Physica A 512 (2018) 698–720

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

The structure of word co-occurrence network for microblogs

Muskan Garg [*], Mukesh Kumar

UIET, Panjab University, India

HIGHLIGHTS

- Comprehensive study of evolution of word co-occurrence network based applications.
- Analytical study of microblog word co-occurrence network using five different key parameters.
- Proposed BArank keyphrase extraction technique using key-parameters.

ARTICLE INFO

Article history:
Received 23 August 2017
Received in revised form 28 June 2018
Available online 14 August 2018

Keywords:
Word co-occurrence network
Microblog analysis
Small world network
Scale-free network
Keyphrase extraction
BArank

ABSTRACT

The study of structure and dynamics of complex networks is seeking attention of academic researchers and practitioners in recent years. Although Word Co-occurrence Networks (WCN) have been studied for different languages, yet there is the need to study the structure of WCN for microblogs due to the presence of ill-formed and unstructured data. In this research article, existing WCN based applications have been explored and microblog WCN have been analysed for multiple key parameters to uncover the hidden patterns. The key parameters studied for microblogs WCN are *scale-free property, small world feature, hierarchical organization, assortativity and spectral analysis*. The twitter FSD dataset has been used for experimental results and evaluation. Different mathematical, statistical and graphical interpretations proved that the microblog WCN are different from the WCN of traditional well-formed text. The robustness of the key parameters of microblogs WCN have been explored for keyphrase extraction from domain specific set of microblogs. The baseline methods used for comparisons are TextRank, TopicRank, and NErank. Extensive experiments over standard public dataset proved that the proposed keyphrase extraction technique outperforms the existing techniques in terms of precision, recall, F-measure, and ROUGE scores.

1. Introduction

Human understandable information is posted by naïve users on internet using social media to communicate and express views. One of the most frequently used social media platform is Twitter [1] for communication. Increasing use of Twitter, has made it popular and significant for social media analysis of views communicated by naïve users. Manual analysis of huge amount of user-generated data on social media is time consuming task. Some of the widely studied application for twitter data analysis are topic detection and tracking [2–4], sentiment analysis [5] and latent pattern mining [6]. The text of twitter feed is usually shorter than that of traditional news and thus, may have different statistical significance. For analysing Twitter content, the major challenges are huge and versatile information, unstructured data, use of multiple human understandable languages, and dissemination of rumours. The traditional approach for twitter data analysis is based on supervised learning

[*] Corresponding author.
E-mail addresses: muskanphd@gmail.com (M. Garg), mukesh_rai9@yahoo.com (M. Kumar).

https://doi.org/10.1016/j.physa.2018.08.002
0378-4371/© 2018 Elsevier B.V. All rights reserved.

Table 1
Keywords identified from three twitter feeds from FSD.

Index	Word	Index	Word	Index	Word
1	Hmm	2	Sunday	3	Mirror
4	News	5	Jasav	6	Tweeting
7	Confirmation	8	Amy	9	Winehouse
10	Died	11	Maybe	12	Rumour
13	Dean	14	Piper	15	Appears
16	Tweeted	17	jlscrazymad	18	Rumours
19	Remains				

which could be inefficient due to huge or versatile information in which the trained dataset is not suitable for time-series analysis; topic modelling [7] based measures which may not be useful for uncertain ill-formed data due to bad semantics of ill-formed text. The idea for this research work is to explore the field of complex networks over WCNs [8] for microblogs.

WCN [9] is the graph with words as nodes and two words occurring together are linked by an edge. The edge weight is the frequency of words occurring together. The WCN evolved from user-generated data is termed as microblog WCN which has been shown in Fig. 1 for Twitter feeds. Three tweets have been used from First Story Detection (FSD) dataset [2] to display the graphical representation of different types of WCN as shown in Fig. 1.

> *Tweets: {1: 'Hmm Sunday Mirror news ed @jasav is tweeting "confirmation" that Amy Winehouse has died.', 2: 'Maybe a rumour but Dean Piper appears to have tweeted that Amy Winehouse has died', 3: '@jlscrazymad rumours amy winehouse has died- remains rumours'}*

The twitter text is pre-processed before identifying keywords from twitter feeds. The irrelevant symbols and stop-words have been removed. The 19 words which are identified as keywords have been recorded and indexed as nodes as shown in Table 1. However, automatically identifying related words from Twitter data [10,4], collocation techniques and other traditional text normalization [11] has been quite popular in recent years. Further, in future, the stemming of words like tweeted and tweeting, rumours and rumour can be accounted as a single node as per requirements.

The microblog WCN have been evolved from co-occurrence of words in twitter feeds as shown in Fig. 1. The pictorial representation of networks in Fig. 1, shows different types of possible WCN for microblogs. The WCN have been broadly classified into three categories [12] namely:

1.1. Nearest neighbour edging (NN) and all pair neighbour (AP) edging

For any phrase/sentence or set of words, the WCN can be mapped using words as nodes and co-occurrence as links among these nodes. The word to word connectivity may follow nearest neighbour edging or all pair neighbour edging. As shown in Fig. 1, the graphs (a), (b), (c) and (d) represent nearest neighbour edging in which the words of phrase which are adjacent to each other are linked together. However, the graphs (e), (f), (g), (h) represent all pair neighbour edging where every word of phrase is linked to every other node to indicate that they belong to same phrase. For instance, in Fig. 1(e), the graph clearly indicates that the set of nodes form three well connected groups for three twitter feeds are [12,13,14,15,16], [10,18,19] and [2,3,4,5,6,7,8]. It can be observed that the three networks are well connected by nodes [8,9,10] which can be decoded as 'Amy Winehouse died'.

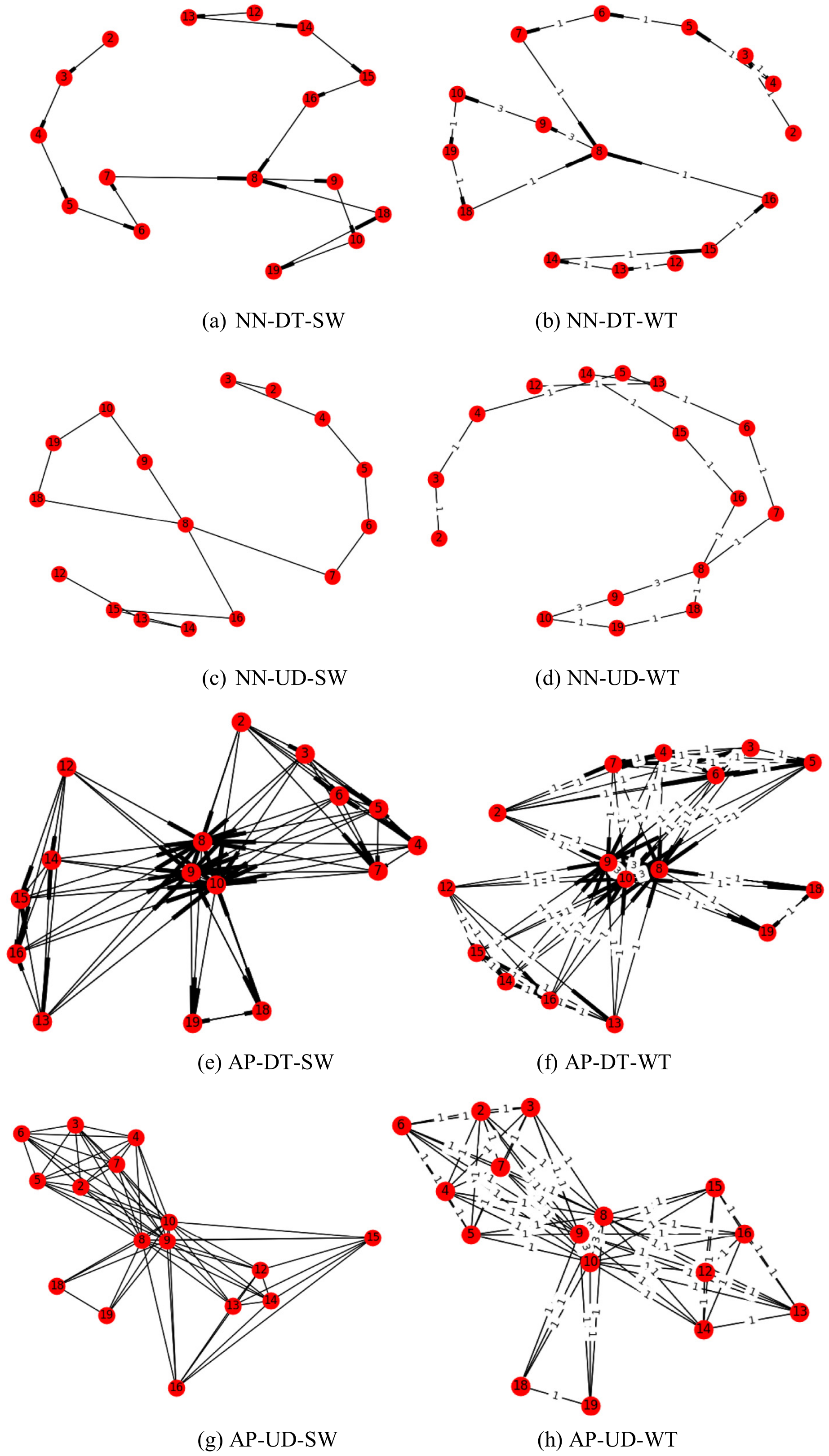
1.2. Same weight (SW) edges and weight as co-occurrence frequency edges (WT)

Edges are the links used for word co-occurrence. To map the word co-occurrence frequency, weight is assigned to an edge. The weight of an edge (u, v) gives the count of number of times the words u *and* v occurs together in the document. On contrary, the weights of all edges is assigned as one if word co-occurrence frequency is not significantly required in WCN. As shown in Fig. 1, the graphs (a), (c), (e), (g) represent same edge weight in the WCN and the graphs (b), (d), (f), (h) represent edge weight as co-occurrence frequency. For instance in Fig. 1(b), it clearly shows that [8,9] and [9,10] are well connected due to high edge weights. This can be decoded as 'Amy Winehouse' and 'Winehouse died' which are important phrases.

1.3. Directed graph (DT) and undirected (UD) graph

The WCN can be directed or undirected. The directed networks as shown in Fig. 1(a), (b), (e), (f) preserve lexical sequence of the phrase. In Fig. 1(a), the edge (2, 3) indicates edge direction from 2 to 3 decodes as 'Sunday Mirror'. However, the undirected networks as shown in Fig. 1(c), (d), (g), (h) contains edges which do not provide direction between two words. Although the direction preserves the lexical sequence of the networks, the undirected network is used to analyse the structure and dynamics of WCN better.

The major contributions of this research article are



(a) NN-DT-SW (b) NN-DT-WT

(c) NN-UD-SW (d) NN-UD-WT

(e) AP-DT-SW (f) AP-DT-WT

(g) AP-UD-SW (h) AP-UD-WT

Fig. 1. Different types of microblog WCNs.

- Comprehensive study of the existing WCN based applications in the literature. The research gaps have been identified for ill-formed, uncertain, and user-generated data from the literature.
- Analytical observations of structure of microblog WCN for key parameters over FSD dataset. The key parameters includes *scale-free property, small world feature, hierarchical organization, assortativity and spectral analysis.*
- To study the robustness of structural analysis of microblog WCN, an application has been proposed as BArank for identifying keyphrase from set of domain specific twitter feeds using structural analysis of microblog WCN. The mathematical and experimental results shows that the convergence of disassortative network to non-disassortative network gives useful insights about topic detection in microblog word co-occurrence network.

The idea behind this research article is to understand the complex network of words adjacency for ill-formed user-generated microblogs which contains slangs and redundant information. In this research work, many statistical key parameters have been analysed and it has been observed that many more can be studied for dynamics of microblogs WCN in near future. The WCNs [13,14] have been studied by many academic researchers and practitioners. The application domain for microblog WCN analysis are: identifying useful insights from social media data; spam detection from emails dataset; and topic detection and tracking from uncertain user-generated data. The dataset used for this comprehensive study is First Story Detection [2] dataset. It has been observed that the co-occurrence of words within user-generated microblogs contains words from active language and active vocabulary of naïve users. The motivation behind this study is the need to study the behaviour and semantics of WCN for microblogs. This has been proved to be useful in many applications including topic detection [15], event detection [3,16], keyword extraction [17–19], keyphrase extraction [20,4] and author attribution [21,22] as discussed in Section 2.

