Online Social Network Information Source Identification Algorithm Based on Multi-Attribute Topological Clustering

Ming Dong 1, Yujuan Lu 2, Zhenhua Tan 2,* and Bin Zhang 2

1 School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China; dongm@mail.neu.edu.cn
2 Software College, Northeastern University, Shenyang 110819, China; luyujuan@mail.neu.edu.cn (Y.L.); zhangbin@mail.neu.edu.cn (B.Z.)
* Correspondence: tanzh@mail.neu.edu.cn

Abstract: This paper focuses on the problem of information source identification in online social networks (OSNs). By analyzing the research situation of source identification problems and challenges (such as the randomness of the information dissemination process and complexity of the underlying network topology), this paper studies the problem of multiple source diffusion and proposes a source identification algorithm based on multi-attribute topological clustering (MaTC). The basic idea of the algorithm is to decompose the multi-source problems into a series of single-source problems by using clustering partitioning to improve accuracy and efficiency. Firstly, it estimates the number of source nodes, which is also the number of network partitions, then characterizes the combination of multiple attribute structures as an attribute index of topological clustering, performs an analysis of the distribution of real source nodes in each partition to evaluate the accuracy of the clustering partition, and finally uses Jordan centrality within each partition for single-source identification. Through comparative experiments, it is verified that the proposed MaTC algorithm is superior to the comparison algorithms in evaluating indicators.

Keywords: OSNs; information source identification; multi-attribute topological clustering; Jordan centrality

1. Introduction

Currently, people live in a society where everyone is connected to various networks, such as social networks, the Internet, the Internet of Things, biological networks, and so on. They acquire, process, and share information in the network. When major events occur in real society, they often lead to extensive discussions in OSNs. Through people publishing or forwarding relevant information, this information can spread at an unprecedented speed and breadth. The emergence of social media platforms such as Twitter, Facebook, Reddit, and Weibo has proved to be very useful in disaster situations such as natural disasters, man-made disasters, and emergencies. The rapid spread of news and information on social networks can attract more users’ attention because social media is a common means to spread trending discussions and breaking news, but the information may contain unconfirmed or false information. The rapid spread of this information in social networks brings many risks, such as exploiting people’s fear of epidemic outbreaks, providing wrong advice in disaster situations, and damaging the reputation of individuals or organizations through anonymous speech. Therefore, it is very important to prevent and control the large-scale spread of unconfirmed or false information on social networks.

Source detection, that is, information source identification, is an important means of rumor control in OSNs. By identifying information sources and controlling them, the risks brought by their diffusion can be reduced. However, due to the complexity of information dissemination processes, real-time data, and dynamic changes in the network, how to
quickly and accurately detect the source of information in social networks has become a very challenging task.

When designing an information source identification method, it is necessary to fully consider the data processing ability in practical applications and the limitations of observation conditions in the process of information dissemination:

(1) The number of observable nodes is limited due to the fast speed of information dissemination, the huge number of nodes involved in the dissemination, and the existence of massive information dissemination, among other reasons. At the same time, due to the limitations of data processing capacity, the number of nodes that can be observed at the same time is limited.

(2) The process of information dissemination in OSNs is random. The forwarding probability of information between users is affected by many factors, such as the content of the information itself and the mutual influence between users and their interests, so the information dissemination process is random.

(3) The complexity of the underlying network structure (such as scale-free characteristics and small-world characteristics) restricts the accuracy of information source identification, which is also the challenge faced by this paper.

Considering the research status of information source identification in OSNs, and that in practice, there is a situation of multi-source diffusion in information publishing in social networks, there is more than one source node that needs to be traced, so this paper studies the information source identification problem of multi-source diffusion in OSNs and proposes an information source identification algorithm based on MaTC.

The main contributions of this paper are as follows:

(1) Using the method of clustering to divide communities, the multi-source problems are decomposed to improve the identification accuracy.

(2) The number of source nodes is estimated by the inflection point method based on the change rate of the clustering loss function, and the experiment proves that the estimated number of source nodes is close to the real number of source nodes.

(3) Based on multi-attribute topological clustering, considering the internal relations between nodes after partition, a topological clustering attribute index with multi-attribute structure is designed, and the traditional clustering method with a single attribute index is optimized.

The paper is organized as follows. Section 2 introduces related work. Section 3 describes our information source identification algorithm based on MaTC in detail, and Section 4 reports some preliminary experiments and performance evaluations, followed by conclusions in Section 5.

2. Related Work

Information source identification is very important, but there is no significant research work in this field. On the one hand, this problem belongs to the reverse process of information diffusion, which is difficult to analyze directly, so it is a kind of NP-hard problem. On the other hand, the dynamic nature of information dissemination and the complexity of network topology introduce challenges to identification. In 2011, Shah and Zaman made a breakthrough in research on information source inference in their work [1]. The authors put forward the concept of a rumor center and proved that the rumor center can maximize the likelihood estimator (MLE) under the condition of a tree graph network. Following this work, detection methods based on rumor centers have been studied further. Luo et al. extended the rumor center to the problem of inference of multi-source diffusion source nodes [2]. In reference [3], the authors studied the problem of information source identification under the condition that every infected node has a probability P of being discovered. References [4,5] studied the detection of information source nodes under a multi-sample observation infection graph and verified that multi-sample observation can significantly improve the correct detection probability. To solve the problem of source node detection, especially in the case of asynchronous propagation and a

For the discrete time information propagation model, some studies have proposed a method based on the most probable sample path to solve the single-source detection problem in the SIR model [7]. Luo et al. [8] improved the reverse infection algorithm (RI) [7] to the single Jordan center algorithm (SJC) and extended it to solve the multi-source estimation problem. In references [9–12], the authors proved that the Jordan center is the optimal source node estimator under the SI, SIS, and SIR models, respectively, for tree graphs. Zhou et al. [13] studied the identification under the SEIR model and also took the Jordan center as the source estimation. In reference [14], the authors proposed an approach for the detection of possible information diffuser backbones among different communities in a social network. The approach is based on a new centrality measure called disseminator centrality. In reference [15], the authors proposed a distance centrality method based on infection potential energy to identify the source. In reference [16], the authors proposed a combined network centrality approach (CNCA) to identify the information source.

