UDC 004.02

DETECTION OF COMMUNITIES IN SOCIAL NETWORKS

Yevheniy Nikitenko
Ph.D., Computer Systems, Networks and Cybersecurity Department, Associate Professor
ORCID https://orcid.org/0000-0002-9222-644X
E-mail:: ev.nikitenko@nubip.edu.ua

Yevhen Ryndych
Ph.D. 05.13.06 Information Technologies, National University "Chernihiv Polytechnic", Associate Professor
ORCID https://orcid.org/0000-0002-2723-4144
E-mail:: zkasterwork@gmail.com

Hoida Ivan
Student, Computer Systems, Networks and Cybersecurity Department
E-mail: vanyahoyda@gmail.com

Abstract. On e of the main threats is malicious programs (bots), fake accounts capable of imitating human behavior.
At the moment, bots create a lot of problems, both for ordinary users and for those who use social networks to conduct a marketing campaign or conduct social research. Using bot profiles in social networks greatly distorts information about the real benefits and interests of portal users. Therefore, it is necessary to determine which users of the social network are programmed, and to be able to divide the flow of data into that generated by bots and by humans.
The threatening scale of the use of social bots requires the creation of effective algorithms for their detection. Internet platforms and social services themselves are not too concerned about this problem. As a result, both ordinary users who organize various communities and companies that promote goods, brands, and services through social networks suffer. Thus, the task of recognizing malicious accounts in social networks and combating them remains relevant in the issue of cyber security.
The solution to the problem will be the development of network analysis methods that are designed to identify and classify communities in social networks, assess their connectivity, degree of trust, as well as develop effective algorithms for detecting malicious accounts. The purpose of the work is to develop an improved algorithm for detecting malicious accounts in social networks, which is based on the study of modern methods.
Keywords: social bots, detection, malicious accounts, social networks, cluster analysis.

Introduction

Currently, social networks are an integral part of most spheres of human life, integrating almost all existing Internet sources. They effectively structure users from political or religious views, interests and passions, affecting almost all segments of the population, and are a powerful tool for self-organization of both individual groups and society as a whole. Social networks, which unite 40% of the planet's population every day, have become not only a means of communication, but also a great source of information, entertainment hosting, a commercial platform with a set of effective tools for distributing services and goods. It is natural that interested persons seek to use such limitless potential for profit and to achieve their far from noble goals.

Over the past decade, social networks have grown significantly along with the speed of Web2.0 application development. Millions of people are registered on social networks such as Facebook, Twitter, LinkedIn, etc. For example, Facebook has about 2,94 billion users as of March 2023. As the number of users only increases, the complexity of detecting social communities also increases. Detecting and clustering data and users in social media communities is an important and challenging problem in creating effective marketing models in the changing and evolving social systems. Such marketing models are based on individual product purchase decisions influenced by friends and acquaintances. This is leading to new models that treat users as part of an online social network rather than traditionally as marketing individuals.

Social connections play an important role in determining user behaviour. For example, a user may buy a product that a friend of theirs has recently purchased. This phenomenon is called social influence and is used to study how strongly the action of one user can provoke others to take certain actions.

Literature review

Network science, which began its journey in the 1700s with Euler's Seven Bridges of Königsberg, has gone through many stages, including the emergence of graph theory, sociograms (graphical representations of social connections), and the emergence of social network analysis, culminating in a boom and establishment as a discipline. Only after some of the most recent important discoveries, such as the development of scale-free networks, did the first social networking sites appear within a decade, with Facebook accounting for 51% of the world's active online users [1].

Social network analysis focuses on connections rather than actors. As a rule, a social network is described by a graph or matrix of relationships (Figure 1) [2].

The main areas of social media research are as follows:

1. Identification of communities in online social networks.
2. Search for key nodes in any society.
3. Building a set of nodes that are used to spread influence on an online social network.

