CHAPTER 1

Mining Misinformation in Social Media

Liang Wu
Arizona State University

Fred Morstatter
Arizona State University

Xia Hu
Arizona State University

Huan Liu
Arizona State University

CONTENTS

A RAPID INCREASE of social media services in recent years has enabled people to share and seek information effectively. The openness, however, also makes them one of the most effective channels for misinformation. Given the speed of information diffusion on social networks coupled with the widespread propagation of fake news [55], phishing URLs [24], and inaccurate information [37], misinformation escalates and can significantly impact users with undesirable consequences and wreak havoc instantaneously. In this chapter, we discuss the generation and diffusion of misinformation in social media, and introduce challenges of identification, intervention, and prevention methods. We use examples to illustrate how to mine misinformation in social media, and also suggest possible future work.

1.1 INTRODUCTION

Social media has changed the way we communicate. More and more people use such platforms to read, release and spread either breaking news [45] or their updates to their friends [52] [20]. The openness of social network platforms enables and motivates them to communicate freely online, but it also brings new problems. Without careful proofreading and fact-checking, fake news and inaccurate information will unintentionally be spread widely by well-meaning users. Such misinformation and disinformation could be devastating in social media, as it corrupts the trustworthiness.

Misinformation is fake or inaccurate information which is unintentionally spread, while disinformation is intentionally false and deliberately spread. In this work, we generally focus on inaccurate information in social media which misleads people, so we refer them as misinformation. Misinformation causes distress and various kinds of destructive effect among social network users, especially when timely intervention is absent. As mentioned in several news[1], misinformation has helped unnecessary fears and conspiracies spread through social media. One such example is Ebola. As some potential cases are found in Miami and Washington D.C., some tweets sounded as if Ebola is rampant and some kept tweeting even after government issued a statement to dispel the rumor [38]. In this work, we survey recent related research results and provide a thorough analysis of misinformation in social media.

[1] http://time.com/3479254/ebola-social-media/

Definition of Misinformation in Social Networks

Toward better investigating misinformation in social media websites, we organize it according to the intention of user spreading misinformation:

> **Unintentionally-Spread Misinformation**:
> Some misinformation is created and forwarded spontaneously. People tend to help spread such information due to their trust of their friends and influencers in a social network, and want to inform their friends of the underlying issue.
> **Intentionally Spread Misinformation**:
> Some rumors and fake news are created and spread intentionally by malicious users to cause public anxiety, mislead people and deceive social network users for improper profit. This kind of misinformation is also called disinformation. In this work we use both words interchangeably.

The suggested categories cover most social media misinformation. But in the real world, misinformation is often more complex and may meet both criteria. For example, some rumors are created by malicious users and people are tricked into amplifying it [40, 49].

Misinformation such as fake news, rumors and inaccurate information can cause ripple effects in the real world. In 2013, a rumor saying that explosions at the White House happened and the president was injured[2]. The rumor sent a shudder through the stock market. Misinformation not only triggers financial panic, it also causes public anxiety [48], and ends careers [35]. In the 2013 World Economic Forum[3], the issue of "misinformation spread" has been voted as one of the top ten globally significant issues of the year [56]. In order to cope with misinformation in social networks, researchers launch crowdsourcing systems to verify social media information[4] [53]; Police arrest users who spread specific rumors to reduce the destructive effects [48]. However, since the spread is faster and wider than ever thanks to social networks, the damage is often beyond control.

Although crowdsourcing helps identify misinformation for journalists, the high velocity of online social media data generation makes it nearly impossible for them to keep up. Traditional penalties toward

[2]http://www.cnn.com/2013/04/23/tech/social-media/tweet-ripple-effect/
[3]http://reports.weforum.org/outlook-14/top-ten-trends-category-page/
[4]https://veri.ly/

malicious information spreaders are still useful for shocking them, but observations from the financial panic [56] and the mass shooting at Sandy Hook Elementary School [48] reveal that misinformation keeps spreading and infecting users even after they are claimed to be false officially. Thus, an effective misinformation intervention method is in need for combating diffusion in social networks. In order to cope with misinformation issue, we first introduce misinformation diffusion, its detection and intervention:

Misinformation Diffusion: In order to know why misinformation is spread wider and faster than ever, we provide readers with a general understanding of information diffusion models. The distinct aspects of misinformation are then discussed. Interesting observations are given about the relationship between misinformation and network structures.

Misinformation Detection: To cope with the high velocity and volume of social media data, we introduce several learning algorithms that detect malicious information or its spreaders.

Misinformation Intervention: We introduce several methods about limiting misinformation spread in an early age or after it is diffused, so as to remove or reduce its impact.

1.2 MISINFORMATION MODELING

Misinformation can be widely spread in a very short time. As denoted in the "financial panic" case, the rumor was forwarded over 3,000 times before Twitter blocked it. To explain how the information is diffused in a networked environment, we first introduce several diffusion models. In addition, we discuss about the distinct aspects of misinformation diffusion and provide readers with several findings between misinformation spread and network structure.

1.2.1 Information Diffusion in Social Networks

Information diffusion in social networks can be viewed as a process by which the news, events and different kinds of information are posted, forwarded and received by social network users. By representing each user as a node and the friendship between users as edges, social networks are transformed into a graph $G = (V, E)$, where V is the set of

nodes and E is the set of edges between nodes. Then the information diffusion process can be viewed as some signal or label being propagated in the network. A formal definition can be found in [8].

There are various models which are designed to abstract the pattern of information diffusion. Here we introduce four diffusion models, i.e., *SIR Model* [29], *Tipping Model* [7], *Independent Cascade Model* [28] and *Linear Threshold Model* [28].

As discussed in [58], there are three essential roles for diffusion: **Senders** who initiate the diffusion process, **Spreaders** who forward such information to their followers, and **Receivers** who receive information being diffused in the social media websites, which are the largest group of people in the whole process and sometimes overlap with spreaders: if people choose to forward news they receive, they become receivers.

The key distinguishing points of different diffusion models are twofold: 1) method of information being diffused between senders, spreaders and receivers; 2) and evolution of individual roles during the process of diffusion. In the following parts, we will describe these aspects of different information and misinformation diffusion models.

SIR Model

The SIR model describes the information diffusion process in network as an infectious disease spread in a community. So the nodes are generally classified into three categories: **S** - the susceptible to be infected, **I** - the infected individuals who are active to infect others, **R** the recovered individuals who recovered and are vaccinated against the disease. In the context of information diffusion in a social network, infected nodes can be regarded as those who were already informed of certain news or events, and are ready to pass them to neighbors; recovered nodes are those who have been informed, but are not passing the information to neighbors; susceptible nodes are the users who are not informed, but may be informed by others.

