News Credibility Evaluation on Microblog with a Hierarchical Propagation Model

Zhiwei Jin[1,2], Juan Cao[1], Yu-Gang Jiang[3], Yongdong Zhang[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]School of Computer Science, Fudan University, Shanghai, China
{jinzhiwei, caojuan, zhyd}@ict.ac.cn, ygj@fudan.edu.cn

***Abstract*—Benefiting from its openness, collaboration and real-time features, Microblog has become one of the most important news communication media in modern society. However, it is also filled with fake news. Without verification, such information could spread promptly through social network and result in serious consequences. To evaluate news credibility on Microblog, we propose a hierarchical propagation model. We detect sub-events within a news event to describe its detailed aspects. Thus, for a news event, a three-layer credibility network consisting of event, sub-events and messages can represent it from different scale and reveal vital information for credibility evaluation. After linking these entities with their semantic and social associations, the credibility value of each entity is propagated on this network to achieve the final evaluation result. By formulating this propagation process as a graph optimization problem, we provide a globally optimal solution with an iterative algorithm. Experiments conducted on two real-world datasets show that the proposed model boosts the accuracy by more than 6% and the F-score by more than 16% over a baseline method.**

Keywords*—Social media credibility; Microblog; news credibility; rumor detection

I. INTRODUCTION

Recent years have seen the rapid growth of many online Microblog services, such as Twitter and the Chinese Sina Weibo. With millions of users acting as sensors, Microblog has become a content-sharing social network and a real-time news source. Traditional news media use Microblog to release instant news; government departments utilize it to publish official announcements; normal users post immediate events around them on Microblog. It is undoubted that Microblog has become one of the most popular platforms for news publishing, spreading and discussing.

Including fresh news and related social opinions, information on Microblog is valuable for opinion mining and decision making. However, before any further analysis, we need to determine how trustworthy are these contents. In fact, at the absence of supervision and self-discipline, many malicious entities use Microblog to spread rumors or fake news. Especially in the case of emergencies, due to the lack or delay of authoritative reports, numerous rumors are aroused immediately and spread through the whole network, which may cause serious damages. Taking the recent emergency event "*Malaysia Airlines Flight MH370 Lost Contact*" as an example, in the first two days, 92 different rumors were spread widely on Sina Weibo[1] , the largest Microblog service in China.

According to their content topics, news on Microblog can be divided into two categories: topic-independent news and topic-related news. Topic-independent news may discuss irrelevant topics while topic-related news covers the same specific topic (for example, the news concentrated on the topic of "*Flight MH370 Lost Contact*"). In this paper, we will experiment with both categories.

Many efforts have been devoted to identify fake news on Microblog. In Sina Weibo, users are encouraged to report suspicious news and a committee composed of reputable users will judge the case.[2] Twitter also tries to enhance trust on their site.[3] However, these manual approaches are inefficient on dealing with ever-increasing contents. Although it is a highly challenging task, automatically verifying the news credibility and filtering the unreliable information have become increasingly crucial.

Recently, several methods have been proposed to handle this challenge. Generally, these methods can be classified into two categories: classification-based approach and credibility propagation approach. The classification-based approach uses supervised learning algorithms to identify an event's truthfulness. The propagation approach, like the one proposed by Gupta et al. [5], has been proved to be quite effective and outperforms the classification-based approach. With initial credibility values learned from a classifier, this method construct a network to propagate credibility values among users, tweets and events. However, it is not easy to assess an event's credibility just from these three aspects. On one hand, they judge an event's credibility with respect to its publisher under the assumption that credible users provide credible tweets with a high probability. In fact, most of users spread fake news on Microblog are credible users, because they may not be able to verify the news and may just spread a fake news unintentionally. The statistic data[4] shows that the top 10 fake

[1] (In Chinese) http://beijing.qianlong.com/3825/2014/03/14/7524@9472229.htm
[2] http://service.account.weibo.com/
[3] https://blog.twitter.com/2010/trust-and-safety
[4] http://www.cssn.cn/zx/201401/t20140106_936433.shtml

1550-4786/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDM.2014.91

news of 2013 released in China were mostly published by authorized Microblog users. On the other hand, an event as a whole contains both truthful and fake information, without deeper analysis of its components, it is hard to get a convincing evaluation. Thus, it is better to minimize users' influence and give more emphasis on an event's deeper semantic relations.

Realizing the limitation of the previous methods, we propose a hierarchical credibility propagation model to evaluate the credibility of news on Microblog. We construct a credibility propagation network for one news with three layers: message layer, sub-event layer and event layer. They are all content-based, and have direct relations with news credibility. We initially introduce the sub-event layer to capture deeper sematic information within an event. Sub-events are various point of views of an event. By clustering messages together, they can represent major parts of an event and minimize the impact of noisy information. As an example, in Figure 1 this idea is illustrated with respect to a fake news in our dataset: "*a kind girl fed a homeless old man*".

In this hierarchical network, on the small scale, a news event contains all related messages; on the large scale, a news event contains several sub-events. Therefore, by evaluating the news credibility at different levels and fusing these scores through the propagation network, more reliable results may be attained.

The main contributions of this paper are as follows:

1) A three-layer hierarchical credibility network is proposed to evaluate the credibility of news on Microblog. The hierarchical structure of message to sub-event, and sub-event to event can reasonably model their relations and the process of credibility propagation. With a sub-event layer, deeper semantic information can be revealed for an event.

2) By formulating credibility propagation on this network as a graph optimization problem, we can provide the globally optimal solution with an iterative algorithm.

3) To validate the effectiveness of the proposed model, two datasets on Microblog are collected: one with random fake news in a year and truthful news at the same time; another with both fake and truthful news related to the same topic: "MH370 lost contacts". Experiments on both datasets show that the proposed model can achieve significant improvements in terms of both accuracy and F-score compared with baseline methods.

The paper is organized as follows. In the next section, we provide a formal definition of the problem. After that, we introduce the proposed hierarchical network, describe its structure and define its links and initialization method. The iterative optimization algorithm is presented in Section IV. Datasets, performance evaluation measures and experimental results are presented in Section V. We discuss related work in Section VI, and, finally, conclude in Section VII.

II. Problem Definition

Given a specific news event and related messages from Microblog, the problem addressed in this paper is how to evaluate the news credibility and identify it as trustworthy or not. In this section, we give formal definitions of entities involved in this problem.