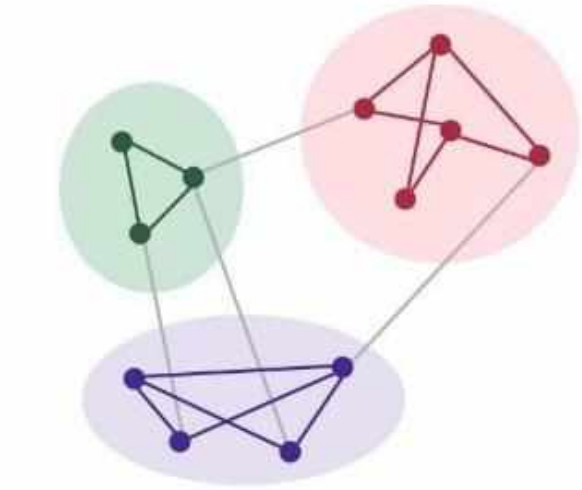

Figure 1. Example of a community structure

The significance of social networks is based on the fact that, on the one hand, they are the subject of socialisation of people, and on the other hand, they are the most powerful and accessible political, ideological and economic tool [3]. The methods of social media analysis include the following:

- methods of graph theory, in particular, the study of oriented graphs and matrices representing them, used to study the structural relationships of a network participant;
- methods of finding local properties of participants, for example, centrality, influence, position, belonging to certain subgroups;

- methods for determining the equivalence of participants, including their structural equivalence;
- probabilistic models, including Markov process models.

Works [4] andi [5] study the problem of spam bot detection using a popular social network as an example. The authors propose to distinguish ordinary users from malware using machine learning classification. Traditional classification algorithms are used for this purpose: decision trees, neural networks, and a naive Bayesian classifier. The number of subscribed and read users, as well as graph-oriented user relationships, were taken as features.

A new methodology for identifying communities was put forward by the authors of the article [6]. It consists in the construction and analysis of scalograms that reflect the interaction of users in a social network. Scalograms reveal a lot of hidden information about the nature of non-stationary processes. They are used in various fields: predicting the consequences of an earthquake, analysing earth movements, analysing the resistance of buildings to hurricanes, analysing the stability of bridges, etc.

Outline of the main material

This article focuses on community detection as one of the most promising tools for world-class marketing research. The task of researchers is to find groups of users who have similar interests or a high level of connections between them. Several existing community detection methods have been studied. The research extended the CNM (Clauset-Newman-Moore) [9] algorithm, to use the Jaccard similarity measure. First, a graph from the original network is output, and as a result, it is labelled as a similarity social network or a virtual social network. The method based on this algorithm shows that by pre-processing the original network, it is possible to obtain higher quality community structures.

Another algorithm for building strong communities online is the ECD-Jaccard algorithm (ECD - «Enhanced Community Detection»), which enriches a virtual social network with edge weights, then applies a quality-optimised version of the CNM algorithm to detect communities. This algorithm has shown that by pre-processing the original network, it is possible to obtain higher quality community structures.

A study was conducted to summarise the state of the environment affecting social influence in online social networks, and different approaches were evaluated, which are combined into an original categorisation system to understand commonalities and distinguish differences.

Clustering was used to separate groups from the total population of users. The advantages of cluster analysis are that it allows you to split objects not by one parameter, but by a whole set of features. In addition, cluster analysis, unlike most mathematical and statistical methods, does not impose any restrictions on the type of objects under consideration and allows you to consider a variety of source data of an almost arbitrary nature. Cluster analysis allows you to consider a fairly large amount of information and dramatically reduce and compress large amounts of information, making them compact and visual. Moreover, clustering can be used cyclically. In this case, the research is carried out until the required results are achieved. In this case, each cycle can provide information that can greatly change the focus and approaches to further application of cluster analysis.

The task of cluster analysis is to divide the set of objects X into m (m – integer) clusters (subsets) $Q_1, Q_2, \ldots, Q_m$ based on the data contained in the set G so that each object Q_j belongs to one and only one subset of the division. And objects belonging to the same cluster were similar, while objects belonging to different clusters were heterogeneous. The solution to the cluster analysis problem is a partition that satisfies some optimality criterion.