Aiming at the problem of multi-source diffusion, the typical research approach is to propose a network partition algorithm to improve the accuracy and time efficiency of identification. In reference [17], the authors proposed a K-Center clustering algorithm to partition the network structure first, then detected the source nodes, and finally combined this with the network partition and the distance centrality to infer the source nodes. The authors of [18] proposed a cluster location algorithm (CL) and a cluster reverse infection algorithm (CRI) to re-detected the source of the tree network and the general network, respectively. In [19], a hybrid approach based on a community detection algorithm and a multi-attribute decision making technique (TOPSIS) was also proposed to partition the network. In addition to the dynamic age centrality defined in the literature [20], the transmission source is considered the oldest node in the infection graph, that is, the node that first joined the infection graph [21]. Comin et al. [22] thought that compared to the original contact network, the node degree of the infected network is more likely to remain unchanged as a local statistic, so it is necessary to remove the deviation caused by the node degree value from the original betweenness centrality index. Reference [23] identified the rumor sources based on the maximum votes on different measures of centrality.

In addition to identification algorithms based on the centrality of source nodes, heuristic algorithms based on a single-sample infection graph are also proposed. In references [24,25], the authors proposed an algorithm named NetSleuth based on the principle of minimum description length and thought that given the disease transmission model, the disease transmission network, and the transmission results, the real source should make the length from the description source to the propagation result shortest. Reference [26] proposed a dynamic message-passing algorithm (DMP) based on an approximate mean field of maximum likelihood estimation. In [27], a belief propagation algorithm (BP) is proposed, which is based on probabilistic models (such as Bayesian networks and Markov random fields) to calculate the source inference. Reference [28] considered the single-source inference problem of the heterogeneous infection time distribution under the SI model (that is, the infection time distribution of each node is different). In reference [29], the authors proposed a rumor source identification method based on the center of the type, which offers a highly efficient source identification with logarithmic approximation error in large networks.

All of the above are aimed at the completely observable infected network. When there is only a part of the observable infected network, it is necessary to try to obtain the status of other unknown nodes before tracing the source. In reference [30], according to some infected nodes observed, the authors used an integral method to detect recovered and unobserved infected nodes in social networks and then used the community cluster algorithm to solve the multi-source identification problem. In reference [31], Louni et al. proposed a two-stage identification method, in which the candidate clusters most likely to
contain sources are identified in the first stage, and source inference is carried out in the candidate clusters in the second stage. Pinto et al. [32] considered tracing the source by introducing observation nodes in any tree network. A summary of the studies in related work is given in Table 1.

Table 1. Summary of the studies in related work.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Core Methodology</th>
<th>Application Scene</th>
<th>Main Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>[13]</td>
<td>SJC</td>
<td>Jordan center</td>
<td>Single source</td>
<td>Jordan center; optimal infection process</td>
</tr>
<tr>
<td>[14]</td>
<td>BID</td>
<td>Disseminator centrality</td>
<td>Single source/multi-source</td>
<td>New centrality measure; information diffuser backbones; disseminator bridges</td>
</tr>
<tr>
<td>[15]</td>
<td>NRSE</td>
<td>Distance centrality</td>
<td>Single source</td>
<td>Network pruning; propagation centrality; Gaussian density values</td>
</tr>
<tr>
<td>[29]</td>
<td>RC</td>
<td>Types center</td>
<td>Single source</td>
<td>Method of types; heuristic-based approach; relative leaf counting</td>
</tr>
<tr>
<td>[19]</td>
<td>GTaCB</td>
<td>Multi-attribute decision making</td>
<td>Multi-source</td>
<td>Community detection; multi-attribute decision-making technique; greedy and hybrid approach</td>
</tr>
</tbody>
</table>

Most existing single-source diffusion studies take the nodes that meet a certain centrality index as the estimated source nodes, and the source nodes can only be obtained by calculating the centrality index values of all nodes in the source identification process. However, the number of source nodes is unknown in the multi-source diffusion. To study the source identification of multi-source diffusion, it is necessary to determine the number of source nodes first; secondly, there is only one information source in single-source diffusion, and the identification accuracy can reach a high level by using the centrality index. However, in multi-source diffusion, it is difficult to determine the information sources (there are multiple source nodes), and the identification accuracy is generally low. Usually, the identification effect is evaluated based on the error distance between the real source node and the estimated source node. Research on the source identification of multi-source diffusion is more complex than that of single-source diffusion. The existing studies on source identification algorithms for multi-source diffusion are relatively few, so this paper chooses to further study the source identification problem in multi-source diffusion. The solution is to partition the nodes in the network and then identify the estimated source nodes in each partition to transform the multi-source problem into a series of single-source problems. Therefore, this paper proposes an information source identification algorithm based on MaTC, which optimizes the accuracy of information source inference under the condition of partitioning the network.

