Community Detection And Link Prediction In Social Networks

Rahul Thorat[#1], Krishna Tiwari[#2], Abhilash Kanaujia[#3], Ms. Seema Redekar[#4]

[#]Department of Information Technology,
South Indian Education Society Graduate School of Technology,
Nerul, Navi-Mumbai

[1]rahultit115@gst.sies.edu.in
[2]krishnatit115@gst.sies.edu.in
[3]bhilash.kanaujia@siesgst.ac.in
[4]seema.redekar@siesgst.ac.in

Abstract

Social Networks have been an important aspect in our lives since the beginning of time. These social networks include physical or digital networks formed over time due to certain similarities. Due to the nature of these networks being so large there can be many communities identified within these networks. Community detection is an important part in various fields such as marketing, recommendation systems, healthcare, detecting terrorist activities, Link prediction and many more. Community detection in social networks helps to understand the network structure and analyse the network properties. Link prediction in social networks is important with respect to identifying future links and connectivity in the network to predict the future of the network. It is used in spam detection, disease prediction, advertising, recommender systems and many more. One way to extract information from communities for the mentioned uses includes techniques like data extraction and data mining. Both these techniques require huge amounts of time and is difficult to perform on large datasets. To overcome this, the use of machine learning algorithms proves essential to saving time and accurately extracting important information from such large datasets. In this paper we aim to identify and predict social communities present in large networks and accurately identify the future links in these networks through various machine learning models and algorithms.

I. Introduction

In this digital age, there is a huge amount of heterogeneous data available which can be put to good use if used carefully. In the analysis of social network data, recognizing groups of similar nodes is a difficult task. Using this data leads to enhance the quality of community discovery. The analysis of social networks, mainly based on graph theories and sociological analysis, aims to study different aspects of these networks. The main factors are network detection, identity of influential actors, and the observe and prediction of the evolution of networks.

Datasets

The most simple and customary sort of datasets are spreadsheet or CSV format. Therein one file is organized as tables of rows & columns. There are other datasets which are stored in other formats, they need many files. Dataset are often a zipper file or folder having numerous data tables with stated data. Datasets are the fastest and most efficient way to work with logically grouped data in your application. We have taken the datasets of social network pages which shows the graph relates the knowledge objects within the save to a series of nodes and edges, the sides representing the relationships between the nodes. The relationships permit information that's saved to couple together directly and, in many cases, retrieved with one operation. The results are obtained on the dataset Fb-pages-food network from the network repository website.

Community Detection Algorithms

1. Girvan-Newman algorithm:

In this algorithm the edges between nodes which have the highest betweenness centrality are removed consecutively until two or more communities are establish. In this paper, this algorithm rule is employed for network community detection.

2. Triangle counting:

In this algorithm, a triangle refers to a set of three nodes or vertices of the triangle and all of the nodes are connected to other nodes. This technique is useful to classify a node into three general communities and is often used for fraudulent website detection.

3. K-1 colouring algorithm:

This algorithm assigns a colour to every node in the graph while making sure that the colours used are as few as possible. This is a NP complete problem and that's why this is a greedy algorithm.

II. Purpose of the Project

Community detection is an important aspect to help detect the structure of the network and identify nodes based on their similarities. With today's digital age it becomes an important aspect to detect communities which can be applied to various fields such as healthcare, marketing et cetera. Link prediction serves an important role to help detect the future of the community and the network. Data mining becomes a tedious and almost impossible task to help detect communities and the links between their nodes. This is where machine learning models and various community detection algorithms come into picture to save time and accurately predict the future of these networks.

III. Literature Survey

The following research papers were referred to gather deeper understanding and knowledge of the subject.

Qi Chen. [1] The primary emphasis of this paper is on the study of identifying patterns of activity and forecasting the potential configuration of advanced networks in overlapping substructures. It examines a few of the best algorithms for detecting overlapping communities in advanced networks.

X.Ma. [2] This paper investigates two evolutionary non-negative matrix factorization mechanisms for detecting complex populations by clustering, mapping, and other techniques. The algorithm used in this paper proved to be more accurate than many state-of-art approaches.

Hao Shao. [3] This paper describes a relation prediction algorithm for unsupervised networks. It focuses on unsupervised machine learning models for detecting node connections accurately.

Junming Shao. [4] For connection prediction and group identification, this paper employs cluster-driven low rank Matrix completion. The proposed algorithm in this paper outperformed many previously studied algorithms in terms of accuracy.