It is possible to define clustering in the context of a real social network by grouping people with high friendships internally and scattered friendships externally. With clustering, you can identify interest groups or communities that share common properties that can be used to learn about these groups and understand their behaviour. For example, Amazon provides users with recommendations based on their purchase history. Twitter also recommends new friends to follow members based on several factors, such as being a friend of the same user.

It is difficult to imagine dividing social media user accounts with a large number of unequal criteria into groups with two degrees of membership of 0 or 1. It is more natural to use a partial membership in the range from 0 to 1, which will allow users whose characteristics are on the boundaries between several clusters to belong to them with different degrees. Therefore, the fuzzy clustering method was chosen as the main method for dividing user accounts into groups. The steps of this algorithm are as follows:

1. The initial information for clustering is an observation matrix *U* of size $n \times k$ (1):

$$U = \begin{bmatrix} u_{11} & \cdots & u_{1k} \\ \cdots & \cdots & \cdots \\ u_{n1} & \cdots & u_{nk} \end{bmatrix}, \quad (1)$$

where *n* – is the number of users, *k* – is the number of features.

2. Using the matrix *U,* we can find the value of the fuzzy error criterion (2):

$$E^2(X, U) = \sum_{i=1}^{N} \sum_{k=1}^{K} U_{ik} \left\| x_i^{(k)} - c_k \right\|^2, \quad (2)$$

where c_k – «centre of mass» of the fuzzy cluster k (3):

$$c_k = \sum_{i=1}^{N} U_k x_i. \quad (3)$$

3. After regrouping the objects to reduce this value of the fuzzy error criterion, it is necessary to return to step 2 until the changes in the matrix *U* become insignificant.

The minimal spanning tree algorithm [8] first builds a minimal spanning tree on the graph and then sequentially removes the edges with the highest weight. The figure shows the minimum spanning tree obtained for nine objects.

By deleting the link labelled CD with a length of 6 units (he edge with the maximum distance), we obtain two clusters: *{A, B, C}* and *{D, E, F, G, H, I}*. The second cluster can be further divided into two more clusters by removing the edge *EF*, which has a length of 4.5 units (Fig. 2).

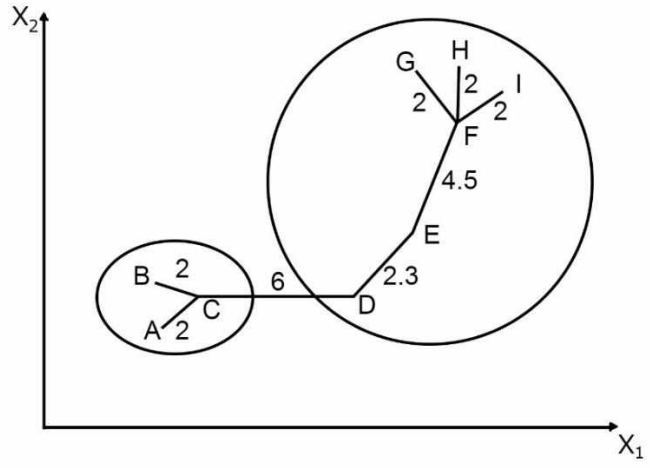

Figure 2. Illustration of the minimum spanning tree algorithm

In the n- dimensional metric feature space, the distance between two objects is considered to be a measure of the «similarity» of two objects.

The Mankowski distance family is a very common class of distance functions and can be represented as follows (4):

$$D(p_i, p_j) = w \sum (p_i - p_j) w, \quad (4)$$

where w - is a parameter with a value greater than or equal to 1. Based on the value of w, different distance functions can be represented, such as the Hamming distance (w = 1), the Euclidean distance (w = 2) and the Chebyshev distance ($w = \infty$). Other similarity measures are the cosine correlation measure and the Jaccard measure.

To investigate the proposed approaches, experimental information was used for both synthetic and popular real-world datasets. The modularity of Q and the normalised mutual information used previously in other experiments were used as evaluation metrics to show the performance and accuracy of the proposed algorithms. The maximum modularity Q_{max} was estimated using the CNM algorithm provided by Clauset et al. [9]. The normalisation of mutual information was also calculated using the formula of Danon et al. [7].

Based on the data in the developed software application, a network was generated and the results were studied.

