# Modeling the Information Diffusion of Overlapped Nodes using SFA-ICBDM

**Mustafa Kamil Mahdi, Huda Naji Almamory**

*Abstract :In recent time, online social networks like, Facebook, Twitter, and other platforms, provide functionality that allows a chunk of information migrates from one user to another over a network. Almost all the actual networks exhibit the concept of community structure. Indeed overlapping communities are very common in a complex network such as online social networks since nodes could belong to multiple communities at once. The huge size of the real-world network, diversity in users profiles and, the uncertainty in their behaviors have made modeling the information diffusion in such networks to become more and more complex and tend to be less accurate. This work pays much attention on how we can accurately predicting information diffusion cascades over social networks taking into account the role played by the overlapping nodes in the diffusion process due to its belonging to more than one community. According to that, the information diffusion is modeled in communities in which these nodes have high membership for reasons that may relate to the applications such as market optimization and rumor spreading. Our experiment made on a real social data, Digg news aggregator network on 15% of overlapped nodes, using our proposed model SFA-ICBDM described in previous work. The experimental results show that the cascade model of the overlapped nodes whether represents seed or node within cascade achieves best prediction accuracy in the community which the node belongs at more.*

*Keywords: predict diffusion cascade, Information diffusion, online social network, overlapping communities, overlapping nodesAbout.*

## I. INTRODUCTION

At the present time, the internet has become the first source of discovering and consuming new information online, this brought out the Social media platforms, such as (Twitter, Google+, and online micro-blogging social network) to be a part of the everyday life of modern society, since it allows groups of an individual to continuously create and share content in various domains, in a very quick and convenient way[1]. Over the years, online social network (OSN) platforms developed to acquire abilities that allow photo, hashtag, or any other pieces of information get reshared several times: a user shares the content with a set of friends, several of these friends share it with their respective sets of friends, in a ‘cascading’ style, uncovering a phenomenon named information cascade or information diffusion [2]. A cascade of resharing can be constructed, potentially reaching numerous individuals. As a summary, a diffusion process of information is usually viewed to has two different phases: (first) the emergence of contents by external source, such as mass media and (second) the cascading spread of the information through internal relations, such as interpersonal communications between connected neighbours in a contact network[3].The distinction between diffusion and cascade has been stated in [4] where the ‘diffusion’ refer to the event of spreading the content among the users of a social networks , while the term ‘cascade’ refers to the result of the diffusion activity, this means that cascades are the structural representation of diffusion on a social network. The diffusion of content over a network and resulting information cascades have been widely investigated on various platforms, indeed it has been shown that predictive and explanatory aspect of information, diffusion is crucial in order both to understand information propagation and to better control it in such huge size interactive media [4][5].For understanding the diffusion processes over OSN may help in better, solving many real-world events and further analysing them regarding different objectives [2]. Where it could be used for investigating and preventing terrorism, observing the trends of election results, optimizing marketing campaigns for businesses, uncover the source of fake news, follow revolutionary waves, and rumor controlling, etc. One of the most notable challenges in understanding the diffusion phenomenon is how can effectively identifying or detecting hidden modular structures referred to as communities.In fact, real-world social networks consist of overlapping communities in which some nodes are common to multiple communities [6][7]. There is no doubt that the overlapping nodes have a big role in diffuse the information from community to other that motivates to focus on overlapping nodes and to model their influence which not been considered in literatures.In this paper, we generally aim to achieve the following two points: first predicting information diffusion cascade. In other words, predicting diffusion process pattern by modeling the influence of node taking into account the behavioral features and structural characteristic of nodes. Second, modeling the information diffusion of overlapped nodes based on the belonging level of such nodes to different communities.

