Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

Detecting misinformation in online social networks using cognitive psychology

KP Krishna Kumar*† and G Geethakumari†

*Correspondence:
kpkrishnakumar@gmail.com
†Equal contributors
Department of Computer Science and Information Systems, BITS-Pilani, Hyderabad Campus, Jawahar Nagar, Hyderabad, India

Abstract

The paper explores the use of concepts in cognitive psychology to evaluate the spread of misinformation, disinformation and propaganda in online social networks. Analysing online social networks to identify metrics to infer cues of deception will enable us to measure diffusion of misinformation. The cognitive process involved in the decision to spread information involves answering four main questions viz *consistency of message, coherency of message, credibility of source* and *general acceptability of message*. We have used the cues of deception to analyse these questions to obtain solutions for preventing the spread of misinformation. We have proposed an algorithm to effectively detect deliberate spread of false information which would enable users to make informed decisions while spreading information in social networks. The computationally efficient algorithm uses the collaborative filtering property of social networks to measure the credibility of sources of information as well as quality of news items. The validation of the proposed methodology has been done on the online social network 'Twitter'.

Keywords: Online social network; Misinformation; Disinformation; Cognitive psychology

Introduction

Internet is a great source of information. It is also called the Web of deception. The use of communication channels of the Internet to propagate false information has become quite common. The advent of social networks has made every user a self-publisher with no editing, checking for factual accuracy and clearly with no accountability. The facts are presented with no authority and for millions of users seeing them on their computer screen is itself a certificate of truthfulness of information being presented to them. In [1], the dangers in the use of Internet like deliberate deception, deliberate misinformation, and half-truths that can be used to divert a user from the real information being sought have been discussed. The use of technologies by people to support lies, deception, misdirection, fraud, spin control, propaganda as discussed in the book have come true with online social networks like Facebook and Twitter being used for purposes for which they were not intended for. Validating the data on the Internet is a challenging proposition and pitfalls by new users and experienced ones are far too often.

Online Social Networks (OSN) have become an important source of information for a large number of people in the recent years. As the usage of social networks increased,

the abuse of the media to spread disinformation and misinformation also increased many fold. The spread of information or misinformation in online social networks is context specific and studies have revealed topics such as health, politics, finances and technology trends are prime sources of misinformation and disinformation in different contexts to include business, government and everyday life [2]. The number of information diffusion models do not take into consideration the type of information while modeling their diffusion process. The information diffusion in social networks due to misinformation or disinformation could follow different patterns of propagation and could be as a result of an orchestrated campaign to mimic widespread information diffusion behaviour. The lack of accountability and verifiability afford the users an excellent opportunity to spread specific ideas through the network.

The detection of misinformation in large volumes of data is a challenging task. Methods using machine learning and Natural Language Processing (NLP) techniques exist to automate the process to some extent. However, because of the semantic nature of the contents, the accuracy of automated methods is limited and quite often require manual intervention. The amount of data generated in online social networks is so huge as to make the task computationally expensive to be done in real time. In this paper, we propose a methodology to detect misinformation content using concepts based in cognitive psychology. We analysed the literature on cognitive psychology to understand the process of decision making of an individual. An individual is seen to make decisions based on cues of deception or misinformation he obtains from the social network. Analysing the social network data using suitable metrics to detect the same cues of deception would enable us to identify patterns of spread of misinformation. This could be used by an online user to take correct decisions about authenticity of information while spreading them. A framework which would enable prevention of spread of inaccurate information would be more effective than one which proposes counter measures after the information diffusion process. We have implemented our proposed framework in Twitter.

Twitter has emerged as one of the more popular micro-blogging sites. Twitter enables propagation of news in real time. Information propagates in Twitter in the form of short messages of maximum 140 characters called 'tweets'. The system enables one to subscribe to another's tweets by following them. It allows quick information dissemination by retweeting the tweets one has received. The ability to post tweets from mobile devices like smart phones, tablets and even by SMS have resulted in Twitter becoming the source of information for many users. These capabilities also make Twitter a platform for spreading misinformation easily.

Background and literature review

Concepts of information, misinformation and disinformation

How they differ?

It is essential to understand the related concepts of information, misinformation, disinformation and propaganda. The definition of information is clear by its very nature to the users. But what needs to be defined is the different forms it can take. We are more interested in the usage of social networks to spread specific kind of information to alter the behaviour or attitude of people. In the cyber space, manipulation of information so as to affect the semantic nature of information and the way in which it is interpreted by users is often called semantic attacks. Semantic attacks in social networks could be a result of

propagation of information in various forms. This could take the shape of misinformation, disinformation or propaganda. The distinction between information, misinformation and disinformation is difficult to be made [3]. The three concepts are related to truth, and to arrive at a universal acceptance of a single truth is almost impossible.

The term information is defined by the Oxford dictionary as 'facts provided or learned about something or someone'. The other forms of information are defined by Oxford dictionary as under:

- *Misinformation* is false or inaccurate information, especially that which is deliberately intended to deceive.
- *Disinformation* is false information that is intended to mislead, especially propaganda issued by a government organization to a rival power or the media.
- *Propaganda* is defined as information, especially of a biased or misleading nature, used to promote a political cause or point of view.

The three definitions have small differences and the most important fact is they involve the propagation of false information with the intention and capability to mislead at least some of the recipients. The advent of social networks has made the speed of propagation of information faster, created large number of sources of information, produced huge amounts of information in short duration of time and with almost no accountability about the accuracy of data. The term 'Big Data' is often associated with the data in social networks. Finding credible information after sifting out the different forms of false information in online social networks has become a very challenging computational task. In this paper, we intend to use the basic tenets of cognitive psychology to devise efficient methods by which the task can be done. Our methodology involves detecting cues of deception in online social networks to segregate false or misleading information with the intention of developing an effective tool for evaluating the credibility of information received by a user based on the source of the message as well its general acceptability in the network.

Conceptual explanation of the distinguishing features

The concept of information, misinformation and disinformation have been differentiated with respect to five important features by Karlova et al. [2]. They are truth, accuracy, completeness, currency and deceptiveness. While all the three are informative in nature, only disinformation is deliberatively deceptive information. The authors have also given a social diffusion model of information, misinformation and disinformation as products of social processes illustrating the way they are formed and disseminated in social networks. The model suggests that people use cues to credibility and cues to deception to make judgements while participating in the information diffusion process.

Accuracy of the information is one of the important measures of quality of information. Honest mistake in the spread of inaccurate information is misinformation, whereas when the intention is to deceive the recipient, it is disinformation. In [4], authors have outlined the main features of disinformation.

- Disinformation is often the product of a carefully planned and technically sophisticated deceit process.
- Disinformation may not come directly from the source that intends to deceive.

- Disinformation is often written or verbal communication to include doctored photographs, fake videos etc.
- Disinformation could be distributed very widely or targeted at specific people or organizations.
- The intended targets are often a person or a group of people.

In order to classify as disinformation, it is not necessary that the disinformation has to come directly from the source of disinformation [4]. In the chain of dissemination of information, most of the people could actually be transmitting misleading information (hence misinformation), though only one of the intermediaries may believe that the information is actually misleading (hence disinformation). This is especially true for social networks where the chain of propagation could be long and quite a few people involved in the process.

Social networks with its freedom of expression, lack of filtering mechanisms like reviewing and editing available in traditional publishing business coupled with high degree of lack of accountability have become an important media for spread of misinformation. Summarily, the propagation of different versions of information, viz misinformation, disinformation and propaganda involves the spread of false or inaccurate information through information diffusion process involving users of social networks where all the users may not be aware of the falsehood in the information. We have used the term misinformation to denote any type of false information spreading in social networks.

Misinformation

The acceptance of misinformation or misleading information by the people depends on their prior beliefs and opinions [5]. People believe things which support their prior thoughts without questioning them. The same is also supported by research in cognitive psychology [6]. The authors have brought out that preexisting political, religious or social views make people accept information without verification if it conforms to their beliefs. Countering such ideological and personal beliefs is indeed very difficult. Another important finding was that countering the misinformation may lead to amplifying the beliefs and reenforcing them.

