FINDING SOURCE OF RUMOR IN LARGE ONLINE SOCIAL NETWORKS

Stage-I Project Report

Submitted by

111508005 Akash Patil

111508075 Jayant Kalani

111508045 Ganesh Landge

B.Tech Information Technology

Under the guidance of

Anish R. Khobragade

College of Engineering, Pune

DEPARTMENT OF COMPUTER ENGINEERING AND

INFORMATION TECHNOLOGY,

COLLEGE OF ENGINEERING, PUNE-5

December, 2018

Table of Contents

1. Abstract

We address the problem of estimating the source of a rumor in large-scale social networks. Previous works studying this problem have mainly focused on graph models with deterministic and homogenous internode relationship strengths.

However, internode relationship strengths in real social networks are random. This uncertainty is modeled by using random, non-homogenous edge weights on the underlying social network graph. Also a novel two-stage algorithm has been proposed that uses the modularity of the social network to locate the source of the rumor with fewer sensor nodes than other existing algorithms.

strengths. Simulations show that the proposed algorithm can determine the actual source within two hops, 69%–80% of the time, when the diameter of the networks varies between 7 and 13.We implement this proposed two stage algorithm and we will propose a new algorithm which will improve the efficiency of finding the source to more than 80%.

2. Introduction

We live in the networked society. Every second we interact with many networks from which we collect, process and transmit a huge amount of information, which increases exponentially each year. Increasing interconnectivity of the world exposes us to world-wide range of pathogens, viruses both physical and virtual, misinformation and rumors with often grievous consequences. A good example is United States presidential election of 2016 when many rumors or fake news became viral on Facebook or Twitter and might have affected elections. Many papers seek finding best conditions for spreading or sets of optimal spreaders but here we investigate an inverse problem. It became clear that one of the major challenges facing network and data scientists is to develop effective methods for detecting and suppressing spread of dangerous viruses, pathogens, misinformation or gossips.

RUMORS spreading in social networks have long been a critical threat to our society. A good example is a fake tweet about explosion in White House in 2013, which caused $130 billion loss on the stock market. This demonstrates that a single rumor can cause great damage to business and life. Nowadays, with the development of mobile devices and wireless techniques, the temporal nature of social networks (time-varying social networks) has a deep influence on dynamical information spreading processes occurring on top of them. The ubiquity and easy access of social networks not only promote the efficiency of information sharing but also dramatically accelerate the speed of rumor spreading Rumors combine the characteristics of the "word-of-mouth" spreading scheme with the dynamic connections between individuals in time-varying social networks

For either forensic or defensive purposes, it has always been a significant work to identify the source of rumors in time-varying social networks. However, the existing techniques generally require firm connections between individuals (i.e., static networks) so that administrators can trace back along the determined connections to reach the spreading sources.

So to solve above all problems and finding the source of rumor need undoubtedly a fast algorithm finding a source of such spread.

3. Software Requirement Specification

3.1 Purpose:

The purpose of this Software Requirement Specification is to give an overview of the functional and non-functional requirements of our project titled "Finding the source of the information in large social network" where we address the problem of estimating the source of rumor in large scale social networks.

3.2 Product Scope:

Nowadays online social networks are good platform to spread the news or information. Studies demonstrate that information spreads much faster in social networks than any other communication media. This ease of information dissemination can also be used to spread rumors. So the purpose of the project is to find the source of the rumor thereby halting the spread of the rumor. The goal is to find the first person who spread the potential misleading information and it is beneficial for government officers to find the culprit.

3.3 Product Perspective:

Recently countries are facing cases like mob lynching, fights and kidnapping due to spread of wrong information or rumor through social media. There comes a need for finding the cause of rumor and stop it. This product is follow-on member of already existing product which is less efficient i.e. it finds node within few hops of actual source node. Besides it provides scope for improvement. So this is dependent product and will be the first release of the product.

3.4 Product Functions

Our software uses the fact that large-scale social networks are modular i.e. made from different small clusters having some weak relationships. We represent the social network in graphical form.

So the major function is to find the first node who initiated the spread of information. We propose a novel two-stage algorithm that uses the modularity of the social network to locate the source of the rumor with fewer sensor nodes than other existing algorithms.

We model the function in following stages:

- **Find the Most Likely Candidate Cluster (First Stage)** : In this stage we find the sub-graph or the cluster where the probability of finding the source is more.
- **Find the Source of Rumor (Second Stage)** : In this stage using the sub-graph of above stage we find the actual source.

3.5 Operating Environment

3.5.1 Hardware Requirements:

- 2.40 GHz processor
- 8 GB memory RAM
- 100 MB of disk space

3.5.2 Software Requirements:

- Linux operating system
- Programming language : Python 3.0
- NetworkX library for network analysis

3.6 Design and Implementation Constraints

- Input Dataset should contain deterministic and homogenous internode relationship strengths.
- The software will support Linux Ubuntu machines with 2.40 GHz CPU and 8 GB memory.
- Specific technologies used are NetworkX, python 3.0, matplotlib, etc
- Varying internode strengths may limit the performance which is major constraint.