David Liben Nowell. [5] This paper focuses on Link estimation using various coefficients such as Jaccard's coefficient, based on a large number of near neighbours, and various paths such as page rank, reaching time, commuting time, and so on. This paper also discusses approaches focused on node neighbourhoods, such as common neighbours. The accuracy of these forecasts was equivalent to 16%, and there is still potential for significant progress in the algorithms used in the paper. Often, the time frame or optimization of these algorithms can be done to increase the time complexity so that they operate much faster on massive data sets.

Mohsen Shahriari. [6] This paper proposes a two-step method for locating overlapping communities in signed social networks. Furthermore, assess the significance of three node classes: additional, overlapping, and intra. The results show that overlapping nodes can predict signals more accurately than intra and extra nodes. In addition, anger was applied as a measure to test errors in fuzzy group identification in signed social networks.

Le Yu. [7] This paper focuses on a population identification thesis based on dynamic network analysis. It suggests a novel algorithm for detecting overlapping communities. The proposed algorithm, as opposed to traditional algorithms relying on node clustering, is based on connection clustering. The connection clustering would reflect groups of links with similar properties. The algorithm employs a genetic operation to cluster on links. An effective coding schema for number of communities can be automatically detected.

Kamal Sutaria. [8] The aim of this paper is to explain the interpersonal interaction between a community of active actors representing various types of structures. Many real-world structures, such as human cultures and various types of components, may be modelled as social networks. Social network research provides key words to a forum for industry to produce product surveys and promote the introduction of new technologies to the public body. This method differs from conventional clustering.

Jaewon Yang. [9] This paper reflects on the basic methods for uncovering operational concepts in networks. To create communities based on edge structure and node attributes, an accurate, scalable, and efficient algorithm for detecting overlapping communities in networks was created. This model integrates with the network configuration and node characteristics, resulting in more precise community identification and greater robustness.

Lei Tang. [10] This paper is written from the standpoint of data mining. It yields graph-based group identification strategies as well as numerous important extensions for dealing with complex, heterogeneous networks in social media.

IV. METHODOLOGY

Step 1: Import the required libraries such as NetworkX, NumPy, pandas, linear models, light GBM models.

Step 2: Using NetworkX library form a network or a graph of the imported dataset and perform exploratory data analysis.

Step 3: Implement the Girvan-Newman algorithm for community detection and visualize the communities using Matplotlib.

Step 4: The most influential nodes for marketing are identified based on betweenness centrality, degree centrality and closeness centrality.

Step 5: For Link prediction on our data set, we first create an adjacency matrix to find all the unconnected pairs.

Step 6: These unconnected pairs form the negative samples; the positive samples are formed by removing the connected node pairs.

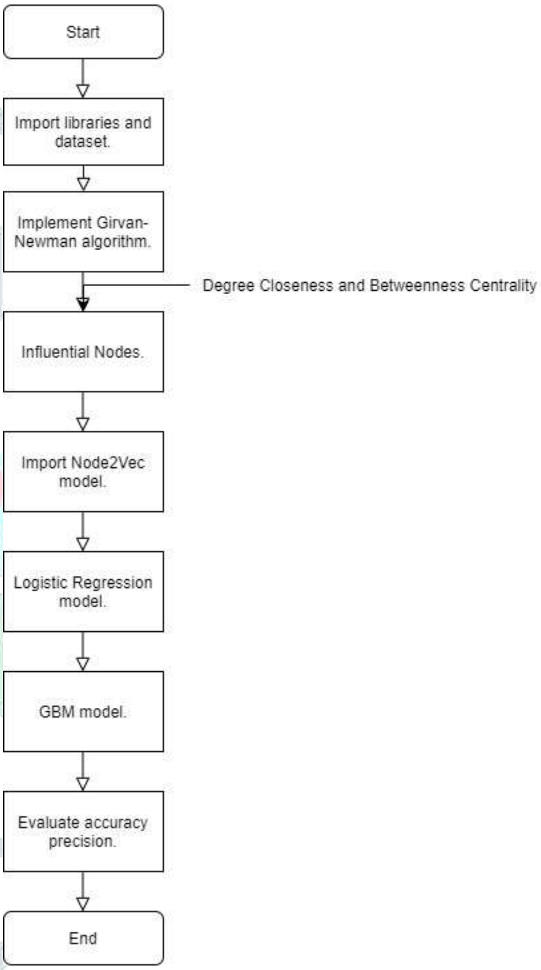
Step 7: We then train our Node2Vec model using these samples. This vector dataset can be used for training our models.

