E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 03-06-2023
Accepted: 10-07-2023

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 1; Issue 1; 2023; Page No. 560-564

Social networks and representation of graph theory

[1]Satish Chandra Chouhan, [2]Dr. Savita Tiwari and [3]Dr. Ashfaq Ur Rahman

[1]Research Scholar, Madhyanchal Professional University, Bhopal, Madhya Pradesh, India
[2-3]Department of Mathematics, Madhyanchal Professional University, Bhopal, Madhya Pradesh, India

DOI: https://doi.org/10.5281/zenodo.13234741

Corresponding Author: Satish Chandra Chouhan

Abstract

In this paper, we introduce the ComTector (Com-munity Detector) algorithm, an improvement over previous methods for detecting communities in massive social networks. Evaluation of the spreading impact, description of the node's location, and identification of interaction centralities may all be accomplished using the identification methods of spreading influence nodes. Recent advances in the field of social network influence node identification algorithms are summarized in this review, with an emphasis on the contributions from physical viewpoints and approaches, such as algorithms based on microstructure, algorithms based on community structure, algorithms based on macrostructure, and algorithms based on machine learning.

Keywords: Social networks, representation, graph theory, ComTector, community structure

Introduction

A group of points on a plane or in space connected by line segments that either meet at two points or join to themselves makes up a graph.

A two-step finite graph G = (V(G), E(G)). There are two sets of elements in a graph: the vertex set (V) and the edge set (E). The former contains non-empty sets of items called vertices, while the latter might include potentially empty sets of elements called edges.

A pair of vertices, known as the end points, are connected to every edge e in E.

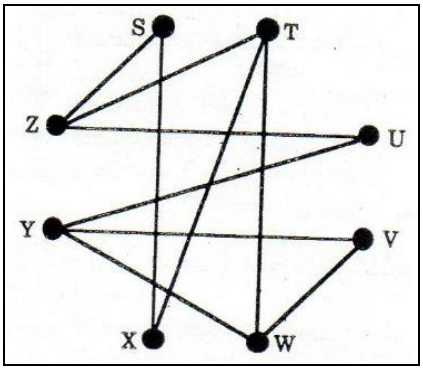

Fig 1: Diagrammatic Representation of a Graph

The set of vertices is represented by V = {S, T, U, V, W, X, Y, Z}, and the set of edges is E, which has ten edges that are attached to the unordered pair of vertices.

The following elements are ordered from most to least significant: (S, X), (S, Z), (T, W), (T, X), (T, Z), (U, Y), (U, Z), (V, W), (V, Y), (W, Y)

Keep in mind that any edge in a graph may have the same two end vertices-that is, a vertex u can be connected to itself by an edge-because this is inherent in the graph definition. The term for this kind of edge is a loop.

To further clarify the above concept, we will now provide an example.

Example: Show that G is equal to (V, E) in cases

$$V = \{a,b,c,d,e\},\ E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}$$

What follows is a formula for the edges' termini:

$$e_1 \leftrightarrow (a,b), e_2 \leftrightarrow (b,c), e_3 \leftrightarrow (c,c), e_4 \leftrightarrow (c,d),$$
$$e_5 \leftrightarrow (b,d), e_6 \leftrightarrow (d,e), e_7 \leftrightarrow (b,e), e_8 \leftrightarrow (b,e).$$

Following this, we may graphically depict G as seen in Figure 2.