**Revised Manuscript Received on 30 July 2019.**

* Correspondence Author
**Mustafa Kamil Mahdi*.** College of Information Technology University of Babylon, Babylon, Iraq,
**Huda Naji Almamory.** College of Information Technology University of Babylon, Babylon, Iraq

This paper is organized as follows; section 2 discusses the related works. In section 3 we present the methods used for the modeling. The results have been discussed in section 4. Finally, Section 5 concludes and outlines future work.

## II. RELATED WORKS

The importance and the popularity of social networking and its related phenomenon such as information diffusion have drawn the attention of many researchers. Several kinds of work have made efforts to develop models and applications considering predictive and an explanatory aspect of information dissemination through social networks[8]. Hu et al. depend on the time series forecasting for proposing a method to predict the short-term popularity of viral topics. They considered three types of features observed in past social activity of the users on a specific topic: "previous-popularity-based, user-comment-based and network-structure-based. The work concludes that the burst topic's popularity is relatively dynamic and changeable. On the other hand and for non-burst topics, historical popularity can still have an impact on later popularity"[9]. The idea of identifying the main cascade of information diffusion has been discussed by Zhu et al in [10]. The work can be summarized in the following steps: first, specify the edge weights, second uncover the most influential nodes in the network and finally, the main diffusion path is the shortest path that gathers the influential users. their work was directed to avoid trace the information in a complete network structure rather focusing only on the most important paths. Analyzing and modeling the properties of users social behaviors attracted a large number of research efforts, in order to improve system performance and develop a variety of social application, as it can be noted in [11][12]. Ruan et al. [11] built social behavioral profiles by studying users' social behavioral features based on the activities on social platforms. The work was aimed to define a method that can be used to differentiate the users in OSN and especially compromised user's accounts. They argue That the constructed behavioral profiles of social user can be applied to accomplish this task. In our study, we also try to find out and represent the features of user behavior considering their previous social activity. The outcomes are utilized more for special user to user similarity detection, which can facilitate the modeling of information diffusion in online social network. In [12] and by proposing a relational latent SVM model, the author focuses on achieving better inference of user attribute performance in social media by considering six types of user attributes, each of which is analyzed and measured to find the roles played by them in an information dissemination process. The work in [13]built a well-known ‘‘following’’ link cascade model which can be used to define the diffusion probabilities in different triadic structures througth information propagation process. Taking into account the time duration, Feng et al. [14] depend on some directed measures in order to better define the content diffusion process, where the main aim of their work was to analyze the efficiency of diffusion in real-world online social media. Kwa et al. Use a total of 106 million tweets to analyze the relationship between the number of followers (fan user) of the initiator of the tweets and their retweets. Depending on cascading tree of retweet activity the study state that the only subset of user's followers actually retweet and in general, individuals are only retweets by few number of users. One more conclusion was discussed in this study and it was that the users have the same average number of retweets for their posts if they have on average less than 1,000 followers [15]. Suh et al. Study the correlation between the number of identified features with the number of retweets of a specific tweet. By evaluating the correlation through a large-scale analysis on over 74 million tweets, the authors have shown that the size of followers, followers, and the account ages have a significant correlation with the retweet size. As large as the number of in and out relation of a node, the more likely his tweets get retweeted is. In addition, the age of the user account has been considered, where they conclude that post initiated by users registered more than 300 days before (senior users), exceeds the average number of retweets . On the other hand, some features does not highly correlates with the number of retweets such as the presence of hashtag or URL in a tweet [16].In [17] the author focus on the interaction between the nodes in online social networks to design and implementation of a flocking based centrality for online social networks. They considered four social network analysis targets, namely: Centrality, link prediction, community detection and trust prediction. Regarding the second objective, a novel link prediction inspired from firefly algorithm has been proposed in this work. Community detection in online social networks is another objective in social network analysis, which has a role in spreading the information over the network. Motivated by the concept of Stability-plasticity dilemma the authors propose and validate an algorithm for overlapping community detection. The Fuzzy adaptive resonance theory inspired algorithm has been tested and validated using benchmark datasets, as well as computer generated data sets.

