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 have had success in producing an accurate awareness of situations by mining traffic in Twitter. Where these systems have been successful, there has been no issue of evaluating Twitter streams for source reliability and information credibility because the situations have not been adversarial. Recent uses of Twitter in political dissent in the Mideast makes the need for computationally tractable approaches to evaluating source reliability and information credibility more acute in order to achieve accurate situation awareness on the basis of Twitter streams in the face of deliberate mis- or disinformation efforts.

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

1 Introduction

Twitter has become the best-known example of a broadcast system for short "status update" messages. Such platforms have become associated with organizing and mobilizing political dissent and disruption [1]. In the recent 2011 uprisings in the Middle East, in Tunisia and Egypt [2], 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 role of social media platforms like Twitter in sparking similar uprisings in Iran and Moldova has been overstated [3]. As unrest continues in the Mideast, regardless of whether Twitter and similar social media are an essential technology for initiating or organizing such dissent or not, it is clear that the use of technologies like Twitter cannot be ignored as a source of situation awareness.

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; currently, approximately 55 million tweets are sent each day 0. 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. Users can send private messages to someone who follows them by by prefixing their message with 'DM' (direct message) and the username. Users typically shorten URLs in their tweets via 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 supports this, and they turn this option on. (Less than 1% of Twitter status updates are geolocated currently). Twitter messages are archived after six months.

Users can provide a short profile message, a profile picture, 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 (called hashtags) or keywords.

Twitter verifies some famous users' identities, and indicates this status on their profile. In general, however, users are not verified, and anyone can tweet under whatever name they like. Twitter suggests that by providing a link to one's Twitter feed on their website, this can provide user authentication as well.

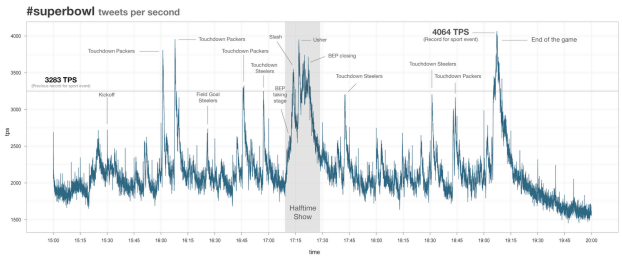
Although, we focus on Twitter here, Facebook and Google Buzz provide similar functionality. Also, the Ushahidi platform (ushahidi.org) combines a map overlay with the ability to post reports by location, via cell phone texts or from Twitter or anonymously from the web, primarily in humanitarian relief 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 provide much insight, aggregated Tweets may convey a strong signal about the situation they reflect. For example, Figure 1, from the Twitter blog, shows tweets per second over time for the hashtag #superbowl during

the 2011 NFL Superbowl game. The spikes in the graph of tweets per second clearly correlate strongly with scoring in the game. Other spikes correlate with moments in the game's half-time show, particularly the surprise appearance of one performer. Armed only with thèse tweets, it is likely that one could recreate an account of what happened in the game and when, by looking for commonalities in the messages at the times corresponding to spikes.

Similarly, Culotta has shown [13] that influenza outbreaks can be tracked in near-real time quite effectively just by looking for simple keywords in tweets. Culotta validated his models by comparing Twitter results with weekly epidemiological reports from the Center for Disease Control.

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



What the Super Bowl and flu situations have in common, is that there is little reason for a Twitter user to publish mis- or disinformation. Therefore, the tweets can be taken at face value. In this paper, our focus will be not primarily on aggregating situation awareness from a multitude of presumably sincere tweets, but 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 equally, because of the adversarial nature of the struggles in which Twitter plays a big part, it may be prudent to treat tweets differentially in terms of their reliability, credibility, and other epistemic properties before constructing a depiction of the situation from them.

2 Information Evaluation

NATO STANAG (Standard Agreement) 2022 "Intelligence Reports" states that [5] 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 a measurement of the reliability of the source of the information and a numeric measurement of the credibility of a piece of information "when examined in the light of existing knowledge".[1]

Reliability of the source 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

Credibility: 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".[2]
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 past reports, and the quality and quantity of previous reports.

[1] 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. JC3IEDM [6] 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.)

[2] 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.