According to the user categorization, the information exchange happens between infectious nodes and susceptible nodes. In order to model the process, a global parameter is introduced as the probability that a susceptible user will be activated if it has a infected friend called β, in addition, a global parameter is also introduced to represent the probability of infected nodes getting recovered called γ. Figure 1.1 depicts the structure of SIR model. In SIR model, if a user has multiple linked infectious neighbors, the links are all independent with each other and the user can be infected by at most one neighbor. The infection process

Figure 1.1: Different user roles and their relationships in SIR model.

is modeled as Equation 1.1:

$$I_\beta(a) = \mathbf{1}(\sum_{\substack{(b,a)\in\mathbf{E},\\ b\in\mathbf{V}\cap\mathbf{I}}} \mathbf{1}(f^{rand} \geq \beta) > 0), \tag{1.1}$$

where $I_\beta(a)$ represents the infection status of a susceptible node a in the next time stamp given β, and $1(\cdot)$ is a function which equals one when its component is true and equals zero otherwise. $\mathbf{E}$ and $\mathbf{V}$ are the set of edges and nodes, respectively, and $\mathbf{I}$ is the infectious node set. We use (b, a) to denote the directed/indirected link between two nodes and use f^{rand} to denote the random probability generation function. Thus, node b is a's infectious neighbor and a will be activated if any of its neighbor actives it.

Tipping Model

As implied by "The Power of Context" [16], human behavior is strongly influenced by its environment. Similar intuition has been adopted by the tipping point model. In tipping model, there are two kinds of users: people who adopted the behavior and people who did not, where spreaders and senders are not explicitly distinguished. Since the adoption process is irreversible, information diffusion only happens between the second class of users and their active neighbors.

As depicted in Figure 1.2, a node will be influenced by its friends about adopting a certain behavior. The behavior can be buying a new cell phone or wearing a specific brand of clothes. In order to model the diffusion process, a threshold probability θ is introduced to judge whether the behavior reaches the "tipping point": If the ratio of a user's friends adopting a certain behavior reaches the tipping point probability θ, the user will also adopt the behavior. The infection process is

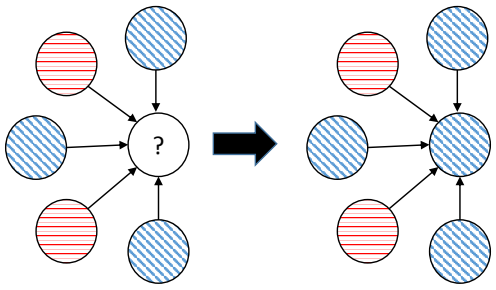


Figure 1.2: An illustrative example of the Tipping model, where the bar's height indicates extent of belief.

modeled as Equation 1.2:

$$I_\theta(a) = \mathbf{1}(\sum_{\substack{(b,a)\in\mathbf{E},\\ b\in\mathbf{V}\cap\mathbf{I}}} f(b,a) \geq \theta), \tag{1.2}$$

where $f(b,a)$ is the activation probability between b and a, and a will be activated when influence from all infectious neighbors exceeds the threshold θ.

Independent Cascade Model

SIR model and tipping model have different definitions of information being diffused, but they all assume a global parameter should work for the whole network. This strong assumption reduces the computational complexity, but it fails to handle the complex situations in the real world. More generalized models with variable parameters are proposed to handle such cases. Independent Cascade (IC) model is the generalized form of SIR model. Similarly, it formulates the diffusion process as disease epidemic, but the infectious probability is associated with different edges. The infection process is modeled as Equation 1.3:

$$I_\beta(a) = \mathbf{1}(\sum_{\substack{(b,a)\in\mathbf{E},\\ b\in\mathbf{V}\cap\mathbf{I}}} \mathbf{1}(f(b,a) \geq \beta) > 0). \tag{1.3}$$

The probability $f(b,a)$ can be achieved based on different applications such as the interaction frequency, geographic proximity and so on. Thus, IC model is able to embed more contextual information.

Linear Threshold Model

Linear Threshold (LT) model is a generalized form of tipping model. Instead of setting a global threshold θ for every edge between users, LT defines a probability distribution over all edges. The infection process is modeled as Equation 1.4:

$$I_\theta(a) = \mathbf{1}(\sum_{\substack{(b,a)\in\mathbf{E},\\ b\in\mathbf{V}\cap\mathbf{I}}} f(b,a) \geq \theta_{(b,a)}). \tag{1.4}$$

Tipping thus can be viewed as a special form of LT model, which employs a uniform distribution. By replacing the deterministic threshold θ with probabilistic thresholds, LT model is more capable of predicting the outcome of a diffusion given the seed set users. A problem of LT is the computational complexity. The probabilistic diffusion process makes the outcome calculation a #P-Hard problem, which cannot be efficiently computed in polynomial time as tipping model. In order to solve the problem, some more scalable methods have been introduced to reduce simulation runs.

One efficient method is called SIMPATH [17]. Normally, the social network linkage structure is complex and contains numerous edges with various probability. When information diffusion starts from the seed set users and goes to all rest individuals, some paths are more influential than the rest. SIMPATH reduces the computational time through filtering out paths without enough confidence. A similar method called Maximum Influence Arborescence (MIA) [9] was also proposed to accelerate the computation of independent cascade model. Various other acceleration algorithms are also available [41] [18].

1.2.2 Misinformation Diffusion

In this section, we introduce the diffusion model of misinformation. The diffusion of misinformation is more related to the trust and belief in social networks. The epidemic models, including SIR and IC, assume the infection occurs between an infectious user and a susceptible user with a predefined probability. As mentioned earlier, the probability may increase with more interactions or other contextual conditions. Although the links of a user are independent, a user who has more infectious friends is more likely to be infected. The tipping and LT models also contain a parameter to estimate probability of a user being infected based on the number of activated friends. Generally, given

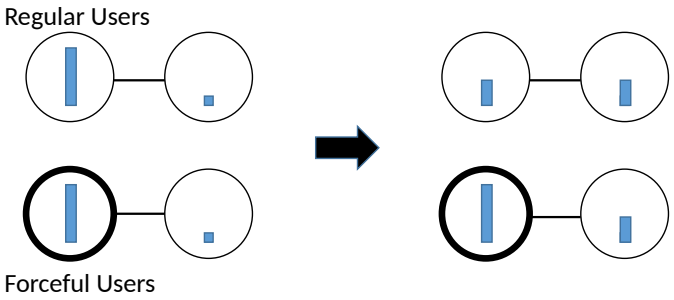


Figure 1.3: An example of belief exchange of misinformation, where heights of bars in circles represent extent of belief.

infinite time and an optimal seed set of senders, they assume all users will be infected.