3. Design of the MaTC Algorithm

Information source identification is an important means to control information dissemination in OSNs. When a rumor spreads on the network, it is often controlled by preventing the source node users from spreading the rumor or asking the source node users to issue a clarification. Therefore, finding these source nodes is the primary task. Existing research has low accuracy and computational efficiency. The spread of rumor information in OSNs is usually caused by multiple source nodes. For example, in social networks, some malicious users use marketing numbers with a large number of fans and spread rumors at the same time. This paper proposes an information source identification algorithm based on MaTC, aiming at solving the identification problem in multi-source diffusion.

Suppose that the diffusion network is $G = (V, E)$, where $V$ represents the set of infected nodes and $E$ represents the set of edges in a diffusion network. Assuming that the
prior probability of each node in the network as the source node is the same, a source node likelihood estimation is constructed based on Bayesian theory. The state information of all nodes in the diffusion network observed at time $t$ is $\mathbf{O}(t)$, and the likelihood estimation of the correct detection probability of specifying node $s$ as the source node is $P(\mathbf{O}(t)|s)$, as the following formula:

$$s = \arg \max_{s \in V} P(s|\mathbf{O}(t)) = \arg \max_{s \in V} P(\mathbf{O}(t)|s)$$ (1)

It is often complicated to calculate $P(\mathbf{O}(t)|s)$ directly. Many studies have changed from other angles to relatively simple algorithms based on node propagation sequences and thought about how to solve identification problems from other angles, such as a propagation center algorithm or a heuristic algorithm.

3.1. Algorithm Framework Design

To improve the identification accuracy of the algorithm, this paper considers using the idea of partition to divide the network, decomposes the multi-source diffusion identification problem into a series of single-source problems, and proposes an MaTC information source identification algorithm to solve the information source identification problem in large-scale networks. The algorithm consists of four parts:

1. Topological multi-dimensional attribute structure design;
2. Estimation of the number of source nodes;
3. Cluster partition network;
4. Identification and estimation of the source nodes.

The overall design of the algorithm framework is shown in Figure 1.

![Figure 1. The overall design of the MaTC algorithm framework.](image)

Given the network snapshot information, that is, that graph $G$ contains node state information and diffusion network topology (multi-source), first, the multi-dimensional clustering attribute structure is designed, and then the number of source nodes $k$ is estimated and calculated in the first step of the algorithm. The second step is to determine the number of clusters according to the estimated number of source nodes $k$, divide the network into $k$ partitions by clustering, and then the set of clustering partitions is $C = \{C_1, C_2, C_3, ..., C_k\}$. The third step is to estimate the source nodes of each partition and estimate that the source nodes set is $S$. By dividing the network, the multi-source problem is
decomposed into a series of single-source problems, which reduces the computational complexity of the algorithm and improves the identification accuracy. In the following, the overall framework design of the algorithm and each component of the algorithm are introduced in detail:

(1) Topological multi-dimensional attribute structure design: The key attributes of nodes are extracted by the information of nodes themselves and their topological network information, including basic attributes (such as node ID, degree, etc.), topological relationship attributes (such as edge ID, edge weight, etc.), combined attributes (such as shortest path, betweenness centrality, etc.), etc., and the multi-dimensional attribute structure of node topological clustering is designed.

(2) Estimation of the number of source nodes: K-Means clustering analysis is carried out on the diffusion network (the number of source nodes starts from 1), and the estimated number of source nodes suitable for the target diffusion network is determined according to the change rate of the loss function for evaluating the clustering effect.

(3) MaTC partition algorithm: According to the multi-dimensional attribute structure design, the node clustering attributes are defined as edge similarity and edge weight. When the two attributes are combined, the node correlation degree is calculated as the clustering attribute index. The improved K-Medoids clustering algorithm is used to partition the network and the number of source nodes estimated by the module is used as the clustering k-value.

(4) Identification and estimation of the source nodes: The Jordan centrality is used to estimate the source nodes, identify the source of each partition after clustering, and calculate the Jordan center in the partition.

The whole process of the identification algorithm in this paper is divided into three stages.

The first stage is the preparation stage of the identification algorithm:
(1) The topological data set of the nodes in popular social networks is selected.
(2) The original relational network data set is simulated by the SI model. After diffusion, the diffusion network topology is generated and the set of facts (including the nodes and the corresponding infection time) is generated at the same time, which is convenient for verifying the accuracy of the identification results.

The second stage is the main execution stage of the identification algorithm:
(1) First of all, the number of source nodes is estimated. The overall idea of the algorithm is to divide the network and then identify the source in each partition, so the number of partitions determines the number of estimated source nodes. Therefore, K-Means clustering analysis is used to calculate the number of partitions that are most suitable for the clustering effect of the target network, and the obtained k-value is the most suitable number of estimated source nodes.

(2) Second, the MaTC algorithm is used to cluster and partition the nodes in the network. To achieve a good clustering partition effect, the clustering attribute index is designed according to the multi-dimensional attribute structure of the OSN topology, including edge similarity and edge weight. The clustering algorithm adopts an improved K-Medoids algorithm, which is mainly improved in three aspects: specifying the k-value, selecting the initial center node, and clustering the attribute index.

(3) Finally, the source nodes are identified. After partitioning the complex initial diffusion network, single-source identification is carried out for each sub-partition of the network. The definition of the source node is the Jordan center, which has been shown to have the best identification effect in most studies and is suitable for different situations and scenarios.

In the third stage, at the end of the identification algorithm, the estimated source nodes set calculated by each partition is integrated and compared with the fact set generated in the preparation stage to verify the identification accuracy of the algorithm.

The overall process design of the identification algorithm based on MaTC is shown in Figure 2.
Figure 2. Overall flow chart of identification algorithm based on MaTC.
3.2. The K-Value Estimation Algorithm for the Number of Source Nodes

In this paper, the number of final source nodes is determined by the number of final partitions, so the problem of estimating the number of source nodes is transformed into the problem of estimating the k-value in the next clustering algorithm.