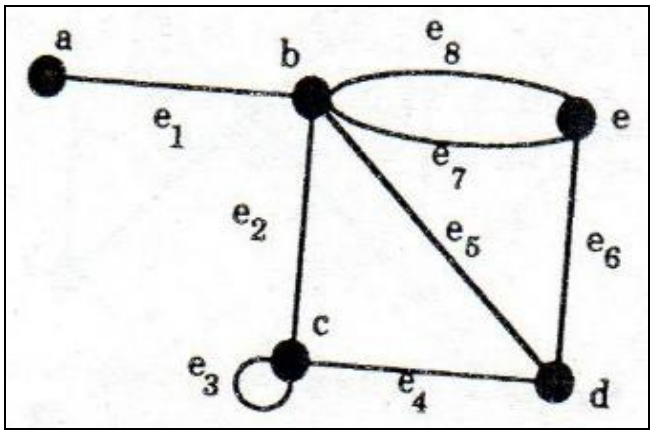

Fig 2: A Graph G with five vertices and eight edges

Literature Review

Manoj Kumar Srivastav *et al*. (2015) [1] In a social network, individuals or groups are "nodes" that are linked to each other through various forms of interdependency, such as friendship, kinship, shared interests, financial transactions, preferences (both positive and negative), or connections based on beliefs, knowledge, or status. The concept of nodes and ties, which are also known as edges, linkages, or connections, allows social network analysis to see interpersonal interactions through the lens of network theory. In a network, nodes are the nodes themselves, while ties are the connections between them. In many cases, very intricate graph-based structures are the end outcome. The connections between the nodes might take several forms. From the level of families to that of countries, social networks are at work, and they have a significant impact on problem-solving, organizational dynamics, and individual achievement, according to studies in a variety of academic disciplines. An elementary definition of a social network would be a diagram showing the relationships (like friendships) between the nodes in the network. Social capital, or the benefit that a person derives from their social network, may also be quantified via the network. The writers set out to find a mathematical model that might account for social networks in this study. The current analysis will point researchers in the right direction on how to enhance the routes of social networks.

Tasleem Arif (2015) [2] Analyzing the relationships and patterns of interaction among the people that make up a social network is where social network analysis really shines. There are two types of online communities: informal (like those seen on social media) and official (like those found in academic institutions). Underlying data defines distinct network properties in each of these networks. It may be impossible to dissect the whole network using traditional methods, considering the scale and variety of these networks. Visualizing social networks allows for a clear and succinct portrayal of complex systems. In order to statistically characterize different network properties, social network visualization technologies depend significantly on quantitative aspects. These characteristics, which are also known as social network metrics, were built on top of commonplace mathematics. The goal of this article is to provide a general introduction to the many metrics used for social network research. Academic social networks also benefit from an explanation of these measures and why they are important.

Chandra Prakash *et al*. (2022) [3] In light of the present situation, some observers are starting to pay more attention to online media. Online media platforms have the capability to create vast quantities of data on the client side. There are a plethora of mining tasks supplied by online media mining services, which helps businesses keep up with the data their users generate. The consumer may build their own local area at their leisure from among several long-distance interpersonal conversation places. The material found on the World Wide Web currently makes up a significant chunk of the whole virtual universe. The reason for this is the high number of users who actively engage in various online groups and keep their own accounts. One defining feature of digital media is the ease with which users may find new communities. This step is similar to the one in data mining called clustering. On the other hand, online media mining may make use of a different identification technique: local area by impact. Much effort has gone into local area recognition using this as a foundation, but it remains crucial. Recognizing locals who are making effective use of Leverage is the main goal of the event. Community detection also cares a lot about scalability and community quality. When compared against competing algorithms, several of these methods outperform the competition and scale well in large networks. With the use of Twitter's social data, we compared and contrasted the algorithms. Hence, the algorithms' scalability in the huge network as measured by the evaluation parameters has been shown. The fact that we have thoroughly tested the algorithm on a large social network is one factor that distinguishes our thesis from others.

Fatemeh Salehi Rizi (2021) [4] With millions of members all around the globe, online social networks are a treasure trove of user data. However, evaluating these networks is fraught with difficulty and expense because of their complicated structure and sparsity. Vector embeddings, which are low-dimensional representations of networked data, have recently been replaced by graph embedding. To make network analysis easier and faster, these representations are fed into pre-built machine learning algorithms. In light of the critical nature of social network research, this thesis seeks to investigate three-way graph embedding for social networks. We begin by encoding the structural feature of users' personal networks, also known as ego networks, with an emphasis on micro-level social networks. In assessment tasks where the success is dependent on the relational data provided by immediate neighbors, these representations are used. For instance, structural information from neighbors in social networks is necessary for both social circle prediction and event attendance inference. Second, we consider topological qualities as a means of evaluating vector embedding content. There are two possible ways to explain this: 1) an approach for learning to rank where the model weights show which qualities at subgraph level are important (ego networks), 2) a regression model that directly estimates the statistical aspects of vertex level networks. Lastly, we suggest enhancing graph embedding to include more information from social networks or signs.