Step 8: We then fit this data set to a logistic regression model, visualize the results and calculate the Roc-Auc score.

Step 9: The accuracy of this model is visualized and another model is used to improve the accuracy and predictions of this logistic regression model.

Step 10: We fit our data set to a light GBM model to increase the efficiency of our predictions.

Step 11: Thus, we have successfully used Girvan-Newman approach for detection of communities and implemented logistic regression model and light GBM model for accurately predicting links between nodes.

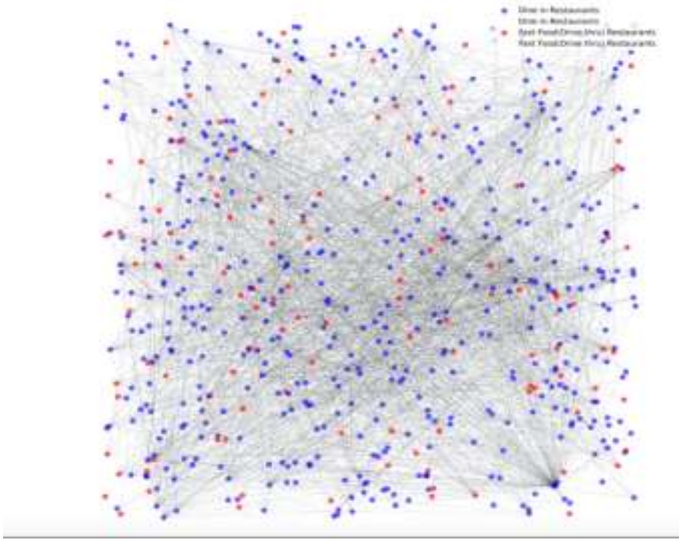

Start
Import libraries and dataset.
Implement Girvan-Newman algorithm.
Degree Closeness and Betweenness Centrality
Influential Nodes.
Import Node2Vec model.
Logistic Regression model.
GBM model.
Evaluate accuracy precision.
End

V. RESULTS

These results are obtained on the dataset Fb-pages-food network from the network repository website.

Community detection:

The two communities created by the Girvan-Newman algorithm are depicted in the diagram below. Dine-in restaurants are represented by blue nodes, while fast-food restaurants are represented by red nodes.


Dine-in Restaurants
Dine-in Restaurants
Fast Food/Drive-thru Restaurants
Fast Food/Drive-thru Restaurants

Link Prediction:

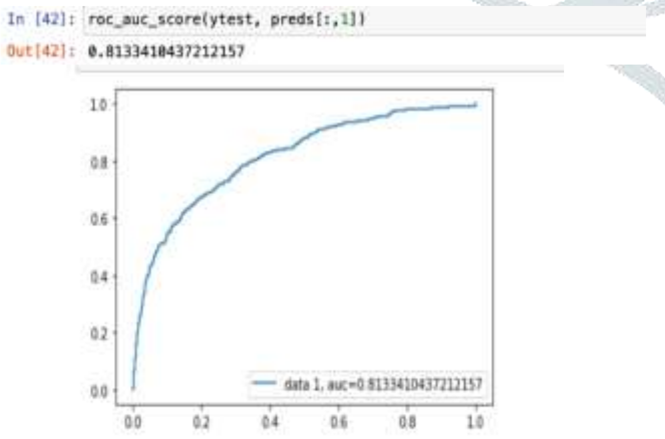
1. Logistic Regression Model

Logistic regression is similar to linear regression in that it forecasts whether something is true or false rather than forecasting something constant, such as scale. In addition, rather than fitting a straight line to the results, logistic regression employs an S-shaped logistic function. And the curve converts it into a value between 0 and 1, indicating the likelihood of nodes in societies containing similar nodes. The power of logistic regression to have probabilities and distinguish new samples using continuous and discrete measurements. It can be used to identify samples, and it can do so using various types of data such as size and/or genotype. It can also be used to determine which nodes in a given dataset are useful.