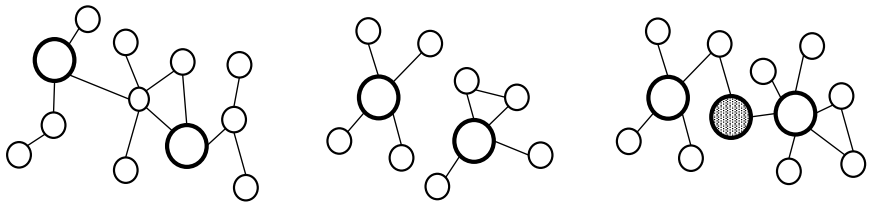
However, the diffusion outcome of misinformation is often the global recovery or immunity. Such phenomena reveal that, no matter how widespread a piece of misinformation is diffused, some nodes will not be affected and will keep intervening such diffusion. In order to model such process, existing approach categorizes people into two genres: regular individuals and forceful individuals [1].

The information diffusion process is reformulated as belief exchange. A parameter $\theta \in \mathbb{R}$ is introduced to represent the probability of learning or accepting some information. Misinformation diffusion process is then defined as the exchange of belief between two nodes.

As depicted in Figure 1.3, the exchange between different users are different. When a regular user interacts with another regular user, they are simultaneously affecting each other. A consensus belief will be achieved by averaging their beliefs. But when a regular node interacts with its forceful counterpart, only the regular node will be affected and the forceful individual will keep the original belief.

In the beginning of misinformation diffusion, all regular nodes are assumed to have a belief probability taken from a certain probability distribution, and forceful individuals may hold some specific scores against the regular. Through iterations of belief exchange, the belief converges to a consensus among all individuals. The higher the consensus belief is, the more widespread misinformation is diffused over the network. More formally, the process is modeled as:

$$\boldsymbol{x}_i(0) = \epsilon_i \boldsymbol{x}_i(0) + (1 - \epsilon_i)\boldsymbol{x}_j(0). \tag{1.5}$$



(a) A graph where forceful nodes may be affected by other forceful nodes and communities.

(b) A graph with two disconnected communities and forceful nodes have infinite impact.

(c) A graph with two separate communities which are bridged by a forceful node.

Figure 1.4: Examples of Network Structures

Here, we assume the social network has n agents. $\boldsymbol{x}(0) \in \mathbb{R}^n$ records the belief level of users at time 0. Equation 1.5 shows how belief exchange takes place between two nodes i and j, where θ measures the belief and ϵ_i measures the extent user i will stick to his original belief. If both user i and j are regular users, then $\epsilon_i = 0.5$ and such change happens to j's belief score ($\epsilon_j = 0.5$). If only user i regular but user j is a forceful individual, then $\epsilon < 0.5$ and user j is not significantly affected by the exchange ($\epsilon_j \approx 1$). When user i is a forceful user, then $\epsilon \approx 1$ and user j will be affected only if he is regular.

Various models have been proposed to predict extent of misinformation diffusion given different network topology and user roles [1]. Authors provide an efficient algorithm to estimate the upper bound of potential for misinformation in a social network. They study properties of different network structures. Three interesting observations are found, which will be described in the following paragraph.

If the nodes in a graph are connected to various forceful nodes as well as connected to many regular nodes, as shown in Figure 1.4a, misinformation diffusion will be attenuated by a large scale of social networks, i.e., the larger the graph is, the smaller the diffusion will be. On the contrary, if the social network is not well connected, as shown in Figure 1.4b, consisting of disconnected communities, the extent of misinformation diffusion will be strengthened. If a forceful node is connected to both the separate communities, which serves like a bridge as shown in Figure 1.4c, the potential extent of misinformation belief will be affected by the bridge: no matter whether being connected to the

bridge or not, members in a same community tend to share a common belief level.

The three observations can be explained by some real world examples. When there are enough people and authorities in a social network, negative effects of misinformation will be reduced after thorough public scrutiny. When nodes in a social network are not well connected and many disconnected subgraphs exist, information aggregation takes place mainly at the intra-group level. The belief level can be easily biased by the local leaders in different communities. Thus the underlying topology is prone to attack of misinformation. The third structure, where a forceful individual bridges different disconnected communities, approximates situations where political leaders obtain different information from different group of individuals. Thus the intra-group belief level converges to the consensus.

Note that the key difference of diffusion models between misinformation and information is the adoption of trust. The information spread can be modeled as an epidemic process, where being contacted increases the probability of infection. On the other hand, being informed of misinformation only changes the extent of trust instead of literally infecting the node. Since diffusion models aim to find the optimal seed set for spreading information, nodes with high centrality scores are better focused. In misinformation diffusion processes, globally influential nodes prove to be less effective than locally influencers in gaining trust of their followers. These research results indicate the importance of locally influential nodes of controlling misinformation.

Different models have been proposed to study behavioral features of people. Rumor spreading has also been regarded as a dynamics model, where social actors are categorized as susceptible, infected and refractory. Karlova et al. categorize users into two genres [27], diffusers and receivers. When receivers receive some information from diffusers, they judge whether to trust and further pass the information based on contextual features. A unified model which jointly considers information and misinformation was also proposed [2].

1.3 MISINFORMATION IDENTIFICATION

In order to provide a brief introduction of misinformation identification techniques, we examine several representative works from two aspects: directly identifying misinformation itself and detecting misinformation through exposing its spreaders. In fact, detecting misinformation from

online information flow has a long history, and has been studied extensively since the last decade when email system entangled with the huge amount of junk mails. Traditionally, the content and network structure are two key factors of telling spam from regular information. Here, content means the information in the email. Since spam is designed to advertise certain websites or products, content is a useful clue for detecting such information. Network structure refers to the graph built by the routes of email transmission. If a user/server has an abnormal pattern of sending emails, such as sending numerous similar emails to a lot of strangers, we can probably predict it is suspicious. Similarly, misinformation and disinformation in a social network also have both features. So we will focus on content and network structure of the two kinds of methods.

1.3.1 Misinformation Detection

Algorithm 1 illustrates the learning framework of misinformation detection algorithms.

Algorithm 1 Misinformation Detection in Social Networks

Input: The raw content $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, the identity label vector $\boldsymbol{y} \in \mathbb{R}^m$,
Output: The optimal misinformation estimator f.
Generate compact representation $\boldsymbol{X'} \in \mathbb{R}^{m \times k}$ based on $\boldsymbol{X}$;
Obtain the label vector $\boldsymbol{y}$
Repeat
Update f to reduce loss $\sum_{i=1}^{m} \mathcal{L}(f(\boldsymbol{x}_i), y_i)$;
until Convergence or Maximum Iterations

Identifying misinformation has been studied in psychology [3] [4] [11] [10]. They refer misinformation as rumors and study the exhibition of rumors and the psychological reasons of spreading rumors. Several detection methods are introduced based on human behaviors. Recent research also reveals how rumor is differently diffused online in microblogging platforms.

More scalable algorithms have been proposed to study the spread of memes and misinformation on the web. Some algorithms directly use the raw features to model the patterns of misinformation. In [47], Ratkiewicz et al. built up a "Truthy" system to automatically identify misleading political memes on Twitter. They directly extract raw features, such as hashtags, links and mentions in their system. By select-

ing such features, a compact representation could be generated. Then different kinds of bayesian classifiers are trained and used to classify misinformation.