The robustness of the key parameters for microblog WCN has been explored for keyphrase extraction from domain specific set of microblogs. Earlier, the keyphrase extraction from textual documents over short text has been studied using random walk measures in WCN. However, the user-generated data may contain many new terms which may be related to new events, topics, or trends which are being discussed, location specific slangs and abbreviations, geo-located popular named entities, and trend based hashtags which may not be grammatically correct or connected. The general discussion about happenings may contain common terms which may not be grammatical but are related and relevant to each other. The related terms shall connect with each other in WCN of microblog. Consequently, this WCN show specific patterns and thus, key parameters have been explored for summarizing user-generated short text. The significance of assortativity over microblog WCN is shown in BArank algorithm as proposed in Section 4. The impact of the BArank can be observed in Section 4.5 where BArank gives highest precision for identifying set of important words as keyphrase.

This article has been structured as follows: Section 2 contains evolution and comprehensive study of existing WCN based applications in literature. The structural analysis of WCNs and analysis for key parameters have been discussed in Section 3. The key parameters includes scale-free property in Section 3.1, the small-world feature in Section 3.2, the hierarchical organization in Section 3.3, the assortativity in Section 3.4, and the spectral analysis in Section 3.5. Based on above investigations, BArank has been proposed using analysis of key parameters. The experimental results and evaluation of BArank, and other baseline keyphrase extraction measures for microblog WCN has been shown in Section 4. Finally, the conclusions and future work have been discussed in Section 5.

2. Historical perspective of word co-occurrence network

The real time WCN for social media data largely depends upon the textual information generated by the naïve user. This information is usually ill-formed and several slangs and short notations have been used due to upper limit on number of characters (140 characters). However, due to character limitation, users tend to share minimal and important information related to topics, trends, and events [16]. Twitter feeds also contain information about sentiments, feedback and opinions. The twitter feeds contain named entities for identifying topics; adjectives, adverbs, semantic information for machine learning, and syntactic information for pattern mining.

Although complex networks [23] have received much attention from researchers in recent years, yet WCNs have been minimally explored. The text authorship has been calculated [21] using dynamics of the WCNs. The study of WCNs has been carried in recent years for English and Chinese [9] languages. Further, the research has been improved over English and Chinese Poems [24], spectral analysis of Chinese language [25] from literary genres and English language [14]. The author [26] compared directed and weighted WCN of six different languages including Arabic, Chinese, English, French, Russian and Spanish. In contrast to this, WCN dynamics have been uncovered for well-formed English book data, Twitter feeds and Facebook data [27]. The author calculated statistical parameters for undirected and weighted edges including average degree, modularity, average clustering coefficient, diameter, path length, edge weights and maximum degree differences of pairing nodes. Further, Liu and Cong [8] considered 12 Slavic and 2 non-Slavic languages. They used WCNs for classifying Slavic and non-Slavic languages using syntactic dependency complex networks. Different analysis have been performed by Beliga et al. [18] for structural analysis of structured well-formed languages, however, there is need to study WCN of unstructured data. It has been observed that degree distribution and small world features have proved to be useful for keyword and keyphrase extraction. Also, Wuchty and Almaas [28] used the co-occurrence network for biological analysis of structure of protein domain network. The major applications where WCNs have been used are keyword identification, automatic extractive summarization and authorship attribution [22].

2.1. Keyword extraction

Keyword extraction is one of the most important segment of research in the domain of natural language processing. *KeyWorld*, an automatic indexing system has been proposed by Matsuo et al. [29] which extracts candidate keywords by measuring their influence on small-world properties. It captures characteristic path length and extended characteristic path length to calculate contribution, the proposed measure. This algorithm has been inspired by small-world phenomenon and keyGraph algorithm proposed by Ohsawa et al. [30]. However, it is useful when used with Inverse Document Frequency (IDF) measure. In 2007, Palshikar [31] proposed hybrid and statistics based approach for keyword extraction using co-occurrence frequency measure. The author described eccentricity, other centrality measure and proximity based keyword identification. Results gives 3 or 4 useful keywords out of 10.

The author [19] proposed Twitter Keyword Graph (TKG) algorithm to extract keywords from Twitter data. The author introduced all neighbour edging and nearest neighbour edging for constructing graph, frequency based and inverse-frequency based weights in graph and different centrality measures. The author used centrality measures for different types of networks as mentioned in Fig. 1. Additionally they also used third type of edge weight given as inverse co-occurrence frequency which has been proved to be superior parameters than same edge weight and co-occurrence frequency edge weight. Another algorithm named selectivity-based keyword extraction (SBKE) has been proposed by Beliga et al. [18] to extract keywords from Croatian news using F1 and F2 measure. In F1 measure, the equal importance of recall and precision is considered. F2 measure is F measure in which precision is two times more important than that of recall. Recently, Batziou et al. [32] used Graph of words (GoW) and centrality measures for keyword extraction. The author also proposed community detection algorithms using GoW and centrality measures to identify the largest community containing key nodes (words) in graph of words. The proposed measures are two parameters namely mapping entropy betweenness (MEB), and mapping entropy centrality (MEC) and have been tested over Jaccard, average precision, and P@10 measures for linking of number of words (N), $N = 2$, and $N = 3$. It has been observed that MEC and MEB outperforms other baseline measures after closeness and performs better for $N = 3$. Also, the author [33] proposed CoreRank to identified keywords from WCN using k core decomposition model [34]. The author used ROUGE-1, and Word Mover's Distance (WMD) performance measure. ROUGE-1 computes similarity based on unigram overlap, while the WMD takes into account semantic similarity between terms, and is therefore more robust to the fact that the abstractive summaries contain words that were never actually spoken. Recently, the graph based keyword extraction model has been proposed using collective node-weight (KECNW) [35]. The proposed KECNW outperformed TKG and SBKE.

2.2. Text summarization

Key-phrase extraction or topic detection is the method of identifying an important segment which describe a set of segments. Further, trend and event detection are significant applications which use text summarization in WCN. In 2004, Erkan and Radev [36] proposed LexRank which is insensitive to noise in text and calculates importance of sentence (or word) using eigenvector centrality. Spectral analysis has been used for community detection of keyword which belong to same class [37]. In 2008, the author [38] proposed HITS based algorithm for keyphrase extraction. In 2009, for event detection and tracking in social streams, [3] used keyGraph algorithm which was proposed earlier by Ohsawa et al. [30]. In 2012, Bellaachia and Al-Dhelaan [39] used graph based approach on text for keyphrase extraction. [40] proposed GRAPHSUM, a novel and general-purpose summarizer based on graph model which represents correlations among multiple terms by discovering association rules. For disaster based event detection techniques, Rudra et al. [41] used bi-gram WCN in which each node contains bi-gram for summarization of disaster based events. Word network topic model (WNTM) is the hybrid approach which use WCN and topic model for topic detection. Similar approaches have been proposed for short text by Chen and Kao [7] and as Enriched-LDA by Shams and Baraani-Dastjerdi [42]. The author used multiple attribute decision making optimization algorithm analytical hierarchical process on proposed attributes based on WCN to identify influential segment [4] for social media data.

The random walk based measures for identifying keyphrase from the textual data has been widely studied in literature. Earlier, PageRank algorithm have been used by the author for topical keyphrase extraction from twitter data [43]. In 2014, the author proposed NErank [39] as random walk based keyphrase extraction technique for twitter data using proposed node score and edge score. The NErank gives better results than TextRank and PageRank for twitter data. In this research work, the baseline measures used for keyphrase extraction from domain specific set of microblogs are TextRank, TopicRank, and NErank.

TextRank is the graph based ranking model for text processing which introduced two successful unsupervised approaches for keyword and sentence extraction. The TextRank has believed to be the baseline measure for most of the algorithms. Another traditional measure, TopicRank [44], outperforms all the existing measures other than NErank. The TopicRank extracted noun phrases and clustered them to use as nodes in WCN. The noun based information gives better results. However, in the microblogs, extraction of noun phrases for ever-changing and dynamic text is still under research. It has been observed from the evolution of keyphrase extraction techniques that the graph based algorithm for identifying keyphrases from twitter data is NErank. The NErank approach of keyphrase extraction has been proposed by using node degree as node score and edge weight as edge score with additional proposed measures. However, the high degree nodes getting connected to high degree nodes add valuable information to high weighted edges. This signifies that assortative or non-disassortative

Table 2
Structure of microblog WCN for different size of corpus.

Dataset	#Tweets	L	N	E	A_n	ASPL	CC
FSD: W/1000	100	748	527	550	7.480	2.574	0.0142
FSD: W/100	1 000	7 359	3 560	5 292	7.359	4.154	0.0142
FSD: W/20	5 000	36 369	11 307	25 022	7.273	3.956	0.0230
FSD: W/10	10 000	73 243	18 433	49 011	7.324	3.696	0.0348
FSD: W/4	25 000	183 466	34 925	116 477	7.338	3.420	0.0493
FSD: W/2	50 000	366 243	56 259	219 950	7.324	3.214	0.0682
FSD: W	100 000	481 982	67 979	281 802	7.320	3.144	0.0765

complex networks could identify important information or topic of discussion. The patterns of WCN for well-formed text using traditional measures proved to be useful. However, the uncertain user-generated data is unstructured, and ill-formed. The existing conventional approaches used random walk but are not suitable for microblogs due to repetition of same word in different forms and thus, multiple nodes. The pattern of WCN as studied in our previous studies [4] has proved to be useful for extracting keyphrases. Thus, non-parametric algorithm, BArank, has been proposed for identifying keyphrases in domain specific microblog WCN. This non parametric reduction of microblog WCN using k – bridge decomposition has proved to be useful for topological sorting based keyphrase extraction.

2.3. Author attribution

During 2013, Segarra et al. [45] used word adjacency network for authorship attribution analysis. Later in 2015, Amancio [46] used fluctuation analysis of network topology and word intermittency for authorship recognition. In 2016, Marinho et al. [22] used network motifs for authorship attribution. The authors shows that motifs, which are the repeating sub-network of set of words, are able to distinguish the writing style of authors. Recently, Akimushkin et al. [21] used dynamics of WCN for text authorship identification.

It has been observed that the existing studies for WCN of well-formed text gives promising results [18]. The structure and dynamics of microblog WCN may vary due to topic and sentiments specific repetitive information in twitter feeds as observed in conversations [33]; domain specific different information as obtained during emergency situation and natural hazards; and large amount of diversity in user-generated information. Words with different forms of presentation like those of spelling mistakes, slangs, and short-hand notations, may give indifferent results with multiple nodes representation. The streaming data may contain topic specific frequent information, and cliques can be mapped with time. Thus, there exist research gap and need to study the structure and dynamics of microblog WCN which may provide useful insights and latent patterns in conversations.

3. The structure of microblog word co-occurrence network

The structure of WCN for social media data could be different. This may vary due to huge amount of short text data which is limited to 140 characters and may contain redundant information. The dataset used for microblog WCN analysis is FSD dataset. The number of tweets used for analysis are 100k and is marked as W. For less number of computations, the reduced the number of tweets W/2, W/4, W/10, W/20, W/100, and W/1000 datasets are used. Although, the behaviour of all datasets are similar, the observations are more acute as the number of tweets increases. Initially, the structure of network science [23] has been studied for microblog WCN. The type of WCN considered for observations are *weighted (WT)* and same weight edges considered as *unweighted (UW)* in this study, *directed (DT)* to preserved lexical sequence of the data, and *nearest neighbour edging (NN) or all pair neighbour edging (AP)* network as shown in Fig. 1. The all pair neighbour edging network is computationally expensive, precisely for large amount of data.

The structural observations for each set of data includes length of the network (L) which indicates the total number of words in the dataset, number of nodes (N), number of edges (E), average number of nodes in a twitter feed (A_n) , average shortest path length ($ASPL$), and clustering coefficient (CC) as shown in Table 2. The ASPL is defined as shown in Eq. (1).

$$ASPL = \frac{2\sum_{i>j} d_{ij}}{N(N-1)} \tag{1}$$

where d_{ij} is the shortest path length between two vertices i *and* j that connects them. To minimize the sampling bias for each $W/s - sized$ set, the experiments have been repeated s times and averaged over these instances. The semantics of the network are considered to be significant and stable only if the observations for variable size and different sets of data have same behaviour [47]. In this research work, although we have observed the results for $W/1000$, this gives 100 tweets which may not be suitable for reliable network analysis due to small size. Thus, for further observations, the network with data size $> W/1000$ is used.

As shown in Table 2, it has been observed that the length of the dataset, number of nodes, and edges, increases with increase in size of corpus. Also, the CC shows slight increase with increase in the number of nodes, and the average number of nodes gives more acute observations and variation decreases with increase in size of the corpus. ASPL decreases with

Table 3
Different values of γ for microblog WCN for different size of corpus.