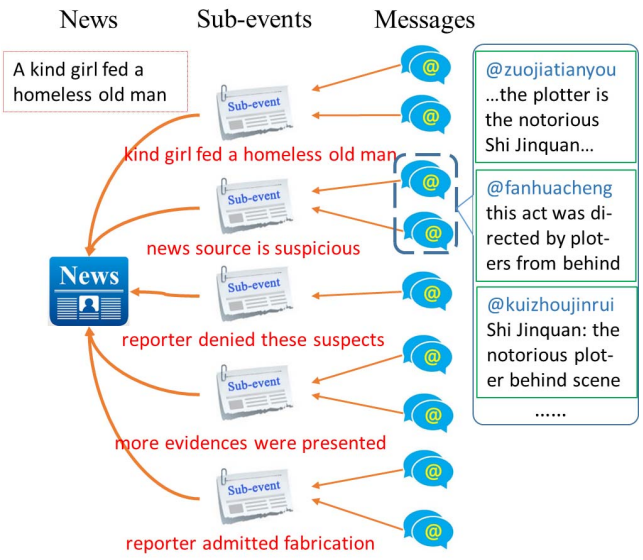

Fig. 1. Factors influencing news' credibility, taking the fake news "A kind girl fed a homeless old man" as an example. The sub-event titles and message examples are translated into English.

Definition (Event) *An event is something that occurs in a certain place at a certain time. On Microblog, an event can be considered as a set of messages containing certain keywords during a certain period of time.*

This definition comes from the version commonly used for the Topic Detection and Tracking (TDT) event detection task over broadcast news [15]. And an event is represented in the context of Microblog. For example, consider the news event "*A kind girl fed a homeless old man on a street in Shenzhen*". We obtain its keywords as "kind girl, Shenzhen", its time span as "2013-03-25 ~ 2013-03-28". Thus messages containing these keywords during this time span represent this event together.

Definition (Sub-event) *A sub-event is a subpart of an event which covers a small topic of this event.*

Among all the messages of a specific news event, there are reports from different views, controversial opinions or extended stories. Sub-events can summarize all these aspects of an event. With this definition, we try to identify sub-events in an event. In this paper, sub-events are detected with single-pass incremental clustering. The details will be discussed in Section IV.

Definition (Message) *In the context of Microblog, a message is a piece of content posted by a user along with social context.*

Compared with traditional news reports, messages on Microblog have some unique features. Figure 2 gives an instance of a message and its publisher on Sina Weibo. Messages on Microblog include three types of features: content features (text content, #hashtag topic, URL links, etc.),

Fig.2. Instance of a piece of message and its publisher on Sina Weibo.

social features (post time, number of forwards, number of comments, etc.) and user features (number of followers, number of followees, etc.). Here user features are integrated into a message under the assumption that the property of a message is partially determined by its publisher.

Credibility of an entity is often treated as a numeric value and a threshold is used to determine it as credible or not. In this paper we define credibility value $\in [-1,1]$ and use 0 as a fixed threshold.

For each entity concerned (event, sub-event and message), its credibility is defined as follows: the credibility of an event is the expected credibility of sub-events that belong to this event; the credibility of a sub-event is the expected credibility of messages that belong to this sub-event; the credibility of a message is a result from a credibility classifier trained with message features.

III. Hierarchical Credibility Network

Given a news event along with its related Microblog messages, sub-events are generated by clustering. Then we built a hierarchical credibility network of three layers: message layer, sub-event layer and event layer. Following that, the semantic and structure features are exploited to adjust the weights of links in the network. Finally, all entities are initialized with credibility values using classification results. We will discuss how to propagate credibility values on this network in the next section.

A. Network Structure

For a news event, Hierarchical Credibility Network (HCNet) is composed of three layers: message layer, sub-event layer and event layer, and the links among them.

There are three types of entities which have been defined in Section II, and four types of links: links from message to sub-event ($g(m_i,s_j)$), links from sub-event to event ($p(s_i,e_j)$), links among all messages ($f(m_i,m_j)$) and links among all sub-events ($h(s_i,s_j)$) (Fig. 3). Links' weights are denoted as functions of the two linked nodes. The edges are created as follows: message m is linked to a sub-event s if it is clustered

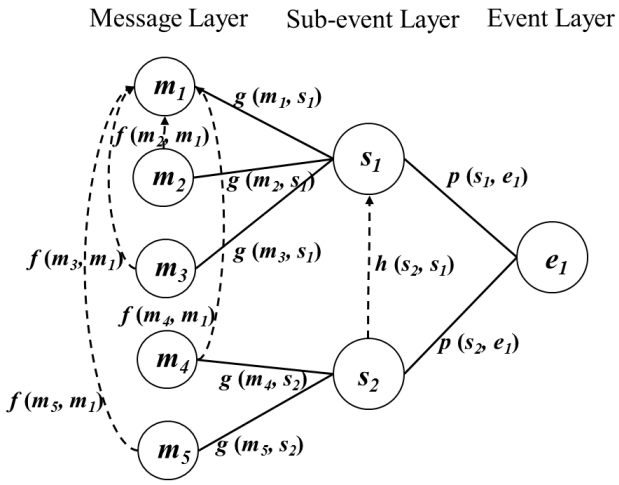

Fig. 3. A three-layer hierarchical credibility network. Some links are omitted for clarity.

into that sub-event; sub-event s is linked to corresponding event e; all the messages are linked to each other, so are the sub-events.

This network structure has several advantages for news credibility evaluation:

- The network hierarchically models a news event from three different scales: the large scale (event itself), the small scale (messages) and the intermediate scale (sub-events). Three levels of representation gives thorough descriptions for an event.
- Four types of network links are fusing results of all entities' to produce a reliable evaluation for this event. The intra-level links reflect the relations among entities of a same type. The inter-level links reflect the impacts from level to level. All these links are properly weighted to indicate entities' relations.
- Sub-event layer is initially introduced to capture deeper sematic information within an event. Sub-events are various point of views for an event. By clustering messages together, they can represent major parts of an event and minimize the impact of noisy information.
- The "one network for one event" design is concise and efficient to perform further propagation. It is also free to cold start: history information of this event or other events is not required for this network defining, compared with [5].

B. Sub-event Detection

Sub-events reveal deeper semantic information of an event and this layer plays an important role in this hierarchical network. We cast the problem of detecting sub-events and their associated messages of a news event on Microblog as a clustering problem. Ideally, each cluster corresponds to one sub-event and contains all messages associated with this sub-event. Although there are various methods to deal with this

problem, the single-pass incremental clustering algorithm is chosen for sub-event detection in this paper.