Political astroturfing in the form of propagation of memes in Twitter was studied by the Truthy team [7,8]. Investigating political election campaigns in US in the year 2010, the research group uncovered a number of accounts sending out duplicate messages and also retweeting messages from the same few accounts in a closely connected network. In another case, 10 different accounts were used to send out thousands of posts, many of them duplicates slightly altered to avoid detection as spam. With URL shorteners available, messages containing links could be altered to give different shortened links to the same source and hence escaping the spam filters.

Decision making out of ignorance is often based on heuristics and the level of confidence on the decision is also low, making correction easier. Such decisions are often correct and are generally not catastrophic. False beliefs based on misinformation are held strongly and often result in greater support for a cause. Such beliefs are also very contagious and the person makes efforts to spread them to others. The persistence of misinformation in the society is dangerous and require analysis for their prevention and early detection [6,9].

Misinformation during an event as it unfolds like casuality figures in a natural calamity, are seldom accurate initially and the figures get updated or changed over a period of time. Such spread of misinformation is often considered benign though media is considered as one of the most important sources of misinformation. The other important sources of misinformation are governments and politicians, vested interests and rumours and works of fiction. Information asymmetry due to new media like social networks play a big role in the spread of misinformation. Social networks spread information without traditional filters like editing. The advent of Web 2.0 has resulted in greater citizen journalism resulting in increase in the speed of dissemination of information using multiple online social media like social networks, blogs, emails, photo and video sharing platforms, bulletin boards etc. The creation of cyber–ghettos has been discussed where the cyber space has become echo chambers and blogs and other social media primarily link to like minded sites propagating similar views than providing contrarian views. This leads to fractionation of the information landscape and consequent persistence of misinformation in large sections of the society for a long period of time. This often result in people holding on to their views on matters of pubic, political and even religious importance due to their misinformed world views and ideology.

In [10], authors have enumerated a number of possible instances of misinformation in the Internet. They include incomplete, out-of-date and biased information, pranks, contradictions, improperly translated data, software incompatibilities, unauthorized revisions, factual errors and scholarly misconduct. However, with the advent of Web 2.0 the list has grown many times and social media is described as one of the biggest sources of information including misinformation. Internet acts as a post modern Pandora's box-releasing many different arguments for information which are not easily dismissible [11].

Countering the spread of misinformation

Misinformation is easily another version of information. Countering spread of misinformation is not an easy task. The simple technique of labelling the other side as wrong is ineffectual. Education of people against misinformation is necessary but not sufficient for combating misinformation. An analysis of the counter measures proposed and modeled in the literature against the spread of misinformation in OSNs are at times not in consonance with the effectiveness of the measures as suggested in studies of cognitive psychology. Theoretical framework for limiting the viral propagation of misinformation has been proposed in [12,13]. The authors have proposed a model for identifying the most influential nodes whose decontamination with good information would prevent the spread of misinformation. The solution to the problem of limiting the spread of misinformation by starting a counter campaign using k influential nodes, called the *eventual influence limitation* problem has been proposed in [14]. The influence limitation problem has also been studied in the presence of missing information. In both the papers, the basic assumption is that when an infected node is presented with correct information, it would become decontaminated. Studies in psychology have proved that removing misinformation from infected persons is not easy [6]. The best solution to the spread of misinformation is early detection of misinformation and launch of directed and effective counter campaigns. In [15], the authors have proposed ranking based and optimization-based algorithms for identifying the top k most suspected sources of misinformation in a time bound manner.

The strategies proposed in [6] for effective counter measures include:

- Providing credible alternative explanation to the misinformation.
- Repeated retractions to reduce the effect of misinformation without repeating the misinformation.
- Explicit warnings before mentioning the misinformation so as to prevent the misinformation from getting reinforced.
- Counter measures be suitably biased towards affirmation of the world view of the receiver.
- Simple and brief retractions which are cognitively more attractive than the corresponding misinformation.

Analysis of work done so far

We have analysed the cognitive process of adoption of information from studies in psychology. The difficulties associated with distinguishing between misinformation, disinformation and true information have been highlighted by most of them [2,3,6]. The cognitive factors which decide the credibility of messages and their consequent acceptance by users can be effectively modulated in OSNs as seen during US elections [7,8]. The inherent beliefs of a user play a very important part in accepting news items and fractionation of cyber space is a consequence of this aspect of human mind. We explore different factors contributing towards deciding the credibility of news items in the next section.

Misinformation has been widely accepted in the society, it becomes extremely difficult to remove. This has been suitably demonstrated during July 2012, when mass exodus of thousands of people took place in India due to a sustained misinformation campaign by vested interests using social media and other telecommunication networks [16]. Preventing the spread of misinformation is a more effective method of combating misinformation, than its subsequent retraction after it has affected the population. Significant contributions towards successful debiasing of misinformation have been made in [6].

While studies in cognitive psychology are sufficient to understand the process of adoption of information by users, we would like to explore the process of diffusion of information. Process of diffusion is a group phenomenon, where we study the process of adoption by different users over a period of time. Patterns arising out of diffusion of information are better studied using algorithms from computer science. We study the process in detail using 'Twitter' in Section "Credibility analysis of Twitter".

A generic framework for detection of spread of misinformation

While formulating a generic framework for detecting spread of misinformation, it is important to understand the cognitive decision making processes of individuals. A study of individual decision making process from a cognitive psychology point of view followed by a generic framework for detection of misinformation using the cues of deception is given in the following subsections.

Identifying cues to deception using cognitive psychology

The presence of misinformation in the society and real world social networks have been studied from psychological point of view extensively. An excellent review of the mechanisms by which misinformation is propagated and how effective corrective measures can be implemented based on cognitive psychology can be found in [6]. As per the authors,

the spread of misinformation is a result of a cognitive process by the receivers based on their assessment of the truth value of information. Acceptance of information is more the norm than otherwise for most of the people. When people evaluate the truth value of any information they take into account four factors. The factors are characterised by asking four relevant questions. These questions are given below and illustrated in Figure 1, where we have summarised all relevant issues of misinformation.

1. **Consistency of message.** Is the information compatible and consistent with the other things that you believe?
2. **Coherency of message.** Is the information internally coherent without contradictions to form a plausible story?
3. **Credibility of source.** Is the information from a credible source?
4. **General Acceptability.** Do others believe this information?

Information is more likely to be accepted by people when it is **consistent** with other things that they believe is true. If the logical compatibility of a news item has been evaluated to be consistent with their inherent beliefs, the likelihood of acceptance of misinformation by the receiver increases and the probability of correcting the misinformation goes down. Preexisting beliefs play an important part in the acceptance of messages. Stories are easily accepted when the individual elements which make them up are **coherent** and without internal contradictions. Such stories are easier to process and easily processed stories are more readily believed. The familiarity with the sender of a message, and the sender's perceived **credibility** and expertise ensure greater acceptance of the message. The acceptability of a news item increases if the persons are subjected to repeated exposure of the same item. Information is readily believed to be true if there is a perceived social consensus and hence **general acceptability** of the same. Corrections to the misinformation need not work all the time once misinformation is accepted by a receiver.