Dataset	K_{in}	K_{out}	K	S_{in}	S_{out}	S	E_w
FSD: W/100	3.320	2.996	2.852	3.300	3.010	2.740	1.361
FSD: W/20	2.376, 4.141	2.349, 4.150	2.145, 3.183	2.309, 4.150	2.240, 4.154	2.120, 2.960	1.340, 4.122
FSD: W/10	2.137, 3.485	2.108, 3.472	2.005, 2.950	2.064, 2.019	2.037, 3.018	2.028, 2.743	1.326, 3.548
FSD: W/4	1.993, 2.828	2.002, 2.843	1.931, 2.491	1.991, 2.529	2.013, 2.550	1.902, 2.224	1.304, 3.094
FSD: W/2	1.921, 2.528	1.906, 2.561	1.825, 2.166, 3.998	1.893, 2.293	1.876, 2.216	1.782, 2.002, 3.095	1.285, 3.063
FSD: W	1.874, 2.401	1.857, 2.373	1.803, 2.096, 3.684	1.837, 2.134	1.827, 2.049	1.774, 1.930, 2.707	1.277, 2.948

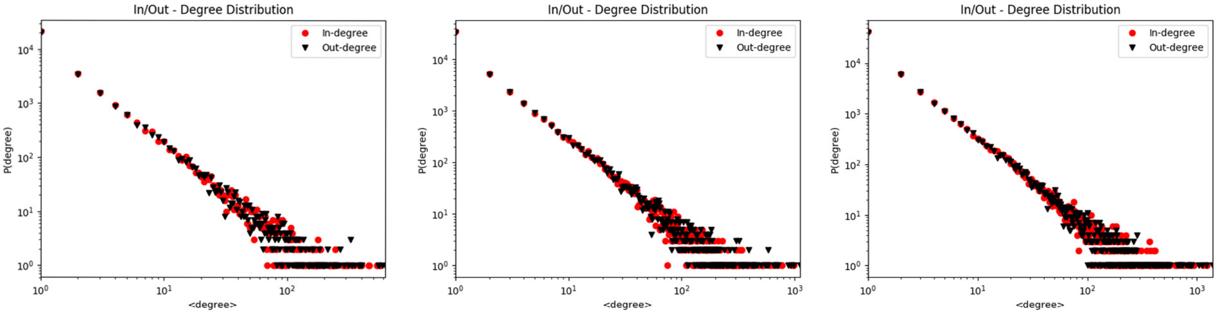


Fig. 2.1. In-out degree distribution for microblog WCNs for W/4, W/2 and W corpus.

increase in size of the corpus. The rate of increase in number of nodes is much lower as compare to increase in corpus size. However, the rate of increase of number of edges shows linear dependency on number of tweets for corpus with size $W/1000$ to $W/2$, and tends to decrease thereafter. The ASPL shows 19% to 30% variation and CC shows 54% to 81% decrease when order of magnitude is reduced by 10^2 *or* 10^3. It has also been observed that the average number of words in the sentences are 7 *or* 8. Thus, the result shows the stability of microblog WCN.

The WCN of social media data contains redundant nodes with syntactic errors of edit distance, slangs, and short-hand notations. However, understanding the structure and semantics of the network formed by user-generated data helps to find solution for identifying keywords from unstructured text using statistical measures. In this section, different key parameters have been studied including scale-free property, small world feature, hierarchical organization, assortativity, and spectral analysis.

3.1. Scale-free property

The degree distribution $p(k)$ is defined as the probability that a randomly chosen node has degree k [24]. The power law is defined as $p(k) \propto k^{-\gamma}$, where $2 < \gamma < 3$ for scale-free networks. The scale-free property indicates that the network does not scale over the parameter chosen for analysis. Here, degree of the node for unweighted networks, and strength of the node for weighted networks, and edge weights are considered as parameters for evaluation. The value of γ may vary as

- $\gamma \leq 2$: Shows anomalous behaviour. This implies that largest hub should have degree $> N$ which is not possible.
- $2 < \gamma < 3$: Gives scale-free regime. This indicates that the first moment is finite and second moment diverge as $N \to \infty$ where N is number of nodes
- $\gamma > 3$: Gives random network regime. This shows that second and third moments are finite.

The WCN is an evolving word web and follows two regime power-law [48] for well-formed text. It has been observed that $\gamma < 1.5$ for value of $k < 10^3$, and $\gamma > 2.7$ for $10^3 < k < 10^5$ where k is the degree. The degree distribution have been explored for in-degree, out-degree, and total degree of the microblog WCN. The probability distribution of node in-degree and out-degree has been shown in Fig. 2.1 and for total degree has been shown in Fig. 2.2. It has been observed from the graphical representation that average number of incoming links for degree k is approximately equivalent to average number of outgoing links. Different values of γ for in-degree (k_{in}), out-degree (k_{out}), total degree (k) of the nodes have been shown in Table 3. It has been observed that in-degree, out-degree, and total degree, follows power law beyond $k > 10^2$. This indicates that the microblog WCN follows power-law precisely unlike traditional WCN which follows two-regime power law. This is due to the fact that the keywords are repetitive, and short text contain repetitive and limited information. Due to power law, words with high degree, medium degree and low degrees have been identified. Most of the high degree words are the common terms which are neither stopwords nor keywords for instance '*today*', '*time*', '*people*', '*good*'. Similarly, medium degree words like '*dead*', '*crashes*' may provide important information. This is probably due to the fact that multiple topics have been discussed in user-generated data and common terms are contained in more number of tweets than specific domain keywords.

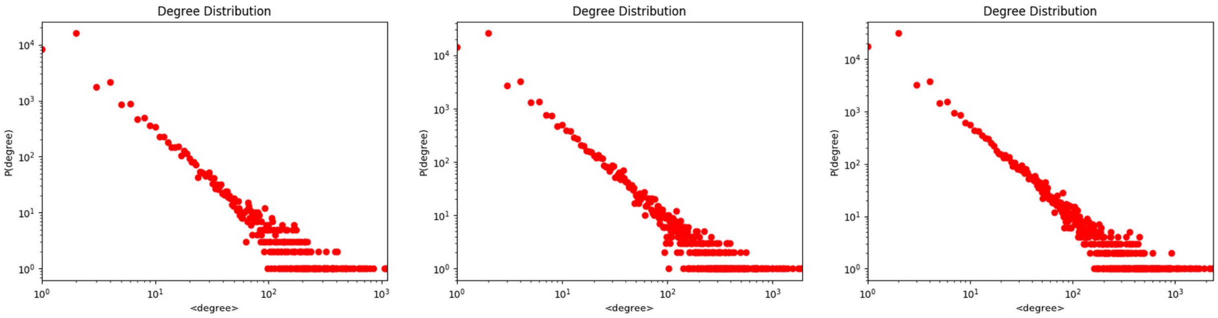


Fig. 2.2. Degree distribution for microblog WCNs for W/4, W/2 and W corpus.

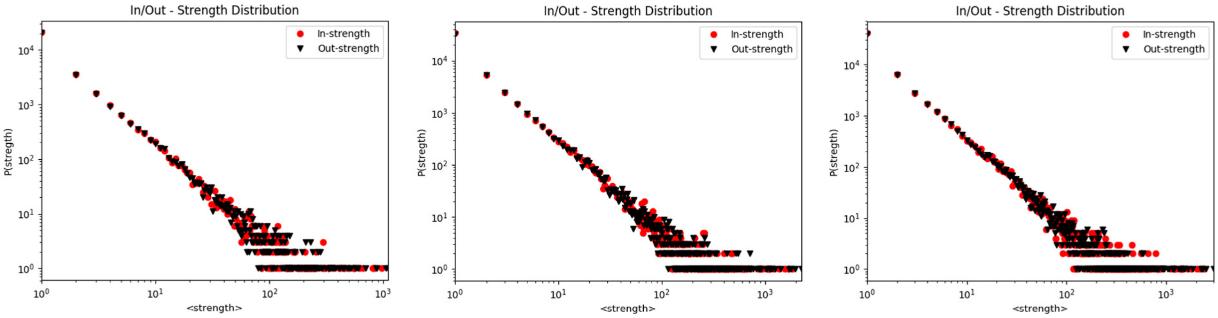


Fig. 2.3. In-out strength distribution for microblog WCNs for W/4, W/2 and W corpus.

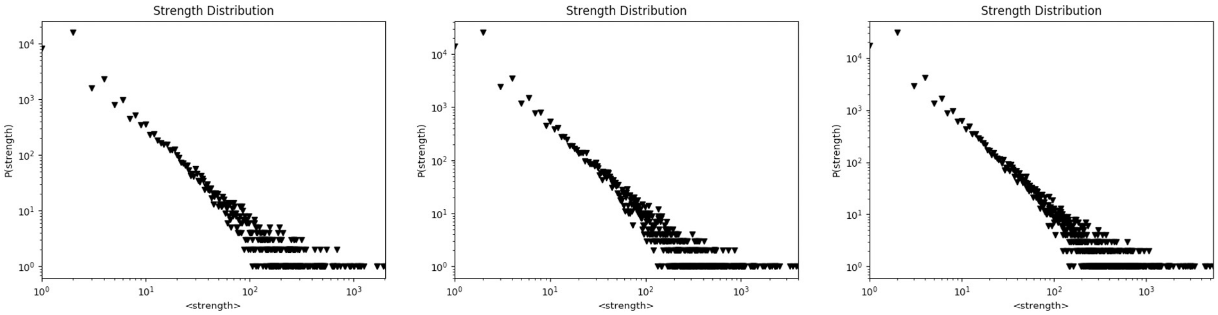


Fig. 2.4. Strength distribution for microblog WCNs for W/4, W/2 and W corpus.

The Figs. 2.3, and 2.4 shows the in-strength and out-strength distribution, and total strength of the node, respectively where strength indicates the total weighted degree of the node. Also, Table 3 shows different values for in-strength, out-strength and total strength distribution which follows power-law and thus, the network is scale-free in terms of node strength as well beyond $k > 10^2$.

Observations have been made for edge weight distribution in addition to node degree. This is because edge-weight plays pivotal role in identifying frequently co-occurring words which might give better results than that of node degree. It can be observed that the network shows scale-free property for the value of $\gamma > 2$ for edge-weight > 10 as shown in Fig. 3.1. This shows high modularity of the network. Thus, community detection can be performed using network motifs [47] and modularity based techniques to identify trending topics. All the results have been obtained for W/4, W/2 and W dataset where W contain 100 000 tweets of the FSD dataset. The edge strength (ES) is the degree of two nodes connecting with each other than being connected to other nodes. It has been defined [49] as shown in Eq. (2).

$$ES = \frac{freq(t_i, t_j)}{freq\,(t_i) + freq\,(t_j) - freq(t_i, t_j)} \tag{2}$$

The observation for edge strength is shown in Fig. 3.2. The log–log graphical representation of edge strength and probability of distribution of edge strength have been observed. The visualization of edge strength distribution signifies that beyond some threshold which is believed to be 10^2 for the observations in 3.2, the distribution follows some pattern of

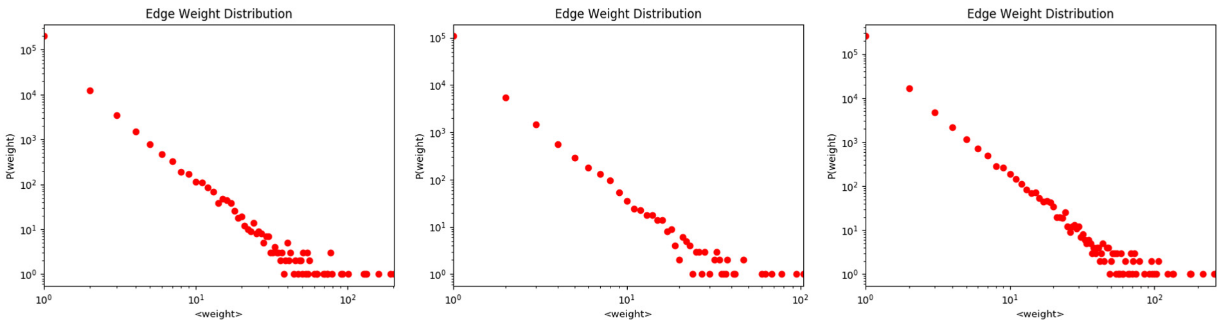


Fig. 3.1. Edge weight distribution for microblog WCNs for W/4, W/2 and W corpus.