Single-pass incremental clustering has been shown to be an effective technique for event detection in textual news documents [14]. For the task of detecting sub-events from a set of Microblog messages, this algorithm sequentially processes the input messages, one at a time, and grows clusters incrementally. A new message is absorbed by the most similar existing cluster if their similarity score is larger than a pre-defined threshold (μ); otherwise the message is treated as a new cluster seed. By adjusting the threshold, one can obtain clusters at different levels of granularity. In this paper, we represent each message with its content term frequency vector and apply the clustering algorithm on these feature vectors. The semantic similarity between two messages is computed via their word vectors.

Some advantages of this clustering algorithm include: 1) it is efficient and scalable on large scale online social media data; 2) it doesn't require *a priori* knowledge of the number of clusters; 3) it has only one parameter (the similarity threshold) which can be easily tuned. In this paper, we represent a message with its text term frequency vector. With these lexical feature expression, the clustering algorithm are applied to extract sub-events.

C. Link Definition

In Fig. 3, links between two entities are denoted as functions. We give formal definitions of these four links and compute them respectively.

1) *Message to Message Link:*

Definition (Message to Message Link) *The link from a message m_i to another message m_j can be defined as a function $f(m_i, m_j) \in [0,1]$. It describes the degree of influence from m_i to m_j. This influence is symmetric, i.e. $f(m_i, m_j) = f(m_j, m_i)$.*

Before computing these links' weights, an assumption is given: similar messages are likely to have similar credibility values. Under this assumption, the more similar two messages are the bigger their link's weight is.

Many NLP techniques can be used to compute the similarity between two Microblog messages ([12][16]). As a message on Microblog is a short text limited to no more than 140 characters, Jaccard coefficient between the unigrams of a pair of messages m_i *and* m_j is used to measure their similarity. Considering a message's sentiment score and avoid the effect of long text, message-message link is refined as follows.

$$f(m_i, m_j) = \begin{cases} 0, \text{ if } Senti(m_i) \times Senti(m_j) < 0 \\ \dfrac{|m_i \cap m_j|}{\min(|m_i|, |m_j|)}, \text{ otherwise} \end{cases} \quad (4.1)$$

In this definition, messages with different sentiment polarities (one is positive while the other is negative) have no influence to each other, and the impact of long message is deduced.

2) *Sub-event to Sub-event Link*

Definition (Sub-event to Sub-event Link) *The link from a sub-event s_i to another sub-event s_j can be defined as a function $h(s_i, s_j) \in [0,1]$. It describes the degree of influence from s_i to s_j. This influence is also symmetric.*

We also give an assumption for computing this link's weight: similar sub-events are likely to have similar credibility values. Under this assumption, the more similar two sub-events are, the bigger their link's weight is. However, unlike messages, it is not straightforward to compare two clusters' similarity. Inspired by clustering algorithms, centroid of a cluster is used to represent it. The centroid for a cluster is defined as the average tf score (term frequency) per term of its messages. Thus these centroid word vectors (C_i, C_j) are used to compute link weight:

$$h(s_i, s_j) = \frac{C_i \bullet C_j}{\| C_i \| \| C_j \|} = \frac{\sum_{k=1}^{N} w_{k,i} w_{k,j}}{\sqrt{\sum_{k=1}^{N} w_{k,i}^2} \sqrt{\sum_{k=1}^{N} w_{k,j}^2}} \quad (4.2)$$

Here, N is the vocabulary size and $w_{k,i}$ denotes the weight of word k of cluster i.

3) *Message to Sub-event Link*

Though sub-event detection clusters related messages together to form a sub-event, different messages could have different influence on this sub-event to determine its credibility. On the other hand a sub-event can also influence the messages in it.

Definition (Message to Sub-event Link) *The link from a message m_i to its corresponding sub-event s_j can be defined as a function $g(m_i, s_j) \in [0,1]$. Conversely, the link from a sub-event s_j to one of its message m_i can also be defined as $g(s_j, m_i) \in [0,1]$. These links describe the influence between them and they are symmetric, i.e. $g(m_i, s_j) = g(s_j, m_i)$.*

Because message is a part of sub-event, it is not appropriate to regard this influence from a similarity angle. We give an assumption: if a message contributes more to a sub-event, it should have more influence on this sub-event. This contribution can be quantized with respect to relevance and importance: if a message is much relevant to the center topic of sub-event, it should have large link weight; if a message is very important (e.g. it raises a lot of propagation on social network), it should have large link weight. The similarity between the message's word vector (W_{m_i}) and the sub-event's centroid word vector (C_{s_j}) can be served as relevance value. And a message's social propagation value (forward times plus comment times, $prop(m_i)$) divided by the largest propagation value in a sub-event is served as importance value. And a parameter $\lambda \in [0,1]$ is set to balance the two factors (Eq. (4.3)).

$$g(m_i, s_j) = \lambda \frac{W_{m_i} \bullet C_{s_j}}{\| W_{m_i} \| \| C_{s_j} \|} + (1-\lambda) \frac{prop(m_i)}{\max\{prop(m), m \in s_j\}} \quad (4.3)$$

The link from a sub-event s_i to a message m_j can be computed by using the symmetric property by computing $g(s_i, m_j) = g(m_j, s_i)$.

4) *Sub-event to Event Link*

Definition (Sub-event to Event Link) *The link from a sub-event s_i to its corresponding event e_j can be defined as a function $p(s_i, e_j) \in [0,1]$. And the link from an event e_j to one of its sub-event s_i can be defined as $p(e_j, s_i) \in [0,1]$. These links describe the influence between them and they are also symmetric.*

Given an assumption: if a sub-event contributes more to an event then it should have more influence on this event, This contribution is defined as relevance and importance. The relevance value is computed as the similarity between the sub-event's centroid and the event's centroid; the importance value is computed as the fraction of propagation; a parameter λ is set to balance the two factors. (Eq. (4.4))

$$p(s_i, e_j) = \lambda \frac{C_{s_i} \bullet C_{e_j}}{\| C_{s_i} \| \| C_{e_j} \|} + (1-\lambda) \frac{prop(s_i)}{\max\{prop(s), s \in e_j\}} \quad (4.4)$$

D. Credibility Initialization

Till now, we have constructed the three-layer hierarchical credibility network with sub-event detection and defined all links in this network. Before further analysis, initial credibility values to all nodes in this network are required. In this part, a classification based method is used to assess the credibility at the message level. As defined in Section II, the initial credibility value for message and then sub-event and event can be acquired from this classifier's result.

