Discovering Communities with SGNS Modelling-based Network connections and Text communications Clustering

Wathsala Anupama Mohotti
School of Computer Science
Queensland University of Technology
Brisbane, Australia,
w.mohotti@qut.edu.au

Richi Nayak
School of Computer Science
Queensland University of Technology
Brisbane, Australia,
r.nayak@qut.edu.au

***Abstract*—By the community discovery, the microblogging services facilitate diverse applications such as viral marketing, disaster management, customized programs, and many more. However, the sparseness and heterogeneity of user networks and text content make it difficult to group users with a similar interest. In this paper, we present a novel method to discover user communities with common interests. The proposed method utilizes both text content and interaction network information where network information is modeled using the concept of Skip-Gram with Negative Sampling for Non-negative Matrix Factorization. Empirical analysis using several real-world Twitter datasets shows that the proposed method is able to produce accurate user communities as compared to the state-of-the-art community discovery and clustering methods.**

***Index Terms*—Community Detection, SGNS Modelling, Network Connections, Text Communications**

I. Introduction

Social media platforms such as Twitter, Facebook and Instagram allow users to disseminate information and obtain social views based on the short-text communication [1]. Community detection in these platforms, for identifying the groups of users with common interests, has been found useful in diverse applications. It creates opportunities for political parties, businesses, and government organizations to target certain user groups for their campaigns, customized programs, and events [2], [3]. Majority of the existing community detection approaches [3]–[7] either (1) require supervision through ground truth labels, or (2) consider only the user network structure of user interactions or (3) analyze only the text content-based communications by users. They highly depend on the targeted applications and domain [8]

Discovering communities in a fully unsupervised manner is an essential requirement in many real-world applications. Disseminating information related to sales promotions, political campaigns and any special event or program need the identification of an interested group of users where prior knowledge on the group is unavailable. The majority of unsupervised community detection work is based on graph-based models [4]–[6] where network connection structures are analyzed to see *how* users are connected through social media. On the contrary, there are content-based text mining methods that consider *what* users communicate in community discovery utilizing the text messages [3], [7]. However, network-based methods face challenges due to sparseness in the network with the heterogeneity of the interactions whereas content analysis methods, which group users based on their written posts, produce inferior outcomes due to the curse of dimensionality in text vectors [9].

The content analysis relies on text messages to identify similar users based on what they have written/shared [1], [3]. Generally, social-media text is short in length that causes extreme sparseness in the data with the lack of co-relational occurrences [1]. Text mining is known to face the curse of dimensionality due to the high number of terms in doc$\times$term matrix representation [9]. This becomes worse in the short text mining due to extreme sparseness. Therefore text-based methods based on distance, density or probability face difficulties [3], [7], [10]. Specifically, the distance difference between near and far points becomes negligible in high-dimensional data [9]. This directly affects the distance-based methods such as k-means [11] in accurately identifying the groups. In addition, the sparseness of this high dimensional matrix representation does not allow to differentiate the user groups based on density differences [7]. The higher-to-lower-dimensional projection-based methods such as NMF [12] and LDA [13] which proposed to handle high dimensional issues also results in poor outcome due to information loss [9].

The network analysis-based methods rely on user connections that can be modeled with graphs, and use graph partitioning approaches to identify user groups [4]–[6]. Similar to text, user$\times$user relationships also result in forming a sparse network representation which requires an accurate modeling technique. In addition, the majority of these methods consider a single piece of network information and apply community discovery methods. However, user connections consider diverse facts other than the actual interest, such as friendships, and make heterogeneous user relationships. The community discovery in multiplex networks [14], [15] addresses this research problem