Sentiment connections, which users create when they share their opinions and thoughts about other people on social media, exist alongside social links. Our goal is to create a single objective function that can simultaneously capture the emotion and social connection semantics. We also provide a stacking autoencoder multi-task learning framework for attribute-label networks, where an adaptive loss weighting layer automatically assigns the weights of the learning tasks. Pankti Joshi and Sabah Mohammed (2020) [5] With the widespread distribution of social media material, social network analysis has emerged as a crucial field of study. Specifying the user's impact by defining the directed connections in social media determines the flow of information. Data management is complicated for a number of reasons, including the sheer volume of data and the lack of organization in most forms of information exchange. Problems like constructing networks from unstructured data, deducing information from the system, and assessing a network's community structure may be effectively tackled with the help of Graph Analytics. The goal of the suggested method is to identify Twitter data influencers using the follower's and retweet's linkages. In order to identify the communities of discourse and influencers inside the Twitter network, a number of graph-based algorithms are applied to the gathered data.

Discovering social network influence nodes

The goal of the social network influence node identification challenge is to locate nodes with the potential to significantly alter the network topology or to significantly increase the rate and breadth of information dissemination. To be more precise, there are two subtasks that may be assigned to the identification of spreading influence nodes: node ranking and influence maximization. Sorting nodes from most to least influential based on evaluation-derived spreading influence ratings is what's known as the "node ranking task." $f(.)$ of nodes that disseminate influence. For a given fixed-size k, the goal of the influence maximization issue is to identify a collection of seed nodes S that maximizes influence.

MSB algorithms

Nowadays, in the age of big data, social networks like Weibo are known for their extensive and complex relationships.

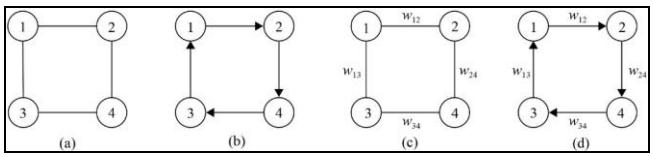

Fig 3: Diagram of four types of networks: (a) undirected and unweighted network; (b) directed and unweighted network; (c) undirected and weighted network; and (d) directed and weighted network

in between those using the service. Finding spreading influence nodes in such networks directly utilizing the complete network's structural information is an inefficient and expensive ordeal. Researchers are increasingly trying to identify spreading influence nodes using just micro-level structural information in order to build efficient identification algorithms that can be used to large-scale social networks. Information will quickly reach the percolation cluster regardless of the global structure of the network if the number of individuals influenced by spreading sources is greater than a small characteristic number; otherwise, it will be contained within a local area. This is the nucleation behavior of the spreading process that they discovered. Many MSB algorithms with low complexity and reasonably good accuracy have been suggested during the last few decades.

Among MSB algorithms, degree centrality (DC) stands out as the most basic. It uses the number of first-order neighbors of a node as its spreading impact, denoted as

$$DC(i) = \frac{k(i)}{n-1},$$

Where is the total number of nodes in the network and k (i) denoting the degree of node i. Everyone knows that when people post things online, it could impact not only their followers but also their friends' friends. It is not uncommon for people to have an equal number of followers on social media websites. Taking into account only the number of nodes that are directly linked might be quite simplistic in certain instances. Figure 3 shows an example where the DC approach will disregard the difference in the number of second-order neighbors, even if the degree centralized Ness of nodes 1 and 2 are equal.

CSB algorithms

According to this theory, Zhao *et al*. (2014) [10] found that combining centralities with an index of the number of communities linked to a node helped find spreading impact nodes, even when using a single centrality alone would have missed them. But this approach can be unstable since various community identification techniques might alter a network's community structure (Palla *et al*., 2005; Newman, 2006; Pan *et al*., 2010; Tang *et al*., 2016). Taking into account the community's size and distribution of neighbors, Zhao *et al*. (2014) [10] suggested community-based centrality (CbC), which is defined as