For the Facebook network, the maximum modularity value for similarity-CNM outperforms the original CNM algorithm by more than 67%. And for the American football network, the maximum value of modularity for similarity-CNM exceeds the value obtained by the original CNM algorithm by more than 23%. The number of steps to achieve Q_{max} is lower for CNM than for similarity-CNM.

For Facebook, the number of steps when running the original CNM is 9879 steps, while in the similarity-CNM it is reduced to 8843 steps. For the American football network, the number of steps is also reduced from 108 steps for the original CNM to 100 steps for the similarity-CNM algorithm.

In the course of the study, software was developed that implements existing and proposed similarity-CNM and ECD-Jaccard algorithms to provide a better community structure. The similarity-CNM algorithm detects similarities between nodes and creates a corresponding virtual social network. Similarly, the ECD-Jaccard algorithm also calculates the similarity of nodes as a preprocessing step, but these values are then assigned as weights to the edge of the network, resulting in a weighted virtual social network, unlike the similarity-CNM approach, which has no weights. The CNM algorithm is then used to detect communities in both approaches. Experimental analysis shows that these preprocessing methods have an advantage over the original CNM algorithm in terms of community modularity.

Conclusions

The article proposes a solution relevant to the cybersecurity of social networks, which involves the development of a new method for detecting social communities. The following analytical and practical results were obtained in the course of solving the task:

1. An analysis of modern detection methods using various pattern recognition approaches and mathematical models. The results of testing existing methods were compared, which showed that most of them work with a certain amount of data. In this regard, there is a need to develop a new algorithm for detecting social communities, which allows to reduce the number of unnecessary user checks.

2. An analysis of clustering methods, taking into account their advantages and disadvantages, is carried out in order to select the most effective ones for solving the tasks set in the paper.

3. A method for improving the quality of account sorting with further analysis of the results and decision-making for individual user groups is proposed.

4. A combined method for tracking the behaviour of suspicious accounts in the network is proposed, which can significantly reduce the number of additional checks.

References.

1. C. K. Leung, F. Jiang, T. W. Poon, P.-É. Crevier. Big Data Analytics of Social Network Data: Who Cares Most About You on Facebook? // Springer, Studies in Big Data, – 2018. - vol. 27, pp. 1-17.

2. Stanford Large Network Dataset Collection [Online resource]. – URL: http://snap.stanford.edu/data/

3. Alaa Alden Al Mohamed, Sobhi Al Mohamed and Moustafa Zino. Application of fuzzy multicriteria decision-making model in selecting pandemic hospital site// Future Business Journal, Springer. – 2023, vol. 9, №14. – pp. 1–22.

4. Vaishali Chawla, Yatin Kapoor. A hybrid framework for bot detection on twitter: Fusing digital DNA with BERT // Multimedia Tools and Applications, – 2023. – vol.82, №20. – pp. 30831-30854.

5. Amit Pratap Singh, Maitreyee Dutta. An Efficient Classifier for Spam Detection in Social Network // International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019. – vol.9, №1. – pp. 2323–2328.

6. Manzoor Hussain. Digital divide in the use of social networking sites: a study of P.G. students (gender-wise) through scalogram analysis// International Journal of Research in Economics and Social Sciences (IJRESS). – Vol. 7. – Issue 9, September – 2017, pp. 527–536.

7. Andreas Kanavos, Nikos Antonopoulos, Ioannis Karamitsos and Phivos Mylonas. A Comparative Analysis of Tweet Analysis Algorithms Using Natural Language Processing and Machine Learning Models// 2023 18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP), 2023.

8. Christopher Braker, Stavros Shiaeles, Gueltoum Bendiab, Nick Savage, Konstantinos Limniotis. BOTSPOT: DEEP LEARNING CLASSIFICATION OF BOT ACCOUNTS WITHIN TWITTER // Internet of Things, Smart Spaces and Next Generation Networks and Systems, – 2021. – pp. 165–175.

9. Martijn Gösgens, Remco van der Hofstad and Nelly Litvak. The projection method: a unified formalism for community detection// Frontiers in Complex Systems, – 2024. – pp. 1-18.