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

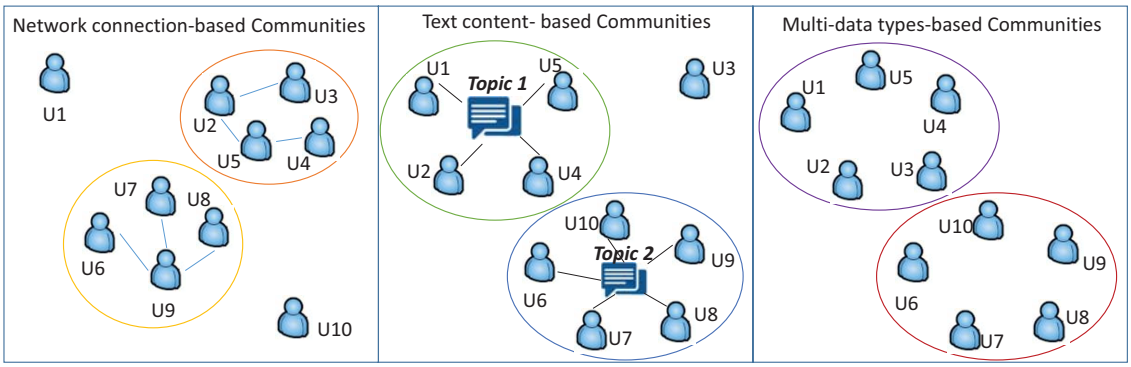

Fig. 1. Example: Multi-type data clustering for meaningful community detection

and tries to identify communities using heterogeneous user relationships without considering single-type data. However, these methods do not take noisy and high dimensional text data as input which makes the community discovery process more challenging and complex.

Fig. 1 shows an example emphasizing why the network or content data individually is unable to identify the community effectively. The network connection based methods will fail to group users U1 and U10 as they did not share connections with other users. Similarly, text content based methods will not find a group for U3 as it did not post similar topic posts as other users in the network. However, a multi-type data method that will be able to utilise both the connections and posts of these users and identify respective communities.

There is a handful of research that attempted to enrich the outcome of community detection with multi-type data (i.e., content and network data). Users' text communications along with the links between users provide rich and complementary sources of information and should be used simultaneously for exploring the user groups. The use of this multi-type data to identify the user groups can produce accurate results compensating sparseness in each data type. Additional information available with social media such as URLs and hashtags can also be incorporated with network representation to identify users with similar interests [16]. Researchers have combined this types of multi information by concatenating them into a single feature matrix and apply a (clustering) method [17].

A few researchers use text information separately with network structures in learning user communities [18]–[20]. These methods [18], [21] use the label information to accurately detect communities following a supervised approach. A probabilistic method [22] based on text and link information was used in community detection. However, communications that occur through short text results in extreme sparseness and user link networks that also create sparse representation would unable to provide high accuracy due to the lack of correlations. Similarly, the latent topic community analysis method based on generative probabilistic model [23] faces the same challenge. In extreme sparse datasets such as short text, matrix factorization based methods have proved to perform better than probabilistic methods [24].

There exist only a few matrix factorization-based approaches that are used in unsupervised community detection [19], [20], [25], [26]. In a latest method [19], to avoid the sparseness in high dimensional data, the nearest neighbors as well as furthest neighbors are used in generating the input matrix to the factorisation process. However, it does not consider the specific contribution of content and network information for community discovery. Methods [25], [26] treat both content and network information with 100% importance . In [20], researchers divide the contribution among different input data matrices within the factorization process for discovering communities. Distinct from these works, we propose to consider the importance of each type individually during matrix factorization. Additionally, we propose to use the modeling technique *Skip-Gram with Negative Sampling (SGNS)* to identify the closely associated users in sparse network representation.

In neural networks, the Skip Gram model is used as a training method to learn neighbors or the context of a word in a corpus [27]. In a fixed window, this technique can predict the surrounding words of a specific word. Considering the most co-occurring words, this concept can obtain a dense document representation for a document similar to word embedding [28]. SGNS modeling is an extended version of this concept, which accurately embeds the contextual information of words to highlight the closely associated words [29], [30]. In [29], [30], applying SGNS is proven as factorizing a word correlation matrix which is symmetric. This paper proposes to use this concept to highlight closely associated users considering the symmetric user affinity matrix.

We propose a novel method, Content and Structure Modelling for Discovering Communities (CSMDC), that combines an accurately modeled sparse network representation that captures the closely interacted users with the content information in social networks. We propose to use the SGNS modeling to identify the closely interacted users in a user network based on their interactions. This network information, represented with a symmetric user$\times$user matrix, and the content information, represented with an asymmetric term$\times$user matrix, are used

in learning the user×community matrix via the Non-negative Matrix Factorization (NMF) framework. The novel hybrid community discovery process proposed in CSMDC considers the specific contribution of each input matrix and accurately produces the final community assignment for each user with the maximum probability to be in a user community.