Although merely using hashtags and links directly leads to a compact representation, the vast information hidden in user text may be ignored. In order to reduce the dimensionality of raw tweet content, Gao et al. [14] propose to cluster tweet content in an unsupervised manner. They try to quantify and characterize campaigns on Facebook by analyzing wall messages spreading between Facebook users. An initial solution is provided [14]. Empirical analysis is employed to test the their proposed model based on real data. The characteristics, impact and sources of malicious accounts are also studied based on their model and empirical analysis. In addition, some third party tools are also employed for validation.

Another similar intuition of compacting social content is to focus on the topics instead of words. When users tweet online, each word can be seen as generating from a certain topic. Then the corresponding tweet can be reduced to a probability distribution over multiple topics. Since the number of topics is normally far smaller than that of words, another low dimensional representation can then be achieved. Existing work has proven the effectiveness of employing LDA to detect deceptive information in a social network [50].

Misinformation detection algorithms normally model the prediction as a descriminant learning problem: directly optimizing the prediction based on training data. Since both the clusters and topics can be viewed as a subset of words, the underlying assumption of such models is that information consisting of similar words or topics tend to have same labels. But this may not hold in the real world applications.

Since content on social media platforms, like Twitter, is shorter, sometimes a single word, a hashtag and even some marks may reverse the whole meaning of a sentence. Misinformation mining focuses on inaccurate and wrong information, which requires a fine-grained analysis of social media. Qazvinian et al. propose to solve the detection problem through Natural Language Processing techniques [46]. They include three kinds of features in their system, including Content-based features, Network-based features and Twitter-based features. Content-based features are taken from the text of Twitter. They extract the words to represent the lexical features, and label all words with their Part-Of-Speech (POS) tags to get the POS features. In order to increase the descriptive power, they also incorporate bigrams of content-based

features. Twitter-based features include hashtags and URLs of a tweet. They further extract features from users for each tweet as Network-based features. This is reasonable since the author is a property of his tweets and articles. The tweets, however, are more oftentimes regarded as a feature of its author in fact. When we concentrate on users instead of tweets, misinformation detection problem is transformed to detecting spreaders of misinformation. Since it is common that online reviews influence customers' decisions on purchase, spreading misinformation has been used for advertising a product or damaging the others' reputation. Jindal and Liu propose to solve this problem based on textual features, user behaviors [25] and review opinions [26]. Links between accounts and reviews are further exploited for collective inference. Mukherjee et al. propose to detect spammer groups through analyzing the textual similarity between reviewers. Li et al. links reviewers based on their prior IP address history [36].

1.3.2 Spreader Detection

Identifying spreaders of misinformation is another way to detect misinformation. As discussed in Section 1.2.2, although some users have the intent to spread rumors, many regular users may be persuaded to believe or re-spread misinformation according to their belief levels. Normally, the majority information of a regular user is accurate. Thus identifying spreaders who are with intent to mislead people is the main focus of spreader detection.

Since information on social media sites is spread from node to node, the spreader of a specific piece of misinformation can be trivially found if provenance paths are known in advance. For example, if roots of a provenance path are found, they are likely to be sources of misinformation. Finding such sources is not only useful for preventing misinformation from being further spread, but is also useful for analyzing truthfulness of rumors. However, since provenance paths are often unknown, Gundecha et al. propose to solve this problem through finding a node subset with maximum utility [19]. Here, utility represents the number of nodes that are ultimately affected, which depends on certain information propagation models.

Since content of misinformation spreaders are different from that of regular users, detecting them can also be reduced to a binomial classification problem. The estimator is trained to predict whether a user is a misinformation spreader based on his social media content

and/or his social behaviors. More formally, here we formulated the optimization objective in Equation 1.6:

$$\min_{f} \sum_{i=1}^{m} \mathcal{L}(f(\boldsymbol{x}_i, \boldsymbol{A}), y_i). \tag{1.6}$$

Equation 1.6 requires a data matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ as input, where $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_m$. Here, m is the number of all users instead of tweets. Since the number of users is normally much smaller than that of tweets, focusing on user-level helps to reduce the computational cost. n is the number of features. y denotes the identity label vector of users. A learning method f is needed to predict the identity label of a user based on the content. The second argument of f is the adjacency matrix $A \in \mathbb{R}^{m \times m}$, where $A_{i,j} = 1$ means there is an edge between user i and j. Since social network users connect to their friends and follow people they are interested in, then it may be difficult for intentional misinformation spreaders to connect with others as regular users. Thus the linkage structure is useful for identifying such spreaders. Algorithm 2 depicts the learning framework of misinformation spreader detection.

Two specific kinds of misinformation spreaders have attracted a great deal of attention in recent years, i.e., spammers and bots. Social spammers generally refer to those accounts who tweet spam, which refers to unwanted information such as malicious content, fraudulent reviews, profanity and political astroturf. Since a large portion of spammers are automatic programs instead of real people, as Italian security researchers Andrea Stroppa and Carlo De Micheli pointed, as many as 9% of the monthly active users in Twitter are bots [13]. These bots are controlled by human beings and can send information according to some commands and predefined programs.

Although it is easy to distinguish between regular users and bots from their definitions, it is almost impossible to accurately tell whether a spammer is bot. Fortunately, bots controlled by the same party behave similarly with each other and differently from regular users. Research on exposing bots focuses on retrieving similar anomalies based on content and graph structure.

Existing work builds up discriminative models directly based on user generated content [39]. Useful features include social tags, advertisements, and URLs in social media content. They try to leverage the content through measuring the similarity between different users' text, i.e., plagiarism. This indirect method avoids the model to be entangled with high dimensional data.

Algorithm 2 Misinformation Spreader Detection in a Social Network

Input: The raw content $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, the identity label vector $\boldsymbol{y} \in \mathbb{R}^{m}$,
Output: The optimal misinformation estimator f.
Generate compact representation $\boldsymbol{X}' \in \mathbb{R}^{m \times k}$ based on $\boldsymbol{X}$;
Obtain the label vector $\boldsymbol{y}$
Repeat
Update f to reduce loss $\sum_{i=1}^{m} \mathcal{L}(f(\boldsymbol{x}_i, \boldsymbol{A}), y_i)$;
until Convergence

In order to directly cope with the content information, a more advanced framework has been proposed. A unified framework Online Social Spammer Detection (OSSD) is put forward by Hu et al., which jointly considers content and structural information [23]. Figure 1.5 displays the framework of dimensionality reduction algorithm.