The credibility classifier is trained at the message level rather than the event level. There are several reasons that the event level classifier fails to work well for it is not entity-aware and ignores relations among entities.

Taking previous proposed features and including several new features, a machine learning classifier model can be trained with some labeled data. In Section III, message is defined with three types of features: content features, social features and user features. Content features are features extracted from message content considering some unique features of Microblog (e.g. hashtag topic). Social Features are features generated during the propagating of messages on social network (e.g. comment/forward times). User Features are features about a message's publisher.

Taking the probabilistic result from the classifier, the creditability of a message is initialized as follows.

$$C(m) = \frac{prob(m) - 0.5}{0.5} \quad (4.6)$$

The probabilistic result describes how likely a message is credible(1 for credible, 0 for non-credible). Then we initialize sub-event's credibility as the average scores of all messages it contains and initialize an event's credibility as the average scores of all sub-events it contains.

With this initialization, each entity has a credibility value $\in [-1,1]$ so that credible entities have negative values and non-credible ones have positive values.

After the network is constructed, all links are computed and all entities are initialized, credibility values of all entities can be propagated over this network to achieve a more reliable result. In the next section, we will formulate this propagation as a graph optimization problem and provides a global optimal solution to it.

IV. Credibility Propagation

In this section, we are dealing with the problem to optimize the credibility of all entities based on the HCNet proposed in the last section. We formulate the credibility propagation as a graph optimization problem, define a loss function to this problem and deduce an iterative algorithm that can produce the global optimal solution with the method of gradient descent.

A. Variable Notation

First of all, formal notations of the variables involved in this problem are provided. There are n messages $\{m_1, ..., m_n\}$, l sub-events $\{s_1, ..., s_l\}$ and one event e_1. We denote each entity's credibility value as $C(m_i)$, $C(s_i)$ and $C(e_i)$, then we get 3 credibility vector: $\mathbf{M} = \{C(m_1), ..., C(m_n)\}$, $\mathbf{E} = \{C(e_1)\}$ and $\mathbf{S} = \{C(s_1), ..., C(s_l)\}$. Every message is linked with other messages, this message to message link matrix is defined as a $n \times n$ matrix $\mathbf{W}_F$, for each $\mathbf{W}_F^{i,j} = f(m_i, m_j)$. Every sub-event is linked with other sub-events, this sub-event to sub-event link matrix is defined as a $l \times l$ matrix $\mathbf{W}_H$, for each $\mathbf{W}_H^{i,j} = h(s_i, s_j)$. Set $f(m_i, m_j) = h(s_i, s_j) = 0$, if $i = j$. The message to sub-event links are defined as a $n \times l$ matrix $\mathbf{W}_G$, for each $\mathbf{W}_G^{i,j} = g(m_i, s_j)$. And the sub-event to event link matrix is a $l \times 1$ matrix $\mathbf{W}_P$, for each $\mathbf{W}_P^{i,1} = p(s_i, e_1)$. With the link functions defined in section III, these four matrixes are all non-negative matrix; $\mathbf{W}_F$ and $\mathbf{W}_H$ are also symmetric matrix.

B. Optimization Formulation

Under the assumption that entities with large link weight between them should have similar credibility values, the credibility propagation problem is formulated as a graph optimization problem. Inspired by the semi-supervised graph learning algorithms([24][25][26]), we choose a loss function in the context of our problem settings. The loss function is defined as Eq. (5.1).

$$Q(\mathbf{M},\mathbf{S},\mathbf{E}) =$$

$$\gamma_f \sum_{i,j=1}^{n} \mathbf{W}_F^{i,j} \left(\frac{C(m_i)}{\sqrt{\mathbf{D}_F^{i,i}}} - \frac{C(m_j)}{\sqrt{\mathbf{D}_F^{j,j}}} \right)^2 + \gamma_h \sum_{i,j=1}^{l} \mathbf{W}_H^{i,j} \left(\frac{C(s_i)}{\sqrt{\mathbf{D}_H^{i,i}}} - \frac{C(s_j)}{\sqrt{\mathbf{D}_H^{j,j}}} \right)^2 + \tag{5.1}$$

$$\gamma_g \sum_{i=1}^{n} \sum_{j=1}^{l} \mathbf{W}_G^{i,j} \left(\frac{C(m_i)}{\sqrt{\mathbf{D}_{GM}^{i,i}}} - \frac{C(s_j)}{\sqrt{\mathbf{D}_{GS}^{j,j}}} \right)^2 + \gamma_p \sum_{i=1}^{l} \sum_{j=1}^{1} \mathbf{W}_P^{i,j} \left(\frac{C(s_i)}{\sqrt{\mathbf{D}_{PS}^{i,i}}} - \frac{C(e_j)}{\sqrt{\mathbf{D}_{PE}^{j,j}}} \right)^2 +$$

$$(1-\gamma_f-\gamma_g)\|\mathbf{M}-\mathbf{M}_0\|^2 + (1-\gamma_h-\gamma_g-\gamma_p)\|\mathbf{S}-\mathbf{S}_0\|^2 + (1-\gamma_p)\|\mathbf{E}-\mathbf{E}_0\|^2$$

Here, $\mathbf{D}_*$ are diagonal matrixes and defined as follows:

$$\begin{aligned} &\mathbf{D}_F^{i,i} = \sum_{k=1}^{n} \mathbf{W}_F^{i,k},\ \mathbf{D}_H^{i,i} = \sum_{k=1}^{l} \mathbf{W}_H^{i,k},\ \mathbf{D}_{GM}^{i,i} = \sum_{k=1}^{l} \mathbf{W}_G^{i,k} \\ &\mathbf{D}_{GS}^{i,i} = \sum_{k=1}^{n} \mathbf{W}_G^{k,i},\ \mathbf{D}_{PS}^{i,i} = \sum_{k=1}^{1} \mathbf{W}_P^{i,k},\ \mathbf{D}_{PE}^{i,i} = \sum_{k=1}^{l} \mathbf{W}_P^{k,i} \end{aligned} \tag{5.2}$$

γ_* are positive parameters which are constrained to ensure $1-\gamma_f-\gamma_h \geq 0, 1-\gamma_h-\gamma_g-\gamma_p \geq 0, 1-\gamma_p \geq 0$.

The loss function has seven parts. The first four terms are the smoothness constraints, which mean that the propagation function should not change too much between entities with large link weight. The last three terms are the fitting constraints, which mean that the propagation function should not change too much from the initial values. Four positive parameters are defined to trade off these constraints.