Empirical analysis using three Twitter datasets show that combining the content and network information accurately provides more accurate communities compared to relevant state-of-the-art methods. More specifically, this paper brings several novel contributions to community discovery.

- An NMF based approach that considers specific contribution of content and network information to identify the user communities.
- A novel modeling technique for network information.

To the best of our knowledge, CSMDC is the first method that uses the SGNS concept to represent user network information in community detection. We propose using a coupled matrix combining content and structure considering their specific contribution for community discovery in factorisation process to accurately represent the communities in an unsupervised setting.

II. Community Detection with SGNS Modelling and Multi-type Data Clustering (CSMDC)

Let there be N users in the network $U = \{u_1, u_2.....u_N\}$ that can be assigned to G communities. Let $S \in R^{N \times N}$ denotes the user interaction matrix between users with each cell representing the number of interactions between two users. This matrix is modeled with the SGNS weighting [30]. Let $C \in R^{M \times N}$ denotes the user content matrix where the short text messages written by N users consist of M distinct terms. CSMDC proposes to combine these two matrices during the NMF process and iteratively learns an optimum coupled matrix representing user community assignment as a factor matrix in a novel fashion. This process of identifying user groups is shown in Fig. 2. Each user is assigned to a community following hard clustering in an unsupervised setting.

A. Data Representation

The user to user relationship denoting their interactions and the user to their written posts relationship are modeled as the network input and the content input respectively. Content information for each user that consists of text, tags, and URLs, appearing in his/her posts, are considered as a single document that represents the user. We use the standard preprocessing steps of stop word removal, lemmatising as well as dictionary-based slang removal and word standardization to deal with the unstructured nature and presence of abbreviations in short text. Let Matrix C be modeled with the unique terms in the processed posts represented with the tf*idf weighting schema to consider both common and rare terms.

Let S be the network input matrix modelling the user interactions including shares, replies, and mentions. For each pair of users, the total number of interactions between them corresponding to sharing, replying and mention interactions are represented as the cell value in S.

1) Skip-Gram with Negative Sampling: CSMDC aims to capture the closer user interactions in S effectively with the SGNS modeling [30] to maximize the probability of highly interactive users while minimizing the probability of less interactive users. SGNS has been used to highlight the word embedding in text data in neural-networks [31]. It is able to capture the context of a word in a corpus as compared to the bag-of-word model [29]. The concept of negative (word, context) sampling is used with the Skip-Gram model to maximize the probability of an observed pair while minimizing the probability of unobserved pairs in distributed word representation [30]. This SGNS weighting is equivalent to factorizing a (shifted) word correlation matrix whose cells are the point-wise mutual information of the respective word and context pairs [29]. We use this concept in CSMDC to weight the closely interacted users considering other user interactions and maximize the probability for the user pairs that show closer interaction presence in comparison to the others while minimizing the probability of user pairs that show fewer interactions.

Let $c_{(u_i,u_j)}$ be the original cell value that represents the interaction relationship between user u_i and u_j in S. The SGNS weighting considers only the cell values that show interactions between u_i and u_j with > 0. The original value that represents the interactions between u_i and u_j is divided by the sum of the values in the u_i row and u_j column. SGNS weights the interaction relationship between u_i and u_j with respect to other user interactions, as in Eq. 1.

$$S_{(u_i,u_j)} = log \left[\frac{c_{(u_i,u_j)} \times T}{\sum_{u_a \in U} c_{(u_a,d_i)} \times \sum_{u_a \in U} c_{(u_a,d_j)}} \right] \quad (1)$$

where T is the total number of user interaction pairs that appear in S. The cell value S_{u_i,u_j} less than 0 is mapped to 0 to minimize the probability of user pairs that show less interaction [29].

B. Matrix Factorization

CSMDC aims to identify the user groups (or the user communities) using both network interactions and text communications as in Eq. 2. It factorizes the high dimensional content matrix C into two factor matrices $W \in R^{M \times G}$ and $H \in R^{N \times G}$ where G is the number of communities. It simultaneously factorizes the SGNS weighted user interaction matrix S into two factor matrices $H \in R^{N \times G}$ and $H_c \in R^{N \times G}$. Matrix H_c reveals the relationship between users and their context represented with respect to other user interactions.