The rest of the paper is organised as follows. In Section "Research design and methodology" we give our research design and methodology where we explain the generic framework for the detection of misinformation in online social networks. As part of its implementation in 'Twitter', we carried out an analysis of the work done in estimating the credibility

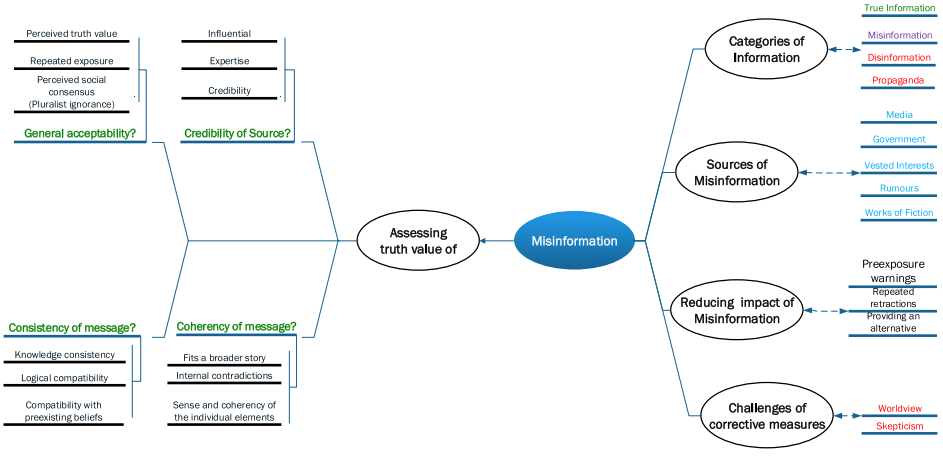

Figure 1 Misinformation landscape. The figure summarises the different aspects of misinformation.

of information propagation in 'Twitter'. In Section "Methods" we explain our methodology and algorithm for speedy detection of spread of misinformation in Twitter to aid a user to recognise misinformation and consequently prevent him from spreading it. In Section "Results and discussion" we show the results obtained using two different Twitter data sets. We outline our future work and conclude in Section "Conclusions".

Research design and methodology

Generic framework for detection of spread of misinformation

Based on the analysis of cognitive process, it becomes clear that the receiver obtains cues of deception or misinformation from the online social network to decide on the accuracy of information. The same cues could be used by a social media monitoring system to detect spread of misinformation, disinformation or propaganda. The proposed framework for such a system is given in Figure 2. The evaluation process of truth value of information begins with identification of suitable credibility metrics of the social network being studied. The metrics would reflect the cues by which a user would have made his decision of estimating the accuracy of information. The subsequent stages of evaluation of truth value of information would involve using the identified metrics to establish the credibility of the source and estimate the general acceptability of the message. A user should be able to make informed decisions regarding the truthfulness of messages with this help from the system and applying his own coherency and consistency values.

We have implemented the proposed framework for Twitter. The details of implementation is given in the subsequent sections. The first step involves the identification of suitable metrics for evaluating the credibility of information propagated in Twitter. For this, we have carried out a detailed analysis of the work done to study the credibility of information propagated in Twitter with a view to identify the most appropriate parameters which would enable detection of misinformation. We classified the parameters using concepts of cognitive psychology and selected the most appropriate metrics.

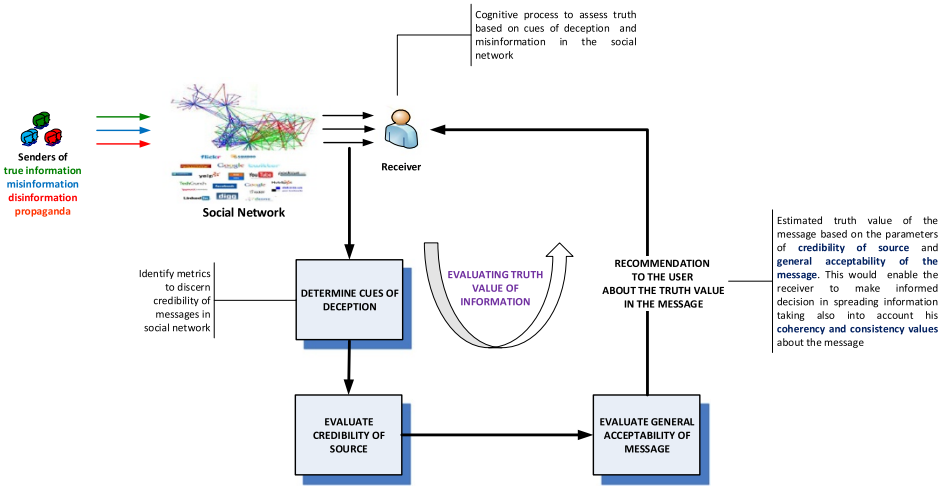

Figure 2 A generic framework for evaluation of truth value of information for a user in an online social network. The figure gives a broad framework for detecting misinformation using cues of deception to help a user to make informed decisions.

Credibility analysis of Twitter

An analysis of the literature to categorize tweets based on the four aspects of *consistency of message, coherency of message, credibility of source* and *general acceptability* as given in the previous section was done. A number of automated techniques using machine learning algorithms have been proposed in the literature to classify tweets based on their characteristic features.

Twitter as a social filter

The credibility of tweets propagated through Twitter has been analyzed in [17]. The authors have used automated methods to assess the credibility of tweets related to trending topics. The features used by the authors include the re-tweeting behaviour, texts of the posts and links to external sources. The authors used supervised learning techniques to build a classifier to estimate the credibility of tweets. The types of features used to characterize each tweet were of four types: message based features, user based features, topic based features and propagation based features. Use of message features included length of the message, positive or negative sentiments, presence of question marks or exclamation marks, and also the use of hashtags and retweets. User based features included number of followers and followees, number of past tweets etc. Topic based features were derived from user based and message based features to include fraction of tweets that contained hashtags, URLs and positive and negative sentiments. The propagation based features included the depth of the retweet tree and number of initial tweets of a topic. Best results of automatic classification of tweets were achieved using J48 learning algorithm. Sentiment features were found to be very relevant for predicting the credibility of tweets. The fraction of tweets with negative tweets was found to be more credible as well as tweets with greater number of retweets. The ability of the Twitter community to act as social filter of credible information has been clearly brought out in the paper. Credible users with large number of followers and followees along with large tweet activity have better reputation score and tend to propagate credible news. While validating the best features to be used for automatic determination of credibility of tweets, the propagation based features were ranked the best. The text and author based features alone are not sufficient to determine the credibility of tweets. The credibility of tweets increases when propagated through authors who have a higher reputation score, having written a large number of tweets before, originate at a single or few users in the network and have many retweets.

Twitter during critical events

Reliability of Twitter under extreme circumstances was also investigated in [18]. The analysis of tweets related to earthquake in Chile in 2010 has revealed that the propagation of rumours in Twitter is different from spread of credible news as rumours tend to be questioned more. The authors selected confirmed news and rumours manually from the set of tweets after the earthquake to analyse patterns of diffusion of information in the form of re-tweets in the network. The use of Twitter as a collaborative-filter mechanism has been proved with the help of this study. Further, the authors have verified the validity of the use of aggregate analysis of tweets to detect rumours.

Credibility of tweets during high impact events was studied in [19]. The authors used source based and content based features to indirectly measure the credibility of tweets and their sources. Content based features in the tweets like number of words, special

symbols, hashtags, pronouns, URLs and meta data like retweets were used. Source based features like number of followers, number of followees, age etc were used to measure the credibility of a user. The features were analysed for credibility using RankSVM and Relevance feedback algorithms. The limitation of their work is the requirement to establish ground truth using human annotation.

Spread of rumours and influence in Twitter

The spread of rumours in micro blogs was investigated in [20]. In particular, the authors investigated the spread of rumours in Twitter to detect misinformation and disinformation in them. The authors have proposed a framework using statistical modeling to identify tweets which are likely to be rumours from a given set of general tweets. They used content based, network based and microblog-specific memes for correctly identifying rumours. NLP techniques in sentiment analysis of the tweets was used for automatic detection of rumours. Content based features like lexical patterns, part-of-speech patterns, features corresponding to unigrams and digrams for each representation were used for classification of tweets. The authors used these techniques for rumour retrieval i.e., identifying tweets spreading misinformation. The belief classification of users to identify users who believe in the misinformation was done using the re-tweet network topology. The importance of re-tweet network topology has been clearly brought out in the paper. The authors have also used Twitter specific features like hashtags and URLs.

The measure of influence as given by the retweet mechanism offers an ideal mechanism to study large scale information diffusion in Twitter [21]. The degree of influence of nodes measured by calculating the number of followers and number of retweets showed different results with little correlation between the two. The relationship between indegree, retweets and mentions as measures of influence have been further analysed in [22]. The authors have supported the claim that the users having large number of followers are not necessarily influential in terms of retweets and mentions. Influential users have significant influence across a number of topics. Influence in terms of retweets is gained only after concerted efforts. Surveys have also shown that the users are poor judge of truthfulness based on contents alone and are influenced by the user name, user image and message topic when making credibility assessments [23].