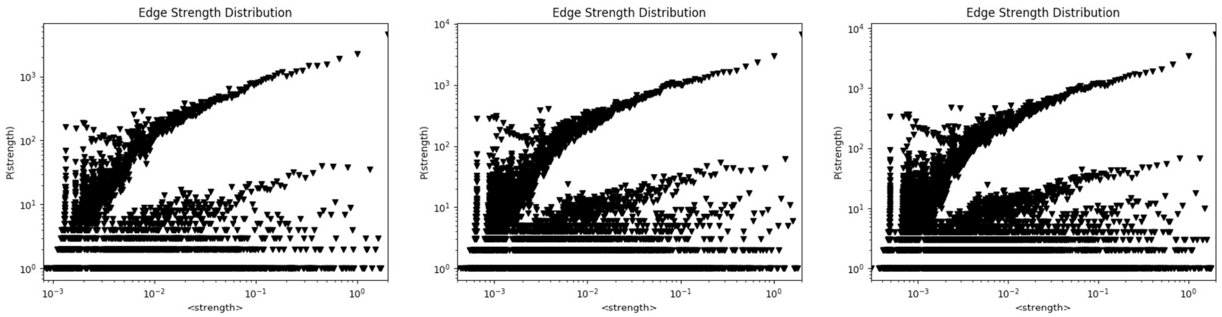


Fig. 3.2. Edge strength distribution for microblog WCNs for W/4, W/2 and W corpus.

dependency. It has been observed that more the value of edge strength closer to 1, more connected are the corresponding nodes and vice-versa.

The observations for scale-free property shows that the network follows core–periphery structure. The core nodes are the nodes having high degree and periphery nodes are the nodes with low degree. Similarly, the core edges are the edges with high edge weight and periphery edges are the edges having low edge weight. It has been observed from Tables 2 and 3 that as the network is stable and for dataset W/4, W/2 and W, the network is similar and thus, further observations have been made for W/4 to reduce computations. Moreover, during data streaming, the data should contain small but reliable set of tweets for making useful insights and summaries. The dataset W/4 contains $> 10^5$ number of nodes in the network which is reliable for analysis of microblog WCN.

In future, the preferential attachment model can be used for identifying communities in temporal data streaming analysis. Moreover, the all-pair neighbour edging network are fully connected networks for each microblog and thus, follows scale-free property. However, the applicability of all-pair neighbour network for scale-free network is computationally expensive, thus, the applications of all-pair neighbour edging network are minimal. In this research work, the degree distribution gives the common terms which are used by users randomly. However, the edge distribution gives the co-occurrence based n-grams which gives more promising keywords than that of degree distribution. Thus, edge weight scale-free property has been used to propose keyphrase extraction algorithm.

3.2. The small world feature

Complex networks can be validated for small world feature [50] using average shortest path length (L) and clustering coefficient (CC) for WCN and Erdos–Renyi (ER) random network [51]. The network is said to possess small world network, iff $L \approx L_r$ and $CC \gg CC_r$ where L and L_r are the average shortest path lengths, CC and CC_r are clustering coefficients of WCN and ER random graph, respectively. As observed in Table 4, The CC and CC_r for all the datasets W/1000, W/100, W/20, W/10 and W/4 has been observed in Table 4. Although, the CC_r decreases with increase in number of tweets and hence, number of nodes beyond dataset W/20, the value of CC decreases with increase in size of the corpus. This might be due to the fact that as observed in Table 2, with increase in number of nodes, the number of edges increases at higher rate than that of number of nodes and thus create more links. More the number of links, more connected will be network and higher will be clustering coefficient. Different application based observations can be deduced from CC of the network. The CC for microblog WCN for dataset W/20, W/10 and W/4 is 61 times, 133 times, and 365 times than that of CC_r, respectively. The huge variation in CC for microblog WCN than that of ER-random network indicates that the network is small-world.

Table 4
Average shortest path length of WCN.

Dataset	CC	CC_r	Directed/Undirected	L	L_r	$L(\max)$	$L_r(\max)$
FSD: W/1000	0.0142	0	Directed	0.938	5.350	2.574	5.367
			Undirected	2.422	1.575	11.431	7.481
FSD: W/100	0.0142	0	Directed	0.916	6.617	4.654	6.611
			Undirected	1.576	1.354	6.090	7.553
FSD: W/20	0.0238	0.00039	Directed	0.710	6.266	3.956	6.253
			Undirected	1.315	2.355	4.829	6.431
FSD: W/10	0.0348	0.00026	Directed	0.643	6.010	3.696	5.992
			Undirected	1.220	2.691	4.524	6.047
FSD: W/4	0.0493	0.000135	Directed	0.645	5.714	3.420	5.703
			Undirected	1.258	3.366	4.240	5.724

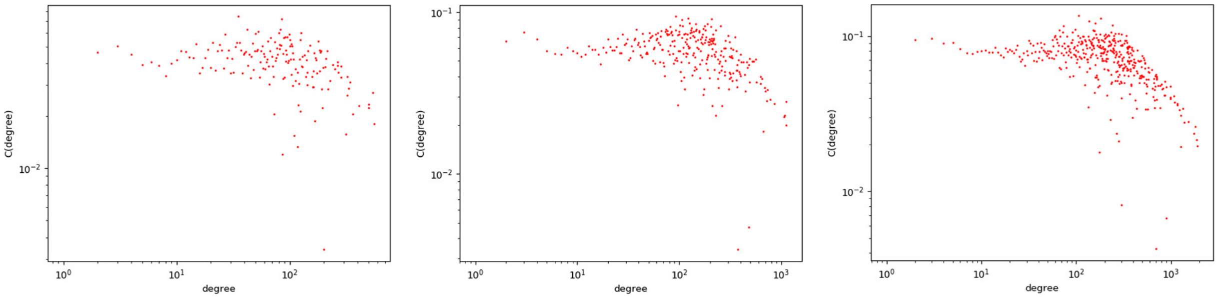


Fig. 4.1. Clustering coefficient of W/10, W/4, W/2 respectively for unweighted microblog WCN.

The Table 4 shows average path length for microblog WCN (L), for ER random network (L_r), for largest cluster of microblog WCN ($L(max)$) and ER random network ($L_r(max)$) for directed and undirected networks of FSD: W/1000, W/100, W/20, W/10 and W/4. It has been observed that for $L(max)$ *and* $L_r(max)$, the values are equivalent for undirected graph but directed graph shows slight decrease in average path length. This indicates that the network have small-world feature. As the number of nodes gets increased, the number of edges are increased with higher rate, thus, the average path length is decreased due to large number of links and more connected graph. The twitter feeds are tokenised to set of words which is considered as a path. However, in random graph, it may or may not contain all edges with respect to specific path. Thus, identifying patterns for biggest connected component is more significant because for all types of network in (L_r), the average path lengths of WCN and ER random graph are equivalent. This shows that as $(L \approx L_r)$ *and* $(CC \gg CC_r)$ microblog WCN follows small world property.

3.3. Hierarchical organization

Hierarchical network model [52] represents the connectivity of nodes of real world networks with each other. Hierarchical organization is the parameter to identify if the network is scale free in terms of hierarchy using clustering coefficient. Hierarchical organization is defined as the probability distribution of average clustering coefficient of all nodes with degree k. The clustering coefficient is defined as shown in Eq. (3)

$$C(k)_i = \frac{2 * E_i}{k_i(k_i - 1)} \tag{3}$$

where E_i denotes the number of edges among the vertices in the nearest neighbourhood of vertex i and k_i denotes the degree of vertex i. It is defined as $C(k) \propto k^{-\beta}$ where asymptotic scaling is required for network to be hierarchically organized with linear dependency using value of β. As shown in Figs. 4.1 and 4.2, it has been observed that the graphical representation for CC does not vary with increase in size of the network for both unweighted and weighted network of nearest neighbour edging microblog WCN, respectively. Also, as observed in Table 2, the clustering coefficient remains constant in magnitude with increase in size of the network. Thus, the network shows hierarchical organization [53]. Thus, for nearest neighbour edging network, the microblog WCN shows anomalous behaviour.

The observations for all pair neighbour edging network have been shown in Figs. 5.1 and 5.2 for dataset W/20, W/10 and W/4 for unweighted and weighted networks, respectively. It has been observed that the microblog network follows scaling properties. It can be observed that as the size of the dataset increases, the curve tends to show varying value of slope. However, the plot follows power-law and the scaling is not linearly dependent. Thus, the microblog WCN is partially hierarchically organized for all-pair neighbour edging network. Also, as the size of the network increases, the bend in the curve plot is increased. The weighted graph shows increase in variations of the plot than that of unweighted graph.

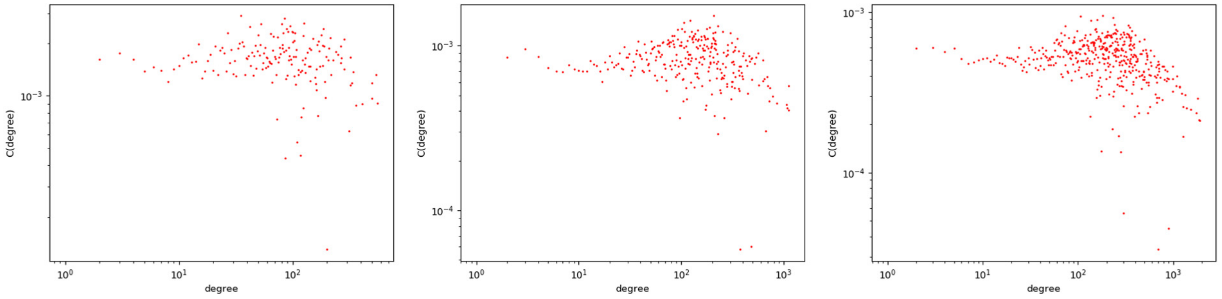


Fig. 4.2. Clustering coefficient of W/10, W/4, W/2 respectively for weighted microblog WCN.

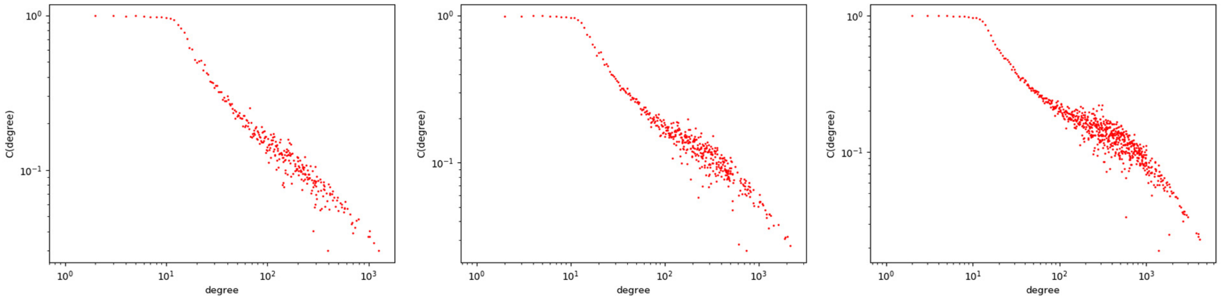


Fig. 5.1. Clustering coefficient of W/20, W/10, W/4 for all pair neighbour unweighted microblog WCN.

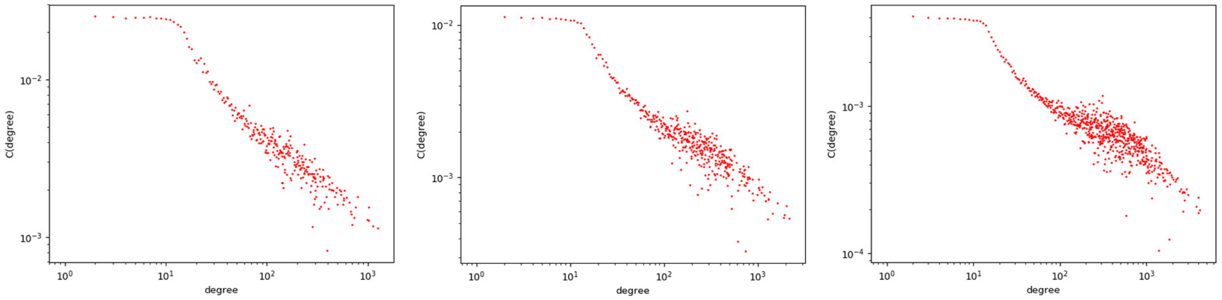


Fig. 5.2. Clustering coefficient of W/20, W/10, W/4 for all pair neighbour weighted microblog WCN.

Thus, the microblog WCN shows anomalous behaviour for hierarchical organization unlike traditional WCN. This may vary from general set of twitter feeds to domain specific set of twitter feeds. This is because the domain specific data shows repetitive information of keywords. However, as discussed in Section 3.1 the node degree may not play pivotal role in microblog WCN whereas edge weight does.