This paper evaluates the clustering effect of each round of a specific diffusion network by an iterative clustering analysis of the diffusion network and finally determines the specific appropriate k-value by finding the most suitable clustering effect for this network. In the k-value estimation module, this paper only needs to cluster the analysis in a certain network size of the appropriate number of clustering partitions, so it was decided to use the K-Means clustering algorithm to achieve the clustering analysis. The loss function is often used to evaluate the clustering effect in clustering algorithms. Among them, the sum of squares due to error (SSE) is a loss function commonly used to measure the clustering effect, and the SSE formula is as follows:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} dist(m_i, p)^2$$

(2)

where $C_i$ is the i-th partition, $dist(m_i, p)$ is the shortest path distance error between the node $p$ of the i-th partition and the centroid node $m_i$, and $SSE$ is the clustering error of all nodes. The smaller the SSE is means that the closer the nodes are to their centers, the better the clustering effect.

The clustering effect evaluation process for determining the k-value for the number of source nodes by the above content is as follows:

Step 1: The k-value of the clustering starts from 1, the K-Means clustering is carried out on the diffusion network, and the SSE value is calculated after one round of clustering.

Step 2: The k-value is increased by 1, and then the SSE value is calculated at this time after a round of clustering, and the change rate of the SSE value between this round and the previous round is calculated.

Step 3: Step 2 is repeated, and when the k-value reaches the preset range, it is time for Step 4.

Step 4: According to the SSE values after multiple rounds of clustering, a graph of the SSE change rate with the k-value is generated and the ‘inflection point’ is calculated, as shown in Figure 3, that is, a point where the SSE change rate is less than the threshold $t$, as the following formula:

$$\Delta SSE \leq t \ (0 < t \leq 0.05)$$

(3)

Step 5: The k-value is output. At this time, the generation of the k-value is the most suitable clustering effect of the target diffusion network, and this k-value is the optimal estimation of the k-value.

![Figure 3. Example of the relationship diagram between the SSE and the k-value.](image)
The optimal value k (k = 5) calculated by the above method is used as the most suitable clustering partition number in the identification algorithm, that is, the estimated number of source nodes, which will be used as the input of the known condition of the next module algorithm.

3.3. The MaTC Partition Algorithm

The purpose of clustering nodes in an information diffusion network is to aggregate a group of nodes with the same attributes to realize the partition of the network. In many studies, nodes in the network are often clustered by a single attribute index. The social network topology data set is usually made up of node information and correlation information between nodes. According to the information about the nodes themselves and the topological relationship between the nodes, the key effective attributes needed by the algorithm design are extracted, such as the basic attributes of the nodes, topological relationship attributes of the nodes, and multiple combined attributes. The basic attribute of a node refers to the basic attribute information of the node itself, such as node ID, node degree, neighbor node set, and so on. The node topological relation attribute is the basic attribute of the connection edge between nodes, including edge ID, edge weight, and so on. Multiple combined attributes refer to the attributes calculated by the combination of multiple basic attributes, such as edge betweenness centrality, edge similarity, and so on. This paper will design the multi-dimensional attribute structure of nodes in the identification algorithm based on multi-attribute clustering according to these attributes.

By analyzing the characteristics of the topological structure of the OSNs and the multiple attribute structure of the nodes, a new attribute index is designed so that the final clustering effect can reach the expectation, that is, the diffusion information received by the nodes belonging to the same partition may come from the same source node. Therefore, a clustering algorithm based on two combined attributes of edge similarity and edge weight is proposed.

First, the influence of the common neighbor nodes on the correlation between nodes is considered, such as the recommendation of common friends, the recommendation of products, etc. For example, if your neighbor recently paid attention to a user, go to your neighbor’s homepage or your homepage, and you will see push messages related to this user. Or, for example, this neighbor recently praised some users’ tweets, and when you frequently view relevant praise tips on this neighbor’s homepage, you will click out of curiosity to see this user’s homepage or the contents of these tweets, which means that it is very likely that you will be praised, paid attention to, or even forwarded. Generally speaking, friends or fans of a person in social networks also determine that person’s circle of interest or even their friends, to some extent.

The research starts from the common neighbor nodes of nodes in the network and considers two nodes in the network that are neighbors to each other (there are directly connected edges in the network). The more common neighbor nodes they have, the stronger the connection between the nodes, and the higher the probability of being affected by the same node, so they are more likely to belong to the same community. In this study, the number of common neighbors (friends) between nodes is regarded as one of the evaluation indices of whether two nodes belong to the same community, and the attribute index determined by the number of common neighbors between nodes is defined as edge similarity.

The neighbor nodes set of \( v \) (including the node itself) is defined as \( \Gamma(v) = \{ w \in V | (v, w) \in E \} \cup \{ v \} \). If there are directly connected edges between \( v \) and \( w \), the edge similarity \( \gamma(v, w) \) can be defined as follows:

\[
\gamma(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{|\Gamma(v)||\Gamma(w)|}}
\] (4)
It is worth noting here that the neighbor nodes set of the node includes the node itself, so the number of common neighbor nodes between the node and its neighbor nodes is at least 2.