After the loss function is defined, the next problem is to minimize it:

$$(\mathbf{M}^*,\mathbf{S}^*,\mathbf{E}^*) = \arg\min_{\mathbf{M},\mathbf{S},\mathbf{E}} Q(\mathbf{M},\mathbf{S},\mathbf{E}) \tag{5.3}$$

This loss function is a convex function (because all matrix $\mathbf{W}_*$ involved are non-negative), so it is ensured to have a unique global minimum solution.

C. Iterative Solution

The objective optimization function (Eq. (5.3)) involves three variables. It is difficult to get an analytical solution out of it. However gradient descent method can be used to get an iterative solution of it. This problem is solved with respect to the tree variables as follows.

1) *Iterative Solution for Message*

Firstly, differentiate $Q(*)$ with respect to $\mathbf{M}$:

$$\begin{aligned} &\left.\frac{\partial Q(*)}{\partial \mathbf{M}}\right|_{\mathbf{M}=\mathbf{M}^*} \\ &= 2\gamma_f(\mathbf{M}^* - \mathbf{T}_F\mathbf{M}^*) + 2\gamma_g(\mathbf{M}^* - \mathbf{T}_G\mathbf{S}) + 2(1-\gamma_f-\gamma_g)(\mathbf{M}-\mathbf{M}_0) \\ &= 2[\mathbf{M}^* - \gamma_f\mathbf{T}_F\mathbf{M}^* - \gamma_g\mathbf{T}_G\mathbf{S} - (1-\gamma_f-\gamma_g)\mathbf{M}_0] \end{aligned} \tag{5.4}$$

Here:

$$\mathbf{T}_F = \mathbf{D}_F^{-1/2}\mathbf{W}_F\mathbf{D}_F^{-1/2} \tag{5.5}$$

$$\mathbf{T}_G = \mathbf{D}_{GM}^{-1/2}\mathbf{W}_G\mathbf{D}_{GS}^{-1/2} \tag{5.6}$$

Then apply the gradient descent to compute $\mathbf{M}$ for the *t*-th iteration:

Algorithm 1 News Credibility Propagation

1: **Input**: Microblog messages about a news event.
2: Clustering to identify sub-events.
3: Initialize credibility of messages using classifier results. ($\mathbf{M}_0$)
4: Initialize credibility of sub-events and event. ($\mathbf{S}_0$, $\mathbf{E}_0$)
5: Set four regulation parameters: γ_f, γ_h, γ_g and γ_p.
6: Compute four link matrixes: $\mathbf{W}_F$, $\mathbf{W}_H$, $\mathbf{W}_G$ and $\mathbf{W}_P$.
7: Compute four normalized matrixes: $\mathbf{T}_F$, $\mathbf{T}_H$, $\mathbf{T}_G$ and $\mathbf{T}_P$.
8: **Repeat**:
9: Update $\mathbf{M}$ using Eq. (5.8).
10: Update $\mathbf{S}$ using Eq. (5.12).
11: Update $\mathbf{E}$ using Eq. (5.14).
12: **Until Converge**.
13: **Return E**.

$$\begin{aligned} \mathbf{M}_t &= \mathbf{M}_{t-1} - \eta\nabla Q(\mathbf{M}_{t-1}) \\ &= \mathbf{M}_{t-1} - 2\eta[\mathbf{M}_{t-1} - \gamma_f\mathbf{T}_F\mathbf{M}_{t-1} \\ &\quad -\gamma_g\mathbf{T}_G\mathbf{S}_{t-1} - (1-\gamma_f-\gamma_g)\mathbf{M}_0] \end{aligned} \tag{5.7}$$

Let $\eta = \frac{1}{2}$:

$$\mathbf{M}_t = \gamma_f\mathbf{T}_F\mathbf{M}_{t-1} + \gamma_g\mathbf{T}_G\mathbf{S}_{t-1} + (1-\gamma_f-\gamma_g)\mathbf{M}_0 \tag{5.8}$$

From this iterative solution for message (Eq. (5.8)), the credibility of a message is determined by three factors: other messages (first term), sub-event it links to (second term) and its initial value (last term). This meets the assumption of a message's credibility.

2) *Iterative Solution for Sub-event*

Take the same steps as above. Firstly differentiate $Q(*)$ with respect to $\mathbf{S}$:

$$\left.\frac{\partial Q(*)}{\partial \mathbf{S}}\right|_{\mathbf{S}=\mathbf{S}^*} = \tag{5.9}$$

$$2[\mathbf{S}^* - \gamma_h\mathbf{T}_H\mathbf{S}^* - \gamma_g\mathbf{T}_G{}^T\mathbf{M} - \gamma_p\mathbf{T}_P\mathbf{E} - (1-\gamma_h-\gamma_g-\gamma_p)\mathbf{S}_0]$$

Here:

$$\mathbf{T}_H = \mathbf{D}_H^{-1/2}\mathbf{W}_H\mathbf{D}_H^{-1/2} \tag{5.10}$$

$$\mathbf{T}_P = \mathbf{D}_{PS}^{-1/2}\mathbf{W}_P\mathbf{D}_{PE}^{-1/2} \tag{5.11}$$

Using gradient descent to deduce its iterative solution:

$$\begin{aligned} \mathbf{S}_t = \gamma_h\mathbf{T}_H\mathbf{S}_{t-1} + \gamma_g\mathbf{T}_G{}^T\mathbf{M}_{t-1} + \\ \gamma_p\mathbf{T}_P\mathbf{E}_{t-1} + (1-\gamma_h-\gamma_g-\gamma_p)\mathbf{S}_0 \end{aligned} \tag{5.12}$$

From this iterative solution for sub-event(Eq. (5.12)), the credibility of a sub-event is determined by four factors: other sub-events (first term), messages it contains (second term), event it links to (third term) and its initial value (last term). This also meets the assumption for a sub-event's credibility.

3) *Iterative Solution for Event*

Similarly differentiating $Q(*)$ with respect to $\mathbf{E}$:

$$\left.\frac{\partial Q(*)}{\partial \mathbf{E}}\right|_{\mathbf{E}=\mathbf{E}^*} = 2[\mathbf{E}^* - \gamma_p \mathbf{T}_P{}^T \mathbf{S} - (1-\gamma_p)\mathbf{E}_0] \quad (5.13)$$

Using gradient descent to deduce its iterative solution:

$$\mathbf{E}_t = \gamma_p \mathbf{T}_P{}^T \mathbf{S}_{t-1} + (1-\gamma_p)\mathbf{S}_0 \quad (5.14)$$

From this iterative solution for event (Eq. (5.14)), the credibility of an event is determined by two factors: sub-events it contains and its initial value.