CSMDC learns the common user community matrix H that coupled content and network information iteratively by minimizing the learning errors in the factorization of the matrix C and S as in Eq. 2. This factorization process gives

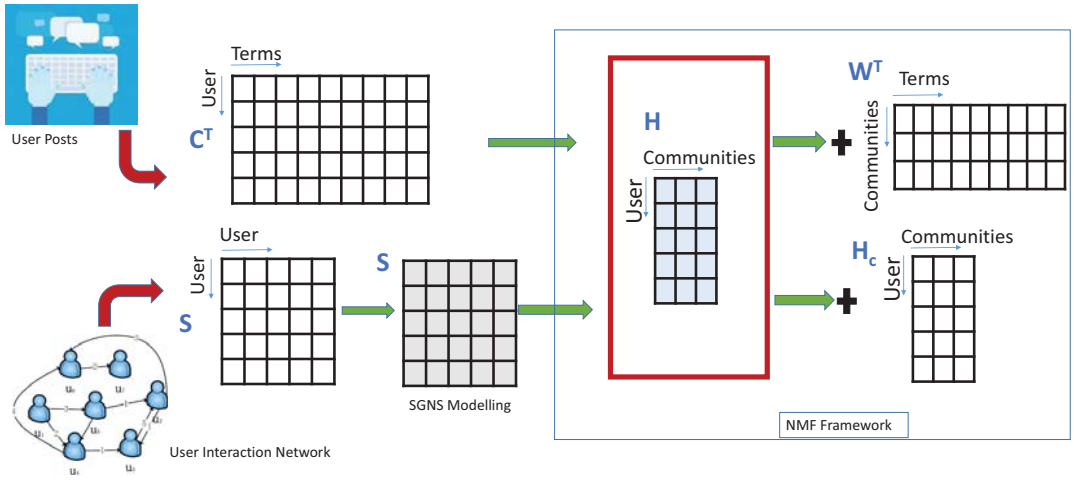

Fig. 2. Overview of the proposed Community Detection Approach

$\alpha, \beta < 1$ importance which learned empirically to each data type. Its objective function can be defined as:

$$\alpha \; min_{W,H\geq 0}\|C - WH^T\|_F + \beta \; min_{H,H_c\geq 0}\|S - HH_C^T\|_F \quad \text{where } \alpha, \beta < 1 \tag{2}$$

1) Solving the Optimization Problem: In each iteration of the optimization process of Eq. 2, entries of the factor matrices W, H and H_c are updated sequentially following the Block Coordinate Descent (BCD) algorithm [32]. The BCD algorithm divides the matrix members into several disjoint subgroups and iteratively minimizes the objective function with respect to the members of each subgroup $g \in G$ at a time.

$$W_{(:,g)} \leftarrow \left[W_{(:,g)} + \frac{(CH)_{(:,g)} - \left(WH^TH\right)_{(:,g)}}{(H^TH)_{(g,g)}} \right] \tag{3}$$

We start updating each matrix W as in Eq. 3 with random initialization of all the factor matrices, and sequentially update H and H_c for each $g \in G$ within each iteration as follows.

$$H_{(:,g)} \leftarrow \left[H_{(:,g)} + \frac{\left(C^TW\right)_{(:,g)} + (SH_c)_{(:,g)}}{(W^TW)_{(g,g)} + (H_c^TH_c)_{(g,g)}} - \frac{\left(HH_c^TH\right)_{(:,g)} + \left(HW^TW\right)_{(:,g)}}{(W^TW)_{(g,g)} + (H_c^TH_c)_{(g,g)}} \right] \tag{4}$$

CSMDC makes use of the most recent values of the associated factor matrices of W for updating H, and the most recent values of H is used for updating H_c.

$$H_{c(:,g)} \leftarrow \left[H_{c(:,g)} + \frac{(SH)_{(:,g)} - \left(H_cH^TH\right)_{(:,g)}}{(H^TH)_{(g,g)}} \right] \tag{5}$$

The user context modelling through SGNS concept is imposed on learning H which represents the user communities as H_c is used in updating H. Finally, CSMDC is able to effectively use multi-type data available with user communicated text messages and interactions for the user community matrix (H) learning process incorporating both C and S with their specific contribution.