## III. PROPOSING METHOD

The aim of this work is to accurately predict the information diffusion cascade initiated by different seed nodes, taking into account the role played by overlapping node in the diffusion process. The overlapped node can either be the initiator of the cascade or it appears in the diffusion cascade tree during the modeling task. Since this node considered as an interface among all the communities it belong to, it should be paid special attention to it. Hence, the belonging degree of the overlapped node have to be measured to all the communities it belongs to, then modeling the influence diffusion of this node only toward the community in which the node scores highest belonging.The claim here is, following this hypothesis, that the influence of that node is more effective in the community which the node has highest membership to. In fact, appearance of overlapped node in a cascade requires modeling the diffusion of that node in the community in which it affects significantly (the same applies if the overlapped node is a seed) due to the possible cost of tracking the cascade in all communities in real life applications. In general, we can describe the framework of our proposal by Fig.1.

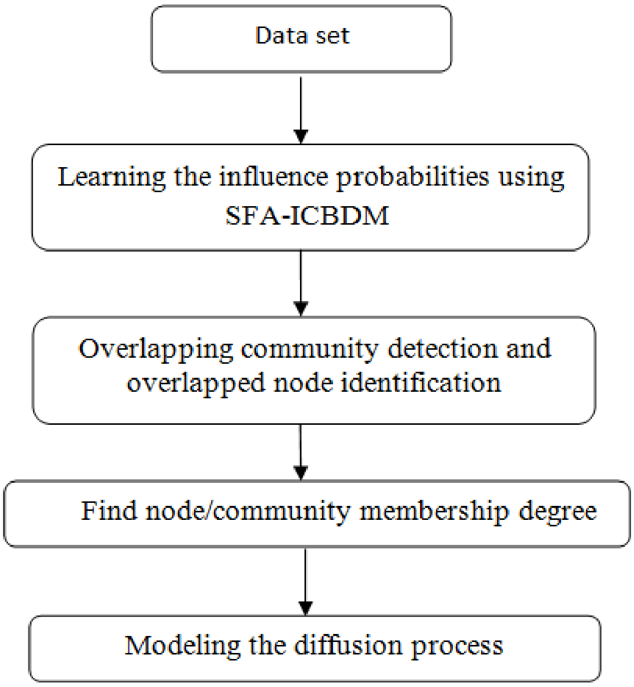
Data set
Learning the influence probabilities using SFA-ICBDM
Overlapping community detection and overlapped node identification
Find node/community membership degree
Modeling the diffusion process

**Fig.1 Schematic Diagram of Proposed Method**

In the following section we will describe the steps of this diagram: Regarding Learning the influence probabilities, the core issue is how to find the influence probabilities $p_{u,v}$ which have to be assigned for every edge in the network. Where the value of $p_{u,v}$, indicate how likely that user u influence on user v, the issue which has been addressed in our paper [18] using SFA-ICBDM model and tried to exploit the structural–based feature and user behavior-based feature. Fig.2 illustrates the performance of the SFA-ICBDM model in term of average F1 measure, the experiment was conducted using 100 deferent random nodes as seed.

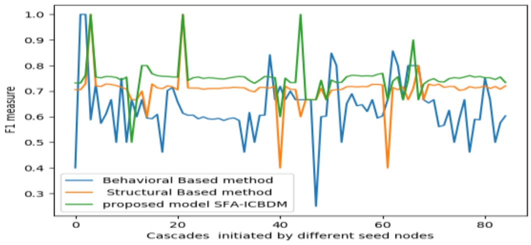

**Fig.2 the effectiveness of SFA-ICBDM hybrid model**

Table (1) show comparison in term of Precision values, between the proposed model and the result from related works:" RUC (Reinforced User-Centric) and DRUC ( Decaying Reinforced User-Centric)"[18], where different cascade length was considered.