Iterative solution to this optimization problem has been deduced now. As it is a convex problem, this solution certainly can be served as the global minima.

4) *Algorithm Summing Up*

Till now, we have presented a credibility network, initialized all entities values, computed weights for all links among them and deduced the credibility propagation iterative algorithm over the network. All these procedures are summed up into an algorithm: News Credibility Propagation (Algorithm 1). The performance of this algorithm will be tested in next section.

V. EXPERIMENTS

In this section, we conduct experiment on two real-world datasets. After describing the datasets and performance measures, we present results to demonstrate the effectiveness of proposed model.

A. Datasets[5]

To verify the effectiveness of proposed model in different situations, two Microblog datasets were collected: SW-2013 and SW-MH370. SW-2013 consists of topic-independent news in the year 2013. It has 18 fake news and 171 true news which are represented by 79296 Microblog messages. SW-MH370 consists of news related to the same topic "Flight MH370 Lost Contact". It contains 32 fake news and 135 true news which are represented by 32526 Microblog messages. Detailed information of these two datasets is listed in Table I.

Both datasets were collected from Sina Weibo, which is the leading Microblog in China. First, news events were collected and their keywords and duration time were extracted. Then we used the search engine of Sina Weibo[6] to collect Microblog messages related to specific event keywords and published on defined time window.

Compared with existing studies, authoritative sources are served as ground truth rather than human labeling. The fake news of SW-2013 were collected from several 2013 year's top fake news rank lists[7] selected by authoritative new agencies, like Xinhua News Agency while the true news came from hot news on the same dates to keep a time consistence. The fake news of SW-MH370 were collected from rumors about this topic on the official rumor reporting service of Sina Weibo[8], while the true news came from a news search engine by using the topic as search query. After that, news which was duplicated or not related to the center topic were manually removed.

TABLE I. DATASET DETAILS

	SW-2013	SW-MH370
Type	Topic-indepedent	Topic-related
#News	189	135
#Fake News	18	32
#True News	171	103
#Messages	79296	31526
#Distinct Users	63604	24775

TABLE II. CONFUSION MATRIX FOR NEWS CREDIBILITY CLASSIFICATION PROBLEM

		Actual	
		True	Fake
Prediction	True	$n_{t\to t}$	$n_{f\to t}$
	Fake	$n_{t\to f}$	$n_{f\to f}$

B. Performance Evaluation

1) *Performace Measures*

To judge the performance of the proposed model, several performance measures are proposed. These measures are defined depend on the confusion matrix given in Table II.

Accuracy is the percentage of correctly identified fake and true news (Eq. (6.1)). Precision is the fraction of fake news predictions that are correct (Eq. (6.2)). Recall examines the fraction of fake news being recognized (Eq. (6.3)). And F-score is the harmonic mean of precision and recall (Eq. (6.4)).

$$Accuracy = \frac{n_{t\to t} + n_{f\to f}}{n_{t\to t} + n_{t\to f} + n_{f\to f} + n_{f\to t}} \quad (6.1)$$

$$Precision = \frac{n_{f\to f}}{n_{t\to f} + n_{f\to f}} \quad (6.2)$$

$$Recall = \frac{n_{f\to f}}{n_{f\to t} + n_{f\to f}} \quad (6.3)$$

$$F-score = \frac{2\times Precision \times Recall}{Precision + Recall} \quad (6.4)$$

1) *Results Comparison*

Performance on both datasets are compared with following methods:

E-Class: a SVM classifier at the event level with 47 features aggregated from messages into the event for training.

[5] Datasets are available at https://www.dropbox.com/sh/9lmy4veobd2oknk/AABEcn77PRHwKJcNJitm7d0Ma?dl=0

[6] http://s.weibo.com/

[7] http://news.xinhuanet.com/zgjx/2014-01/08/c_133024019.htm; http://opinion.haiwainet.cn/n/2013/1220/c232601-20062341.html;

[8] http://service.account.weibo.com/

TABLE III. PERFORMANCE EVALUATOIN

	Accuracy	Precision	Recall	F-score
E-Class	0.820	0.289	0.611	0.393
M-Class	0.868	0.4	0.778	0.528
CP-Initial	0.878	0.424	**0.778**	0.549
NewsCP	**0.889**	**0.448**	0.722	**0.553**

(a) Performance result on dataset SW-2013.

	Accuracy	Precision	Recall	F-score
E-Class	0.793	0.583	0.438	0.5
M-Class	0.822	0.605	0.719	0.657
CP-Intial	0.830	0.621	0.719	0.667
NewsCP	**0.851**	**0.657**	**0.781**	**0.714**

(b) Performance result on dataset SW-MH370.

M-Class: a SVM classifier at the message level. The event's credibility are generated from the average of its messages' prediction values.

CP-Initial: the initial result for credibility propagation network without further iterations.

NewsCP: News credibility propagation, the method proposed in the paper.

All these methods are tuned to choose best parameters, and a 4-fold cross validation is used for two classification methods. There are several parameters influencing the performance of NewsCP: the clustering threshold parameter μ, the four regulation parameter γ_f, γ_g, γ_h and γ_p. We take an empirical setting for regulation parameters as $\gamma_f = 0.3$, $\gamma_g = 0.06$, $\gamma_h = 0.06$ and $\gamma_p = 0.5$. We set $\mu = 0.6$ for SW-2013, and set $\mu = 0.8$ for SW-MH370 as it has smaller granularity of sub-events.

Some conclusion can be drawn from the results in Table III:

- NewsCP achieves best F-score and accuracy performance on both datasets. For topic-independent dataset SW-2013, NewsCP provides about 7% boost in accuracy and 16% boost in F-score over the E-class method. For topic-related dataset SW-MH370, NewsCP provides about 6% boost in accuracy and 21% boost in F-score. It also outperforms M-Class and CP-Initial in accuracy and F-score.
- The performance of M-Class is better than that of E-class, this proves the limitations of the event-level classification method. The performance of CP-Initial is better than that of two classification methods this proves the rationalization of the network's structure and its initialization method.
- NewsCP is more effective on topic-related dataset than topic-independent dataset. On SW-MH370, this method provides 21% f-score improvements over E-class and 5% f-score improvements over CP-Initial, while the corresponding improvements on SW-2013 are only 16% and 1%.