Secondly, the influence of the connection edges between partitions on the correlation between nodes is considered. In social networks, the connection between users in the same community is relatively close, while the connection between users in different communities is less. The bridge edges are the key bridges connecting the two communities, and the number of the bridge edges is relatively small but very important. The dissemination of information between partitions must pass through these bridge edges. Therefore, the edge betweenness centrality of this kind of edge in the network is relatively high. Therefore, the edge betweenness is defined as the number of times passing through an edge \((v,w)\) in the set of the shortest path (Dijkstra) \(\sigma_{ij} = \{\text{dist}(i,j)|i \in V, i \neq j\}\) between any two nodes \(i\) and \(j\) in the network, as shown in the following formula:

\[
EB(v,w) = \sum_{i,j:v \neq i} \sigma_{ij}(v,w)
\]

(5)

The \(\sigma_{ij}(v,w)\) is the number of times that an edge \((v,w)\) acts as a betweenness edge on the shortest path between two nodes.

Assuming that these key bridge edges are disconnected, the information between partitions cannot be transferred from one partition to another, which is equivalent to cutting off the connection between two partitions connected by these bridge edges, and it is more likely to split into two independent partitions.

Starting from the bridge edge with high edge betweenness in the network, considering that the higher the edge betweenness, the greater the possibility that the two nodes connected by the edge belong to different communities, the weaker the connection between the two nodes may be, so the edge betweenness centrality is regarded as another attribute index of clustering. The edge weight is defined as the reciprocal of edge betweenness centrality, so that the higher the edge weight between two nodes, the smaller the edge betweenness, the stronger the connection between nodes, and the greater the probability that they are affected by the same source node (from the same partition). Therefore, the edge weight \(\omega(v,w)\) is defined as shown in the following formula:

\[
\omega(v,w) = \frac{1}{\sum_{i,j:v \neq i} \sigma_{ij}(v,w)}
\]

(6)

In the above, the relationship between nodes in social networks is analyzed from two angles: the relationship between common neighbor nodes and nodes between different partitions. In this paper, two clustering attributes proposed in Formulas (4) and (6), edge similarity and edge weight, are combined as attribute indices of the clustering partition algorithm, and the new combined attribute is defined as the node correlation degree, which represents the close degree of correlation between two nodes.

For two nodes with directly connected edges in \(G = (V,E)\), that is \((v,w) \in E\), given the balance coefficient \(\alpha \in (0,1)\), combined edge similarity, and edge weight attributes, the formula to define the node correlation degree between \(v\) and \(w\) is shown in the following formula:

\[
l(v,w) = \alpha \cdot \gamma(v,w) + (1-\alpha) \cdot \omega(v,w)
\]

(7)

The formula of the node correlation degree is generalized. For the node correlation degree of any two nodes (not necessarily adjacent) in \(G = (V,E)\), the node correlation degree value of each pair of nodes corresponding to each edge of the path in the shortest path between the two nodes is multiplied cumulatively (if there is no connection to the path between the two nodes, the correlation degree is 0), specifically as shown in the following formula:

\[
L(v,w) = \max_p \{\prod_{(i,j) \in p} l(i,j)\}
\]

(8)
where $p = \{(v,m),(m,n) \ldots (t,w)\}$ represents the set of all edges in the shortest path from $v$ to $w$, and $p'$ is the shortest path set of all paths like $p$ between $v$ and $w$, because there may be more than one such shortest path in the network.

The MaTC partition algorithm proposed in this paper is similar to K-Medoids in its basic idea of clustering. To improve the accuracy of clustering, the K-Medoids algorithm is improved. In the MaTC partition algorithm, the k-value is not given randomly, but the most appropriate estimated k-value is determined by calculation, which is introduced in detail in the previous section. At the same time, the clustering attribute index adopts the node correlation degree combining edge similarity and edge weight, while the attribute index based on Euclidean distance between nodes is commonly used in the K-Medoids algorithm. Furthermore, the research in this paper is based on the SI model, so even after multiple rounds of diffusion, nodes will always have the ability to infect other nodes as long as they are infected. The node with the largest degree in the network, that is, the node that infects the most neighbors, is more likely to be infected earlier and closer to the source node.

Therefore, in the MaTC partition algorithm, k nodes are not randomly selected as the initial central nodes, but the largest k nodes in the network are selected in turn to join the initial central node set.

The MaTC partition algorithm is described in Algorithm 1.

**Algorithm 1: MaTC partition algorithm ($G,k$)**

**Input:** $G$: social network; $k$: size of clusters

**Output:** $C$: clusters of the social network $G$

1. **Begin**
   1. initialize $V_{center} \leftarrow \emptyset$, $V_{surplus} \leftarrow V$ //\(V\) means all nodes from $G$
   2. $C = \{C_1, C_2, C_3 \ldots C_k\}$ //initialize each subset as null
   3. //MaTC partition algorithm
   4. calculate node similarity $\gamma(v,w)$ and node weight $\omega(v,w)$
   5. calculate $L(v,w)$ for each pair of nodes
   6. choose the node with top degree centrality in $G$ join in $V_{center}$, then remove it from $V_{surplus}$, and also remove this node and all connections related to it from $G$
   7. **for** $i = 1$ to $k-1$ **do**
   8. choose the node with top degree centrality in $G$ join in $V_{center}$, then remove it from $V_{surplus}$, and also remove this node and all connections related to it from $G$
   9. **end for**
   10. $C_i = \{i| i \in V_{center}\}$ //each node in $V_{center}$ initialize a cluster of $C$
   11. **for** each node $v$ in $V_{surplus}$ **do**
   12. find the node $u$ ($u \in V_{center}$) which has the maximum $L(v,u)$ with $v$
   13. $C_u = C_u \cup v$
   14. **end for**

**return** $C$

3.4. Single-source Identification Algorithm in Partition Based on Jordan Center

In the centrality identification algorithm, centrality algorithms based on network topology include degree centrality, betweenness centrality, tight centrality, and so on. The identification algorithm based on the communication center has the rumor centrality and
Jordan centrality. The algorithm based on Jordan center has been proven to have higher accuracy in the tree graph network structure. There is a high probability that the corresponding node is only a fixed distance from the real source node, and it has been extended to multiple scenarios. It is shown that the Jordan center is the best source node estimator in the SI, SIS, and SIR models of disease transmission. Therefore, in this paper, the node corresponding to the Jordan center in the network is defined as the source node. After clustering and partitioning the nodes in the network, the single Jordan center identification is carried out for the partition, and the Jordan center in each partition is calculated as the estimated source node.