1 / (1 + e^-value)

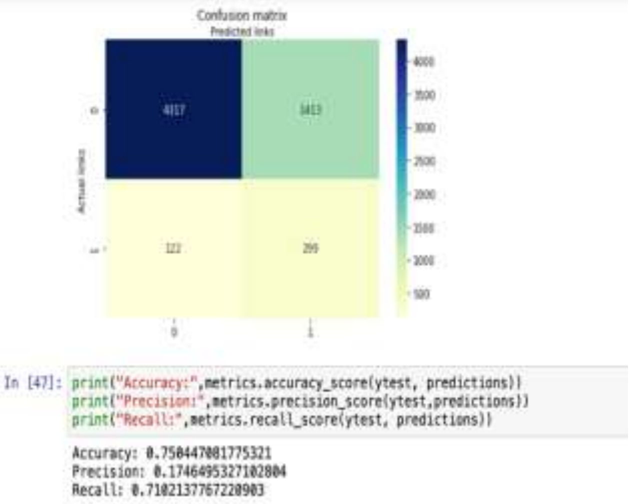
In natural logarithms where e is base (Euler's number or the EXP () function in your spreadsheet) furthermore, esteem is the actual numerical value that you need to change.

The figures below represent the roc-auc score after fitting the logistic regression model for our dataset. It is roughly 81.33%. The curve of the auc score is visualized below.



```
In [42]: roc_auc_score(ytest, preds[:,1])
Out[42]: 0.8133410437212157
```

The confusion matrix for actuals links vs predicted links is represented below, based on which the accuracy, precision and recall is calculated. The accuracy is roughly 75.04% and the precision and recall stand at 17.46% and 71.02% respectively.



```
In [47]: print("Accuracy:",metrics.accuracy_score(ytest, predictions))
         print("Precision:",metrics.precision_score(ytest,predictions))
         print("Recall:",metrics.recall_score(ytest, predictions))

         Accuracy: 0.750447081775321
         Precision: 0.1746495327102804
         Recall: 0.7102137767220903
```

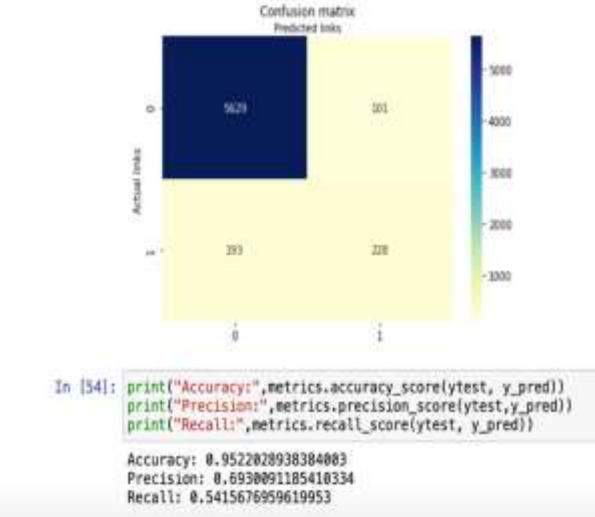
2. Light GBM model

Light GBM is another variant of gradient boosting, and the light stands for the light iteration, which supposedly makes this faster spread, more effective at maximizing the energy, and perhaps a little more reliable, and we checked this on the two data sets of social network sites that we took. It's a gradient boosting algorithm, which means it's built on the decision tree algorithm, similar to XGboost or even random forest. It divides the tree leaf by leaf, and there are several ways to divide these trees.

Benefits of Light GBM

1. Faster preparation speed and better proficiency
2. Lower memory consumption
3. Greater precision than most enhancing calculations
4. Compatibility with Large Datasets
5. Parallel learning help

The figure below represents the confusion Matrix for predicted links versus actual links when fitted to a light GBM model. We can clearly see that the accuracy increased to 95.22% while the precision jumped to 69.30% and the recall stands at 54.15%.



```
In [54]: print("Accuracy:",metrics.accuracy_score(ytest, y_pred))
         print("Precision:",metrics.precision_score(ytest,y_pred))
         print("Recall:",metrics.recall_score(ytest, y_pred))

         Accuracy: 0.9522028938384083
         Precision: 0.6930091185410334
         Recall: 0.5415676959619953
```

- Influential nodes

1. Influential nodes dependent entirely on degree centrality: Degree centrality is a simple tally of the cumulative number of connections attached to a vertex. It is regarded as a type of prominence metric, though an unrefined one that does not distinguish between quantity and consistency. Degree centrality does not distinguish between a relationship to the Indian president and a connection to a high school dropout. The degree is a proportion of the total number of edges associated

with a given vertex. There are two-degree proportions for organized groups. The number of connections that point internally at a vertex is known as in-degree. The number of associations that originate at a vertex and point outward to other vertices is known as out-degree.