C. Community Assignment

The overall algorithm of CSMDC for community detection is provided in Algorithm 1. The matrix $H \in R^{N\times G}$ shows the probabilities of a user $u_i \in U$ being in each community $g \in G$ with the coefficients in the users' row. CSMDC follows a hard cluster assignment policy for community detection. The final community assignment (i.e., a $g \in G$) for a user is determined according to the maximum coefficient that a user shows within its' respective row in H as in Eq. 6 and assign to the vector h^F.

$$h^F = g \leftarrow argmax \sum_{j=1}^{g\in G} \left(H_{(:,j)}\right) \tag{6}$$

Algorithm 1: Content and Structure Modelling for Discovering Communities (CSMDC)

Input : Term-User matrix C
User-User matrix S
Number of Communities G
Output: Final User-Community vector h^F
Init: $W \geq 0$, $H \geq 0$, $H_c \geq 0$, random real numbers

while *Convergence of Eq. 2* **do**
  foreach *g=1: G* **do**
    Compute W using Eq. 3
    Compute H using Eq. 4
    Compute H_c using Eq. 5
  end
end
Convergence: *old error − new error < 1e-3 OR number of iterations < 100*
Compute h^F using Eq. 6

TABLE I
DATASETS DESCRIPTION

Dataset	Users	Interactions	Tweets	Terms	Conductance	Classes
DS1:Cancer	1585	1174	8260	2975	0.152	8
DS2:Health	2073	2191	19758	5444	0.274	6
DS3:Sports	5531	19699	12044	3558	0.098	6

III. EMPIRICAL ANALYSIS

Three Twitter datasets obtained from TrISMA[1] [19] are used in experiments. A set of groups under the Cancer, Health and Sports domains were identified to collect posts and interaction information from the identified Twitter accounts. These information is used as the datasets and those groups in each domain are taken as the ground-truth communities. Table I provides the dataset description. The conductance is a property that shows the fraction of total edge strength that points outside the ground-truth communities in a user network. It indicates the level of inter-community interactions for a network. The sports dataset shows the lowest inter-community interactions while health dataset showing the highest inter-community interactions among the considered datasets.

Bench-marking has been done with three types of methods: (1) Popular content-based clustering methods, *k*-means [11], NMF [12] and LDA [13]; (2) Popular network-based community discovery methods, Louvain [4], InfoMap [5] and Clauset-Newman-Moore [6]; and (3) Multi-type community detection methods, MTRD-NMF [19], GenLouvain [15] and PMM [14].

The state-of-the-art MTRD-NMF takes multi-type data, content and structure matrices C and S as two input data types. However, multiplex network methods, GenLouvain and PMM that work with heterogeneous user relationships in network data, can only handle the symmetric input matrices. Therefore, we modify the matrix C into a symmetric user$\times$user matrix based on the number of overlapping terms between user pairs. The modified C and S become the input to GenLouvain and PMM to obtain the user communities.

The performance evaluation is done with standard pairwise F1-score which calculates the harmonic average of the precision and recall and, Normalized Mutual Information (NMI) which measures the purity against the number of clusters [7] to evaluate the accuracy of the community discovery (i.e. clusters of users).

A. Accuracy Analysis

Results in Table II show CSMDC, that uses both network and content information, produces more accurate communities as compared to the methods that use single type data on all data sets. It can be noted that NMF is able to identify user communities using content information with comparable performance on DS1 and DS2 to CSMDC. However, it (and all other content types methods) performs exceptionally inferior on DS3, where the number of inter-community interactions is comparably low. This dataset shows highly cohesive communities within the network representation (indicated by high conductance in Table I) comparatively. It is ascertained by the higher performance of the single type methods that use interaction network as input. In this dataset (DS3), the network-based methods achieve higher accuracy than content-based methods. The use of the interaction network information in community discovery results in a large accuracy difference between CSMDC and NMF in DS3. This confirms that CSMDC is able to learn the more accurate and meaningful user communities in the high-dimensional data by using both of this complementary information.

It should also be noted that CSMDC performs superior in comparison to all benchmarked multi-type input methods. CSMDC and MTRD-NMF both use NMF framework to learn multiple types of data in factorisation. However, due to inclusion of SGNS modeling technique that emphasises on including closely interactive users, CSMDC significantly outperforms MTRD-NMF. The performance gap is significant with the larger size datasets as in DS2 and DS3. The multiplex network-based GenLouvain and PMM methods, that use modified symmetric matrix using C and S as inputs, fail to bring a satisfactory result compared to CSMDC due to negligence of contribution of each data type for deciding the communities. CSMDC is able to incorporate the specific contribution of interaction network and content information in the community discovery process.