Firstly, the node eccentricity is defined. Generally speaking, node eccentricity is calculated to calculate the diameter of the diffusion network with a certain node as the endpoint, that is, the distance between the nodes farthest from the node in the network. The expression of the node eccentricity is shown in the following formula:

\[
ecc(v) = \max_{u \in V} \text{dist}(v, u)
\] (9)

The \(\text{dist}(v, u)\) is the distance from node \(v\) to other nodes.

The Jordan center is defined as the node with the smallest eccentricity in the network. Calculate the eccentricity of all nodes in the network, and the Jordan center is the node with the smallest maximum value of the shortest path with other nodes, so the expression formula of the Jordan center as the estimated source node is shown in the following formula:

\[
\hat{s} = \arg\min_{v \in V} \ecc(v)
\] (10)

The single-source identification algorithm in the partition based on the Jordan center is described in Algorithm 2.

**Algorithm 2 Source Identification (C)**

**Input:** \(C\): clusters of social network \(G\)

**Output:** \(S\): set of estimated source nodes

1. **Begin**
2. initialize \(S\leftarrow\emptyset\) //set of the initial source node
3. \(C = \{C_1, C_2, C_3 \ldots C_k\}\) //initialize those subsets of \(C\) as \(\emptyset\)
4. //source identification
5. for \(C_i\) in \(C\) do
6. calculate eccentricity for every node in cluster \(C_i\) as follows,
7. \(e\ecc(v) = \max_{u \in C_i, u \neq v} \text{dist}(v, u)\)
8. find the Jordan center for \(C_i\)//\(C_i\) means the ith cluster as follows,
9. \(\hat{s} = \arg\min_{v \in C_i} e\ecc(v)\)
10. \(S = S \cup \hat{s}\)
11. end for
12. **Return** \(S\)

**4. Experimental Analysis**

To verify the effectiveness and feasibility of the proposed algorithm, in this paper, different identification algorithms from existing research are selected to compare with identification algorithms based on MaTC. At the same time, the performance of these identification algorithms in performance evaluation indicators is analyzed.
4.1. Introduction of the Experimental Data Sets

This paper chooses three data sets of different scales for experiments, namely the scale-free network data set, the power grid data set, and the Wiki-Vote data set. The scale-free network is a random network with uneven distribution. In this kind of network, there are a few nodes with a large number of edge connections, while most nodes have only a few edge connections. For example, a few users in social networks have a large number of fans, so it is a data set close to the distribution of real social networks. The power grid data set and the Wiki-Vote data set are real-world social network data sets. The power grid data set is an undirected and unweighted network; it is a topological data set of power grid distribution in the western states of the United States, which comes from the website (http://www-personal.umich.edu/~mejn/netdata/), and the Wiki-Vote data set comes from the website of Stanford University SNAP Library (http://snap.stanford.edu/). It is a directed unweighted graph itself, but in the experiment, this paper transforms it into an undirected graph, removing repeated edges for reuse. This paper accessed the data sets on 6 October 2023. The specific data information for these three networks is given in Table 2.

Table 2. Experimental data sets.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Nodes</th>
<th>Edges</th>
<th>Maximum Node Degree</th>
<th>Network Diameter</th>
<th>Average Path Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale-free</td>
<td>500</td>
<td>996</td>
<td>48</td>
<td>7</td>
<td>3.837</td>
</tr>
<tr>
<td>Power grid</td>
<td>6599</td>
<td>19,778</td>
<td>6595</td>
<td>3</td>
<td>1.999</td>
</tr>
<tr>
<td>Wiki-Vote</td>
<td>7066</td>
<td>100,736</td>
<td>1065</td>
<td>7</td>
<td>3.248</td>
</tr>
</tbody>
</table>

4.2. Introduction of the Comparison Algorithm

The identification algorithm based on MaTC proposed in this paper is implemented based on the above data sets, and five comparison algorithms are selected for comparative experiments, namely BC, CC, DC, SJC, and RD:

(1) The BC (betweenness centrality) algorithm calculates the betweenness centrality of the nodes in the network, that is, the number of times a node acts as an intermediate node on the shortest path between any pair of nodes in the network, so that the higher the betweenness value, the closer the correlation between the nodes and other nodes, and the greater the possibility of being a source node. Therefore, the idea of the BC identification algorithm is to calculate and select the first k nodes with the highest betweenness centrality as the estimated source nodes, as shown in the following formula:

\[ BC(v) = \sum_{i,j \in V, i \neq j} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \]  \hspace{1cm} (11)

\[ \hat{v} = \text{arg max}_{v \in V} BC(v) \]  \hspace{1cm} (12)

(2) The CC (closeness centrality) algorithm takes the average distance between nodes as the measurement index, and the sum of the distances from a node to all other nodes in the network is the smallest, indicating that the node is closest to all other nodes, so the more likely it is to be the source node. Therefore, the idea of the CC identification algorithm is to calculate and select the first k nodes with the highest tight centrality as the estimated source nodes, as shown in the following formula:

\[ CC(v) = \frac{1}{\sum_{i,j \in V, i \neq j} \text{dist}(i,j)} \]  \hspace{1cm} (13)

\[ \hat{v} = \text{arg max}_{v \in V} CC(v) \]  \hspace{1cm} (14)