**Table 1. Compare SFA-ICBDM with related work base on precision values**

| Cascade length | 1 | 2 | 3 | >= 4 |
|---|---|---|---|---|
| RUC | 0.63 | 0.50 | 0.63 | 0.67 |
| DRUC | 0.63 | 0.50 | 0.62 | 0.68 |
| SFA-ICBDM | 0.830 | 0.731 | 0.757 | 0.781 |

Overlapping community detection and overlapped node identification. In our experiment, we use a link-density based technique of locating overlapping communities, in directed networks which was introduced by extending the clique percolation method (CPM) originally proposed for undirected networks[19]. The next step is to find node/community membership degree. The key point here is how to find the membership degree of an overlapped node to each community it belongs to. Given a network G (V, E) with N and M representing the number of nodes and edges in the network respectively, the weight of the edge $W_{u,v} = 0$ if nodes u and v are disconnected . For a node u and a community c , the belonging degree B(u,c) between u and c is introduced in (1):

$$B(u,c) = \frac{\sum_{v\in c} W_{u,v}}{K_u} \qquad (1)$$

$$K_u = \sum_{v\in V} W_{u,v}$$

From (1), "B(u,c) reflects how tight the node u is with community c " [20]. If all neighbours of node u are in the community c, then B(u,c) = 1; otherwise, B(u,c) < 1.
Modeling the diffusion process. In this work we use the well known Independent Cascade model(IC model), which focus on the influence from an adopted node u to next inactive node v. the active node u can independently activate the not activated node v with a probability of $P_{u,v}$ [0, 1] at time t. If node u seceded then node v will not be deactivated in the subsequent time steps. Regardless of whether the node u success to activate v at time t, node u will not permitted to activate node v in the future. The IC model is much sutable for prediction and influence research. Although we specifically use the (IC model), we did not employ this model directly. Rather we use our proposed method described in [18] to find the influence probabilities $P_{u,v}$ of network edges. The following flowchart (Fig.3) describes the modeling process over several random nodes considered as seed nodes in our experiment, including the overlapping nodes.

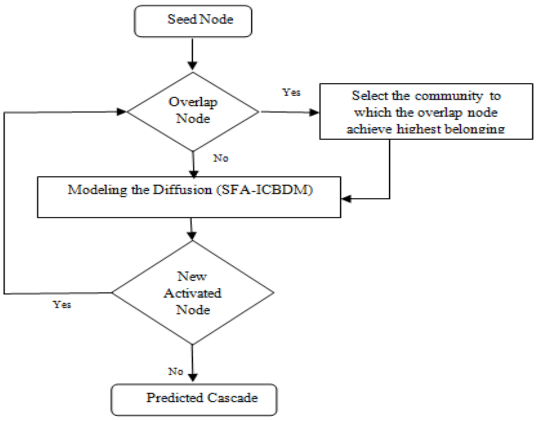
Seed Node
Overlap Node
Yes
Select the community to which the overlap node achieve highest belonging
No
Modeling the Diffusion (SFA-ICBDM)
New Activated Node
Yes
No
Predicted Cascade

**Fig.3 descriptive flowchart of modeling process through overlapping nodes**

## IV. EXPERIMENT AND RESULTS

We conduct our experiment on a real-world social platform, Digg dataset which consist data regarding  stories promoted to Digg's front page in 2009  over a period of 30 days[21]. The constructed network has 2008 nodes and 10333 edges in which we consider the largest strongly connected components of the dataset relation. The learning procedure which we employ find for each relation the proper influence probability based on the structural and behavioural information extracted from the dataset.We adapt the CPMd algorithm with k (clique) value was set to 5, to uncover the overlapping community structure. The algorithm discover 398 community in the dataset. Out of 2008 nodes in the network, 584 nodes were marked as overlapping nodes with different number of community per overlap node. Fig.4 show the membership distribution of the nodes among network communities, for instance there are 1424 nodes belong to only one community while there are 56 nodes belong to four communities and so on. Fig.5 is visualizing examples of community overlapping, in which the overlapping nods and ordinary nodes are listed along with their belonging degree to different communities.