2.1 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 [7], and even though perhaps less than 25% of thèse 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 [8], 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 [7]. Moreover, Barracuda Labs reports that in 2009, 12% of new Twitter user accounts were shut down by Twitter for violating their policies [8]. So, while Twitter does police itself to some extent, a potentially large number of Twitter users may be unreliable.

In a networked environment like the contemporary operating environment, an analyst is exposed to many novel sources of information across PMESII-PT categories and has very little ability to check their reliability directly [11]. 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 [9] 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 can be understood as computing the likelihood that a random web surfer will end up on a particular document given that, for each document, the web surfer tends to jump to a new document some percentage of the time (standardly, a 15% likelihood to jump is employed as the so-called damping factor, defining the propensity to continue to a new page). This algorithm is recursive and typically computed for only a small number of iterations, 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. "Link-farming" to inflate PageRank is ferreted out.

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 open-source and unclassified data 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 [10].

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 as the weighted cosine similarity 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. Social search metrics such as those incorporated by Vark are surely 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 identify highly central figures in a network. A highly central figure has more authority, and is probably more likely to be reliable than a marginal figure in a social network, at least with respect to information that relevant to its participants. We conclude, then, that network-theoretic centrality metrics used in civilian information retrieval applications, should be investigated for systematically estimating source reliability where tracking source reliability directly is impractical or unfeasible, such as in the Twitter network.

2.2 Current Approaches to Credibility

STANAG 2022 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 previous reports. STANAG 2022's highest credibility ranking goes to information that is independently confirmed. The lowest credibility ranking goes to those reports that contradict previous information, presumably that has been well-confirmed.

Many information portals on the Web address the credibility of the information they provide by either limiting the information they provide to well-regarded sources (e.g. Wolfram Alpha [18]) 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 doesn't edit what is said on the site as long as it doesn't violate their policies, nor is there a common version of every assertion that can be edited, as in a Wiki.

In information retrieval, text-based question-answering systems have used sameness of text in search snippets to identify consensus answers to factual questions in a textual corpus. The AskMSR system [15], 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 [12], the authors provide a sophisticated method for estimating the proportion of texts of the same type 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[3]. 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 [17].

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 (reliability) via some centrality metric. While social question-answering systems incorporate metrics for reliability or 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 for most part 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 [19].

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

3 Applying the STANAG 2022 Rubric 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 provided by Mark Borkum's "Shredded Tweet" service.[4] Borkum's service converts Twitter search results into RDF/XML, using a variety of namespace and properties from well-known ontologies, including Dublin Core Metadata Initiative[5], the SIOC (Semantically-Interlinked Online Communities) Core Ontology[6], and the FOAF (Friend of a Friend) vocabulary spécification.[7]

Figure 2 depicts the RDF graph associated with a single tweet from Twitter user @FortWayneHub reporting on a fire in Ft. Wayne, IN, and providing a link to a news story about the fire. This simple tweet produces 24 RDF (Resource Description Framework) triples: five with the user as subject, and ten with the tweet as the subject.

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

[4] http//shreddedtweet.org/

[5] http://dublincore.org/documents/2010/10/11/dcmi-terms/

[6] http://rdfs.org/sioc/spec/

[7] http://xmlns.com/foaf/spec/

Figure 2 RDF Graph of a Single Tweet, Produced by ShreddedTweet.org

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 as a proxy for source quality, and since Twitter is a network structure, this is an attractive option here as well.

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 followers those users have. Indegree centrality can easily be synthesized. Since a Twitter account can be purchased at the cost of a valid email address, it would be easy to automatically create an account with many followers. Therefore, simple indegree centrality (follower counts) is not a good proxy for reliability.

Daniel Tunkelang's TunkRank metric [20] can be adapted as a measure of eigenvalue centrality for Twitter. Tunkelang's algorithm recursively produces a TunkRank (Influence) score based on the expected number of people that will see a message that X tweets and the (assumed) constant probability p that someone 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 but who are not themselves followed by anyone would receive a TunkRank of zero. Thus, TunkRank is not easily faked by providing a Twitter user with many fake followers.

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

Equation 1 TunkRank equation

The top Twitter users by TunkRank are listed at http://tunkrank.com/score/top. At first glance, the results seem to be empirically 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 algorithm, it is worth considering that what TunkRank really measures is the pass-along value of a Twitter user to his or her followers.