B. Sensitivity Analysis

The success of CSMDC relies on the data representation technique used for the input data. We analyze the different representations for text content as in Fig. 3 (a). The analysis with the use of popular terms weighting models *tf*, *idf* and *tf*idf* for text content show that *tf*idf*, which considers common terms as well as rare terms, can model the text communication of users in social network accurately to identify the communities.

The SGNS weighting used for matrix S that represents the interaction network information is one of the major strengths in CSMDC. We validate this network representation of using SGNS with just having interaction frequency as in Fig. 3 (b). It shows that the use of SGNS weighting for S consistently provides superior results for all the datasets capturing the closely interactive users with high probability.

Another important contribution of CSMDC is the consideration of each data type (i.e., text content and interaction network) as per their importance for community detection. We analyse through Eq. 2 using two weights α and β for content and interaction network matrix approximation respectively. We fixed either α or β to 0.1 and analyze how the performance

[1] https://trisma.org/

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT COMMUNITY DISCOVERY METHODS

Method	DS1		DS2		DS3	
	NMI	F1-score	NMI	F1-score	NMI	F1-score
Combined Multiple types - Content and Interaction Network						
CSMDC	**0.76**	**0.79**	**0.61**	**0.72**	**0.54**	**0.62**
MTRD-NMF	0.67	0.69	0.37	0.45	0.27	0.39
GenLouvain	0.52	0.50	0.46	0.62	0.48	0.53
PMM	0.26	0.38	0.18	0.33	0.14	0.37
Using only a Single type - Content						
k-means	0.72	0.74	0.50	0.59	0.07	0.36
LDA	0.41	0.47	0.38	0.60	0.0	0.32
NMF	0.76	0.79	0.60	0.69	0.09	0.37
Using only a Single type - Interaction Network						
NMF	0.21	0.38	0.14	0.44	0.39	0.56
Louvain	0.32	0.4	0.24	0.4	0.44	0.49
InfoMap	0.32	0.4	0.26	0.43	0.45	0.47
Clauset-Newman-Moore	0.32	0.42	0.25	0.43	0.53	0.58