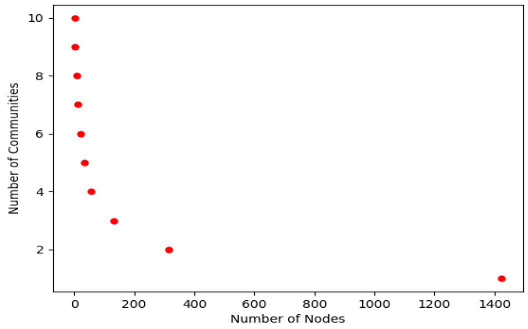

**Fig.4 overlap node - community distribution**

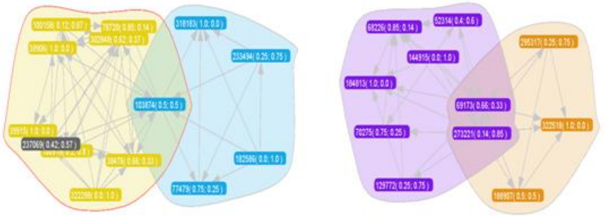

**Fig.5 Two example of discovered overlapped communities**

In our experiment, over 15 % of the overlapping nodes were randomly selected to be considered as seed nodes through modeling process. The result show that over 75% of the tried seeds the prediction is more accurate (in term of F measure) when the diffusion has been modeled toward the community in which the overlapped node scores highest belonging. Fig.6 show how the prediction accuracy positively correlate with the belonging degree. In other words, the predicted cascade is more realistic in community which node has highest membership to.

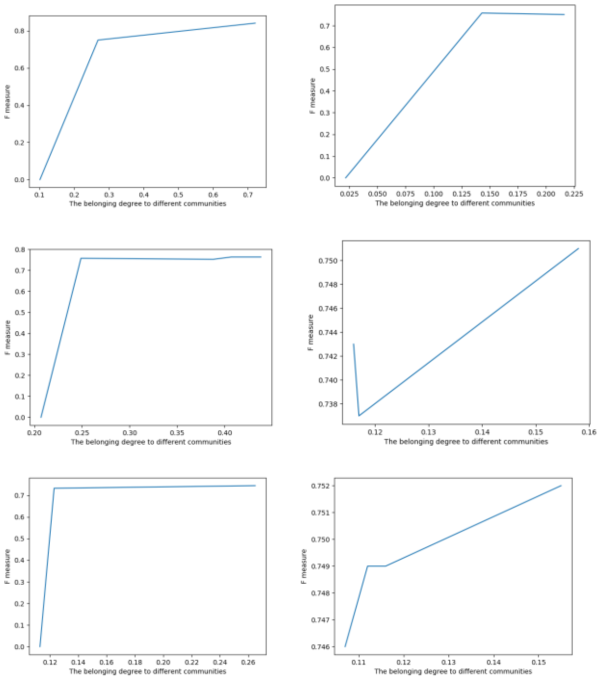

**Fig.6 Membership of overlapped nodes versus F-measure.**

## V. CONCLUSION

In addition to conclusions that have been recorded  in previous work related to SFA-ICBDM model, the experiment uncovers a fact of existing a relation between cascade prediction accuracy and the membership of overlapped nodes in different communities .

We conclude that prediction  accuracy is  positively  correlate with the degree of belonging of overlapped node to different communities.