3.4. Assortativity network analysis

As observed in Table 5, the average of average degree of neighbours of each node is represented as $\langle \boldsymbol{k} \rangle$ is termed as average degree correlation. The degree assortativity $p(k)$ is defined as the probability that a randomly chosen node with degree k is connected to the nodes with higher degree comparative to degree k [24,54]. The significance of the assortativity is that it defines the pattern of the network and is defined as shown in Eq. (4).

$$\tau = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2} \tag{4}$$

where j_i and k_i are the degrees of the two endpoints of the ith edge, and M is the total number of edges in the network. If $\tau > 0$, the network is claimed to be assortative mixing; while if $\tau < 0$, the network is called disassortative mixing. The network is assortative if high degree nodes tend to connect with high degree nodes and low degree nodes tend to connect with low degree nodes. The network is disassortative if high degree nodes avoid connecting with high degree nodes and

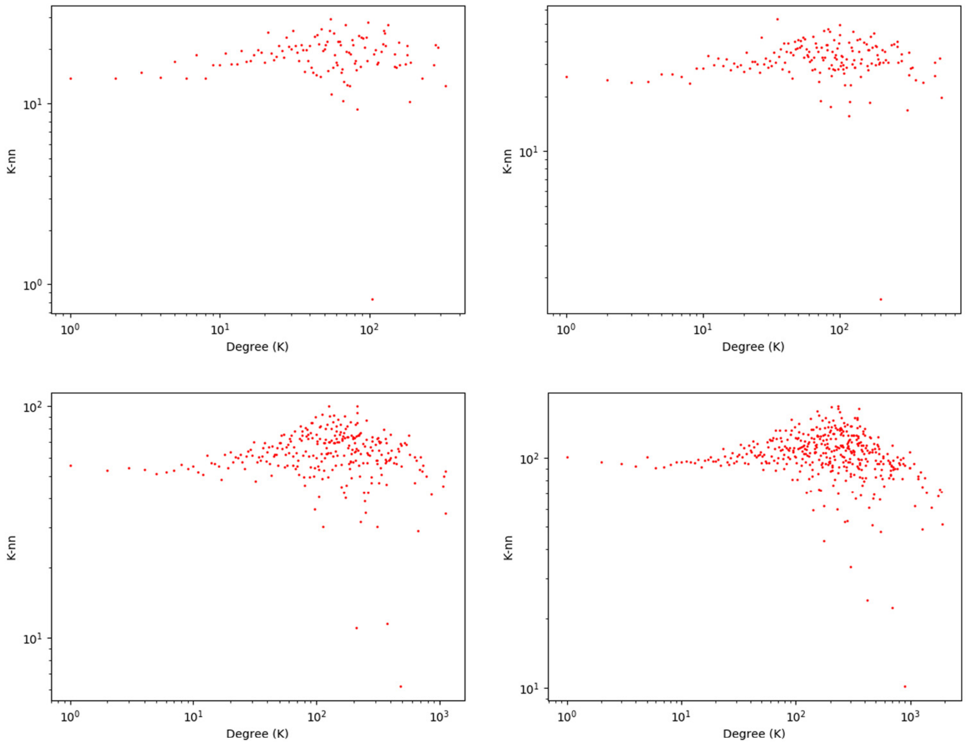


Fig. 6.1. Degree correlation of W/20, W/10, W/4 and W/2 respectively for nearest neighbour unweighted microblog WCN.

Table 5
Average shortest path length of WCN.

Dataset	$\tau(\boldsymbol{UW})$	$\tau(\boldsymbol{WT})$	$\mu(\boldsymbol{UW})$	$\mu(\boldsymbol{WT})$
FSD: W/1000	−0.385	−0.421	4.188	4.188
FSD: W/100	−0.268	−0.267	3.946	3.946
FSD: W/20	−0.319	−0.288	2.226	2.409
FSD: W/10	−0.323	−0.278	1.846	1.756
FSD: W/4	−0.329	−0.265	1.539	1.481
FSD: W/2	−0.334	−0.261	1.482	1.386
FSD: W	−0.334	−0.256	1.392, 3.724	1.354, 2.615

low degree nodes avoid connecting with low degree nodes. The assortativity is analysed using two types of analysis namely k-nearest neighbour and to compute the assortative mixing index [55] using Pearson correlation coefficient. The k-nearest neighbour based analysis is defined as $k_{nn(k)} \propto k^{-\mu}$, where k_{nn} is k-nearest neighbour of the network and if $\mu > 0$, the network is disassortative as stated by Liang et al. [24].

The graphical log–log plot as shown in Figs. 6.1 and 6.2 for W/20, W/10, W/4 and W/2 dataset of unweighted and weighted nearest neighbour network, respectively. It has been observed that the pictorial representation shows two-regime law for unweighted nearest neighbour microblog WCN. As the size of the network is increased, the network is assortative for degree $k < 10^2$ and disassortative beyond degree $k > 10^2$. The value of $\boldsymbol{\tau}(\boldsymbol{UW})$ in Table 5 indicates the Pearson correlation coefficient which constantly shows negative values for $\boldsymbol{\tau}$ indicating that the network is disassortative. However, the magnitude of $\boldsymbol{\tau}$ does not provide any significant information. Also, the positive value of $\boldsymbol{\mu}$ $(\boldsymbol{UW})$ ***and*** $\boldsymbol{\mu}(\boldsymbol{WT})$indicates that the network is disassortative.

Although the weighted nearest neighbour edging microblog WCN shows the same behaviour as that of unweighted network, the plot is sparser than that of unweighted nearest neighbour network. This is because nodes having same degree may not have same weighted degree. The high degree nodes tend to connect with nodes of lower degree. This is because the high degree nodes are the commonly used terms as mentioned earlier. The commonly used terms can be clubbed with multiple domains, multiple keywords, hashtags, entities and mentions. It has been observed that all pair network show similar behaviour and follows disassortativity. However, the set of domain specific tweets may contain keywords fully connected to each other which can be used for trending topic, tweet summarization, keyphrase extraction, topic detection, and event detection. The fully connected subgraph is assortative and equally connected subgraph tends to be non-disassortative. In this research work, this property has been used for as terminal state for BArank.

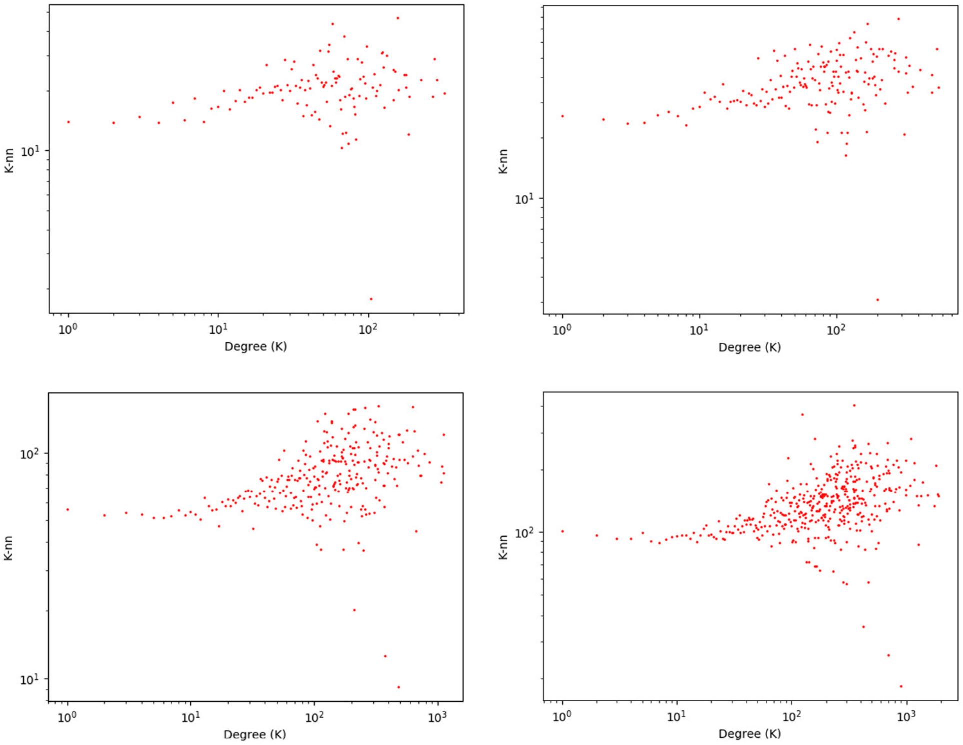


Fig. 6.2. Degree correlation of W/20, W/10, W/4 and W/2 respectively for nearest neighbour weighted microblog WCN.

Table 6
The eigenvalues of microblog WCN.

Dataset	λ_1	λ_2	λ_3	λ_4	λ_N	N_λ
FSD: W/1000	4.169	4.012	3.667	3.530	−4.134	513
FSD: W/100	14.510	11.478	10.833	9.466	−11.434	3 472
FSD: W/20	51.974	29.209	25.222	23.561	−28.914	11 056
FSD: W/10	98.194	51.745	45.032	40.539	−48.809	18 035

3.5. *Spectral analysis*

The spectral analysis [25] of microblog WCN can be made by observed spectral distribution of eigenvalues as shown in Fig. 7. The eigenvalues have been calculated from matrix of the microblog WCN. The eigenvalues λ can be observed from Table 6. The highest eigenvalue λ_1, second λ_2, third λ_3, fourth λ_4, the lowest eigenvalue λ_N and number of eigenvalues N_λ have been recorded in Table 6 for nearest neighbour edging network. The largest eigenvalue is significantly high than the second largest eigenvalue. As the size of network is increased, the eigenvalue and the number of eigenvalues are increased. It has been stated that the principal eigenvector of a graph is used to compute the centrality of the nodes which is also knows as PageRank in context of world-wide web [56]. Similarly, the second eigenvector component is used for graph clustering. The eigenvalue computations are expensive for large amount of data. Thus, the computations have been observed for dataset with W/1000, W/100, W/20 and W/10. It has been observed that W/20 and W/10 datasets can be used for online learning of latent patterns in data. The spectrum is the set of all the eigenvalues of its adjacency matrix.

The graphical plot of eigenvalues as shown in Fig. 7.1, shows that the slope of the plot of eigenvalues is decreased with increase in size. This is because the number of eigenvalues are increased. The plot shows that there is significance difference between highest and second highest eigenvalues as the size of the graph is increased. The plot shows that the variations at two ends of the curve varies much more than those in between. The plot shows similar behaviour as that of existing WCN for English and Chinese languages [25].

The spectral distribution is mapped using adjacency matrix of microblog word co-occurrence model. The spectral distribution for W dataset for nearest neighbour edging can be observed in Fig. 7.2. As observed the spectral distribution is concentrated near values −1 and 1. It contains multiple values for eigenvalues near 1 and −1. The spectral density for all other values is uniform and is given as 1. The spectral distribution shows the triangle formulation of the plot for spectral distribution during acute observations near eigenvalue −1 and 1. More study and analysis of the shape can be used for different applications.

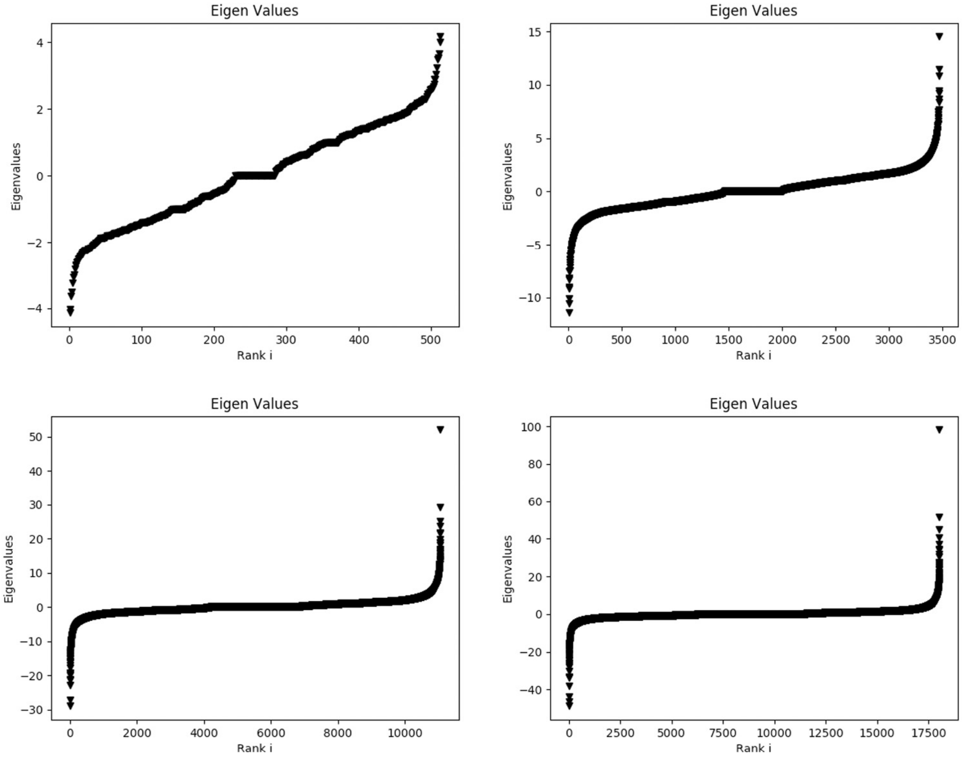


Fig. 7.1. Eigenvalues of W/1000, W/100, W/20, and W/10, respectively for nearest neighbour microblog WCN.