Entertainers who tweet have a different kind of value to their followers than the 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 will continue to 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 ratio of Twitter users that have a lower TunkRank. We thus map these percentiles to STANAG 2022 rubrics as follows: if a Twitter user has a TunkRank percentile greater than 90%, we consider them A: Completely Reliable. Over 80%, B: UsuallyReliable. Over 75%, Fairly Reliable. Over 50%, Somewhat Reliable. Less than 50%, Not Usually Reliable. Less than 10%, Unreliable. If a TunkRank cannot be computed, then the user will be considered F: Reliability Cannot Be Determined. This stands in contrast to previous work in which we counted any Twitter user that was a news organization as A: Completely Reliable and all other users as F: Reliability Cannot Be Determined [21]. As in that work, the reliability of a Twitter user is asserted as an RDF triple in RDF graphs containing tweets by that user.

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

We can directly determine that a Twitter user is saying something false in a certain class of cases, however, through application of formal rules. One such type of case in when a Twitter user misrepresents the provenance of a (purportedly) retweeted message. That is, 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 Ebola outbreak in Scranton, PA!
```

Such a tweet has the appearance of retweeting a report by the US President's staff that there is an Ebola virus outbreak in Scranton, PA. The user gets to make the assertion while attributing it (falsely) to someone else. False rumors can be promulgated this way, by leveraging

the popularity and perceived reliability of the retweeted user to start a rumor cascade.

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 [22], 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 not((?a rdf:type b:MicroBlogPost) & (?a sioc:has_creator ?user1) & (?a sioc:created ?t) & (?a sioc:content ?c) & (?c sioc:body ?d) & (?d matches "^RT ?user2 ?text") & (?e rdf:type b:MicroBlogPost) & (?e sioc:created ?t2) & (?t1 > ?t2) & (?e sioc:has_creator ?user2) & (?e sioc:content ?f) & (?f sioc:body ?g) & (?h 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 true retweet pattern, then the retweet is bogus and 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.

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.

3.2 Twitter Credibility

Twitter poses many challenges for the STANAG 2022 rubric with respect to credibility: a message is credible to the extent that multiple, independent users assert the same thing. Automatic evaluation of this metric requires automatically determining (i) the independence of Twitter users and (ii) identifying messages that assert the same propositional content.

Suppose an analyst sees two 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 updates are ipso facto independent and therefore that either report confirms the other. Both Twitter updates might merely be retweeting what a mutual contact, C, had said previously, without the conventional retweet attribution. On social media platforms, it is often possible to trace how information flows from one user to another directly by means of hypertext trails, shortened URLs, retweets or hat tip citations, timestamps and other mechanisms.

In a network of sources, independent confirmation must require independence of sources. Almost all users on Twitter would fail to qualify as independent if independence requires that no path exists from one source to another through the Twitter social graph. In fact, the average path length between any two users on Twitter has been determined empirically to be only 4.12 links [16]. Since a relatively short path exists between most users, source independence must be taken to mean that if A and B both report the same thing, and A and B do not have a shortest path between them closer than the average shortest path length between any two nodes in the social network and there is no source C in the network who reports the same thing that has a shorter path between both A and B than A and B have to one another, then A and B's reports independently confirm one another. As such, they can be annoed with has_credibility 1: Independently Confirmed.

Because the average path length between any two users on Twitter is fairly short, it is not computationally intensive to calculate the relatedness of two users based on common users they follow and common interests.

We compute the relatedness between any two Twitter users as his function produces a metric of similarity between two Twitter users by computing the Dice coefficient of shared friends (those who both users follow, shared followers, and topic terms, where these include @tags, hash tags (#tags), urls, and capitalized phrases, all of which are obtainable via the Twitter API[8]. Our Twitter relatedness measure doesn't only identify a path between two users, determined by following relations. It determines the overlap between common followed users and followers, as well as distinctive terms shared by the two users. The Dice coefficient s is defined in Equation 2 as the ratio of the two times the number of shared features over the combined number of features assigned to each element, sets X and Y.

$$s = \frac{2|X \cap Y|}{|X| + |Y|}$$

Equation 2 Dice coefficient