3.7 Assumptions and Dependencies

- Input dataset is assumed to be in specific format which gives relationship between nodes and corresponding strengths of edges among them.
- Social networks are assumed to be modular which can be divided into clusters.
- Dependent libraries are : NetworkX, matplotlib, etc.
- Information spread is assumed to be just like epidemic spread (SI, SIS or SIR models).
- The user has basic knowledge of computer and command line.

3.8 Functional Requirements

System Features:

1. **Estimating the source of a rumor on a dataset provided by Twitter**
    a. Parsing the essential attributes from JSON dataset of twitter handles/accounts.
    b. Converting the dataset into a discrete graph.
    c. Finding the sensor nodes and clusters in the graph.
    d. Implementing a two-stage algorithm on the graph to predict the source of a given tweet.

2. **Adding Machine Learning capabilities to the algorithm**
    a. Improvising the existing algorithms to improve speed and accuracy using various ML techniques.
    b. Designing a training dataset containing a set of tweets and their source in the form of a graph.
    c. Creating a model that can be trained using the designed training dataset to predict the source of a tweet on a test dataset.

3. **Documenting the improvised algorithm in the form of an IEEE research paper**

3.9 Nonfunctional Requirements

1. **Performance Requirements:**

- **Scalability** : System should be able to handle large-scale networks. For e.g. handling datasets containing around thousands of node
- **Speed** : The application should be fast. It should not slow down exponentially with increase in the size of database. Search functionality should be fast to enable better user experience.
- **Reliability** : System should be reliable enough to enable great experience

2. **Software Quality Attributes**

- **Usability** : User interface should be simple and clear to beak to understand by any users.
- **Availability** : System should be available at all time. It should be ensured that there should be minimum no downtime to ensure better user experience.

- **Testability** : Software should be testable. A separate test environment should be set up where testers and Quality Assurance engineers can test the application for bugs and/or incomplete or missed requirements.
- **Maintainability** : The software should be developed in such a way that it is extensible. It should be easy to incorporate new features, requirements or accommodate a change in existing requirements.

4. Literature Review :

4.1 History

The basic component of such a system is undoubtedly a fast algorithm finding a source of such spread. The first widely discussed research on this subject has been done by Shah and Zaman[21] and Pinto, Thiran and Vetterli[22]. In social networks, Shah and Zaman introduced *rumor centrality* of a node as the number of distinct ways a rumor can spread in the network starting from that node. They showed that the node with maximum *rumor centrality* is the Maximum Likelihood Estimator of the rumor source if the underlying graph is a regular tree. They studied also the detection performance for irregular geometric trees, small-word networks and scale-free networks.. Pinto *et al*. relaxed some of these constraints since their algorithm requires information about state of not every node, but only about some fraction of nodes called *observers*. After these two publications, the topic of the source detection became popular and many other variants of this problem have been studied. We can distinguish two main approaches to this issue: the snapshot-based[21,23,24,25] and the detector-based[22,26,27] source detection. The first one requires the snapshot of an entire network at a certain time instance, the second needs to monitor only a small subset of nodes but all the time. Regardless of the above division, researchers considered also different epidemic models[25,28], spreading at weighted or time-varying graphs[29,30,31,32] and multi-source detection problems[33,34]. In 2014 Jiang *et al*. described state-of-the-art and conducted comparative studies[35]. One of their conclusions is that current methods are too computationally expensive and they can not be use for a quick identification of the propagation source. The main goal of our research was finding the method which executes in reasonable time on large complex networks and delivers high quality of localization results at the same time.

Some important Research timeline :

Time Period	Summary
2012	Earliest idea on source localization given by Pedro C. Pinto, Patrick Thiran, Martin Vetterli . They presented a strategy that is optimal for arbitrary trees, achieving maximum probability of correct localization.
2013	Different Epidemic Spread model were used to study the dissemination of rumor in social networks.Kai Zhu, Lei Ying followed the popular Susceptible-Infected-Recovered (SIR) model to study the problem of detecting the information source in a network.
2014	Studies were done on static-internode relationship networks. Jiang *et al*. described state-of-the-art and conducted comparative studies35. One of their conclusions is that current methods are too computationally expensive and they can not be use for a quick identification of the propagation source

4.2 Who Spread That Rumor: Finding the Source of Information in Large Online Social Networks With Probabilistically Varying Internode Relationship Strengths :

- By Alireza Louni and K. P. Subbalakshmi

Resource : https://ieeexplore.ieee.org/document/8299476

This is two stage algorithm for finding source of Rumor in the Large Scale social network

First Stage : Find the Most Likely Candidate Cluster

we first need to discover the clusters/communities existing in the network.