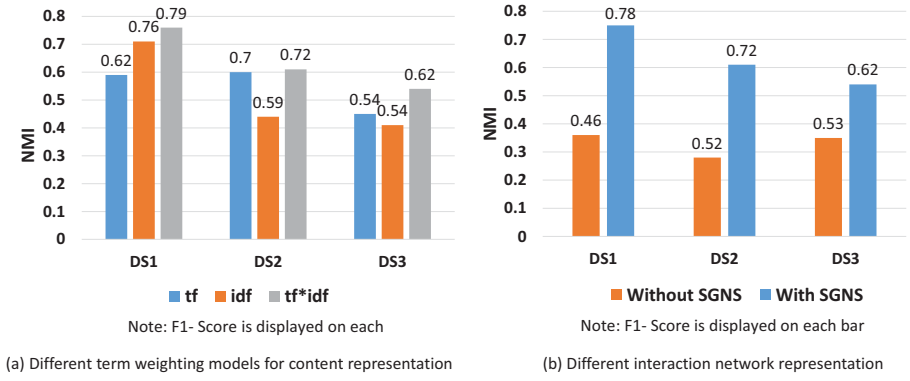

Fig. 3. Different data representation techniques and Performance

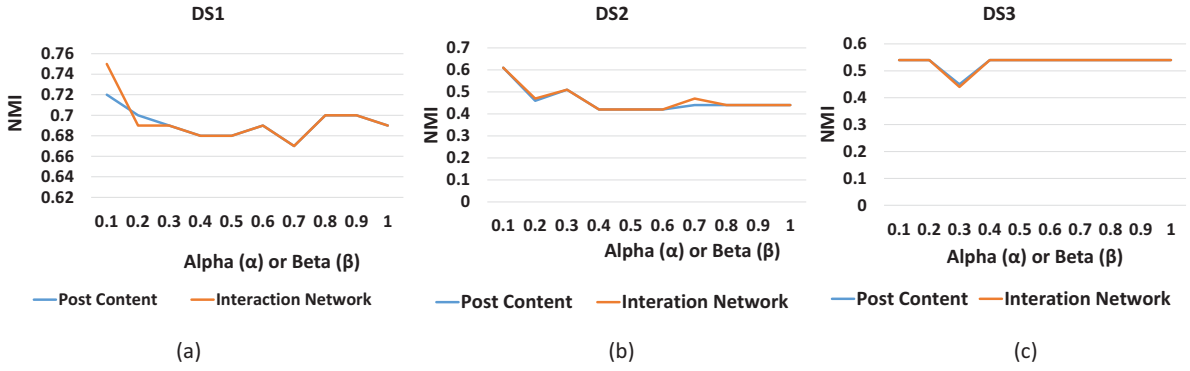

Fig. 4. The level of importance of content and interaction network data

varies for different values of the other variable. Figure 4 shows how the performance varies with the change of contribution from text post content in blue color line while orange color shows performance variation with the change of contribution from interation network. The maximum performance with the highest NMI is obtained when both α and β are set to 0.1 for all the datasets as in the Fig. 4.

Furthermore, we separately analyse the performance with NMI changing the β for different α as in Fig. 5 (a),(c) and (e) for DS1-DS3. Similarly, we analyse the performance changing the α for different β as in Fig. 5 (b),(d) and (f). In DS1 and DS2, the use of higher contribution of content information (> 0.1) for community learning gives lesser performance as shown by the lower NMI for increased α values (Fig. 5 (a)

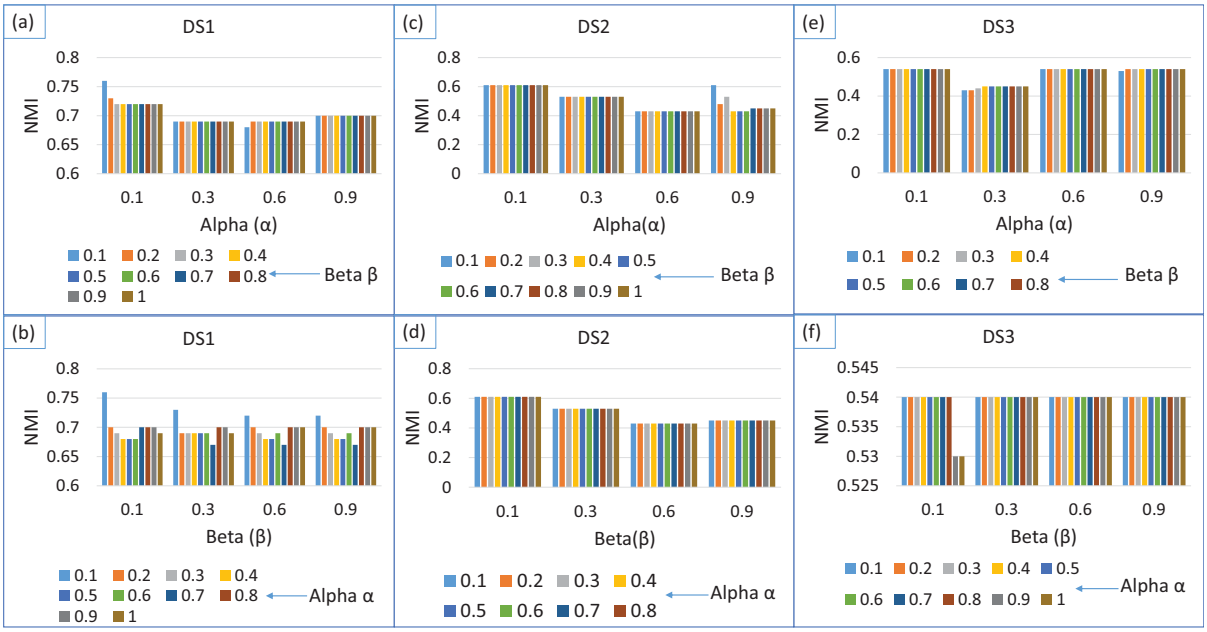

Fig. 5. The contribution of Multi-type data for community detection

and (c)). For these two datasets, almost similar outcome was obtained when the higher contribution of network information is used with increased β values as shown in Fig. 5 (b) and (d). However, DS3 shows a slightly different results by giving similar performance to the lowest (0.1) α and β and increased values as well (Fig. 5 (e) and (f)). In this dataset, the use of higher contribution from content and network information also allows to identify the communities with same accuracy due to tightly cohesive groups in network (as shown by the high conductance value). However, Fig. 5 confirms that the maximum NMI can be obtained for the lowest (0.1) α and β for all the datasets. Thus we set the objective function of CSMDC with α and β as 0.1 in Eq. 2.