## REFERENCES

1. N. V. Anh, D. N. Son, N. T. T. Ha, S. Kuznetsov, and N. T. Q. Vinh, "A method for determining information diffusion cascades on social networks," Eastern-European J. Enterp. Technol., vol. 6, no. 2 (96), pp. 61–69, 2018.
2. T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter – Analysis of predictive features," J. Comput. Sci., vol. 28, pp. 257–264, 2018.
3. J. Cheng, L. A. Adamic, P. A. Dow, J. Kleinberg, and J. Leskovec, "Can Cascades be Predicted?," 2014.
4. N. Alrajebah, M. Luczak-Roesch, L. Carr, and T. Tiropanis, "Deconstructing diffusion on Tumblr: Structural and temporal aspects," WebSci 2017 - Proc. 2017 ACM Web Sci. Conf., pp. 319–328, 2017.
5. B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," 2018.
6. Z. Zhang, Y. Gong, K. Wang, and J. Gu, "A survey of overlapping community detection based on multi-label propagation," Proc. 2017 12th IEEE Conf. Ind. Electron. Appl. ICIEA 2017, vol. 2018-Febru, pp. 995–999, 2018.
7. J. Leskovec, A. Rajaraman, J. D. Ullman, J. Leskovec, A. Rajaraman, and J. D. Ullman, "Mining Social-Network Graphs," Min. Massive Datasets, pp. 325–383, 2014.
8. X. Zhou, B. Wu, and Q. Jin, "User role identification based on social behavior and networking analysis for information dissemination," Futur. Gener. Comput. Syst., vol. 96, pp. 639–648, 2019.
9. Y. Hu, C. Hu, S. Fu, and P. Shi, "Author ' s Accepted Manuscript Time Series Forecasting Reference : To appear in : Neurocomputing," Neurocomputing, 2015.
10. H. Zhu, X. Yin, J. Ma, and W. Hu, "Identifying the main paths of information diffusion in online social networks," Physica A, no. xxxx, pp. 1–8, 2016.
11. X. Ruan, Z. Wu, H. Wang, S. Member, and S. Jajodia, "Profiling Online Social Behaviors for Compromised Account Detection," vol. 6013, no. c, pp. 1–12, 2015.
12. Q. Fang, J. Sang, C. Xu, M. S. Hossain, and S. Member, "Relational User Attribute Inference in Social Media," vol. 17, no. 7, pp. 1031–1044, 2015.
13. J. Zhang, Z. Fang, W. Chen, and J. Tang, "Diffusion of ' Following ' Links in Microblogging Networks," vol. V, no. JANUARY 2014, 2015.
14. Y. Feng, S. Member, B. Bai, W. Chen, and S. Member, "Information Diffusion Efficiency in Online Social Networks," pp. 1138–1142, 2015.
15. H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter , a Social Network or a News Media ? Categories and Subject Descriptors," pp. 591–600, 2010.
16. B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be Retweeted ? Large Scale Analytics on Factors Impacting Retweet in Twitter Network," pp. 177–184, 2010.
17. Ebin Deni Raj and Dhinesh Babu, these on" Effective strategies for information spreading in online social networks", 2016 Available:https://shodhganga.inflibnet.ac.in/handle/10603/168198
18. M. K. Alasadi and H. N. Almamory, "Diffusion model based on shared friends-aware independent cascade." international scintific confirence at Al-qadisiyah university-colloge of scince-2019
19. G. Palla, "Directed network modules",New journal of physics, vol. 9, pp. 1–21, 2007.
20. D. Chen, M. Shang, Z. Lv, and Y. Fu, "Detecting overlapping communities of weighted networks via a local algorithm," Physica A, vol. 389, no. 19, pp. 4177–4187, 2010.
21. T. hogg and K. lerman, "Social dynamics of digg " Springer ,EPJ DATA SCIENCE pp. 1–26, 2012.

## AUTHORS PROFILE

**Mustafa Kamil Mahdi** Received the M.Sc. degree in computer science from the Hamderd University, Delhi, India, in 2012. Ph.D. student, *College of Information Technology University of Babylon, Babylon, Iraq*, He has co-authored the papers, "Diffusion model based on shared friends-aware independent cascade" "Finding overlapped communities in directed network based on improved CONGA" His current research interests include the modelling, analysis, and performance information diffusion modelling and community structure in online social network.