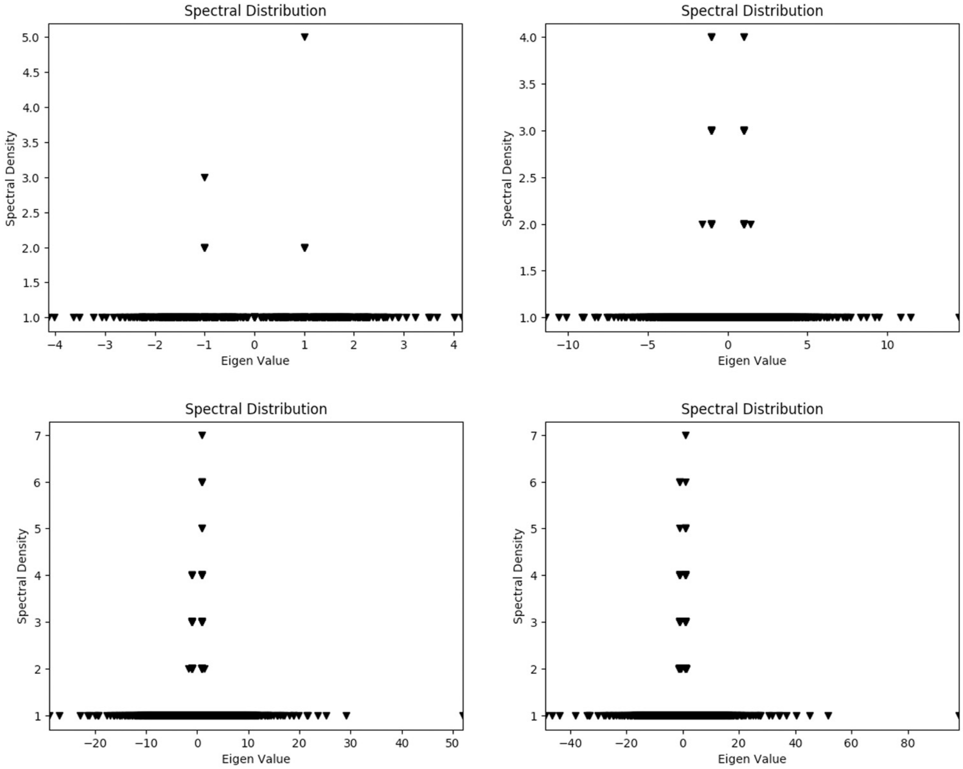


Fig. 7.2. Spectral distribution of W/1000, W/100, W/20, and W/10, respectively for nearest neighbour microblog WCN.

Table 7
Summarization of structural analysis of WCN for microblogs.

SN	Key parameter	Results	Inferences
1	Scale-free property	$2 < \gamma < 3\, for\ k > 10^2$	The value of γ shows that the network is scale free for degree distribution, strength distribution and edge weight distribution
2	Small world feature	$(L\,(max)) \approx (L_r\,(max))$ $and\ \ (C \gg C_r)$	The microblog WCN follows small world property.
3	Hierarchical distribution	Constant magnitude for CC(k) with respect to degree k	The nearest neighbour edging network shows anomalous behaviour and is thus, hierarchical organization not obvious
		CC remains constant with increase in number of nodes	The all-pair neighbour edging network shows partial hierarchical organization.
4	Assortativity	$\mu > 1\, for\ all\ networks$ $\tau < 0\, for\ all$	The network is disassortative.
5	Spectral distribution	$\lambda_1 \gg \lambda_2\, for\ all\ networks$	The value of λ shows importance of graph
		The spectral density shows triangular shape of the distribution of eigenvalues	The value of highest eigenvalue is very high from second highest eigenvalue.

4. Keyphrase extraction: An application of the analytical study

The results as observed from the empirical study and analysis of microblog WCN have been summarized in Table 7. It has been observed that the different application domains for which the study of microblog WCN can be used are keyword extraction, keyphrase extraction, topic detection and tracking, and sentiment analysis. For keyphrase extraction, the community detection in WCN has been widely studied in recent years. The real-time streaming data analysis provides valuable and useful insights from uncertain user-generated data. The time series based information shows current trends. As shown in Table 7, the structural analysis of microblog WCN can be summarized as scale-free, small-world, and disassortative. The properties have been studied and applied over application of keyphrase extraction technique BArank.

Although there are many keyphrase extraction techniques which have been proposed in literature, as described in Section 2, the need for identifying keyphrase for domain specific set of microblogs is still the major area of research. This is due to the presence of user-generated, unstructured, and ill-formed text. Thus, the WCN of the well-formed languages shows different behaviour than that of the microblog WCN.

The BArank is the extension of our previous studies: identifying influential segment, for keyphrase extraction in which the k – bridge decomposition [4] is the technique proposed for extracting important information from WCN. It is used for reducing the number of edges by retaining edges with high edge weight only. In the previous technique, the subgraphs have been obtained using threshold parameters during k – bridge decomposition. The keyphrase extraction has been performed using position of the nodes in the twitter feeds. This gives different phrases from single subgraph which provide brief and detailed information as marked by ranking process. The Analytical Hierarchical Process (AHP) technique is used for ranking phrases obtained from the sub-graph. The ranking is done for intra-subgraph phrases. There is need to rank different phrases obtained from different subgraphs. Also, the parameter free keyphrase extraction as manual analysis of identifying the termination value of the parameter for domain specific microblog WCN is tedious task for large amount of data.

4.1. *BArank: Keyphrase extraction technique*

In this research paper, the non-parametric keyphrase extraction technique has been proposed for microblog WCN. The BArank keyphrase extraction technique used k-**B**ridge decomposition [4], **A**ssortativity and **rank**ing measure for identifying top-n keyphrases. The k – bridge decomposition algorithm has been proposed for extraction of phrases and phrases have been ranked [4]. Following contributions have been made in BArank.

4.1.1. *Terminal state*

Different parameters have been used for terminal state in existing technique such as fixed maximum number of nodes for every subgraph after decomposition. However, BArank is a parameter free approach. The microblog WCN is disassortative. As per observations from Section 3.4, during k-bridge decomposition, the subgraph may terminate if the resulting subgraph is not disassortative anymore, as explained in Section 4.5. Assortative networks indicates that higher degree nodes get connected to high degree nodes and low degree nodes get connected to low degree nodes. The edge decomposition of the network is performed till the network gets non-disassortative. This provides important words from discussion which are topic specific.

4.1.2. *Topological sorting*

In the previous, the node positions have been used as first node and last node by considering zero in-degree and zero out-degree, respectively after k – bridge decomposition. Thereafter the paths have been calculated from every first node to every last node. The BArank used topological sorting for extracting keyphrase from directed graph. This gives single keyphrase and

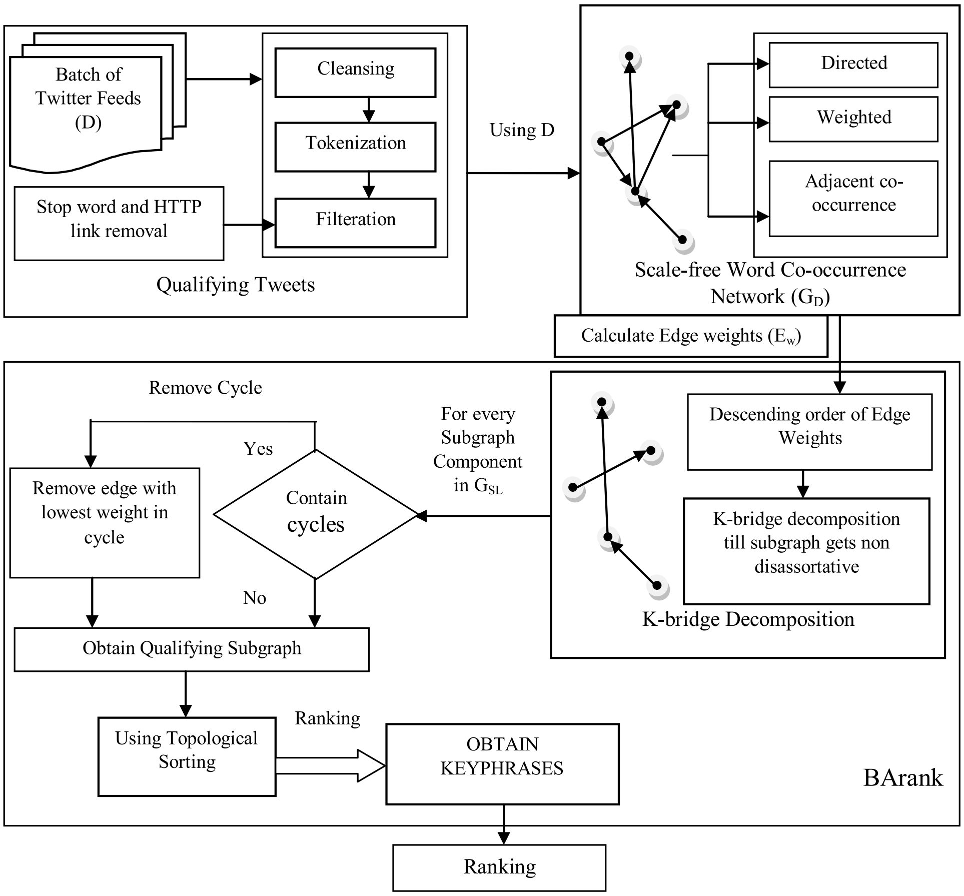

Batch of Twitter Feeds (D)
Stop word and HTTP link removal
Cleansing
Tokenization
Filteration
Qualifying Tweets
Using D
Directed
Weighted
Adjacent co-occurrence
Scale-free Word Co-occurrence Network (G_D)
Calculate Edge weights (E_w)
Remove Cycle
Yes
Remove edge with lowest weight in cycle
Contain cycles
No
For every Subgraph Component in G_SL
Descending order of Edge Weights
K-bridge decomposition till subgraph gets non disassortative
K-bridge Decomposition
Obtain Qualifying Subgraph
Using Topological Sorting
Ranking
OBTAIN KEYPHRASES
BArank
Ranking

Fig. 8. Framework of BArank for identifying keyphrases.

all the nodes are covered within single phrase for single subgraph. It has been observed that the directed microblog WCN preserve lexical sequence.

4.1.3. *Ranking*

In previous studies, the AHP for inter-ranking of influential segments has been proposed for single subgraph. Many different sub-graphs were obtained after k – bridge decomposition of domain specific microblog WCN. However, as only single keyphrase is obtained for every subgraph, the inter-keyphrase ranking is obtained using density of the subgraph. The density [57] is measured as the ratio of the sum of edge weights in the subgraph to the number of edges as shown in Eq. (2).

The work-flow for proposed methodology, BArank, have been evaluated using First Story Detection dataset which used 27 topics related data to identify keyphrases from the set of microblogs. The BArank have been proposed for set of microblogs as input and resulting keyphrases have been obtained from domain specific twitter feeds which have been shown in Fig. 8.

4.1.4. *Qualifying tweets*

Consider the tweet text collected using tweet id as one document $d\epsilon D$, where $D = \{d_1, d_2, \ldots, d_n\}$ is the collection of n documents or n tweets. Every document $d_i\epsilon D$ is stored in database. The set of tweets have been obtained to pre-process the tweet such that it qualify for further processing.

Step 1: The set of microblogs is fed as set of documents D as the corpus. Consider the tweet text collected using tweet id as one document $d\epsilon D$, where $D = \{d_1, d_2, \ldots, d_n\}$.

Step 2: Preprocessing: The twitter feeds undergo cleansing, tokenization, and filtering. The cleansing of twitter feed means white space character removal, striping, and trimming the data. Further tokenization is performed by acquiring words as tokens. The text of tweets is filtered by removing stop-words, and HTTP links. It has been observed that URL plays minimal significance while identifying keyphrase from set of twitter feeds. The pre-processing has been further described in Algorithm 1.