The general aim of all dimension reduction algorithms is to map the raw data onto a low dimensional space. In order to smooth the mapping, OSSD defines two constraints: 1) If two users' social media contents are similar in the raw feature space, their projections on the reduced low dimension should keep the similarity; 2) If two users are connected, their similarity should be high after being mapped into the latent space. The first constraint focuses on the content and is widely accepted by dimension reduction algorithms, while the second one incorporates the social graph structure.

$$\mathcal{R} = \frac{1}{2} \sum_{i,j:\boldsymbol{A}_{ij}=1} \boldsymbol{\pi}(i)\boldsymbol{P}(i,j)||\boldsymbol{H}_{i*} - \boldsymbol{H}_{j*}||. \tag{1.7}$$

The social graph constraint is formulated as in Equation 1.7. Since in the training dataset, links exist not only between regular users, but also between spammers. Spammers pretend to be regular users through following numerous other people. Thus the links that spammers follow regular users are taken out from the adjacency matrix. Links created by regular users are reserved. Since the relationship between spammers may indicate they have similar intentions, such links are also reserved.

Due to the different levels of activeness, some users may follow many people while some others may follow very few. In order to normalize the impact of different links, a probability matrix $\boldsymbol{P} \in \mathbb{R}^{m \times m}$ is introduced. The correlation between two user i and j are denoted as $\boldsymbol{P}_{ij} = \boldsymbol{A}_{ij}/\boldsymbol{d}_i^{out}$, where $\boldsymbol{d}^{out}$ is a vector contains out degree of all nodes. Since users have

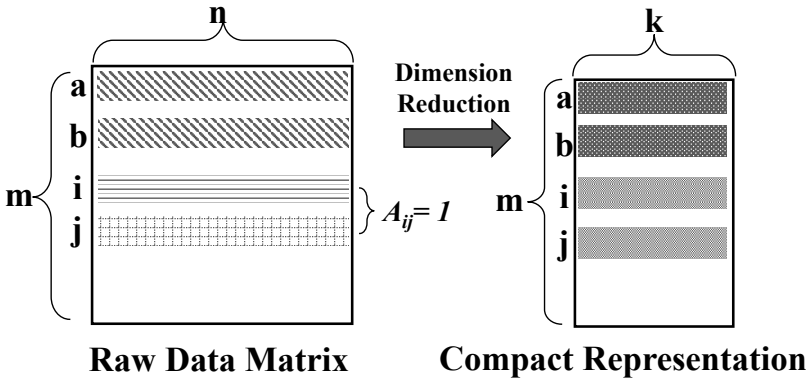


Figure 1.5: The dimension reduction framework of OSSD.

different global impacts, Hu et al. introduces a random walk model to measure it. $\boldsymbol{\pi}$ is the stationary distribution of the transition matrix $\boldsymbol{P}$. Then $\boldsymbol{\pi}_i$ represents the influence of the corresponding node. Due to the explosive increase of users, efficient and incremental update of such models is also a key concern for real applications. Authors also present an algorithm for incremental optimization.

Various variants have been proposed to solve the problem from different perspectives. Hu et al. study how sentiment information can be leveraged to expose spammers from social media [22]. Their experimental results show that significant differences exist between sentiment posed by spammers and regular users. In order to cope with the huge amounts of social media users, Zhao et al. propose a distributed learning system which is able to detect anomalies from graph based on the linkage structure [60]. Nguyen et al. proposed to find the sources of certain diffusions to detect malicious users [43].

A problem of most existing spammer and bot detection algorithms is their dependency on third party data. In order to enable scalable analysis, they often leverage the labels released by either independent third parties or service providers, e.g., Twitter suspends a lot of accounts and such accounts are often considered to be malicious users. In [32], Lee et al. propose a passive way to wait social network anomalies to "label" themselves. Figure 1.6 illustrates their proposed framework. They first create several social honeypots, which are social network accounts with legitimate profiles. Once the honeypot detects suspicious behaviors, such as sending deceptive information and malicious links, such users are labeled as a spammer candidate. Then they extracted

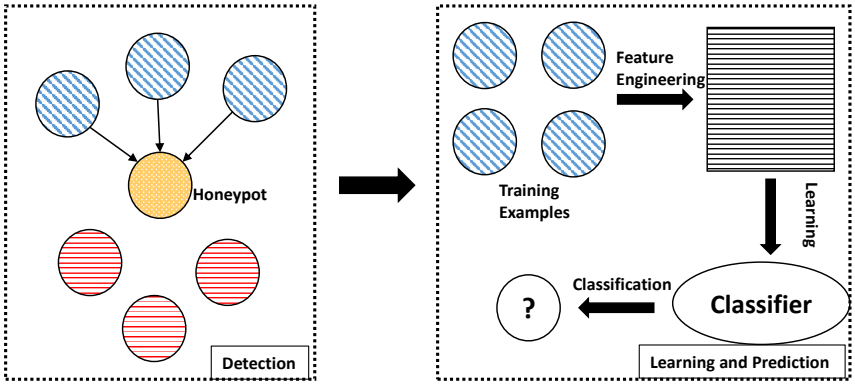


Figure 1.6: Misinformation spreader detection framework based on social honeypot.

features from such spammers and actively detect malicious accounts in social networks in a supervised manner.

In order to obtain labels to build training datasets, various methods have been used. Most approaches fall into three categories, i.e., manual labeling, honeypot, and suspension list. Manual labeling is the most direct way to get labels. Normally tweets about a topic are collected, and human annotators are hired to tell whether each tweet is misinformation. For example, A Twitter dataset consisting of more than 10,400 tweets based on five topics are labeled by two annotators [46]. Since manual labeling is often expensive, it is unaffordable to obtain ground truth on a large scale. An automatic alternative is to set up honeypot accounts [57], where several accounts are created with specific profiles tempting attackers to send misinformation. Spreaders can then be identified accurately. However, since honeypot accounts wait to be attacked, the process usually takes time. In addition, honeypot accounts can only help identify malicious accounts (positive data), where labels of regular accounts are inaccessible. In order to remedy it, some accounts are randomly selected to be used as negative data. The third kind of method is using the suspension list. Since social media websites regularly ban malicious accounts, such suspended user list can be used as a gold standard [23, 54]. The list is usually publicly available and contains accounts on a large scale, which is proper for evaluation. However, it can only identify malicious accounts and thus fails to reveal the real cutoff between malicious and regular information. We summarize

Table 1.1: Features of Different Processes of Obtaining Ground Truth.

Method	Expenses	Time	Distribution
Manual Labeling	High	Medium	Real
Honeypot	Low	Long	Fake
Suspended Users	High	Short	Fake

Table 1.2: Statistics of Datasets.

Country	Suspended	Regular	Ratio
Libya	7,061	87,474	7.47%
Syria	40,109	474,741	8.45%
Yemen	3,239	41,331	7.84%

features of different process of obtaining ground truth in Table 1.1 in terms of expenses of time, cost and truthfulness of distribution.

In order to economically obtain datasets which keep in line with real distribution, we introduce a method which is based on given topics. We first set a topic and compile a list of related keywords, and extract all tweets containing the keywords [30]. In particular, we collected Twitter accounts who post with Arab Spring hashtags from February 3^{rd}, 2011 to February, 21^{st}, 2013. The employed hashtags are **#gaddafi**, **#benghazi**, **#brega**, **#misrata**, **#nalut**, **#nafusa** and **#rhaibat**. Several geographic bounding boxes are used to capture tweets from Libya[5], Syria[6] and Yemen[7]. The statistics of the three datasets are illustrated in Table 1.2.

