The measure is not symmetric.

(A) Just went for a run
(B) I went for a run after work

In our calculation, we measure the similarity of messages as the ROUGE-S metric of the longer message compared

to the shorter message, after first removing special terms (e.g. RT, DM), @names and hashtags. Messages that are at least 80% similar by ROUGE-S count as saying the same thing, for our purposes.

Messages that would count as saying the same thing if a negation is removed count as contradictory reports. In the sample dataset, for example, there are no contradictory reports that Hetherington was "not killed" or "not dead" (or "alive"). However, we identify 435 tweets that report that Chris Hondros, a Getty Images photographer, was killed (or dead or died) along with Hetherington in the same incident, but 7 contradictory tweets that report that he is still alive, as was first reported.

Thus, from these tweets, reports that Hetherington had died would be marked 1:IndependentlyConfirmed, but the reports that Hondros had died would be marked 2:Probably True since the vast majority of the reports indicate do not contradict it, but some do.

4 Discussion

In this paper, we have shown that although evaluating information in Twitter is called for, because of the adversarial uses to which Twitter is increasingly used in organizing and mobilizing political dissent, there has been little attempt to apply approaches to information evaluation, along the lines of STANAG 2022, to the Twittersphere.

We have shown that Twitter streams can be converted to RDF graphs upon which formal rules for reasoning about source reliability and information credibility can be applied. We then motivated an eigenvector centrality measure (TunkRank) as being most appropriate to the Twitter situation and mapped it to the STANAG reliability metric. We also discussed tractable ways in which source independence and message consistency can be calculated in the Twittersphere and showed how special cases could be incorporated. We illustrated our approach with example tweets about a tragic incident in Libya.

References

[1] Eric Schmidt and Jared Cohen, "The Digital Disruption", Foreign Affairs, November/December 2010.

[2] Steven A. Cook and Jared Cohen, "Q&A on Tunisia", ForeignAffairs.com, January 24, 2011. http://www.foreignaffairs.com/discussions/interviews/qa-with-steven-a-cook-and-jared-cohen-on-tunisia

[3] Malcolm Gladwell, Small change: Why the revolution will not betweeted. The New Yorker, 4 October 2010.

[4] Biz Stone. Tweet Preservation. Twitter Blog. April 14, 2010. http://blog.twitter.com/2010/04/tweet-preservation.html

[5] Carolyn Penner. #superbowl. Twitter Blog. February 9, 2011. http://blog.twitter.com/2011/02/superbowl.html

[6] STANAG 2022 (Edition 8) Annex. North Atlantic Treaty Organization (NATO)

[7] Multilateral Interoperability Programme. THE JOINT C3 INFORMATION EXCHANGE DATA MODEL (JC3IEDM Main). Version 3.0.2. May, 2009.

[8] Claire Cain Miller. "Why Twitter's CEO Demoted Himself". New York Times. P. BU1 October 30, 2010.

[9] Barracuda Labs. Annual Report 2009. http://barracudalabs.com/downloads/BarracudaLabs2009AnnualReport-FINAL.pdf

[10] Page, L. Brin, S., Motwani, R., Winograd, T., (1998) The pagerank citation ranking: Bringing order to the web. Report, Stanford Digital Library Technologies Project.

[11] Ulicny, B., Matheus, C., Kokar, M., Metrics for Monitoring a Social-Political Blogsophere: A Malaysian Case Study. IEEE Internet Computing, Special Issue on Social Computing in the Blogosphere. March/April 2010.

[12] Flynn, MG M. T., Pottinger, Capt. M., USMC, Batchelor, P. D., "Fixing Intel: A Blueprint for Making Intelligence Relevant in Afghanistan", Center for a New American Security (CNAS) Working Paper. Jan 4, 2010.

[13] Hopkins, D., King, G., A Method of Automated Nonparametric Content Analysis for Social Science, Am. J. of Political Science 54, 1 (January 2010): 229-247

[14] A. Culotta. Detecting influenza outbreaks by analyzing Twitter messages. http://arxiv.org/abs/1007.4748

[15] Horowitz, D., Kamvar, S., Anatomy of a Large Scale Social Search Engine. WWW2010, Raleigh, NC. 2010.

[16] Banko, M. et al. AskMSR: Question answering using the worldwide web. *2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*

[17] Kwak, Haewoon and Lee, Changhyun and Park, Hosung and Moon, Sue. "What is Twitter, a Social Network or a News Media?". Proc. of WWW'10. Raleigh, NC. Pp. 591—600.

[18] Leskovec, J.; Backstrom, L.; Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *KDD '09*.

[19] Talbot, D., "Search Me: Inside the launch of Stephen Wolfram's new "computational knowledge engine"." *Technology Review*. July/August 2009

[20] Galen Pickard et al. Time Critical Social Mobilization: The DARPA Network Challenge Winning Strategy. arXiv:1008.3172

[21] Daniel Tunkelang. A Twitter Analogy to PageRank. The Noisy Channel Blog. http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/

[22] B. Ulicny, C. Matheus, M. Kokar. A Semantic Wiki Alerting Environment Incorporating Credibility and Reliability Evaluation. Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS 2010), Fairfax, VA, October 27-28, 2010

[23] C. Matheus, The Practical Use of Rules with Ontologies. Presentation at Semantic Technology Conference, San Francisco, CA, June 22-25, 2010

[24] Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

[25] "Happy Rockefeller" UPDATED: The HB Gary Email That Should Concern Us All. DailyKos blog. http://www.dailykos.com/story/2011/02/16/945768/-UPDATED:-The-HB-Gary-Email-That-Should-Concern-Us-All

[26] US Air Force Air Mobility Command. 2010. Solicitation RTB220610 "Online Persona Management". June 22, 2010. FedBizOpps.gov. Cached version at http://embedit.in/KIcSGwa6xL

[27] Ryan Singel. 2011. Burning Question: Why am I being followed by Twitter robots? Wired Magazine. 19.06. June, 2011. P. 78.

[28] G. Linehan. 2011. Bin Laden and The IT Crowd: Anatomy of a Twitter Hoax. BBC Magazine. May 23, 2011