IV. Conclusion

In this paper, we present a novel approach of coupled matrix-based NMF to identify user communities using both the network and content information inherent in social media. This paper explores how the content information present on social media (i.e., media posts) as well as user interaction contribute to the community discovery process. The proposed method CSMDC uses the concept of Skip-Gram with Negative Sampling to model network interaction information by meaningfully weight a user×user interaction considering other user interactions and incorporating that information in matrix factorization process. The experimental results on Twitter datasets and benchmarked with several state-of-the-art methods show the importance of considering multiple data types for achieving meaningful community groups. In the future, we will explore the applicability of this approach in learning overlapping community discovery problems.

References

[1] X. Hu and H. Liu, "Text analytics in social media," in *Mining text data*. Springer, 2012, pp. 385–414.

[2] R. Iyer, J. Wong, W. Tavanapong, and D. A. Peterson, "Identifying policy agenda sub-topics in political tweets based on community detection," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017, pp. 698–705.

[3] A. Park, M. Conway, and A. T. Chen, "Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach," *Computers in human behavior*, vol. 78, pp. 98–112, 2018.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[5] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[6] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[7] W. A. Mohotti and R. Nayak, "Corpus-based augmented media posts with density-based clustering for community detection," in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2018, pp. 379–386.

[8] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics reports*, vol. 659, pp. 1–44, 2016.

[9] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[10] W. A. Mohotti, D. C. Lukas, and R. Nayak, "Concept mining in online forums using self-corpus-based augmented text clustering," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2019, pp. 397–402.

[11] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[14] L. Tang, X. Wang, and H. Liu, "Community detection via heterogeneous interaction analysis," *Data mining and knowledge discovery*, vol. 25, no. 1, pp. 1–33, 2012.

[15] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science*, vol. 328, no. 5980, pp. 876–878, 2010.
[16] Q. Li, A. Nourbakhsh, S. Shah, and X. Liu, "Real-time novel event detection from social media," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 2017, pp. 1129–1139.
[17] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1089–1098.
[18] M. Qin, D. Jin, K. Lei, B. Gabrys, and K. Musial-Gabrys, "Adaptive community detection incorporating topology and content in social networks," *Knowledge-Based Systems*, vol. 161, pp. 342–356, 2018.
[19] T. M. G. Tennakoon, K. Luong, W. Mohotti, S. Chakravarthy, and R. Nayak, "Multi-type relational data clustering for community detection by exploiting content and structure information in social networks," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2019, pp. 541–554.
[20] C. He, Y. Tang, H. Liu, X. Fei, H. Li, and S. Liu, "A robust multi-view clustering method for community detection combining link and content information," *Physica A: Statistical Mechanics and its Applications*, vol. 514, pp. 396–411, 2019.
[21] M. Akbari and T.-S. Chua, "Leveraging behavioral factorization and prior knowledge for community discovery and profiling," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 71–79.
[22] M. Kim and J. Leskovec, "Latent multi-group membership graph model," *arXiv preprint arXiv:1205.4546*, 2012.
[23] Z. Yin, L. Cao, Q. Gu, and J. Han, "Latent community topic analysis: Integration of community discovery with topic modeling," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 4, pp. 1–21, 2012.
[24] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between lda and nmf based schemes," *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019.
[25] P. Lahoti, K. Garimella, and A. Gionis, "Joint non-negative matrix factorization for learning ideological leaning on twitter," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 351–359.
[26] Y. Pei, N. Chakraborty, and K. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.
[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[28] P. Lee, L. V. Lakshmanan, and E. E. Milios, "Incremental cluster evolution tracking from highly dynamic network data," in *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 2014, pp. 3–14.
[29] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.
[30] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations," in *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 2018, pp. 1105–1114.
[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
[32] J. Kim, Y. He, and H. Park, "Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework," *Journal of Global Optimization*, vol. 58, no. 2, pp. 285–319, 2014.