As shown in the Table 1.2, real misinformation datasets are often skewed. Such skewness has been overlooked by existing misinformation detection algorithms. Since rare category analysis [21] has been proven to be effective in solving such problems, it will be interesting to apply related methods to help expose misinformation. An alternative way is to apply ranking algorithms, which has been used for fraud detection for mobile applications [61] and system faults detection [15], where the dataset is also skewed.

[5]Southwest Lat/Lng: 23.4/10.0; Northeast Lat/Lng: 33.0, 25.0
[6]Southwest Lat/Lng: 32.8/35.9; Northeast Lat/Lng: 37.3, 42.3
[7]Southwest Lat/Lng: 12.9/42.9; Northeast Lat/Lng: 19.0, 52.2

1.4 MISINFORMATION INTERVENTION

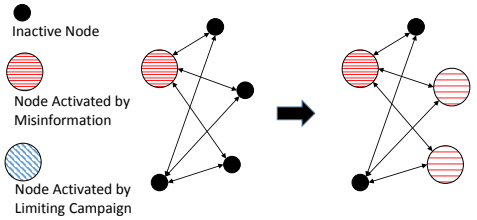
As mentioned in Section 1.1, rumors about Ebola kept spreading widely even after some official intervention took place. Effectively reducing negative effect of misinformation is of great interests for governments and social network service providers. In this section, we introduce two measures toward combating misinformation. The first part examines techniques to detect and prevent misinformation from spreading in an early stage. The second one introduces how a competing campaign could be employed to fight against misinformation.

1.4.1 Malicious Account Detection in an Early Stage

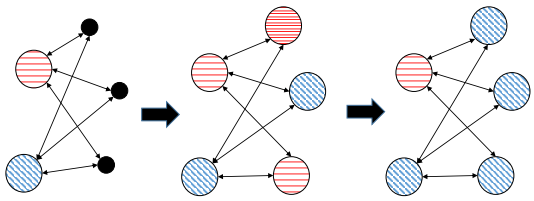
Misinformation spreads quickly on social networks. Although spreaders are blocked once they are found to forward or send rumors, but misinformation is already spread by then. It will be favorable if such accounts can be found before they start spamming. In order to mislead social network users in a large scale, malicious programmers create a plethora of accounts which can send information according to certain policies. Profiles of automatic generated accounts may look similar, since they may be created using the same template. Investigation toward spam social profiles reveals that duplication of profiles widely exists between malicious accounts [57]. Instead of using a template, malicious programmers tend to directly copy information from others, some URLs are thus shared by many bots. Through studying geographic distribution of malicious accounts, they also found that such accounts gather in some specific states. More behavioral features of profiles are taken into consideration by Egele et al. [12].

Profile features of malicious accounts are extremely sparse. Many items of a spam profile are empty. Since at least a name has to be available for each account, Lee and Kim propose to reveal the patterns hidden behind malicious accounts [34], so as to filtering them in an early age. A problem of modeling account names is the sparseness, since they are even sparser than profiles. To cope with the sparsity issue, they cluster all account names and then analyze on the cluster level. Agglomerative hierarchical clustering is adopted, where the likelihood of two names being generated by an identical Markov chain is used for measuring distance, and characters are used as features. After obtaining clusters of similar account names, a supervised method is adopted to classify whether a name cluster is a group of malicious

accounts. Lee and Kim leverage several cluster statistics, such as the length distribution and average edit distance within a group, as features for their Support Vector Machines. Account names have also been quantitatively examined [59]. More features including behavioral ones are further incorporated in such algorithms [5, 33, 51, 62].



(a) An example of general information diffusion models in social networks, where only misinformation is spread to infect inactive nodes.



(b) An example of intervention models of misinformation in social networks, where a campaign is used to limit spread of misinformation.

Figure 1.7: Illustrative examples of information diffusion and misinformation intervention

1.4.2 Combating Rumors with Facts

The perfect plan of limiting misinformation is to identify and filter malicious accounts before they spread misinformation. However, a potential of such algorithms in reality is the cost of the false positive rate. Since social networks cannot emphasize more on involvement of users, mistakenly blocking regular accounts is unacceptable. A more practical method will be controlling the spread and limiting the influence of misinformation after it's spread.

An effective method of combating rumors is to spread truth. By

sending the accurate information, people in a social network can be either saved (if they have been "infected" by the rumor) or be immunized. The key issue of such methods is to choose an optimal subset of users to start it. Budak et al. propose a new diffusion algorithm to model the process. They introduce the Multi-Campaign Independence Cascade Model (MCICM) where two campaigns may coexist [6]. Concretely, as mentioned in Section 1.2.1, most information diffusion models consider the situation where only one topic is being spread. In MCICM, besides misinformation campaign, another "good" campaign tries to reduce the effect of misinformation, which is illustrated in Figure 1.7. As shown in Figure 1.7a, general information diffusion models aim to predict the optimal seed set which can result in the largest scale of diffusion, where only one campaign is available. But for misinformation limitation, there are two campaigns at the same time. The aim becomes identifying optimal set of individuals to be convinced to adopt the competing campaign, which minimizes the effect of the bad campaign eventually. An important assumption needs noting is that once a node has been cured by the positive campaign or being immunized, they can no longer be infected again.

Similar efforts have been paid by Nguyen et al. [44]. They aim to find the smallest set of highly influential nodes whose decontamination with good information helps to contain the viral spread of misinformation. Different from MCICM, which only considers the Independent Cascade Model, they also incorporate Linear Threshold model. Several interesting patterns have been observed in the investigation. When the good campaign budget is limited, meaning that we can only select small number of users, influential nodes in larger communities are more effective in competing misinformation. On the other hand, when we can choose more nodes to start with and misinformation has been widespread, choosing influential nodes from small communities can efficiently immunize the whole community. The bridge influencers, which connect between different communities and are expected to have a large impact factor, are not useful for combating misinformation. It is because communities are often with high intra-group propagation probabilities, such bridges are of slim chance to influence them.

Simulation is often used to quantitatively measure rumor intervention methods. Rumors are first assumed to be spread in social networks, and the intervention algorithm is then used to select optimal node set to spread truth. Thus the number of nodes which are inactive for the rumor indicates the effectiveness of intervention. Benchmark datasets

include both social networks from online websites[8], synthetic network and academic collaboration graph[9].