Orchestrated semantic attacks in Twitter

Detection of suspicious memes in microblog platforms like Twitter using supervised learning techniques has been done in [7,8]. The authors have used supervised learning techniques based on the network topology, sentiment analysis and crowd-sourced annotations. The authors have discussed the role of Twitter in *political astroturf* campaigns. These are campaigns disguised as popular large scale grassroots behaviour, but actually carried out by a single person or organization. As per the authors, orchestrating a distributed attack by spreading a particular meme to a large population beyond the social network can be done by a motivated user. The paper discusses methods to automatically identify and track such orchestrated and deceptive efforts in Twitter to mimic the organic spread of information. The authors have described *Truthy*, a Web service to track political memes in Twitter to detect astroturfing and other misinformation campaigns. The importance of the use of retweets to study information diffusion in Twitter has been highlighted by the authors. Network analysis of the diffusion of memes followed by sentiment analysis

was used by the system to detect coordinated efforts to spread memes. The importance of detection of the spread of memes at an early stage itself before they spread and become indistinguishable from the real ones was also highlighted in the paper.

Being in the first page of the search results of any search engine is often regarded as an indicator of popularity and reputation. Search engines have introduced real time search results from social networking sites like Twitter, blogs and news web sites to appear in their first pages. A concentrated effort to spread misinformation as in political astroturf campaigns could have far reaching consequences if such search results are displayed prominently by search engines like Google. While studying the role of Twitter in spread of misinformation in political campaigns, Mustafaraj et al. have concluded that one is likely to retweet a message coming from an original sender with whom one agrees [24]. Similarly repeating the same message multiple times indicates an effort to motivate others in the community to accept the message. The authors described an attack named *Twitter-bomb* where the attackers targeted users interested in a spam topic and send messages to them, relying on them to spread the messages further. The authors have highlighted the ability of automated scripts to exploit the open architecture of social networks such as Twitter and reach a very wide audience. Measuring hourly rate of generation of tweets seems to be a meaningful way of identifying the spam accounts.

Analysis of measuring credibility of tweets

A summary of the analysis of the literature on measuring the credibility of information propagation in Twitter along with the pointers towards detection of misinformation is given in Table 1. The present efforts to detect the spread of misinformation in online social networks can be broadly classified with relation to the questions of *consistency of message, coherency of message, credibility of source* and *general acceptability* as given in Section "A generic framework for detection of spread of misinformation".

The analysis has brought out the following aspects:

- Automated means of detecting tweets are accurate, but computationally intensive and manual inputs are required.

Table 1 Comparison of metrics for measuring credibility of tweets

Criteria	Metrics	Authors	Accuracy	Complexity	Usefulness for fast detection	Remarks
Consistency of message	Retweets, mentions	[7,8,17,19, 22,24]	Retweets are better than mentions	No	Yes	
Coherency of message	Questions, affirms, denial, no of words, pronouns, hashtags, URLs, exclamation marks, negative and positive sentiments, NLP techniques	[7,8,17-20]	Decision tree algorithms with a combination of various factors are accurate	Yes	Computationally intensive, requires ground truth	Content analysis required. Metrics are an indirect measure
Credibility of Source	Tweets, retweets, mentions, indegree, user name, image, followers, followees, age	[7,8,17,19, 21,24]	Retweets are more accurate	No	Yes	
General acceptability	Retweets	[7,8]	Good	No	Yes	

- Retweets form a unique mechanism available in Twitter for studying information propagation and segregating misinformation.
- Analysing the information propagation using models in Computer Science and concepts of Cognitive Psychology would provide efficient solutions for detecting and countering the spread of misinformation.

Methods

We want to examine the information propagation pattern in Twitter to detect the spread of misinformation. The cognitive psychology analysis revealed that source of information is an important factor to be considered while evaluating the credibility. Moreover, the difference between misinformation and disinformation is in the intention of the source in spreading false information. The retweet feature of Twitter would enable us to understand the information propagation and grouping tweets based on their sources would reveal patterns which would enable us to estimate the credibility of the information being propagated.

Out of the four parameters stated above, *consistency of message* and *coherency of message* are internal to the user. These would be the first tools the user would employ to confirm the authenticity of the received information. A user who is convinced of the authenticity of the message or lack of it by these two parameters would not be bothered about using the other parameters to come to a decision. The external parameters of *credibility of source* and *general acceptability* would be used when the user has greater suspicion of the news item. We would assume that most of the news items spreading misinformation would fall in this category and a user accepts and forward them after evaluating the source of the message as well as the perceived general acceptance of the message.

Data sets

We carried out experimental evaluation on data sets obtained from the online social network 'Twitter'. We collected data using Twitter API for two different topics. The spreadsheet tool TAGS v5 used for collection of tweets using the Search API was provided by Martin Hawskey [25]. The topics enabled us to define the context for the study of spread of information. We had carried out our studies on a number of data sets obtained from Twitter using different keywords. Tweets were obtained for events like natural calamities, acts of terrorism, political events etc. Here we describe two of those to bring out our results. The keywords refer to different types of news items as explained below and the statistical details are given at Table 2:

- **Egypt.** Heavy political unrest and massive protests spread in Egypt during the months of Aug-Sep 2013. The news related to the these events were captured using the keyword 'egypt' for the period from 13 Aug 2013 to 23 Sep 2013.
- **Syria.** The use of chemical weapons in Syria in the month of Aug 2013 attracted worldwide criticism. The reflection in Twitter of the events was tracked using the keyword 'syria' for the period from 25 Aug 2013 to 21 Sep 2013.

Table 2 Details of the data set

Data set	#Tweets	#Retweets	#Senders	#Re-tweeters	Period
Egypt	141682	51723	10850	27532	13 Aug 2013 to 23 Sep 2013
Syria	104867	44708	11452	25415	25 Aug 2013 to 21 Sep 2013

Methodology

We now describe the methodology we adopted to analyse the data sets to detect misinformation in them. We aim to detect non credible information which have the potential to spread and possible collusion between users in spreading them. We outline the steps and then give the complete algorithm in the form of a flow chart.

Step 1: Consider only the retweets

We use the Twitter user as the first stage of social filter by not considering any tweets which are not retweets. The effectiveness of retweets to determine the credibility of information in Twitter has been verified already by our review of work done in the field. Retweets are the easiest means by which tweets are propagated and any person retweeting a tweet personally has validated the information content in the tweet. Further, this would also remove personal chats, opinions and initial misinformation not considered credible. The fact that a tweet has not been retweeted also indicates that the probability of it spreading to a sizeable proportion of the population is minimal.

Step 2: Evaluate the source of retweets

The credibility of the source is the next important factor to be considered. For this we identified and segregated the retweets as per the source. This step would enable us to estimate the unique tweets of the source which are being retweeted and the unique number of users who are retweeting the same. The pattern of retweeting was analysed. Human annotation was done to determine the credibility of the information. We found greater unevenness amongst the users retweeting tweets of a source, when the credibility of the source was poor. The sources spreading misinformation were being retweeted heavily by a limited proportion of users who have at least retweeted the source once. This points towards collusion between the users and deliberate attempts being made to spread misinformation. Favouritism in the retweeting behaviour is most often due to questions of credibility as has been validated in all the data sets studied.

The disparity in the retweeting behaviour was measured using Gini coefficient. Gini coefficient is a measure of inequality of a distribution. The metric is more often used to measure the disparity in income of a population. In the retweet graph, we wanted to measure the pattern of distribution of retweets of the tweets of a particular source. A value of Gini Coefficient nearer to zero would indicate a more even retweet behaviour and a value above 0.5 and nearer 1 would indicate greater disparity in that a few users are involved in retweeting a large number of tweets from the source and hence the possibility of misinformation in the contents and reduced credibility of the source.

