Source Detection in Networks using the Stationary Distribution of a Markov Chain

Extended Abstract

Yael Sabato
Ariel University
Ariel, Israel
yael.sabato@msmail.ariel.ac.il

Amos Azaria
Ariel University
Ariel, Israel
amos.azaria@ariel.ac.il

Noam Hazon
Ariel University
Ariel, Israel
noamh@ariel.ac.il

ABSTRACT

Nowadays, the diffusion of information through social networks is a powerful phenomenon. One common way to model diffusions in social networks is the Independent Cascade (IC) model. Given a set of infected nodes according to the IC model, a natural problem is the source detection problem, in which the goal is to identify the unique node that has started the diffusion. Maximum Likelihood Estimation (MLE) is a common approach for tackling the source detection problem, but it is computationally hard.

In this work, we propose an efficient method for the source detection problem under the MLE approach, which is based on computing the stationary distribution of a Markov chain. Using simulations, we demonstrate the effectiveness of our method compared to other state-of-the-art methods from the literature, both on random and real-world networks.

KEYWORDS

Source detection; Maximum likelihood estimation; Markov chains; Independent cascade model

ACM Reference Format:
Yael Sabato, Amos Azaria, and Noam Hazon. 2024. Source Detection in Networks using the Stationary Distribution of a Markov Chain: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

In the age of social media, the spread of information and infection through networks is a significant phenomenon. Understanding the dynamic of information spread and its origin are important for a wide range of applications, including marketing, public health, and identification of fake news. The Independent Cascade (IC) is a common model of the spread of information in a social network [2]. In the IC model, the process of diffusion concerns a message that is propagated through the network. Every connection between two friends is associated with a probability; this value determines the probability that if the first user shares the message, the second user will share the message with her friends as well. As commonly occurs in the spread of fake news, the diffusion process starts with a single initial source. A natural goal is that given a set of users who shared a specific message, to seek the unique source that started the diffusion.

There are many approaches to finding the source of a diffusion in the literature, each assuming different spreading models and various amounts of knowledge of the network parameters (for example, [3–5, 9]). When the probabilities associated with the connections are known or can easily be estimated, a natural mathematical approach for finding the source of the diffusion is the Maximum Likelihood Estimation (MLE) principle. According to the MLE principle, one should compute the likelihood of each user being the source, and output the user with the maximum likelihood.

The first to formalize the computational problem of finding the source of a diffusion in a network in the IC model are Lappas et al. [5]. They show that for arbitrary graphs, the source detection problem is not only NP-hard to find but also NP-hard to approximate. Therefore, they propose an efficient heuristic, but it does not utilize the MLE principle. Zhai et al. [8] present a heuristic that utilizes the MLE principle, and they further show that their heuristic outperforms the heuristic of [5]. However, their heuristic requires extensive computation and does not perform very well. In addition, they note that “although the IC model is popular in social network research, finding source in the IC model is rarely studied”. Recently, Amoruso et al. [1] provide a strong heuristic for finding the source of a diffusion in a network in the IC model.

In this paper, we propose an efficient method that uses the MLE principle for source detection in the IC model. Our method is based on computing the stationary distribution of a Markov chain and is inspired by [4]. Specifically, we recognize that if we represent the social network as a weighted directed graph, the diffusion in the IC model induces a tree that spans the set of users who shared the message, and the root of the tree is the user that initiated the diffusion. In addition, the tree is associated with a weight, which equals to the product of the weights of its edges. In order to estimate the probability of a specific user to be the source, we would like to sum the weights of all spanning tree rooted at this user. However, directly considering all trees is computationally expensive. Therefore, we propose converting the social network into a Markov chain, and the Markov chain tree theorem [6] allows us to compute the sum of the weights of all spanning trees rooted at each user in polynomial time. We consider two approaches for converting the social network to a Markov chain—the *self-loops* and the *no-loops* methods. We show that when using a direct calculation of the stationary distribution, both methods compute the exact value of the sum of the weights of all spanning trees rooted at each user, but this is not guaranteed when using a random walk to estimate the stationary

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

	Advogato	Digg	Epinion	Facebook	Google	Slashdot	Twitter	Youtube	**Average**
			trust	friendships	plus			links	
Self-loops (direct calc.)	222	451	428	354	125	471	241	486	**347.25**
10 steps	68	216	161	146	89	246	167	141	154.25
100 steps	86	237	205	170	98	289	230	168	185.375
1000 steps	111	300	270	233	116	358	230	270	236
10000 steps	175	383	363	308	118	418	247	341	294.125
No-loops (direct calc.)	222	451	428	354	125	471	241	486	**347.25**
10 steps	88	248	216	179	82	301	167	203	185.5
100 steps	132	311	286	252	121	389	253	287	253.875
1000 steps	191	398	384	307	128	444	236	367	306.875
10000 steps	214	441	423	344	126	468	236	441	336.625
Naive	139	172	146	176	78	174	149	131	145.625
Max weight arbo.[1]	136	353	329	278	125	380	184	358	267.875
Random	52	130	98	89	71	137	115	79	96.375
Max out-deg	39	115	82	76	79	130	132	47	87.5
Min in-deg	70	162	125	117	72	155	115	128	118
Max (out/in)-deg	63	132	115	95	86	141	161	154	118.375
IM based	94	309	230	196	120	302	218	273	217.75