Social media data is useful for decision-making when some events happen, such as terrorist attacks and disasters. In order to avoid being influenced by false rumors during events, several systems have been developed. The crowdsourcing system[10] depends on active participants. In addition, some Twitter analytic tools are also proposed to help information management during crisis. Kumar et al. try to detect relevant and credible information sources during disasters [31], locations and topics of discussion are extracted as features to analyze their affinity to certain events. Through actively selecting optimal information sources, more and better information can be generated using social media. A problem of their system is that location information is missing for some users. To cope with missing data, they further propose an approach to estimate user location by jointly analyzing prior posts and user profiles [42].

1.5 EVALUATION

In this section, we discuss the evaluation of misinformation detection and intervention, including available benchmark datasets and evaluation metrics.

1.5.1 Datasets

Although there is no agreed way of obtaining labels or benchmark datasets, several datasets are available for evaluating misinformation detection and intervention algorithms, which are obtained from social media sites such as Twitter[11] and Facebook[12]. We list several representative ones below.

Rumor Dataset: Some fact-checking websites have collections of rumors. Crowdsourcing and expert ratings are used for judging the truthfulness of a rumor. In Verily[13], stories and pictures of questionable veracity are posted, and registered users then discuss and judge whether

[8] http://socialcomputing.asu.edu/
[9] http://research.microsoft.com/en-us/people/weic/projects.aspx
[10] https://veri.ly/
[11] http://twitter.com/
[12] http://www.facebook.com/
[13] https://www.veri.ly/

it is true or false. In PolitiFact[14], truthfulness of political statements is evaluated, where a "Truth-O-Meter" is assigned to each statement and the rating ranges from "True" to "Pants on Fire". Data on both websites is publicly available. However, connections between these rumors and social media contents are not directly available. Qazvinian et al. propose to employ human annotators to label tweets manually [46]. In Weibo[15], the content management team[16] will regularly attach labels of false rumors, and such labels, posts and the corresponding diffusion information are available.

Spreader Dataset: As shown in Table 1.1, three methods are used for obtaining misinformation spreader accounts on social media sites. Lee et al. set up several honeypot accounts and captured 23,869 content polluters from Twitter. The corresponding user posts and their social links are available[17]. Authors found 23% of detected polluters are also suspended by Twitter after less than a month, which indicates both methods are effective in finding spammers. Suspension of a Twitter account can be found by using Twitter's API [30].

Misinformation Diffusion Dataset: Simulation is often used for evaluating effectiveness of misinformation intervention. Misinformation and the corresponding fact are simultaneously simulated to spread on social networks, so only social links are needed for evaluation. Nguyen et al. adopted network structures of the author collaboration graph and Facebook users [44], and Budak et al. propose to use regional user graph on Facebook. Besides social graphs, networks of customers, videos, web pages and Wikipedia articles[18] could be used for simulation.

1.5.2 Evaluation Metrics

Since misinformation and spreader detection problems are often modeled as binary classification tasks, measures used by supervised learning algorithms can be used as evaluation metrics.

Accuracy: Accuracy measures the similarity between prediction results and real labels. A common practice is to describe accuracy by considering how many errors are made. A widely used measure is Mean

[14]http://www.politifact.com/
[15]http://weibo.com/
[16]http://service.account.weibo.com/
[17]http://infolab.tamu.edu/data/
[18]http://snap.stanford.edu/data/index.html;http://socialcomputing.asu.edu/pages/datasets

Absolute Error (MAE), which is defined as:

$$MAE = \frac{1}{|U|}\sum_{i \in U}|p_i - l_i|, \tag{1.8}$$

where U is the user set and p_i is the prediction result and l_i is the true label. A smaller MAE value represents better performance, meaning that less errors are made.

Precision, Recall and F-measure: Accuracy measures number of prediction mistakes regardless of positive or negative examples. Since capturing misinformation is more important, precision is often used, which is defined as follows:

$$Precision = \frac{\#TP}{\#TP + \#FP}, \tag{1.9}$$

where $\#TP$ means the number of true positives, representing the number of correctly identified spammers/misinformation; While $\#FP$ means the number of false positives, representing the number of mistakenly identified spammers/misinformation. Since misinformation datasets are often skewed, a high precision can be easily achieved by making less positive predictions. In order to avoid this, recall is used to measure sensitivity:

$$Recall = \frac{\#TP}{\#TP + \#FN}, \tag{1.10}$$

where $\#FN$ the number of means false negatives, representing the number of unidentified spammers/misinformation. F-measure is used to combine both precision and recall, which is defined as follows:

$$F_\beta = (1 + \beta^2)\frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}. \tag{1.11}$$

where β controls the importance of recall. $\beta = 1$ is often normally used, where precision and recall are equally weighted. If β is 2, recall weights twice higher than precision; And if β is 0.5, precision weights higher than recall.

Outcome of Simulation: Misinformation intervention is often modeled as a information diffusion problem, where a subset of nodes are selected to send factual claims. Thus, the final number of nodes immune from rumors can be viewed as effectiveness of the method.

1.6 CONCLUSION AND FUTURE WORK

As the world becomes increasingly connected, social media platforms have made everyone a news source. Misinformation gets issued and repeated more quickly and widely than ever due to the connectivity of social networks, which impact the real world. A false tweet has a negative impact on community or a family, triggers financial panic and even strains diplomatic relations. To cope with the spread of misinformation, we must first understand it. In this chapter, we discuss about the distinct aspects of misinformation diffusion in social media, and elaborate existing work of identifying misinformation, its spreaders and intervention methods.

This chapter has discussed some essential issues of misinformation. Benchmark datasets and evaluation metrics are also introduced for misinformation identification and intervention. As remedies for weaknesses of existing approaches, we propose a method to obtain ground truth for spreader detection based on the suspension list, where data distribution is in line with real world. Since mining misinformation in a social network is an emergent field of study, we also list a number of interesting potential problems for future exploration:

How to seek the provenance of misinformation in social media? The information spread in social media follows a path, i.e., from one user to the other and from one site to other social media sites. For example, a terrorist attack may first be reported on a social media site, then reported by news media and is finally tweeted by more users. Such linking nature enables information to be traceable. Though Centrality measures have been studied to find the spreading trace of misinformation within a social media site [19], a global analysis based on different web sites can further facilitate recovering the trace and seeking the real provenance of misinformation.

How to predict the potential influence of misinformation in social media? The verification practices of online information is time-consuming, which makes it almost impossible for service providers and newsrooms to catch up with with the speed of social media. Since budget of competing misinformation is often limited, efforts should be paid on the most destructive rumors. An effective way to estimate potential impact of misinformation will be very useful to control

the negative influence.

How to find the vulnerable individuals from social network users? Social network platform consists of various people. Misinformation has different effects on people who are of different levels of vulnerability. Such vulnerability provides possibility for us to actively avoid weak individuals from being infected. To estimate the level of vulnerability in terms of user profile, misinformation topics and the network structure will be very challenging but useful.

How to exploit data from multiple sources to facilitate misinformation analysis? There are more and more social media platforms. People are usually simultaneously involved in different social network websites. The generation and diffusion may separately start from different platforms, and information exchange also takes place between different websites. It provides a complete picture of misinformation to integrate data from different sources.