We calculate the Gini coefficient, G using the approximation of the Lorenz curve, where it is approximated on each interval as a line between consecutive points. Let X_k be the cumulative proportion of the 'users' variable, for k = 0, ... n and $X_0 = 0$ and $X_n = 1$ and let Y_k be the cumulative proportion of the retweeets, for k = 0, ··, n and $Y_0 = 0$ and $Y_n = 1$. If X_k and Y_k are indexed such that $X_{k-1} < X_k$ and $Y_{k-1} < Y_k$, the Gini coefficient, G is given by

$$G = 1 - \sum_{k=1}^{n} \left(X_k - X_{k-1}\right)\left(Y_k + Y_{k-1}\right) \tag{1}$$

The patterns of retweet behaviour of sample users from the two data sets are shown in Figures 3 and 4. The graphs in Figures 3 and 4 depict the distribution of retweets

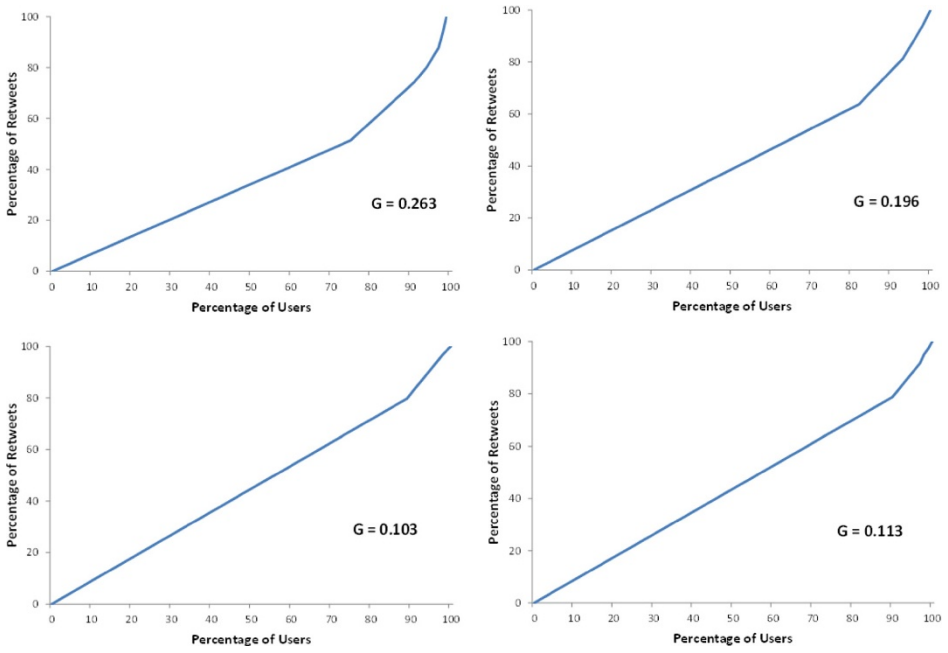

Figure 3 Distribution of retweets from four sample sources not involved in the spread of misinformation. The figure shows the retweeting pattern of tweets of non misinforming sources of information. The Gini coefficient is nearer to zero for all such source nodes.

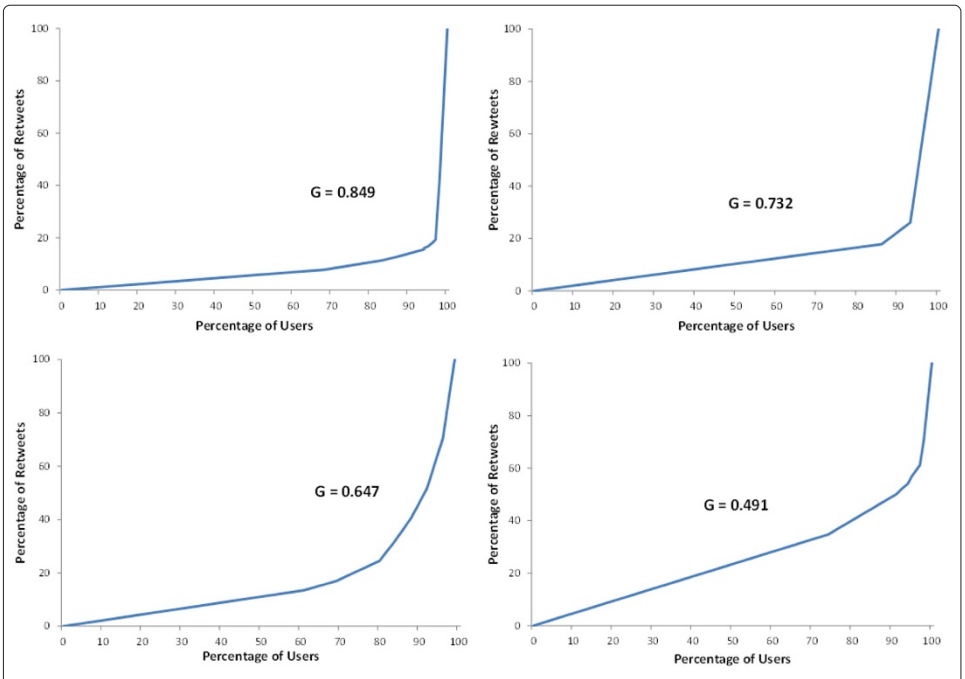

Figure 4 Distribution of retweets from four sample sources involved in the spread of misinformation. The figure shows the retweeting pattern of tweets of misinforming sources of information. The Gini coefficient is normally high and always above a minimum value of 0.4.

amongst the users for four different sources. The differences in patterns of distribution of retweets of the sources are quite obvious in the graphs. The graphs in Figure 3 are for sources who are not spreading misinformation. The equitable distribution of retweets is indicated by the low value of Gini coefficient G. A value of G equals zero would indicate perfectly even distribution. A value of unity for G would indicate just the opposite with one user completely taking all the share. The graphs in Figure 4 depict the distribution for sources involved in the spread of misinformation. We measured a threshold of G as 0.5, above which the sources were classified as spreading misinformation. The high values of G and the consequent different shapes of the graph are easily identifiable. All the sources involved in deliberate spread of misinformation had a high G value. But all the sources having a high G value were not found to spread misinformation. We conclude that the user would forward only tweets which he feels as possible misinformation and hence the high G value would indicate sources of misinformation for all the messages which had to be classified by automated techniques.

Step 3: Construct a retweet graph

We construct a retweet graph as in Figure 5. In the figure, the source 'S' has made three tweets RT1, RT2 and RT3, which have been subsequently retweeted by user nodes U1, U2, U3 and U4. Hence there are directed edges from the tweets to the source 'S' and from the 'retweeters' to the tweets. While U1, U3 and U4 have retweeted only one tweet of source 'S', U2 has retweeted all the three tweets.

Step 3: Evaluate the general acceptability of the tweet

We constructed a retweet graph similar to Figure 5 but involving all the sources and the tweets which have been retweeted. The graph is a bipartite graph with two types of nodes - user nodes and retweet nodes. This graph would clearly depict who tweeted what and the nodes responsible for their propagation in the network. The general acceptability of the tweet was measured using PageRank algorithm in this retweet graph [26]. The PageRank

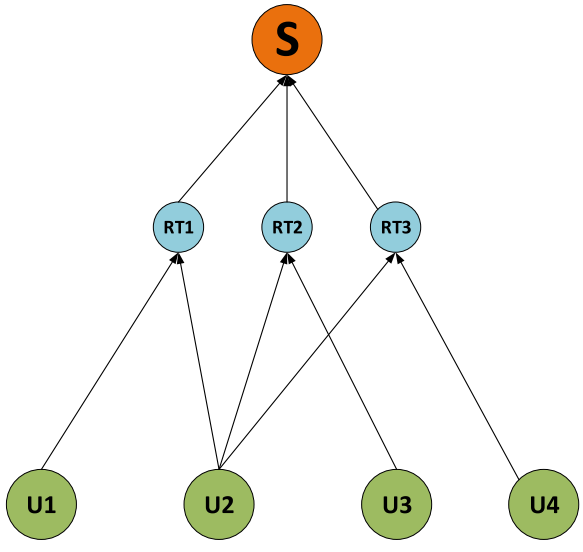

Figure 5 A retweet graph for a source 'S'. A sample retweet graph to show the construction of the retweet graph for a source 'S'.