$$CbC(i) = \sum_{q=1}^{c} k_{iq} \frac{N_{Com_h}}{n},$$

Where k_{iq} indicates how many nodes in the community are neighbors of node i *q*, represents the overall count of towns, and N_{Com_h} is proportional to the population h. Researchers Tulu *et al*. (2018) presented the community-based mediator (CbM), which uses the Shannon entropy to quantify the spreading impact of nodes and defines the relations between a target node and nodes in its own community as well as nodes in other communities as its spreading influence, and uses this value as

$$CbM(i) = H(i) \times \frac{k(i)}{\sum_{j=1}^{n} k(j)},$$

$$H(i) = \left(-\sum p_i^{\text{in}} \log\left(p_i^{\text{in}}\right)\right) + \left(-\sum p_{ih_1}^{\text{ex}} \log\left(p_{ih_1}^{\text{ex}}\right)\right),$$

Where $H(i)$ represents the node's internal entropy as well as its exterior edge density i, and h_1 is the community. $p_{ih_1}^{\text{ex}}$ and p_i^{in} stand for the densities of the edges around node i and inside it, respectively. By including the node's and its neighbors' community structure information, Zhao *et al*. (2014) [10] enhanced the accuracy of closeness centrality (CC). We may express the enhanced CC (ICC) mathematically as

$$ICC(i) = CC(i)\frac{N_{Com_i}}{n} + \sum_{w \in W_i} \max\{CC(j)\}\frac{N_{Com_w}}{n},\ (j \in W),$$

Where $CC(i)$ denotes the CC of node i, and W_i is the group of neighborhoods that node i's linked to, with the exception of the community that contains the node i.

Performance evaluation metrics

In order to compare the ranking outcomes achieved by identification algorithms and diffusion models, evaluation metrics are required. In order to ensure that the algorithms used to identify spreading impact nodes are accurate, this procedure is necessary. Here we provide an overview of eight common metrics used for assessment in this area of study.

Spreading effect on average One way to gauge the effectiveness of various identification techniques is to compare the average impact of the top p×n (p ∈ [0, 1]) nodes. One such algorithm's average spreading effect (AvgSI) on the nodes it finds is

$$AvgSI = \frac{\sum_{v \in S} \sigma(v)}{p \times n},$$

Where S stands for the collection of seed nodes found using a particular technique, is the total count of nodes, and σ (v) is the spreading impact of node v.

Influence scale

An algorithm's ability to detect nodes and track their effect over time is reflected in the influence scale F (t), where t is defined as

$$F(t) = \frac{N_{I(t)} + N_{R(t)}}{n},$$

Where $N_{I(t)}$ and $N_{R(t)}$ stand for the total number of infected nodes and the total number of recovered nodes at time t.

Imprecision function

The loss-of-precision procedure $\varepsilon(p)$ comes into play to measure the disparity between typical spreading scales from top $p \times n$ routers that were detected by the detection method and the $p \times n$ the most effective dispersers found using diffusion models $(p \in [0, 1])$. How well the word gets out the amount of infected nodes, abbreviated as M_i, i, $\delta_{\text{eff}}(p)$ signifies the group with highest $p \times n$ nodes chosen based on the effectiveness of spreading, and δ_x (p) denotes the collection of highest $p \times n$ nodes that the spreading effect identification algorithm x finds. When x is specified with an imprecision, it is

$$\varepsilon_x(p) = 1 - \frac{M_x(p)}{M_{\text{eff}}(p)},$$

in which $M_x(p)$ and $M_{\text{eff}}(p)$ symbolize the mean impact of $\delta_x(p)$ and $\delta_{\text{eff}}(p)$, in that order. The value of imprecision becomes closer as $\varepsilon_x(p)$ The closer to zero the value is to the average effect of the set of nodes found using diffusion models, as measured by, and the less influential the set identified by.

Relative difference of spreading scales

The dissimilarity in spreading scales between two groups of very $p \times n$ defining the most significant nodes found using two distinct techniques for identifying node significance as

$$\Delta_y(p) = \frac{S_y - S_x}{S_x},$$

Where S_y represents the overall impact of the seed nodes found by algorithm y. Algorithm y's seed node identification has a larger overall impact than algorithm when $\Delta_y(p) > 0$.