(3) The DC (degree centrality) algorithm takes the degree of nodes as the measure standard, that is, the more neighbors a node has, the wider the relationship, indicating that the greater the centrality of a node, the more likely it is to be the source node.
Therefore, the idea of the DC identification algorithm is to select the first k nodes with the largest degree in the network as the estimated source nodes in turn, as shown in the following formula:

$$\hat{s} = \arg \max_{v \in V} d(v)$$  \hspace{1cm} (15)

(4) The SJC (single Jordan centrality) algorithm directly calculates the node eccentricity in the network and then calculates the Jordan center in the network according to the node eccentricity. The idea of the SJC identification algorithm is to calculate and select the first k nodes with the largest Jordan centrality in the network as the estimated source nodes in turn, as shown in Formulas (9) and (10).

(5) The RD (random) algorithm randomly selects k nodes in the network as estimation source nodes.

4.3. Performance Evaluation Index

For the identification algorithm of multi-source diffusion, the common performance evaluation indicators used to test the identification effect of the algorithm include the correct detection rate and the average error distance.

The final result of the identification algorithm proposed for the multi-source diffusion problem in this paper is to obtain a set of source nodes containing k estimated source nodes. To verify whether these estimated source nodes are real diffusion source nodes, this paper uses the correct detection rate as an evaluation index to test and analyze the identification accuracy of the identification algorithm more intuitively. The correct detection rate takes the intersection of the estimated source nodes set $S^*$ and the real source nodes set $\hat{S}$ as the evaluation object, that is, it is defined as the ratio of the number of the real source nodes contained in the estimated source node set to the number of real source nodes. The specific calculation publicity is shown in the following formula:

$$Cdr = \frac{|S^* \cap \hat{S}|}{|S^*|}$$  \hspace{1cm} (16)

It is worth noting that, at present, the probability of directly detecting the real source nodes is generally low in existing research. In this paper, the correct detection rate is used to evaluate the performance of the algorithm, mainly to verify whether the proposed algorithm has efficient identification detection capability and to verify the advantages of the proposed algorithm compared with other algorithms.

The average error distance evaluates the identification performance of the algorithm by calculating the average distance error between the nodes in the estimated source nodes set $\hat{S}$ and the nodes in the real source nodes set $S^*$. It is worth noting that, in this paper, the distance refers to the shortest path length in the network topology. The specific calculation publicity is shown in the following formula:

$$Aed = \frac{\sum_{u \in \hat{S}} \min_{v \in S^*} \text{dist}(u,v)}{|\hat{S}|}$$  \hspace{1cm} (17)

The smaller the average error distance between the estimated source node and the real source node, the closer the estimated source node is to the real source node, and the smaller the range in which the real source node is determined.

4.4. Experimental Results and Analysis

In this paper, the real data set is used to simulate the generated data set after diffusion of the SI model. According to the identification algorithm based on MaTC in this paper, the whole experiment is divided into three parts, namely, the number of source nodes estimation experiment, the MaTC clustering partition experiment, and the source nodes estimation experiment. The experiment used the data set in Table 2: the diffusion data set was generated by three real data sets at different infection rates $(r = 0.3/0.5/0.7)$, the number of diffusion source nodes was set to 5, and the initial value of $\alpha$ is 0.5.
Firstly, the estimation algorithm for the number of source nodes is verified by experiments. On different data sets (Table 2), the K-Means clustering analysis is used to estimate the number of source nodes in the diffusion network. The test k-value changes from 1 to 10, and the total test times are 100 times. The experimental results are shown in Figure 4.

![Figure 4. Estimation result of source nodes in three networks: (a) the experimental results of estimating the number of source nodes on the data sets generated by the scale-free data set; (b) the experimental results of estimating the number of source nodes on the data sets generated by power grid; (c) the experimental results of estimating the number of source nodes on the data sets generated by Wiki-Vote networks.](image)

The experimental results show that, according to the test results of data sets generated by three different networks under different diffusion infection rates, the probability that the algorithm in this paper can accurately estimate the number of source nodes is relatively high, basically staying above 50%, and the rest of the estimated values fall close to the real source values. For the diffusion network with the same data set under different infection rates, the experimental results of the number of source nodes estimation also show similar proportional changes. The estimation algorithm for the number of source nodes proposed in this paper has certain references for estimating the number of source nodes in the network.

Then, the network clustering partition experiments are carried out based on the MaTC clustering algorithm. On three data sets with $r = 0.5$, the MaTC partition algorithm is used to cluster the diffusion network. The effects of the three diffusion networks under the MaTC clustering partition are shown in Figure 5. It can be seen that after the clustering partition, the nodes in the network are generally modular.

![Figure 5. MATC clustering partition results: (a) MATC clustering partition results of scale-free data set; (b) MATC clustering partition results of power grid data set; (c) MATC clustering partition results of Wiki-Vote data set.](image)
Finally, the identification performance of the identification algorithm based on MaTC is verified by experiments. On different data sets (Table 2), the Jordan center is used to carry out identification experiments on each network partition based on the MaTC partition results, and the performance of the algorithm under two indicators of a correct detection rate and average error distance is evaluated in three data sets.

(1) Experimental results on the scale-free network data set

Figure 6 shows that when $r = 0.3$, the correct detection rate of the MaTC algorithm can reach about 20%, and the correct detection rate of other identification algorithms is 0, except for the RD algorithm. When $r = 0.5$, the correct detection rate of all the algorithms is 0. As mentioned earlier, it is difficult to find the real source nodes through the identification algorithm, and it can only be as close to the real source nodes as possible, so the correct detection rate in the identification algorithm is generally very low. When $r = 0.7$, the nodes and edges in the network are relatively dense. Both the MaTC algorithm and the SJC algorithm proposed in this paper have good identification detection effects, and the correct detection rate is about 20%.