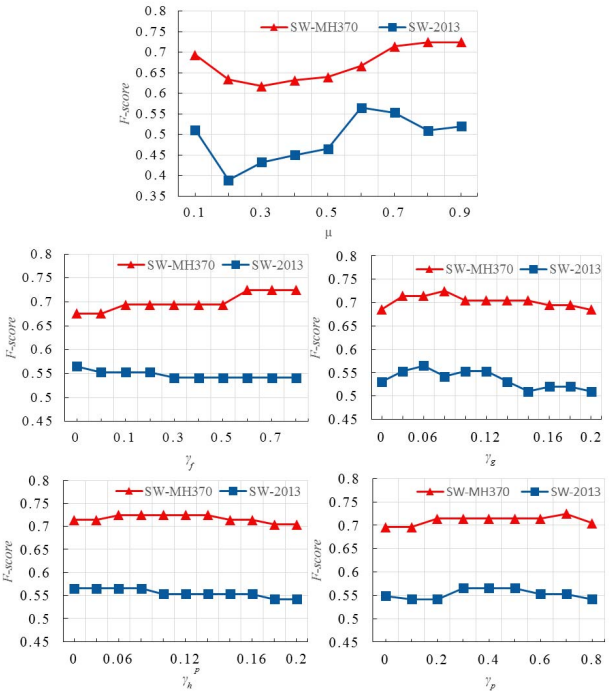

Fig. 4. F-score results with respect to different parameters.

These conclusions prove the effectiveness of the proposed model under different situations.

2) *Varying Parameters*

To examine each parameter's effect, five parameters are varied one at a time and fixed the others to perform NewsCP. For each parameter, F-score is examined for evaluating fake news identification performance.

Some observation can be made from Fig. 4:

- The performance of NewsCP is most sensitive to the clustering threshold μ .As μ controls the sub-event's granularity, this means the introduction of sub-event layer plays an important role in this model: with an appropriate choice of sub-event granularity, NewsCP gains significant improvement.
- NewsCP's performance is also sensitive to the message to sub-event link regulation parameter γ_g. This means the message to sub-event inter-layer relationship is vital and proves the importance of sub-event layer from another angle.
- NewsCP's performance is influenced slightly by the other regulation parameters (γ_f, γ_h and γ_p). This means the other three types of relations are not so important, though they still have some impacts for overall performance.

3) *Iteration times*

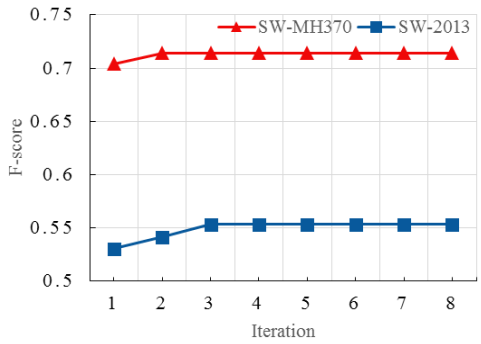

Fig. 5. F-score results with respect to iteration times.

As illustrated in Fig. 5, the f-score performance for both datasets stabilized in a few iterations: experiment conducted on dataset SW-MH370 reaches best f-score in 2 iterations and on dataset SW-2013 in 3 iterations. This shows that the proposed propagation model can converge to a best performance in a few iteration steps.

4) *Case Studies*

Sub-events of a news reveals detail aspects of a news event, they can give evidences to explain the judge of an event's truthfulness explicitly. Here, several cases of these evidences are listed: for news identified by NewsCP as fake, the sub-event with the smallest credibility value of it are presented to explain this evaluation.

- Case 1 (SW-MH370)

News: CIA reports the flight has crashed in Penang.

Evidence: The findings by CIA indicated that the lost Flight MH370 was shot down by Malaysia military aircraft and crashed in Malaysia Penang at 2:21 on March 8, 2014. A Royal Malaysian Air Force commander gave the final orders. (CNN CORRESPONDENT: ONEJOYO sent from Kuala Lumpur).

- Case 2 (SW-MH370)

News: MH370: All people are survived, the co-pilot is terrorist.

Evidence: Can't wait any more, please spread this. [repost] Britain has just finished broadcasting the news, saying that people on the plane are alive, all alive, and the co-pilot was a terrorist. The plane is on an island in India. The co-pilot destroyed navigation and signal system. This news has not been broadcast in China yet.

- Case 3 (SW-2013)

News: A kind girl fed a homeless old man in Shenzhen

Evidence: Video: "The kindest girl fed a homeless old man in Shenzhen" is fake? Kindest or Not [Figure] - Social - Shun Net News http://t.cn/zT7SoGC.

From these case examples, it can be concluded that NewCP is not only effective for news credibility evaluation, but also explanatory, thus this approach has practical utility.

VI. Related Work

There is an extensive body of related works on information credibility evaluation for online content. In this section, we provide a brief review of the research that is most closely related to ours in three main areas: spam detection on social media, credibility evaluation on Microblog and truth discovery.

A. *Spam Detection on Social Media*

Although large-scale social systems have gained huge popularity across the world, they also lead to a lot of spam information spreading. For example, there are 3 million spam tweets per day and 25 million spammers on Twitter.

Many studies have been done to identify these spams/spammers. Supervised spam detection is the most popular approach. By training classifier with labeled data, this approach has been proved effective in several domains (e.g. [27], [28]). Some interesting social features are extracted for this approach. Another approach for identifying spammers is ranking users based on their social graph ([29]). Recently, crowd wisdom methods are also utilized to identify fake accounts on social networks([4]). In [9], spatiotemporal groundings for claims on social networks are made to help assess content credibility on social networks.

B. *Credibility Evaluation on Microblog*

To address the problem of automated information credibility evaluation on Microblog, some methods have been proposed. Classification based approach is widely used to identify untrustworthy information. Castillo et al. [1] compare some supervised learning algorithms to determine an event as credible or not. They construct the classifier with features extracted from four aspects: the message, user, topic, and propagation. [2][8] exploit this idea to detect rumors on Sina Weibo with several new features. M. Gupta et al. [5] analyze event's credibility by constructing a three-layer network to propagate credibility between users, tweets and events. They initial the network with values learned with classification approach and optimize it after each propagation iteration. Unlike [1], where the features can be assigned only for events, they exploits inter-entity relationships. Although they use an iterative algorithm, they provide no guarantee of convergence, and no description of the objective function being optimized. In this paper , we overcome limitations of previous approaches and formulate the credibility propagation approach as an optimization problem and provide a global optimal solution to it with an iterative algorithm. A sub-event layer is also initially introduced for deeper sematic mining.