Algorithm 1: $Preprocess(d)$

```
Given: $d$ as a tweet
$//Cleansing$
$d = removeWhiteSpaceCharacters(d)$
$d = d.strip()$
$d = d.trim()$
$//Tokenization$
$d = d.split()$
$//Filtering$
$For\ eachWord\ in\ d.split()$:
    $if\ eachWord\ in\ nltk\_stopwords()$
        $remove\ eachWord$
    $else$
        $continue$
$Remove\ HTTP\ link\ URLs\ from\ d\ using\ regular\ expressions$
```

4.1.5. Scale-free word co-occurrence network

Consider the set of words $w\epsilon D$, where $w = \{w_1, w_2, \ldots, w_t\}$ is collection of words in a pre-processed document $d\epsilon D$ for N documents. Every word w_i is considered as a node n_i. Every word w_i in document d_k is connected to adjacent word w_j in document d_k. This generates directed graph G_k for words occurring within same document d_k and every node w_i points to next adjacent node w_j. This directed graph of all pre-processed documents $D = \{d_1, d_2, \ldots, d_n\}$ is called Graph G. Every edge e is given a weight nw which represents number of times the edge is connecting n_i *and* n_{i+1} in documents D. The same word n_i *and* n_j are merged to overlap and create one node. This represent word co-occurrence network (WCN).

Algorithm 2: $CreateWeightedDiGraph(D)$

```
Given:  $D$
$G = nx.DiGraph()$
$Extract\ edges\ E_i from\ d_i joining\ adjacent\ words$
$For\ (u, v)\ in\ E$:
    $G.add\ edge(u, v, weight = nw)$ //u and v are nodes which are connected by an edge
```

The weight for each edge is added as one and each time the edge gets repeated, the edge weight is increased by 1 as shown in Algorithm 2. Further, it has been observed from Section 3.1 that the WCN for microblogs is scale – free and thus, as per observations and analysis over the scale free property, k – bridge decomposition have been used for identifying keyphrases.

4.1.6. BArank

Different types of microblog WCN can be generated as shown in Fig. 1. The BArank has been proposed for scale-free WCN indicating that the decomposition of disassortative network with respect to edges gives significant information. This information is obtained by topological sorting of the resulting sub-graph so obtained. The pseudo-code for BArank has been shown in Algorithm 3.

Algorithm 3: $BArank$

```
Given: Set of Tweets $D$
For each document $d$ in $D$:
    $T_d = Preprocess(d)$
    For each word $w_i$ in $T_d$:
        $w_i = Preprocess(w_i)$
$G = CreateWeightedDiGraph(D)$
$If\ G\ is\ Disassortative()$:
    For $eachSubgraph\ in\ G.subgraph$:
        $L_e = Ascending(G.subgraph.edges())$
        $i = 0$
        While $eachSubgraph! = Assortative()$:
            Remove $L_e[i]$
            $i = i + 1$
For $eachSubgraph\ in\ G.subgraph$:
    $keyphrase = topologicalSorting(eachSubgraph)$
$RankingKeyphrases(G)$
```

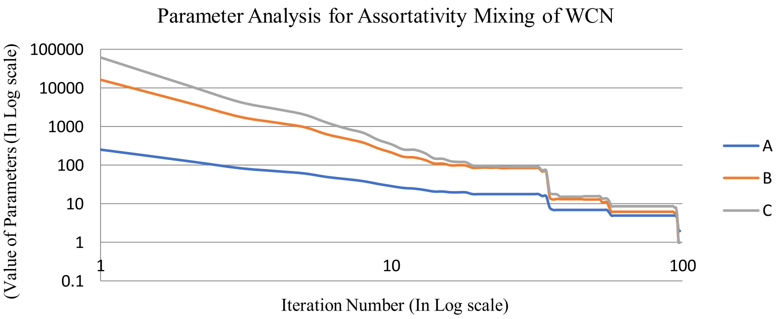


Fig. 9. Parameter analysis of assortativity mixing of WCN for microblogs.

The k – bridge decomposition is the edge based network reduction which provides a probe to study the hierarchical properties of large scale networks, focussing on the network's regions of increasing centrality and connectedness properties [58]. As per the hierarchical properties of microblog WCN, it has been observed that for high node degree, the value of clustering coefficient is low. Although the network is disassortative from Eq. (1), the complex network can be reduced by removing low edge weights as shown in Algorithm 2.

Step 1: The assortativity is defined as shown in Eq. (5) and is denoted by τ.

$$\tau = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2} \tag{5}$$

where j_i and k_i are the degrees of the two endpoints of the ith edge, and M is the total number of edges in the network. If$\tau > 0$, the network is claimed to be assortative mixing; while if$\tau < 0$, the network is called disassortative mixing.

Step 2: This assortative mixing coefficient is then equivalent to Eq. (6)

$$\tau = \frac{\sum_i j_i k_i - M^{-1}\left[\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2}{\sum_i \frac{1}{2}(j_i^2 + k_i^2) - M^{-1}\left[\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2} \tag{6}$$

Step 3: For negative value of τ, disassortative mixing is obtained as shown in Eq. (7)

$$\frac{\sum_i j_i k_i - M^{-1}\left[\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2}{\sum_i \frac{1}{2}(j_i^2 + k_i^2) - M^{-1}\left[\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2} < 0 \tag{7}$$

Step 4: This implies that for disassortative mixing, the resulting coefficients should follow Eq. (8).

$$either\ (a < b < c)\ or\ (c < b < a), \tag{8}$$

where $a = \sum_i j_i k_i$, $b = M^{-1}\left[\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2$ *and* $c = \sum_i \frac{1}{2}(j_i^2 + k_i^2)$. This implies that b should lie in between a *and* c.

Step 5: On the contrary, as per analysis, the resulting coefficients for assortative mixing should follow Eq. (9)

$$either\ (a > b\ and\ c > b)\ or\ (b < a\ and\ b < c)\ , \tag{9}$$

where $a = \sum_i j_i k_i$, $b = M^{-1}\left[\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2$ *and* $c = \sum_i \frac{1}{2}(j_i^2 + k_i^2)$. This implies that either b is smaller than both a *and* c, or b is greater than both a *and* c. Thus, the breaking point for disassortative network is when $a = b$ *or* $c = b$. The observations have been shown in Fig. 9. The assortativity mixing for different parameters a, b, and c, have been observed for iterative process of k-bridge decomposition for WCN of domain specific microblogs. The network is decomposed as low weighted edges are removed and high weighted edges are kept. It has been observed that initially the parameter b lies in between a *and* c and shows disassortativity as mentioned in Eq. (8), till the value gets equal to 4 where the condition for disassortativity is violated. The results are obtained for the graph having number of nodes 4 and topological sorting gives the keyphrase "*rip amy dead winehouse*" which gives informative results. Similarly, this process is carried for 27 different topic related tweets as observed from FSD dataset.

The network is assortative if high degree nodes tend to connect with high degree nodes and low degree nodes tend to connect with low degree nodes. The network is disassortative if high degree nodes avoid connecting with high degree nodes and low degree nodes avoid connecting with low degree nodes. However, as observed in Section 3.4, the WCN for microblogs is observed as disassortative network. The lowest weighted edges are removed until subgraph becomes non-disassortative. The cycles and self-loops are removed from the resulting subgraph. The keyphrases are obtained using topological sorting.

4.1.7. Ranking keyphrases

Ranking keyphrases is the final decision to identify those keyphrases which are important. The importance of keyphrase with respect to single event depends upon the occurrence of nodes and the extent of closely knitted nodes in the network. Thus, the keyphrases obtained using topological sorting are ranked using density (Den) of the subgraph defined as shown in Eq. (10).

$$Den\ (S) = \frac{\sum_{(i,j)\epsilon M} w_{i,j}}{M} \tag{10}$$

where M is the number of edges and $w_{i,j}$ indicates the weight of the edge in the subgraph (S). The resulting ranked keyphrases represent the important information about data. The keyphrases are ranked on the basis of density as shown in Algorithm 4.

```
Algorithm 4: $RankingKeyphrases(G)$

$Den = dictionary()$
For $eachSubgraph\ in\ G.Subgraph()$:
    $EdgeWeightSum = 0$
    For $eachEdge\ in\ eachSubgraph.edges()$:
        $w_{i,j} = frequency(node_i, node_j)$
        $EdgeWeightSum = EdgeWeightSum + w_{i,j}$
    $Den(eachSubgraph) = \frac{EdgeWeightSum}{Count(eachSubgraph.edges())}$
$L_k = list(desending(Den.values()))$
$Return\ L_k$
```

The keyphrases of the domain specific set of microblogs are ranked and thus, important information about the topic is obtained on the basis of ranked keyphrases.

4.2. Experimental setup

The key-parameters scale-free property for edge weight distribution and assortativity have been used for identifying keyphrase from domain specific set of tweets. The twitter dataset used for the experimental analysis is FSD dataset [2]. In this dataset, 51,879,318 tweet ids have been mentioned which are related to different topics. The number of tweets marked as relevant tweets are 3034 for 27 different topics. For each tweet, corresponding topic id has been marked. Tweets have been collected with given tweet id using Tweepy module for Python. The results are evaluated using the gold standard (ground truth) information in relevance_judgement.txt file available in the folder. The best keyphrase extracted is evaluated based on the words as mentioned in ground truth data.

4.3. Performance measures

The performance evaluation metrics used for experiments is ROUGE score (Recall-Oriented Understudy for Gisting Evaluation). The ROUGE-N score indicates the common n-grams over total number of n-grams as obtained from sequence of the topic. This measure is used for measuring recall for the keyphrase extraction and text summarization [36] problem. The ROUGE-N score has been defined as shown in Eq. (11).

$$ROUGE - N = \frac{(n - grams\ in\ extracted\ keyphrase) \cap (n - grams\ in\ reference\ text)}{Total\ number\ of\ n - grams\ in\ reference\ text} \tag{11}$$

Thus, ROUGE-1 and ROUGE-2 are used to calculate the performance for unigram and bi-grams, respectively of the keyphrase so obtained using BArank. The ROUGE-L is the measure of Longest Common Subsequence (LCS) which measures the LCS between the automatically extracted keyphrases and reference text. The highest value of common sub-sequence is returned by ROUGE-L. The ROUGE-L indicates the maximum words occurring together with reference to the words co-occurrence in reference summary after removing stop words. The ROUGE-L score precisely returns maximum common n-grams which co-occur in both extracted and reference text. It has been observed that the BArank outperforms all the existing techniques using performance measures recall, precision and F-measure. Recall is the ratio of the number of relevant records retrieved to the total number of correct relevant records in the database as mentioned in Eq. (12).

$$recall = \frac{Number\ of\ words\ (Resulting\ phrases \cap Ground\ truth\ topic)}{Number\ of\ correct\ words\ in\ ground\ truth} \tag{12}$$

Recall represents the percentage of correct phrases recall with respect to the topic and is thus valuable information for measuring and comparing performance of all the techniques. The precision shows the ratio of relevant records to total number of words obtained. Precision is defined as shown in Eq. (13).

$$precision = \frac{Number\ of\ words\ (Resulting\ phrases \cap Ground\ truth\ topic)}{Total\ Number\ of\ words\ obtained} \tag{13}$$

F measure is defined as the harmonic mean of precision and recall and is represented as shown in Eq. (14).

$$precision = \frac{2 * precision * recall}{precision + recall} \tag{14}$$

The ROUGE-1 score, ROUGE-2 score, ROUGE-L score, precision, recall, and F-measure performance measures are used to evaluate the results of the BArank keyphrase extraction algorithm, and compare it with the existing keyphrase extraction technique for microblog WCN.

4.4. Baseline measures

The baseline techniques considered for comparison are TextRank, NErank, and TopicRank. ***TextRank***: The TextRank is the keyphrase extraction technique [59] which has been proposed using random walk measure as introduced by [60] in WCN. The PageRank score of any vertex V_i is defined as given in Eq. (15)

$$S\left(V_i\right) = (1-d) + d * \sum_{j \epsilon In(V_i)} \frac{1}{Out\left(V_j\right)} S\left(V_j\right) \tag{15}$$

where $G = (V, E)$ be a directed graph with the set of vertices V, and set of edges E where E is the subset of $V * V$, $In\left(V_i\right)$ is the number of nodes pointing to V_i and $Out(V_j)$ is the number of links pointing out of node V_j. Here d is the damping factor which is usually set as 0.85 [60]. The vertex score is calculated after convergence of random walk over WCN. The convergence is achieved when the error rate for any vertex in the graph falls below a given threshold which is given as 0.0001.

Initially, the words have been defined as vertex and relation between them is defined as an edge between two vertex. The graph based ranking algorithm is iterated over WCN until convergence. The vertex are sorted based on their final score. The author performed two tasks namely keyword extraction task, consisting of the selection of keyphrases representative for a given text, and sentence extraction task which identifies most important sentences in the text. However, in this research paper, the former problem has been studied for comparison of the proposed keyphrase extraction technique. The selection of keywords in TextRank is dependent upon presence of word and random walk on it, in the network. However, there is the need to evaluate the relation among words in the network. Top k vertices are considered as keywords. The parameter N chosen for declaration of the window of the document is considered as 2 to keep the size of WCN minimal.