![Correct detection rate of the scale-free network](image)

**Figure 6.** Correct detection rate of the scale-free network.

Figure 7 shows that under different infection rates, the average error distance between the source nodes and the real source nodes estimated by the identification algorithm proposed in this paper is smaller than that of other algorithms: the error distance is about 2. It shows that the MaTC algorithm can reduce the range of real source nodes and approach the real source nodes compared with other algorithms. Second, the average error distance of the CC algorithm is also better than that of other algorithms, and the error distance is kept around 3.
(2) Experimental results on the power grid data set

Figure 8 shows that when $r = 0.3$ and $r = 0.7$, the identification performance of the MaTC algorithm is relatively stable, and the correct detection rate reaches 20%. At the same time, the correct detection rate of the SJC algorithm is also better, but when $r = 0.5$, the correct detection rate of the MaTC algorithm and other algorithms is 0, and the correct detection rate of the SJC algorithm reaches 20%.

Figure 9 shows that the average error distances of the BC, CC, and DC algorithms are almost the same under different spread infection rates. When $r = 0.3$, the error distances of these three algorithms are lower than those of $r = 0.5$ and $r = 0.7$. The MaTC algorithm presented in this paper performs better and the basic error distances are kept within 2, followed by the SJC algorithm.
(3) Experimental results on the Wiki-Vote data set

Figure 10 shows that the correct detection rate of almost all identification algorithms on the Wiki-Vote data sets generated by different diffusion infection rates is 0. As mentioned above, when the number of nodes in the network is large, the probability of directly detecting real source nodes is almost 0, and identification can only be infinitely close to real source nodes, thus narrowing the identification range. However, when $r = 0.3$, the correct detection rate of the MaTC algorithm can reach about 10%, and when $r = 0.7$, the correct detection rate of the RD algorithm can reach 20%, but the performance of the RD algorithm is unstable.

Figure 11 shows that the average error distance of the MaTC algorithm is slightly better than other comparison algorithms under different infection rates on the whole, and when $r = 0.7$, the average error distance of the MaTC and the central identification
algorithm is lower than that of \( r = 0.3 \) and \( r = 0.5 \). The performance of the BC, CC, DC, and RD algorithms is similar to that of the MaTC algorithm, while the performance of the SJC algorithm is relatively poor.

![Figure 11. Average error distance of the Wiki-Vote network.](image)

In summary, under the performance index test of the correct detection rate and average error distance, the MaTC identification algorithm proposed in this paper is slightly better than other comparison algorithms. The correct detection rate index of MaTC has outstanding advantages on the scale-free and power grid data sets, that is, it is very possible to identify the real source nodes in the identification process, and its performance on the Wiki-Vote data set is not much different from other good algorithms. In the performance index of average error distance, the identification result is close to the distance of the real source nodes and has a high probability of keeping in a fixed range, which is better than all other comparative identification algorithms. On the power grid data set, the performance of the SJC algorithm and MaTC algorithm is better than other algorithms, but on the other two data sets, the performance of the SJC algorithm is poor. The other BC, CC, and DC algorithms are not outstanding on different data sets, and the overall performance of these algorithms is not much different for diffusion data sets with different infection rates. The performance of the RD algorithm can occasionally perform well, but the performance of the algorithm is not stable. Therefore, in comparison, the proposed MaTC identification algorithm is slightly better than other comparison algorithms in terms of the correct detection rate and average error distance.

5. Conclusions

In this paper, an identification algorithm based on MaTC is proposed. The algorithm aims to solve the identification problem of multi-source diffusion in OSNs. The multi-source problem is decomposed into a series of single-source problems. The algorithm is divided into three stages. Firstly, the number of source nodes is estimated according to the network scale. Secondly, the nodes in the network are clustered by using edge similarity and edge weight as clustering attribute indices, and the network is clustered to form multiple sub-network partitions. Finally, Jordan centrality is used for the single-source identification of each network partition divided by clustering.

Based on the scale-free network data set, the power grid data set, and the Wiki-Vote data set preprocessed by the SI model, identification experiments are carried out for two performance evaluation indices, namely, the correct detection rate and average error distance, and they are compared with the identification comparison algorithms commonly
used in existing research. The experimental results show that the identification algorithm based on MaTC proposed in this paper is slightly better than other comparison algorithms in performance.

Identification research relies on real data sets and may not apply to another scene. Therefore, the complexity of different data sets and networks often hinders the development of identification research, and identification research usually needs to obtain the fact set of information diffusion to verify the effectiveness of identification research. For the multi-source diffusion problem, the real data sets used in this paper are limited, and only three types of network distribution scenarios are considered, which fails to cover the general identification scenarios. Therefore, the identification effect of the algorithm in another case may not show advantages. In addition, in the research of multi-source problems, the estimation of the number of source nodes is also a research focus, because the number of source nodes will directly affect the accuracy of the identification results.

**Author Contributions:** M.D., Y.L., and Z.T. contributed to the study conception, the design of the experiments, and the paper’s structure. M.D. and Y.L. performed the experiment analysis and wrote the first draft of the manuscript. M.D., Z.T., and B.Z. participated in the revision and proofreading of the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the National Key Research and Development Program of China under Grant No. 2023YFC3306201, the Fundamental Research Funds for the Central Universities No. N2317004, and the National Natural Science Foundation of China No. 61772125.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**References**


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.