Bibliography

[1] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. Spread of (mis) information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

[2] Divyakant Agrawal, Ceren Budak, and Amr El Abbadi. Information diffusion in social networks: observing and affecting what society cares about. In *Proceedings of CIKM*, pages 2609–2610, 2011.

[3] Floyd H Allport and Milton Lepkin. Wartime rumors of waste and special privilege: why some people believe them. *The Journal of Abnormal and Social Psychology*, 40(1):3, 1945.

[4] Gordon W Allport and Leo Postman. The psychology of rumor. 1947.

[5] Fabricio Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, Chao Zhang, and Keith Ross. Identifying video spammers in online social networks. In *International Workshop on Adversarial Information Retrieval on the Web*, pages 45–52. ACM.

[6] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674. ACM, 2011.

[7] Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

[8] Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013.

[9] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, pages 1029–1038. ACM, 2010.

[10] Nicholas DiFonzo and Prashant Bordia. Rumor and prediction: Making sense in the stock market. *Organizational Behavior and Human Decision Processes*, 71(3):329–353, 1997.

[11] Nicholas DiFonzo, Prashant Bordia, and Ralph L Rosnow. Reining in rumors. *Organizational Dynamics*, 23(1):47–62, 1994.

[12] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Compa: Detecting compromised accounts on social networks. In *NDSS*, 2013.

[13] J Elder. Inside a twitter robot factory. *Wall Street Journal*, 821400, 2013.

[14] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 35–47. ACM, 2010.

[15] Yong Ge, Guofei Jiang, Min Ding, and Hui Xiong. Ranking metric anomaly in invariant networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):8, 2014.

[16] Malcolm Gladwell. *The tipping point: How little things can make a big difference*. Little, Brown, 2006.

[17] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 211–220. IEEE, 2011.

[18] Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.

[19] Pritam Gundecha, Zhuo Feng, and Huan Liu. Seeking provenance of information using social media. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1691–1696. ACM, 2013.

[20] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. WWW '13 Companion, pages 729–736, 2013.

[21] Jingrui He. *Rare category analysis*. PhD thesis, Carnegie Mellon University Pittsburgh, PA, 2010.

[22] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Social spammer detection with sentiment information. In *Proceedings of the IEEE International Conference on Data Mining*, ICDM'14. IEEE, 2014.

[23] Xia Hu, Jiliang Tang, and Huan Liu. Online social spammer detection. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[24] Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[25] Nitin Jindal and Bing Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM, 2007.

[26] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM, 2008.

[27] Natascha Karlova and Karen Fisher. Plz rt: A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*, 18(1), 2013.

[28] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[29] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 115, pages 700–721, 1927.

[30] Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter data analytics*. Springer, 2014.

[31] Shamanth Kumar, Fred Morstatter, Reza Zafarani, and Huan Liu. Whom should i follow?: identifying relevant users during crises. In *Hypertext*, pages 139–147. ACM, 2013.

[32] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[33] Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*. Citeseer, 2011.

[34] Sangho Lee and Jong Kim. Early filtering of ephemeral malicious accounts on twitter. *Computer Communications*, 54:48–57, 2014.

[35] Todd Leopold. In today's warp-speed world, online missteps spread faster than ever. *CNN.com*, 2012.

[36] Huayi Li, Zhiyuan Chen, Bing Liu, Xiaokai Wei, and Jidong Shao. Spotting fake reviews via collective positive-unlabeled learning. In *ICDM*, pages 899–904. IEEE, 2014.

[37] Bruce R Lindsay. Social media and disasters: Current uses, future options, and policy considerations. 2011.

[38] Victor Luckerson. Fear, misinformation, and social media complicate ebola fight. *Tech report*, 2014.

[39] Benjamin Markines, Ciro Cattuto, and Filippo Menczer. Social spam detection. In *International Workshop on Adversarial Information Retrieval on the Web*, pages 41–48. ACM, 2009.

[40] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[41] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrak, and Andreas Krause. Lazier than lazy greedy. *arXiv preprint arXiv:1409.7938*, 2014.

[42] Fred Morstatter, Nichola Lubold, Heather Pon-Barry, Jurgen Pfeffer, and Huan Liu. Finding eyewitness tweets during crises. *arXiv preprint arXiv:1403.1773*, 2014.

[43] Dung T Nguyen, Nam P Nguyen, and My T Thai. Sources of misinformation in online social networks: who to suspect. In *Military Communications Conference, MILCOM*, pages 1–6, 2012.

[44] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 213–222. ACM, 2012.

[45] Onook Oh, Kyounghee Hazel Kwon, and H Raghav Rao. An exploration of social media in extreme events: Rumor theory and twitter during the haiti earthquake 2010. In *ICIS*, page 231, 2010.

[46] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*, pages 1589–1599, 2011.

[47] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonffalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the WWW'11, Companion Volume*, pages 249–252. ACM, 2011.

[48] Christine Roberts. Social media users spreading false information about sandy hook massacre could face charges, say police. *New York Daily News*, 2012.

[49] Scott Shane and Ben Hubbard. Isis displaying a deft command of varied media. *New York Times*, 31, 2014.

[50] Long Song, Raymond YK Lau, and Chunxiao Yin. Discriminative topic mining for social spam detection. 2014.

[51] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9. ACM, 2010.

[52] Jeannette Sutton, Leysia Palen, and Irina Shklovski. Backchannels on the front lines: Emergent uses of social media in the 2007 southern california wildfires. In *Proceedings of the 5th International ISCRAM Conference*, pages 624–632. Washington, DC, 2008.

[53] David Talbot. Preventing misinformation from spreading through social media. *MIT Technology Review*, 2013.

[54] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 243–258. ACM, 2011.

[55] Farida Vis. Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 uk riots. *Digital Journalism*, 1(1):27–47, 2013.

[56] Farida Vis. To tackle the spread of misinformation online we must first understand it. *Guardian Comment Network*, 2014.

[57] Steve Webb, James Caverlee, and Calton Pu. Social honeypots: Making friends with a spammer near you. In *CEAS*, 2008.

[58] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[59] Reza Zafarani and Huan Liu. 10 bits of surprise: Detecting malicious users with minimum information. CIKM'15. ACM.

[60] Yao Zhao, Yinglian Xie, Fang Yu, Qifa Ke, Yuan Yu, Yan Chen, and Eliot Gillum. Botgraph: Large scale spamming botnet detection. In *NSDI*, volume 9, pages 321–334, 2009.

[61] Hengshu Zhu, Hui Xiong, Yong Ge, and Enhong Chen. Discovery of ranking fraud for mobile apps. *Knowledge and Data Engineering, IEEE Transactions on*, 27(1):74–87, 2015.

[62] Yin Zhu, Xiao Wang, Erheng Zhong, Nathan Nan Liu, He Li, and Qiang Yang. Discovering spammers in social networks. In *AAAI*, 2012.