C. *Truth Discovery*

Truth discovery refers to the problem on finding the truth with conflicting information. Several approaches have been proposed to handle this problem. V. V. Vydiswaran et al. [17] study the problem of truth discovery with semi-supervised graph learning. They use a small set of ground truth data to help distinguishing true facts from false ones. Semi-supervised graph learning has been studied by Zhu et al. [24][25] and Zhou et al. [26]. The main purpose of these approaches is to make predictions consistent with both labeled data and the

graph structure. Inspired by this approach, we give a formal deduction of credibility propagation iterative algorithm on a three layer credibility network with certain consistent assumptions.

VII. CONCLUSION

As the fake news on Microblog can lead to very serious consequences in our society in recent years, it is crucial to evaluate news credibility automatically. In this paper, we have proposed a hierarchical credibility propagation approach to tackle this challenging problem. A three-layer hierarchical credibility network was presented, which consists of messages, sub-events and events, with links built with semantic and social relations among these entities. In the network, a sub-event layer was initially introduced in this paper to reveal deeper semantics of a news. Through formulating the credibility propagation process on this network as a graph optimization problem, we have provided the globally optimal solution with an iterative algorithm. With experiments on two real-world datasets we collected, the proposed approach has been validated to be signficantly better than the baseline methods in terms of both accuracy and F-score.

ACKNOWLEDGEMENTS

This work was supported by the National High Technology Research and Development Program of China (2014AA 015202), National Nature Science Foundation of China (61172153,61100087), National Key Technology Research and Development Program of China (2012BAH39B02), and Beijing New Star Project on Science & Technology (2007B071).

REFERENCES

[1] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In Proc. of the Intl. Conf. on World Wide Web (WWW), pp. 675–684. ACM, 2011.

[2] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on Sina Weibo. In Proc. of the ACM SIGKDD Workshop on Mining Data Semantics (MDS). Article 13, pp. 1-7. ACM, 2012.

[3] J. Allan. Introduction to topic detection and tracking. In J. Allan, editor, Topic Detection and Tracking - Event-based Information Organization, pp. 1-16. Kluwer Academic Publisher, 2002.

[4] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. J. Metzger, H. Zheng, and B. Y. Zhao, Social Turing Tests: Crowdsourcing Sybil Detection. In NDSS, 2013.

[5] M. Gupta, P. Zhao, and J. Han. Evaluating Event Credibility on Twitter. In Proc. of the 2012 SIAM International Conference on Data Mining (SDM), pp. 153-164. SIAM / Omnipress, 2012.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In Proc. of the Intl. Conf. on World Wide Web (WWW), pp. 161–172. ACM, 1998.

[7] S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang. Prominent Features of Rumor Propagation in Online Social Media. 2013 IEEE 13th International Conference on Data Mining (ICDM), pp. 1103-1108. 2013.

[8] S. Sun, H. Liu, J. He, X. Du. Detecting Event Rumors on Sina Weibo Automatically. In Proc. of the 15th Asia-Pacific Web Conference (APWeb), pp. 120-131. Springer, Heidelberg, 2013.

[9] L. Derczynski, K. Bontcheva. Spatio-temporal grounding of claims made on the web, in Pheme. Proceedings of the 10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation (ISA 10). 2014.

[10] V. V. Vydiswaran, C. Zhai, and D. Roth. Content-Driven Trust Propagation Framework. In Proc. of the Intl. Conf. on Knowledge Discovery and Data Mining (SIGKDD), pp. 974–982. ACM, 2011.

[11] R. Balakrishnan. Source Rank: Relevance and Trust Assessment for Deep Web Sources based on Inter-Source Agreement. In Proc. of the Intl. Conf. on World Wide Web (WWW), pp. 227–236. ACM, 2011.

[12] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In Proceedings of the third ACM international conference on Web search and data mining (WSDM '10), pp. 291-300, ACM, 2010.

[13] J. Pasternack and D. Roth. Knowing What to Believe (When You Already Know Something). In Proc. of the Intl. Conf. on Computational Linguistics (COLING). pp. 877-885. Tsinghua University Press, 2010.

[14] Y. Yang, T. Pierce, and J. Carbonell. A study on retrospective and on-line event detection. In Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR'98), pp. 28-36, ACM, 1998.

[15] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T.Archibald, and X. Liu. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 14(4):32-43, 1999.

[16] K. Lee, J. Caverlee, , Z. Cheng, and D. Sui, Campaign Extraction from Social Media. In ACM TIST, Vol. 5, No. 1, pp. 1-28. ACM. Dec. 2013.

[17] V. V. Vydiswaran, C. Zhai, and D. Roth. Content-Driven Trust Propagation Framework. In Proc. of the Intl. Conf.on Knowledge Discovery and Data Mining (SIGKDD), pp. 974–982. ACM, 2011.

[18] X. Yin andW. Tan. Semi-Supervised Truth Discovery. In Proc. of the Intl. Conf. on World Wide Web (WWW), pp. 217–226. ACM, 2011.

[19] X. Yin, P. S. Yu, and J. Han. Truth Discovery with Multiple Conflicting Information Providers on the Web. IEEE Transactions on Knowledge and Data Engineering (TKDE), 20(6):796–808, 2008.

[20] L. Bao, J. Cao, Y. Zhang, J. Li, M. Chen, and A. G. Hauptmann. Explicit and implicit concept-based video retrieval with bipartite graph propagation model. In Proceedings of the international conference on Multimedia (MM '10). pp. 939-942. ACM, 2010.

[21] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida.Detecting Spammers on Twitter. In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), July 2010.

[22] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd.Detecting spam in a Twitter network. First Monday, 15(1),January 2010.

[23] X. Yin, J. Han and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. KDD'07.

[24] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. pp. 19-26. CMU Technical Report CMU-CALD-02-107, 2002.

[25] X. Zhu, Z. Ghahramani and J. Lafferty. Semi-supervisedlearning using Gaussian fields and harmonic functions. pp. 912-919. ICML'03.

[26] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with Local and Global Consistency. In NIPS, Vol. 16, pp. 321-328, 2003.

[27] S. Lee, and J. Kim, WarningBird: Detecting suspicious URLs in Twitter stream. In NDSS, 2012.

[28] K. Lee, B. Eoff, and J. Caverlee, Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In ICWSM, 2011.

[29] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and P. Gummadi K.Understanding and combating link farming in the twitter social network. In WWW. pp.61-70. ACM, 2012.