2. Influential nodes primarily based on closeness centrality: The concept "closeness centrality" refers to a node that is closest to all other nodes with respect to the average distance between all other nodes. Closeness centrality, like degree centrality, can be measured on a network where we don't know the location of the node, but closeness of priority can also be broken down into nearest in terms if you know the direction of the node broken down into closest in terms of people closest to you, so closest in terms of incoming node at you or closest in terms of outgoing node from you.

3.Influential nodes solely dependent on Betweenness Centrality: Betweenness centrality is a metric of how often a node acts as a conduit between other nodes. So, in this, we calculate all of the shortest paths between the nodes, then we calculate the percentage of the shortest paths between every two nodes, and finally we add the percentages over all pairs of nodes. When looking at leverage or how information is disseminated across the network and between the centrality is a linchpin metric for looking at stuff like brokerage in a network, because for example, if you believe that essential information is dispersed across various regions in a network, individuals with high betweenness centrality united the network.

VI. Conclusion

Data mining techniques are used to derive valuable information from large databases, which can be a time-consuming process. It is absolutely avoided with the aid of machine learning algorithms and optimized crowd detection algorithms. The algorithms used in this paper allow us to accurately predict future relations between nodes in a network using logistic regression, light GBM models, and the Girvan-Newman algorithm, as well as predict the creation of communities in networks. This can be applied to various fields such as marketing with the help of influential nodes, detecting communities for analysis of social networks. The link prediction model can be used in social media to recommend people and pages one may connect with and like respectively.
The identification of future nodes which may link can be well utilized to predict the future growth of a network and to study how a network might grow and to understand constantly evolving networks in an easier way. In the future, the accuracy and predictability can be further improved to detect communities and link predictions more precisely.

Acknowledgment

We would like to express our appreciation to our honorable Prof. Ms. Seema Redekar, who provided us with the opportunity to work on this project and guided us during the process with her vast expertise in the field and often helpful spirit. We would like to take this opportunity to thank our HOD, Dr. K Lakshmi Sudha, and Principal, Dr. Atul Kemkar, for their invaluable advice and tremendous assistance in providing the necessary facilities. We would also like to thank the entire faculty of the IT Department for their insightful suggestions and prompt assistance with this effort. Last but not least, we would like to thank our college's teaching and non-teaching staff for their assistance in encouraging timely delivery. Throughout the course, we acquired a great deal of understanding and expertise in the field of group identification and relation prediction. We also gained organizational skills and a team-oriented mindset, which paved the way for our personal growth.

References

[1] Qi Chen; Lingwei Wei, “Overlapping Community Detection of Complex Network”, 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Gold Coast, QLD, Australia, DOI: 10.1109/PDCAT46702.2019.00102.

[2] X. Ma; D. Dong, “Evolutionary Nonnegative Matrix Factorization Algorithms for Community Detection in Dynamic Networks”, 2017 IEEE Transactions on Knowledge and Data Engineering, Pages: 1045 – 1058, DOI: 10.1109/TKDE.2017.2657752.

[3] Hao Shao; Lunwen Wang; Jian Deng, “A Link Prediction Algorithm by Unsupervised Machine Learning”, 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), DOI: 10.1109/CISCE.2019.00145. Pages 3382-3388. https://doi.org/10.24963/ijcai.2019/469.

[4] Junming Shao; Zhong Zhang; Zhongjing Yu; Jun Wang; Yi Zhao; Qinli Yang, “Community Detection and Link Prediction via Cluster-driven Low-rank Matrix Completion”, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Pages 3382-3388, https://doi.org/10.24963/ijcai.2019/469.

[5] David Liben-Nowell; Jon Kleinberg, “The Link Prediction Problem for Social Networks”, CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management, November 2003, Pages 556–559, https://doi.org/10.1145/956863.956972.

[6] Mohsen Shahriari; Ralf Klamma “Signed social networks: Link prediction and overlapping community detection”, 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Pages: 1608-1609, DOI Bookmark: 10.1145/2808797.2809357.

[7] Le Yu; Bin Wu; Bai Wang “LBLP: link-clustering-based approach for overlapping community detection”, Tsinghua Science and Technology, DOI:10.1109/TST.2013.6574677.