If two users share all followers, followed users, and special terms, then those users would count as completely related by our metric (s=1). If two users share no common features, then they are completely unrelated (s=0).

We calculate the relatedness of any two users as the maximum sum of relatedness coefficients along a path connecting the users via following relations, divided by the length of that path. That is, if two users have only one friend in common, but they are completely similar to that friend, then they are as similar to each other as they are to

[8] http://dev.twitter.com

the common friend. Their messages cannot be taken to be independent.

On the other hand, if no path between two users of length less than four has any more than a trivial amount of relatedness, then the two users are completely independent.

If two messages are related by a chain of retweets from one to another, then the messages are not independent. If two users tweet the same or similar messages, and the two users are less than 10% related, according to this metric, then the message is 1: Independently Confirmed. If less than 25% related, then 2: Probably True. If less than 50% related, then 3: Possibly True. If roughly as many independently confirmed reports assert both p and not-p, then each report is labeled 4: Doubtful. If at least twice as many independent messages assert p, then independently confirmed messages asserting not-p, then the not-p messages should be marked 5: Improbable. Finally, if these calculations can't be computed, then the message should be marked 6: Credibility Cannot Be Determined.

This leads to a question of how to identify messages that say the same thing. Messages that are string identical, modulo 'RT' and other special terms, are not likely to be independently produced. We use the Rouge-S metric, developed for automatically computing the similarity of document summaries [23], 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 of over 4 words that are completely string identical do not count as independent.

Again, some special cases apply. Messages related by retweet chains do not count as independent messages. Nor do messages that contain the same shortened URL, because shortened URLs are unique to their originator. On the other hand, messages that cite distinct, dereferenced URLs cannot be saying the same thing.

4 Discussion

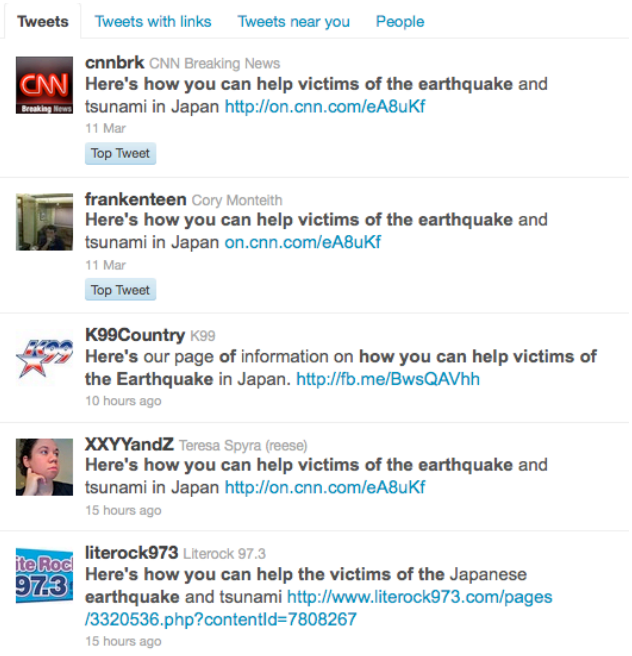
In Figure 3, we see several tweets from different information sources. The source of the first tweet, cnnbrk (CNN Breaking News) has a TunkRank in the 80% percentile. Thus, we would assign it a reliability of B: Usually Reliable. The message it asserts, about where to find information about helping earthquake victims in Japan, would be assigned a default Credibility rating of 3: Possibly True because it is a "freshly reported" piece of information not contradicted by other reports.

The messages from frankenteen (Cory Monteith) and XXYYandZ (Teresa Spyra (reese)), on the other hand, have whatever TunkRank their originators have, but they do not count as independently confirming the message of cnnbrk because their content is string identical and over four words. These messages are essentially unacknowledged retweets, so they have the credibility of the original message.

Finally, the messages of K99Country and literock973 do not independently confirm the message of cnnbrk because they cite different URLs, and therefore, do not assert the same message. They are not retweets, so they also have the default credibility rating of 3: Possibly True.

Thus, despite the similarity of the messages, we have no reason to be more confident of what they assert than the original message.

Figure 3 Sample Tweets



5 Conclusion

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 abour source reliability and information credibility can be applied. We then motivated an eigenvector centrality measure (TunkRank) as being most appropriate to Twitter situation. We also discussed tractable ways in which source independence and message similarity can be calculated in the Twitter, and showed how special cases could be incorporated. We illustrated our approach with example tweets.

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.

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

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

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

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

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

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

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

[10] 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.

[11] 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.

[12] 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

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

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

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

[16] 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.

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

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

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

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

[21] 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

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

[23] 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.