of the source node is dependent on the PageRanks of its retweet nodes and then PageRank of the retweet nodes are in turn dependent on the number and type of user nodes retweeting them. A higher PageRank would indicate greater acceptability of the tweet - due to more number of retweets and being retweeted by more credible user nodes. A tweet can be considered generally acceptable if it has a higher PageRank. The threshold value is taken as the value at which the tailed distribution begins. As seen in the data sets most of the tweets have very low PageRank indicating that only a small proportion of the tweets are getting retweeted more number of times. The PageRank of a node n_i - retweet node or a user node, $PR(n_i)$ was calculated based on the equation:

$$PR\left(n_i\right) = \frac{1-d}{N} + d \sum_{n_j \varepsilon S(n_i)} \frac{PR\left(n_j\right)}{L\left(n_j\right)} \tag{2}$$

Here, $n_1, n_2, \cdot\cdot, n_N$ are the nodes in the retweet graph. $S(n_i)$ is the set of nodes that have a link to node n_i. $L(n_j)$ is the number of outgoing links from the node n_j. N is the total number of nodes in the bipartite retweet graph. We used the standard value of damping factor d as 0.85. Like all social computing strategies, PageRank is also susceptible to manipulation. This would happen when there is collusion between the users where each of them retweet the others' tweets. Such collusion would invariably result in greater communication edges between the nodes involved. The resulting favoritism in retweet behavior can be detected to a large extent using the Gini coefficient explained earlier.

Step 4: Content analysis of the finally filtered items

The output of the previous steps is given to the user. Based on his evaluation of the consistency and coherency of the message, and the additional quantified inputs of the credibility of the source and the general acceptability of the tweet, the user would be able to make an informed decision on the authenticity of the tweet. This would prevent more number of people from retweeting misinformation. Preventive measures such as these are bound to be more effective than any counter measures launched after the misinformation has spread to a large section of the population.

Proposed framework for speedy detection of misinformation in Twitter

The proposed framework would evaluate the credibility of the source and the general acceptability of the news items using collaborative filtering techniques to enable the user to make informed decisions and thus possibly avoid spreading misinformation. We have already outlined the steps for detection of sources involved in the spread of misinformation. The decision tree we constructed to implement the proposed sequence of actions is given at Figure 6.

Summary of the steps involved

The algorithm would do the following:

- Identify the original source of information (tweets) in the network.
- Evolve a methodology to rate the credibility of the source based on the acceptance of the tweets by the receivers.
- Construct a retweet graph to evaluate and measure the 'misinformation content' of a tweet and determine its credibility by the level of its acceptance by all the affected users using Gini coefficient.

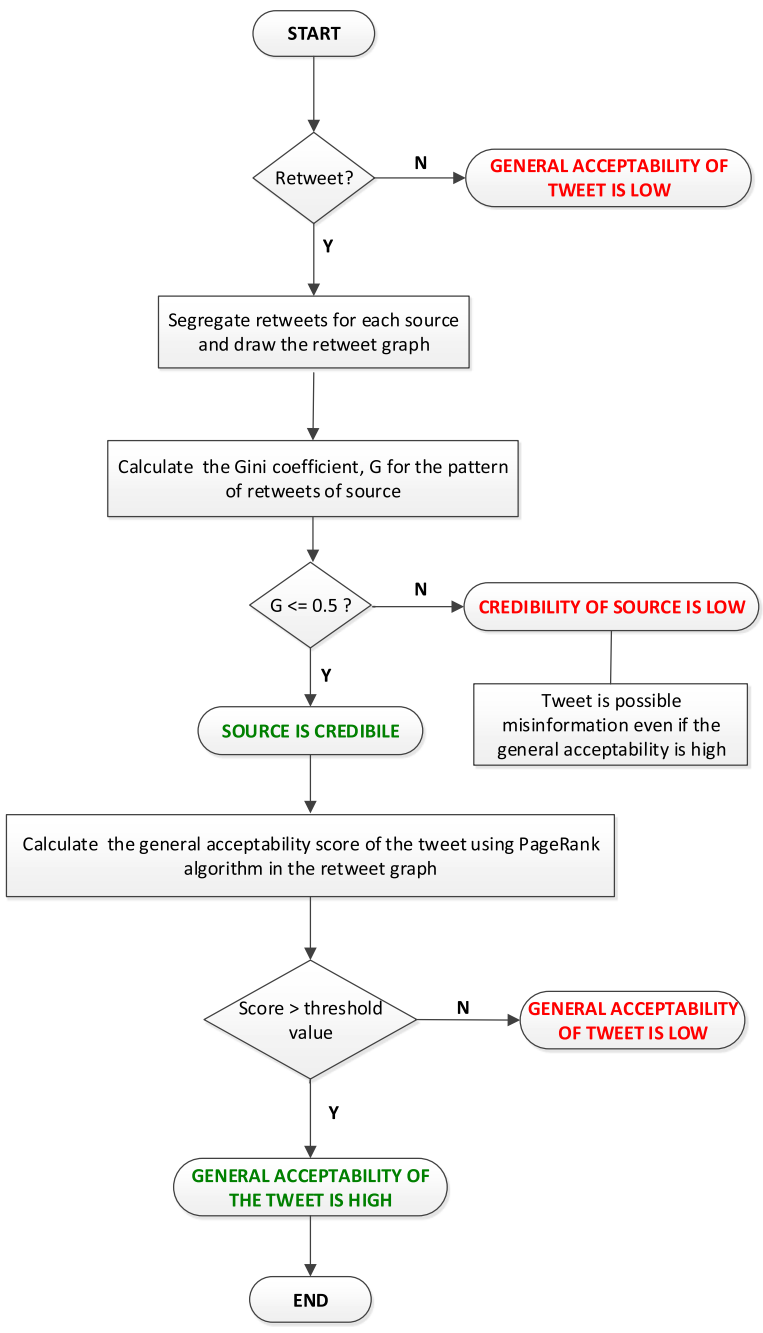

Figure 6 Algorithm for determination of misinformation in tweets. The flowchart giving the algorithm to detect misinformation using a retweet graph to estimate the general acceptability of tweets and credibility of sources.

- Segregate the possible sources of misinformation as non credible users and the corresponding tweets.
- Evaluate the general acceptance of tweets from credible users using PageRank algorithm in the retweet graph.

- Present the credibility of the source and the general acceptance of the tweet to the user to help him evaluate the information contents of the tweet.

Results and discussion

We evaluated the proposed algorithm in all the data sets. The retweet graphs of the data sets were visualised using the software Gephi [27]. In Figures 7 and 8 we show the retweets graphs of the Egypt data set. While Figure 7 gives the broader view of the retweet graph, an exploded view of a section of the retweet graph showing the internal details like name, PageRank and modularity class of the nodes is shown in Figure 8. The nodes with names starting with 'RT' are the retweets and the others are the user nodes in the bipartite graph. Similar results were observed for the Syria data set also.

Measuring credibility of tweets

A plot of the Gini coefficients obtained for the Egypt and Syria data sets are given at Figures 9 and 10 respectively. They show a heavy tailed distribution where the Gini coefficients of most of the user nodes are 0. The figure corroborates the fact that most of the communication are not misinformation. The user nodes with higher Gini coefficients are fewer in number and hence they can be monitored effectively. A threshold value of 0.5 is