[8] Kamal Sutaria; Dipesh Joshi; C.K. Bhensdadia; Kruti Khalpada, “An Adaptive Approximation Algorithm for Community Detection in Social Network”, 2015 IEEE International Conference on Computational Intelligence & Communication Technology, DOI: 10.1109/CICT.2015.103.

[9] Jaewon Yang; Julian McAuley; Jure Leskovec, “Community Detection in Networks with Node Attributes”, 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, DOI: 10.1109/ICDM.2013.167.

[10] Lei Tang; Huan Liu, Community Detection and Mining in Social Media, Morgan & Claypool, 2010.

[11] Q. Liu; G. Liu and X. Chu, "Comparison of different spatial resolution bands of SPOT 5 to plant community patch detection," 2012 5th International Congress on Image and Signal Processing, 2012, pp. 1029-1033, doi: 10.1109/CISP.2012.6469748.

[12] M. Girvan; M.E.J. Newman, Community structure in social and biological networks, Proc. Nat. Acad. Sci. USA, 99 (2002) 7821-7826.

[13] E. Bütün; M. Kaya; R. Alhajj, ‘‘Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks,’’
Inf. Sci., vols. 463–464, pp. 152–165, Oct. 2018.

[14] Y.-X. Zhu, X.-G. Zhang, G.-Q. Sun, M. Tang, T. Zhou, and Z.-K. Zhang, ‘‘Influence of reciprocal links in social networks,’’ PLoS ONE, vol. 9, no. 7, Jul. 2014, Art. no. e103007.

[15] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, ‘‘Link prediction in complex networks: A local naive Bayes model,’’ Europhysics Lett., vol. 96, no. 4, p. 48007, Nov. 2011.

[16] X. Ma, L. Gao, X. Yong, L. Fu, Semi-supervised clustering algorithm for community structure detection in complex networks, Physica A, 389 (2010) 187-197.

[17] K. Dutta, M. Sharma, U. Sharma, S. K. Khatri and P. Johri, "Information Gain Model for Efficient Influential Node Identification in Social Networks," 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 146-150, doi: 10.1109/AICAI.2019.8701344.

[18] M. U. Ilyas and H. Radha, "Identifying Influential Nodes in Online Social Networks Using Principal Component Centrality," 2011 IEEE International Conference on Communications (ICC), 2011, pp. 1-5, doi: 10.1109/icc.2011.5963147.

[19] K.-K. Shang, M. Small, X.-K. Xu, and W.-S. Yan, ‘‘the role of direct links for link prediction in evolving networks,’’ Europhysics Lett., vol. 117, no. 2, Jan. 2017, Art. No. 28002.

[20] H. A. Deylami and M. Asadpour, "Link prediction in social networks using hierarchical community detection," 2015 7th Conference on Information and Knowledge Technology (IKT), 2015, pp. 1-5, doi: 10.1109/IKT.2015.7288742.

[21] M. Anjerani and A. Moeini, "Selecting influential nodes for detected communities in real-world social networks," 2011 19th Iranian Conference on Electrical Engineering, 2011, pp. 1-6.

[22] Greene, Derek & Doyle, Dónal & Cunningham, Padraig. (2010). Tracking the Evolution of Communities in Dynamic Social Networks. Proceedings - 2010 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2010. 2010. 176-183. 10.1109/ASONAM.2010.17.

[23] Kempe D., Kleinberg J., Tardos É. (2005) Influential Nodes in a Diffusion Model for Social Networks. In: Caires L., Italiano G.F., Monteiro L., Palamidessi C., Yung M. (eds) Automata, Languages and Programming. ICALP 2005. Lecture Notes in Computer Science, vol 3580. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11523468_9

[24] S. Soundarajan and J. Hoperoft, "Using community information to improve the precision of link prediction methods,” in proceedings of the 21st International Conference Companion on World Wide Web, New York: ACM Press, 2012: 607-608

[25] J. Xie and B. K. Szymanski, “Towards linear time overlapping Community detection in social networks,” in Proc. 16th Pacific-Asia Conf. Adv. Knowl. Discovery Data Min., 2012, vol. 2, pp. 25–36.

[26] V. Fionda and G. Pirrò, "Community Deception - Or: How to Stop Fearing Community Detection Algorithms (Extended Abstract)," 2018 IEEE 34th International Conference on Data Engineering (ICDE), 2018, pp. 1789-1790, doi: 10.1109/ICDE.2018.00252.