Kendalls' τ correlation coefficient

A family known as Kendalls' τ One common way to evaluate the efficacy of identification algorithms is by calculating the correlation coefficient, which compares the degree of similarity between two sorted lists. Making the assumption that there are two ordered lists with n entries each, we have and. and (A_i, B_i) represents the ith element pair of A and B. When there are no pairings of A and B elements with different rankings, such $A_i > A_j$ and $B_i > B_j$ or $A_i < A_j$ and $B_i < B_j$, in this case, the two sets of elements are concordant; in the other case, they are discordant. The number of concordant and discordant pairings of two sorted lists determines the calculation of the Kendalls' τ correlation coefficient.

$$\tau = \frac{2(C-D)}{k(k-1)},$$

Where C and D are the numbers of pairings that are in agreement and in disagreement, respectively, and k is the sum of the numbers of items in both sets of orders. In terms of similarity between the two sorted lists, the closer the Kendall's τ coefficient is to 1, the better. The Jaccard correlation coefficient, which is provided as an additional

assessment measure, serves a purpose similar to that of the Kendalls' τ coefficient.

$$J_c = \frac{|X(c) \cap Y(c)|}{|X(c) \cup Y(c)|},$$

X (c) stands for the seed nodes chosen by the identification method Y (c), while represents the most influential nodes obtained by stimulating the diffusion model.

Monotonicity

A uniqueness metric for the ranks of nodes determined by the identification methods of spreading influence nodes is the monotonicity, which is expressed as

$$M(X) = \left(1 - \frac{\sum_{i \in I} n_i (n_i - 1)}{n(n-1)}\right)^2,$$

in which n_i is the total number of nodes, X is a technique for identifying nodes with a spreading impact, I includes all the unique values produced by applying X, and reflects the number of nodes allocated to rank i. M (X) may take on values between zero and one. As the value approaches 1, a smaller number of nodes are given the same rank.

Similar to monotonicity, the complementary cumulative distribution function (CCDF) characterizes the distribution of nodes in various ranks. Here is the mathematical definition:

$$CCDF(Z) = \Pr(Z > z) = 1 - CDF(z),$$

Where CDF(z) is the cumulative distribution function that represents the likelihood that the rank of the node is equal to or less than z.

Conclusion

A social network is a system that links several social components. Elements of society might be linked or unrelated. Elements and things in the environment may be either linked or unconnected, and they can be both similar and distinct. It is always possible to choose the most direct route when establishing relationships between various social factors and things. When trying to establish a link between seemingly unrelated items, the linear span of a social network could be useful. The shortest route method, as well as union and intersection procedures connecting many social networks, will be the focus of our future efforts. Additionally, the writers will make an effort to investigate how social networks might be impacted by finite state automation.

Reference

1. Srivastav M, Nath A. Study on mathematical modeling of social networks. International Journal of Emerging Technology and Advanced Engineering. 2015;5:611-618. ISSN 2250-2459.
2. Arif T. The mathematics of social network analysis: metrics for academic social networks. International Journal of Computer Applications Technology and Research. 2015;4:889-993. doi:10.7753/IJCATR0412.1003.
3. Prakash C, Agrawal C, Meena P. Graph theory based data extraction method for community detection on social media. International Journal of Innovative Research in Technology and Management. 2022;6(5):16-23.
4. Salehi Rizi F. Graph representation learning for social networks; c2021.
5. Joshi P, Mohammed S. Identifying social media influencers using graph based analytics; c2020.
6. Wang D, Li J, Xu K, *et al*. Sentiment community detection: exploring sentiments and relationships in social networks. Electronic Commerce Research. 2017;17:103-132. doi:10.1007/s10660-016-9233-8.
7. Fox W, Everton S. Using data envelopment analysis and the analytical hierarchy process to find node influences in a social network. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology. 2014;12:157-165. doi:10.1177/1548512913518273.
8. Singh S, Mishra S, Kumar A, Biswas B. Link prediction on social networks based on centrality measures. 2021. doi:10.1007/978-981-16-3398-0_4.
9. Nandini Y, Lakshmi T, Krishna Enduri M, Sharma H. Link prediction in complex networks using average centrality-based similarity score. Entropy. 2024;26:433. doi:10.3390/e26060433.
10. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, *et al*. JASPAR an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic acids research. 2014;42(D1):D142-147.