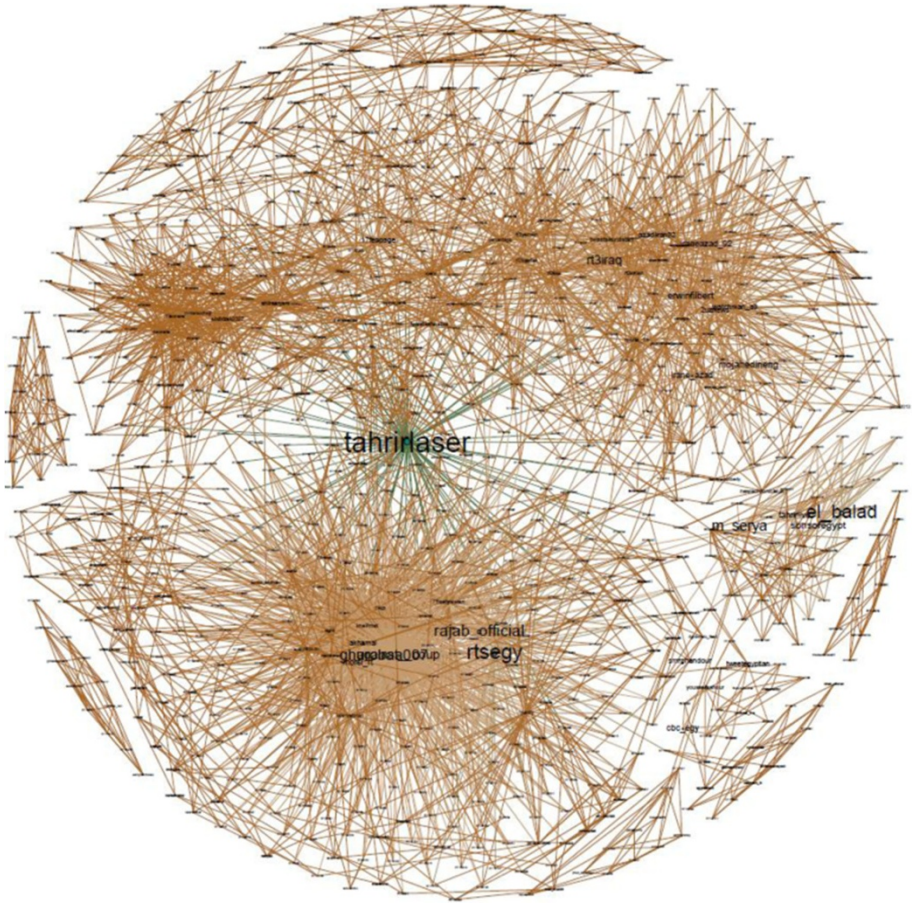

Figure 7 The Retweet graph of Egypt data set. Part of the retweet graph obtained from the Egypt data set showing the retweet nodes and user nodes. The graph has been made using the software 'Gephi'.

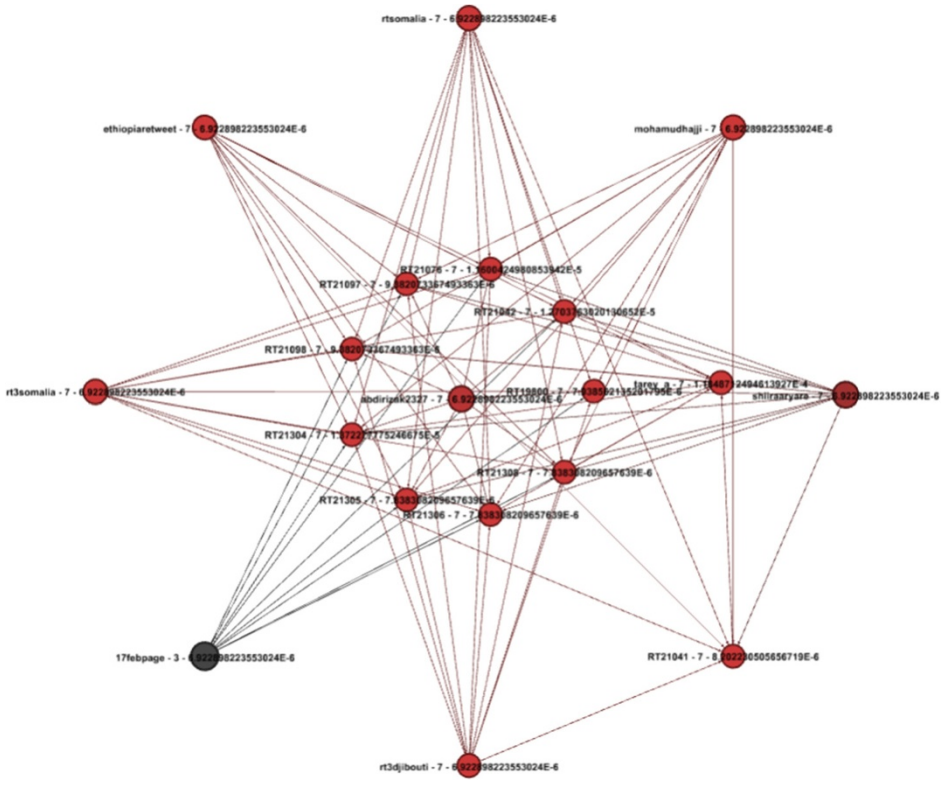

Figure 8 The magnified view of a section of the retweet graph of the Eqypt dataset. A section of the retweet graph has been magnified to show the two types of nodes and the modularity number and PageRank of the nodes.

meaningful to separate out the misinforming users, which has been validated by verifying them. A plot of the PageRank scores obtained for the Egypt and Syria data sets are given at Figures 9 and 10 respectively. As expected, they also show a heavy tailed distribution, with few retweets and user nodes having higher PageRank values. This again supports the fact that a few tweets would be retweeted more heavily than others.

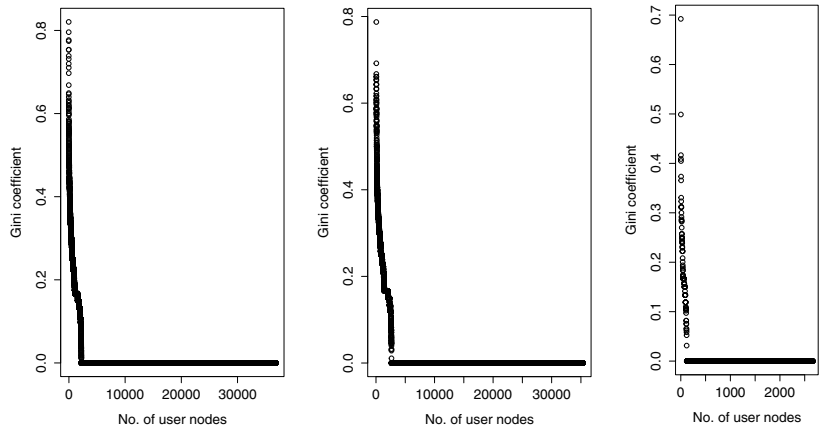

Figure 9 Distribution of Gini coefficients in the Egypt data set. The figure shows the tailed distribution of Gini coefficients of all the user nodes in the retweet graph in the Egypt data set. Majority of the user nodes have Gini coefficients nearer to zero, indicating that spreading misinformation is limited to a very small set of nodes.

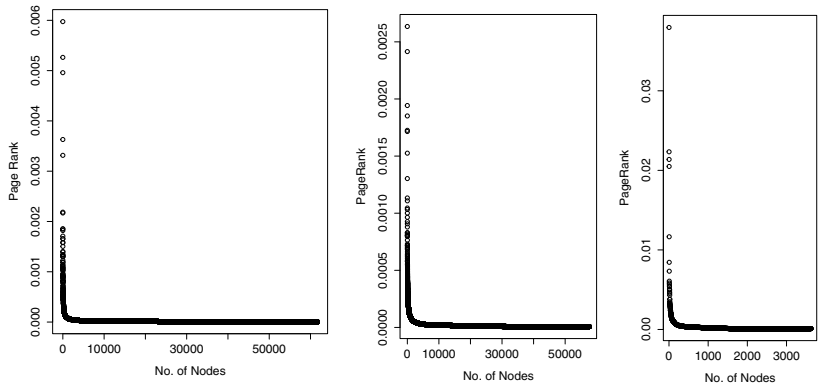

Figure 10 Distribution of Gini coeffcients in the Syria data set. The figure shows the tailed distribution of Gini coefficients of all the user nodes in the retweet graph in the Syria data set. The results are similar to the Egypt data set and majority of the user nodes have Gini coefficients nearer to zero, indicating that spreading misinformation is limited to a very small set of nodes.

Analysis

Using Gini coefficient, the True positive rate to correctly identify all sources of misinformation was over 90%. The False positive rate, where users were wrongly identified as sources of misinformation, was less than 10%. The False negative rate, where the algorithm failed to identify sources of misinformation was less than 10%. The tailed distribution of the spread of Gini coefficients in the data sets revealed that the sources involved in the spread of deliberate misinformation were few as expected.