Table 1: The number of times in which each method finds the correct source node in the *real-world networks*. The values are out of 1000 cases in which the number of active nodes is at least 20 and there is no trivial solution.

distribution. For evaluating the effectiveness of our approach, we use 14 types of random graphs, and sample 1000 graphs from each type. In addition, we evaluate the effectiveness of our approach on 8 real-world networks, including a portion of Digg, Facebook, and Twitter. We show that our methods outperform several baseline methods, including the method proposed by [8] and [1]. We further show that the *no-loops* method outperforms the *self-loops* method when using a random walk to estimate the stationary distribution. That is, the no-loops method requires fewer random walk steps to approach the performance of our methods when using a direct calculation of the stationary distribution.

2 EXPERIMENTS

For the evaluation of the performance of the self-loops and the no-loops methods, and comparing it to other baselines heuristics, we use 14 types of directed random graphs that have diffusion probabilities on their edges, as well as 8 real-world directed networks from the "social" category of the Konect database [1].

We evaluate the performance of the self-loops and the no-loops methods, using a direct calculation of the stationary distribution. In addition, we evaluated these methods when the stationary distribution is estimated by random walks with 10, 100, 1000, or 10,000 steps. Finally, we evaluate the performance of the following baseline methods:

- **Naive:** The Markov chain approach with the naive conversion method.
- **Random:** A random selection of a node.
- **Max out-degree:** The node with the maximal weighted out-degree is selected.
- **Min in-degree:** The node with the minimal weighted in-degree is selected.
- **Max (out/in) degree:** The node with the maximal weighted out-degree divided by its weighted in-degree is selected.
- **IM based:** For each node, we simulate 1000 diffusions, and the node with the maximal average size of the active set is selected.
- **Maximum arborescence [1]:** The node which is the root of the maximum weight spanning out-tree (arborescence) is selected.

We also evaluate the performance of the method proposed by [8]. However, unfortunately, this method takes extensive time to run and does not perform as well as our simple heuristics (max out-degree and min in-degree).

Table 1 presents the results for the real-world graphs. As can be seen in the table the self-loops and no-loops methods using a direct calculation of the stationary distribution outperform all other methods (on average). As expected, these methods achieved the exact same results. We note that, at times, our methods using random walks may perform slightly better than when using the direct calculation; however, this is only due to the inherent randomness of the problem.

In addition, we found that the no-loops method using random walks requires fewer steps than the self-loops method using random walks to approach the performance of our methods when using a direct calculation of the stationary distribution.

ACKNOWLEDGMENTS

This research has been partly supported by the Ministry of Science, Technology & Space (MOST), Israel. The full version of the paper is available at [7].

[1] http://konect.cc/networks/

REFERENCES

[1] Marco Amoruso, Daniele Anello, Vincenzo Auletta, Raffaele Cerulli, Diodato Ferraioli, and Andrea Raiconi. 2020. Contrasting the Spread of Misinformation in Online Social Networks. *Journal of Artificial Intelligence Research* 69 (2020), 847–879.

[2] Jacob Goldenberg, Barak Libai, and Eitan Muller. 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-mouth. *Marketing Letters* 12, 3 (2001), 211–223.

[3] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. 2012. Inferring Networks of Diffusion and Influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5, 4 (2012), 1–37.

[4] Ankit Kumar, Vivek S Borkar, and Nikhil Karamchandani. 2017. Temporally Agnostic Rumor-source Detection. *IEEE Transactions on Signal and Information Processing over Networks* 3, 2 (2017), 316–329.

[5] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. 2010. Finding Effectors in Social Networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1059–1068.

[6] F Leighton and Ronald Rivest. 1986. Estimating a Probability Using Finite Memory. *IEEE Transactions on Information Theory* 32, 6 (1986), 733–742.

[7] Yael Sabato, Amos Azaria, and Noam Hazon. 2024. Source Detection in Networks using the Stationary Distribution of a Markov Chain. *arXiv preprint arXiv:2401.11330* (2024).

[8] Xuming Zhai, Weili Wu, and Wen Xu. 2015. Cascade Source Inference in Networks: A Markov Chain Monte Carlo Approach. *Computational Social Networks* 2, 1 (2015), 1–17.

[9] Le Zhang, Tianyuan Jin, Tong Xu, Biao Chang, Zhefeng Wang, and Enhong Chen. 2017. A Markov Chain Monte Carlo Approach for Source Detection in Networks. In *Proceedings of the 6th National Conference on Social Media Processing*. 77–88.