We use the Louvain method to find the clusters. The method is a greedy algorithm with

Time Complexity : O(|V | log |V |)

Input : Graph of sensor nodes (Each sensor node indicates a cluster)

Output : A sensor node indicating most likely candidate cluster

Second Stage : Find the Source of Rumor

This stage uses the Betweenness Centrality Scores (BCS), the node that maximizes the likelihood value is chosen as the source of the rumor

Time Complexity : O(a.|V|^3) where

|V| = number of nodes in candidate cluster

'a' = the ratio of of the nodes acting as the sensor in each stage

Input : Most likely candidate cluster from stage one

Output : source node (node having maximum likelihood)

Accuracy : Simulations show that the proposed algorithm can determine the actual source within two hops, 69%–80% of the time, when the diameter of the networks varies between 7 and 13.

4.3 Fast and accurate detection of spread source in large complex networks

- By Robert Paluch, Xioyan Lu and Janu Holyst

Resource:https://www.nature.com/articles/s41598-018-20546-3#article-info

Pinto, Thiran and Vetterli[22] proposed a general framework for the localization of the spread source in which some of the nodes in network act as observers and report from which neighbor and at what time it received the information. It uses Gradient Maximum Likelihood Approach.

Complexity : Using the symbols K_0 and N_0 we reformulate the time complexity of GMLA as $O(N_0(K_0{}^3+N^2))$ in the worst case. Assuming N_0~log(N) and $K_0 \ll N$, which is true for our method, the complexity can be further simplified into $O(\log(N)N^2)$.
Accuracy : The accuracy of a single realization is {a}_{i}=1/|{V}_{top}| if $s* \in V\ top$ or $a\ i = 0$ otherwise, where $s*$ is the true source and $V\ top$ is a group of nodes with the highest score (top scorers). The total accuracy a is an average of many realizations $a\ i$, therefore $a \in [0,1]$. This measure takes into account the fact that there might be more than one node with the highest score (ties are possible).

Rank : The rank is the position of the true source on the node list sorted in descending order by the score.

Distance Error : The distance error is the number of hops (edges) between the true source and a node designated as the source by an algorithm. If $|V\ top\ | > 1$, which means that an algorithm found more than one candidate for the source, the distance error is computed as a mean shortest path distance between the real source and the top scorers.

4.4 Rumor Source Identification in Social Networks with Time-Varying Topology

- By Jiaojiao Jiang, Sheng Wen, Shui yu, Yang Xiang, Walnei Zhou

Resource : https://ieeexplore.ieee.org/document/7393814/figures#figures
This algotrithm is divided into three steps :
First, we reduce the time-varying networks to a series of static networks by introducing a time-integrating window.
Second, instead of inspecting every individual in traditional techniques, we adopt a reverse dissemination strategy to specify a set of suspects of the real rumor source. This process addresses the scalability issue of source identification problems, and therefore dramatically promotes the efficiency of rumor source identification.
Third, to determine the real source from the suspects, we employ a novel microscopic rumor spreading model to calculate the maximum likelihood (ML) for each suspect.

Accuracy : The evaluations are carried out on real social networks with time-varying topology. The experiment results show that, this method can reduce 60% − 90% of the source seeking area in various time-varying social networks. The results further indicate that this algorithm method can accurately identify the real source, or an individual who is very close to the real source. To the best of our knowledge, the proposed method is the first that can be used to identify rumor sources in time-varying social networks.

5. Timeline Required for Overall implementation :

MONTH	GOAL	STATUS
October	● Completion of Abstract ● Completion of Requirement specification. ● SRS Documentation.	Completed
November	● Literature Summary	Completed
	● Study of Dataset and formatting it into appropriate way.	Pending
December	● Study of required Libraries. ● Study of machine learning algorithm which needed to be implemented.	Pending
January	● Implementation of existing algorithm and analysis of result.	Pending
February	● Improve existing algorithm and increase its efficiency.	Pending
March	Reserved	-------
April	Submission.	-------

NOTE : The above schedule may change as we explore and go deep into our project.

6. References :

IEEE paper :

- Alireza Louni and K. P. Subalakshmi, “Who Spread That Rumor: Finding the Source of Information in Large Online Social Networks With Probabilistically Varying Internode Relationship Strengths” IEEE transactions on computational social systems, vol. 5, no. 2, June 2018.
- D. Shah and T. Zaman, “Detecting sources of computer viruses in networks: Theory and experiment,” SIGMETRICS Perform. Eval. Rev., vol. 38, no. 1, pp. 203–214, Jun. 2010.
- K. Zhu and L. Ying, “Information source detection in the SIR model: A sample path based approach,” in Proc. Inf. Theory Appl. Workshop, Feb. 2013, pp. 1–9.
- C. H. Comin and L. da Fontoura Costa, “Identifying the starting point of a spreading process in complex networks,” Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 84, p. 056105, Nov. 2011.