The Process of diffusion of information is more effectively understood using metrics like Gini coefficient. If we consider retweets as a measure of adoption of information, differences in adoption behaviour would indicate differences in perceived credibility of information. Misinformation or disinformation are context specific and hence responses of users assume great significance. If information from a certain source is accepted as credible uniformly by a large number of users, quite possibly that source is credible. On the other hand, if there are variations in the acceptance levels, the simple explanation is apparent non credibility of messages of the source. Similarly if most of the users receiving users do not repropagate information from a source also, his credibility is low. However, gini coefficient value would be low as the the variations in acceptance are not pronounced. This result is also acceptable, as we are unable to detect misinforming tweets which have been decided by the collective intelligence of the network users to be non credible. Our algorithm would segregate messages which are repropagated differently by a significant section of users, which has the potential to create a certain perceived level of social consensus. By deploying our proposed framework at the client end, we give better inputs regarding social consensus and credibility of sources.

The PageRank algorithm could correctly identify the tweets which were being retweeted in greater numbers. The threshold value would decide the level of classification of a news item as generally acceptable or not. We had taken a threshold value of the PageRank score before it starts to even out to near zero levels. In all cases, the acceptability of the tweets were correctly evaluated.

The algorithm is proposed to be used by an user to detect the credibility of sources and general acceptability of the tweets. It would also use the cognitive powers of the user to

carry out the initial screening of messages. With that, the false positives of the algorithm would be minimum and the algorithm would provide valid inputs to the user in an accurate manner. The users who have a high degree of communication with the segregated sources could also be now identified along with the tweets involved in the spread of false information.

Conclusions

The paper has explored the application of principles of cognitive psychology in evaluating the spread of misinformation in online social networks. We have proposed an effective algorithm for speedy detection of spread of misinformation in online social networks taking Twitter as an example. Analysing the entire content of a social network using linguistic techniques would be computationally expensive and time consuming. The aim was to propose an algorithm which would use the social media as a filter to separate misinformation from accurate information. We were also interested only in misinformation which were likely to spread to a large section of the social network. Analysing the problem from a cognitive psychology point of view enabled us to understand the process by which a human mind determines the credibility of information. The literature review of the work done in detection of misinformation in Twitter brought out the critical features of Twitter which would help us identify the cues to deception in tweets. Our proposed algorithm is simple and effective in limiting the computation required to identify the users involved in spread of misinformation and estimate the level of acceptance of the tweets.

Prevention is better than cure. The spread of misinformation can be prevented if users are enabled to make correct decisions while retweeting the messages they receive. The algorithm would enable the user to make informed decisions while spreading information in OSNs. The implementation of the algorithm at the client end can be done in the form of a browser plug-in or a Twitter app. The proposed plug-in or the app would help the user to make correct decisions while forwarding messages and thus prevent large scale misinformation cascades. The important feature of the algorithm is that it does not make use of any specific features of Twitter. The proposed methodology would be applicable for other online social networks also which support easy re-propagation like Facebook, with its 'share' feature, Diggs with its 'voting' mechanism or even e-mail networks with their 'forwarding' features.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The work was carried out as part of research on online social networks and is a result of equal contribution by each of the authors. Both authors read and approved the final manuscript.

Acknowledgements
The authors want to acknowledge BITS-Pilani, Hyderabad campus for the facilities provided. We also want to thank Agrima Srivastava, our co-researcher for her excellent inputs during various brainstorming sessions.

Received: 22 January 2014 Accepted: 19 July 2014
Published online: 24 September 2014

References

1. Mintz AP (2002) Web of deception: Misinformation on the Internet. Information Today, Inc., New Jersey, USA
2. Karlova NA, Fisher KE (2013) "Plz RT": A social diffusion model of misinformation and disinformation for understanding human information behaviour. Inform Res 18(1):1–17
3. Stahl BC (2006) On the difference or equality of information, misinformation, and disinformation: A critical research perspective. Inform Sci: Int J Emerg Transdiscipline 9:83–96

4. Fallis D (2009) A conceptual analysis of disinformation. iConference, Chapel Hill, NC, California, USA
5. Libicki MC (2007) Conquest in cyberspace: National security and information warfare. Cambridge University Press, New York, USA
6. Lewandowsky S, Ecker UK, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction continued influence and successful debiasing. Psychol Sci Public Interest 13(3):106–131
7. Ratkiewicz J, Conover M, Meiss M, Goncalves B, Patil S, Flamini A, Menczer F (2010) Detecting and tracking the spread of astroturf memes in microblog streams. arXiv preprint arXiv:1011.3768
8. Ratkiewicz J, Conover M, Meiss M, Goncalves B, Patil S, Flammini A, Menczer F (2011) Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th International Conference Companion on World wide Web. ACM, Hyderabad, India, pp 249–252
9. De Neys W, Cromheeke S, Osman M (2011) Biased but in doubt: Conflict and decision condence. PLoSONE 6(1):e15954
10. Fitzgerald MA (1997) Misinformation on the internet: Applying evaluation skills to online information. Emerg Libr 24(3):9–14
11. Kata A (2010) A postmodern pandora's box: Anti-vaccination misinformation on the internet. Vaccine 28(7):1709–1716
12. Nguyen NP, Yan G, Thai MT, Eidenbenz S (2012) Containment of misinformation spread in online social networks. In: Proceedings of the 3rd Annual ACM Web Science Conference. ACM, Illinois, USA, pp 213–222
13. Nguyen NP, Yan G, Thai MT (2013) Analysis of misinformation spread containment in online social networks. Comput Netw 57(10):2133–2146
14. Budak C, Agrawal D, El Abbadi A (2011) Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web. ACM, Hyderabad, India, pp 665–674
15. Nguyen DT, Nguyen NP, Thai MT (2012) Sources of misinformation in online social networks: Who to suspect?" In: Military Communications Conference, MILCOM 2012. IEEE, Orlando, USA, pp 1–6
16. Reuters IANS (2013) Ethnic riots sweep assam, at least 30 killed. [Online]. Available: http://in.reuters.com/article/2012/07/24/india-assam-riots-floods-idINDEE86N04520120724. Last accessed 21 August 2014
17. Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web. ACM, Hyderabad, USA, pp 675–684
18. Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: Can we trust what we RT? In: Proceedings of the First Workshop on Social Media Analytics. ACM, 2010: Washington DC, USA, pp 71–79
19. Gupta A, Kumaraguru P (2012) Credibility ranking of tweets during high impact events. In: Proceedings of the 1st Workshop on Privacy and Security in Online Social Media. ACM, Lyon, France, p 2
20. Qazvinian V, Rosengren E, Radev DR, Mei Q (2011) Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburg, UK, pp 1589–1599
21. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. ACM, Raleigh, USA, pp 591–600
22. Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user in uence in twitter: The million follower fallacy. In: Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM) 14. AAAI, Washington DC, USA, pp 10–17
23. Morris MR, Counts S, Roseway A, Hoff A, Schwarz J (2012) Tweeting is believing?: understanding microblog credibility perceptions. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. ACM, Raleigh, USA, pp 441–450
24. Mustafaraj E, Metaxas PT (2010) From obscurity to prominence in minutes: Political speech and real-time search. In: WebSci10: Extending the Frontiers of Society On-Line. The Web Science Trust, Raleigh, USA p 317
25. Hawksey M (2013) Twitter Archiving Google Spreadsheet TAGS v5. JISC CETIS MASHe: The Musing of Martin Hawksey (EdTech Explorer). http://mashe.hawksey.info/2013/02/twitter-archive-tagsv5/, [Online: Last accessed: 21 August 2014]
26. Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. Stanford InfoLab, California, USA
27. Bastian M, Heymann S, Jacomy M (2011) Gephi: an open source software for exploring and manipulating networks. 2009. In: International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, pp 361–362

doi:10.1186/s13673-014-0014-x
Cite this article as: Kumar and Geethakumari: **Detecting misinformation in online social networks using cognitive psychology.** *Human-centric Computing and Information Sciences* 2014 **4**:14.