TopicRank: Further, the author proposed TopicRank [61]. The $G = (V, E)$ complete and connected graph is constructed with the set of vertices V, and set of edges E where E is the subset of $V * V$. The WCN is the graph G which is used for random walk. The TopicRank is defined as shown in Eq. (16)

$$S\left(t_i\right) = (1-d) + d * \sum_{t_j \epsilon V_i} \frac{w_{j,i} * S\left(t_j\right)}{\sum_{t_k \epsilon V_j} w_{j,k}} \tag{16}$$

where V_i are the topics voting for node t_i and d is the damping factor generally factor defined to 0.85 (Brin et al., 1998). The $w_{i,j}$ weight of the edge is defined as shown in Eq. (9) and calculated using the reciprocal distances $dist(c_i, c_j)$ between the offset positions of the candidate keyphrases c_i *and* c_j in the document where pos (c_i) represents all the offset positions of the candidates keyphrase c_i. The weights and reciprocal distances are shown in Eqs. (17) and (18), respectively.

$$w_{i,j} = \sum_{c_i \epsilon t_i} \sum_{c_j \epsilon t_j} dist\left(c_i, c_j\right) \tag{17}$$

$$dist\left(c_i, c_j\right) = \sum_{p_i \epsilon pos(c_i)} \sum_{p_j \epsilon pos(c_j)} \frac{1}{|p_i - p_j|} \tag{18}$$

This needs the manual definition of the parameters such as window of words which is defined as 2 in this study. This is done to keep evaluate the existing techniques over similar parameters for comparison.

NErank: It has been observed that the Twitter have uncertain and user-generated data and thus, a novel graph based keyphrase extraction techniques have been proposed for Twitter data using edge weights and node weights [39]. The formula for calculating NErank has been given as shown in Eq. (19).

$$S\left(V_i\right) = (1-d) * W\left(V_i\right) + d * W\left(V_i\right) * \sum_{j:V_j \rightarrow V_i} \frac{w_{ji}}{\sum_{k:V_j} w_{jk}} S\left(V_j\right) \tag{19}$$

where $W\left(V_i\right)$ is the weight of the current node and d is damping factor. The weight of the node has been calculated as TFIDF measure as shown in Eq. (20).

$$W\left(V_i\right)_{TFIDF} = tf\left(V_i\right) * \log_2 \frac{N}{df\left(V_i\right)} \tag{20}$$

Table 8
Performance analysis of TextRank, TopicRank, NErank, and BArank for microblog WCN.

Algorithm	ROUGE-1	ROUGE-L	ROUGE-2	Recall	Precision	F-measure
TextRank	19.13	0.67	5.55	19.13	8.61	11.87
NErank	13.73	0.77	1.38	13.73	25.76	17.91
TopicRank	26.54	0.88	11.11	26.54	18.11	21.53
BArank	**37.80**	**1.62**	**12.66**	**37.80**	**39.95**	38.84

The NErank has been proposed for Twitter data and outperformed the existing techniques namely PageRank and TextRank in terms of *Precision* and *Binary preference measure (Bpref)*. The precision is defined as the ratio of the number of correct keywords and number of extracted keywords. The*Bpref* measure has been defined as shown in Eq. (21).

$$Bpref = \frac{1}{R}\sum_{r\epsilon R} 1 - \frac{|n \text{ ranked higher than } r|}{M} \tag{21}$$

where R is the correct keywords within M extracted keywords in a method, where r is a correct keyword and n is incorrect keyword. It has been observed that using hashtag boosting, the NErank gives better precision. However, the proposed technique BArank has not been compared with hashtag based boosting of NErank to keep the parameters uniform. Also, the precision of the BArank is better than all the existing techniques as discussed in Section 4.5.

4.5. Experimental results and evaluation

The BArank used modification and improvement over our previous studies [4]. New insights have been obtained for microblog WCN from Section 3. The threshold parameter has been set for termination state in k-bridge decomposition [4]. However, the BArank is improved over previous studies using parameter free keyphrase extraction approach. Similarly, AHP for ranking phrases obtained from single subgraph [4]. The ranking of different keyphrases has been obtained using density based score. The graph based keyphrase extraction techniques have been used as baseline measures for comparison namely TextRank, TopicRank, and NErank.

The microblog WCN is stable as observed in Section 3, and shows that it follows the small-world property. The evaluations have been performed for different dataset. As observed in Section 3.1, it has been observed that the graph follows scale-free property for degree distribution and edge distribution. The degree distribution gives common terms for high degree nodes and edge distribution provides bi-gram which indicates high co-occurring frequency. As observed in Section 3.1, the edge distribution provides high co-occurrence among words and shows more important words than high degree words. Thus, the edge based analysis has proved to be better and effective for keyphrase extraction. Also, NErank used node scores and edge scores for random walk based keyphrase extraction algorithm and provides useful results. The microblog WCN is disassortative as observed in Section 3.4 for large amount of data. The value of assortative mixing coefficient as described in Section 3.4 is negative. The value of absolute disassortative mixing coefficient is decreased as the edges with lowest weights are removed from the sub-graph. It has been observed that after few iterations, the value of assortative mixing coefficient is observed as zero. This indicates the minimal disassortativity of the sub-graph and is marked as breaking point. The nodes which have been obtained from sub-graph after breaking point are significant words.

The TextRank, TopicRank, NErank and BArank keyphrase extraction algorithms have been applied over FSD dataset for 27 different topics. The implementations have been made using Python 2.7.11 and 'pke' library [62]. The observations have been obtained based on performance measures, ROUGE score, which have been recorded in Table 8. The BArank clearly outperforms all the existing techniques. Although TopicRank outperformed the existing TextRank techniques, yet it has been used for identifying keyphrases for microblog WCN. Earlier keyphrase extraction have been performed over twitter data using NErank algorithm. The NErank is based on random walk and BArank is based on sampling of microblog WCN using k-bridge decomposition and assortativity.

As observed from Table 8, the precision, recall and F-measure is obtained using automatic evaluation of 1-gram based tokenization. The BArank outperforms the existing techniques with 38.84 F-measure and showing an edge for recall and precision measures as 37.80 and 39.95, respectively. Thereafter, TopicRank gives second best performance for recall whereas NErank gives second best performance for precision. For ROUGE scores, the TopicRank gives better results than that of NErank in terms of recall but have low precision score. The NErank used degree score and edge score and thus, gives better performance over all existing techniques in terms of precision for microblog data. As the microblog WCN follows power-law for edge distribution, the consideration of edge scores has proved to be useful. Thus, BArank outperforms the existing techniques by identifying important and relevant words which indicates the topic. It has been observed that BArank gives better results for large amount of data. There are some topics which contain small number of tweets and hence, does not give reliable results. The value for precision and recall increases by 18%–30% if large amount of data based results are considered for evaluation of experimental results.

As observed from Table 9, the keyphrases obtained from TextRank are less likely to be significant words and more likely to be repetitive and popular terms. NErank gives good performance over few topics by using nodes score and edge weight score. TopicRank gives improved results over many topics. However, the results for TopicRank do not include much of the

Table 9
Keyphrase obtained using TextRank, TopicRank, NErank, and BArank for microblog WCN.

Technique\Topic	Plane carrying Russian hockey team Lokomotiv crashes, 44 dead	Terrorist attack in Delhi	Betty Ford dies
TextRank	Breaking news	Ox blood ruffin	Isolabella former
TopicRank	Plane crash	High court Delhi	Betty Ford
NErank	Russia Lokomotive hockey	Reuters India	Isolabella former
BArank	Plane crash Hockey team	Blast outside Delhi high court	RIP lady Betty Ford

words which signify the event and thus, shows low precision. The BArank gives considerable improvement for precision among all the existing techniques. This is because it gives the nodes which are related to high degree nodes and thus, gives community of words having assortative nature.

It has been observed that the complexity of convergence of the TextRank, TopicRank and NErank is $O(n+m)$ where n is the number of nodes, and m is the number of edges. Also, additional complexity includes the calculation of node and edge scores, if any, in the technique. However, in BArank, edge removal follows the convergence from disassortative to non-disassortative network with complexity of $O(m)$. The baseline measures have been used for wide range of application for summarization of retrospective data, however, the BArank has been proposed for handling streaming data by setting values for controlling parameters during trend and event detection. In future, this analysis can be explored for identifying keyphrases for domain specific set of streaming twitter feeds. It has also been observed that the BArank gives information about multiple topics which are being discussed in the set of microblogs. The summaries can be obtained using community detection for temporal analysis.

5. Conclusion and future work

The WCN for microblog datasets has been studied for scale-free property, small world feature, hierarchical organization, assortativity, and spectral distribution. It has been observed that although the existing WCN evolved from well-formed and structured large documents follows two-regime power law, the microblog WCN is created from unstructured and short-text and follows power-law. Thus, the microblog WCN is scale-free. This network contains the small world feature with relatively reduced path length and very high clustering coefficient than that of random networks. The microblog WCN is disassortative in nature which indicates that high degree nodes avoid connecting to each other. Hierarchical organization gives more acute observations for large dataset than considering small dataset. Although the network is inclined towards hierarchical organization due to scale-free nature of the network, the network is partially hierarchically organized. The robustness of key parameters have been studied for identifying key-phrases from the network and BArank has been proposed using analytical observations. The BArank shows an edge over existing techniques with 38.84 F-measure. Also, the BArank gives ROUGE-1, ROUGE-2, and ROUGE-L scores as 37.80, 12.66, and 1.62, respectively. In future, the dynamics of WCN could be studied for processing of the twitter feeds for various real-time applications of topic detection and tracking.

References

[1] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 591–600.

[2] S. Petrović, M. Osborne, V. Lavrenko, Streaming first story detection with application to twitter, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 181–189.

[3] H. Sayyadi, M. Hurst, A. Maykov, Event detection and tracking in social streams, in: Icwsm, 2009.

[4] M. Garg, M. Kumar, Identifying influential segments from word co-occurrence networks using ahp, Cogn. Syst. Res. 47 (2018) 28–41.

[5] H. Wang, C. Zhai, Generative models for sentiment analysis and opinion mining, in: A Practical Guide To Sentiment Analysis, Springer International Publishing, 2017, pp. 107–134.

[6] S.A. Moosavi, M. Jalali, N. Misaghian, S. Shamshirband, M.H. Anisi, Community detection in social networks using user frequent pattern mining, Knowl. Inf. Syst. 51 (1) (2017) 159–186.

[7] G.B. Chen, H.Y. Kao, Word co-occurrence augmented topic model in short text, Intell. Data Anal. 21 (S1) (2017) S55–S70.

[8] H. Liu, J. Cong, Language clustering with word co-occurrence networks based on parallel texts, Chin. Sci. Bull. 58 (10) (2013) 1139–1144.

[9] W. Liang, Y. Shi, K.T. Chi, J. Liu, Y. Wang, X. Cui, Comparison of co-occurrence networks of the chinese and english languages, Physica A 388 (23) (2009) 4901–4909.

[10] S. Maity, A. Chaudhary, S. Kumar, A. Mukherjee, C. Sarda, A. Patil, A. Mondal, WASSUP? LOL: Characterizing out-of-vocabulary words in twitter, in: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, ACM, 2016, pp. 341–344.

[11] T. Baldwin, Y.B. Kim, M.C. De Marneffe, A. Ritter, B. Han, W. Xu, Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition, in: ACL-IJCNLP 126, 2015.

[12] W.D. Abilhoa, L.N. De Castro, A keyword extraction method from twitter messages represented as graphs, Appl. Math. Comput. 240 (2014) 308–325.

[13] M. Choudhury, D. Chatterjee, A. Mukherjee, Global topology of word co-occurrence networks: beyond the two-regime power-law, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 162–170.

[14] W. Liang, Spectra of english evolving word co-occurrence networks, Physica A 468 (2017) 802–808.

[15] Y. Zuo, J. Zhao, K. Xu, Word network topic model: a simple but general solution for short and imbalanced texts, Knowl. Inf. Syst. 48 (2) (2016) 379–398.

[16] M. Garg, M. Kumar, Review on event detection techniques in social multimedia, Online Inf. Rev. 40 (3) (2016) 347–361.

[17] T. Bian, J. Hu, Y. Deng, Identifying influential nodes in complex networks based on AHP, Physica A 479 (2017) 422–436.

[18] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, Selectivity-based keyword extraction method, Int. J. Seman. Web Inf. Syst. (IJSWIS) 12 (3) (2016) 1–26.

[19] W.D. Abilhoa, L.N. De Castro, A keyword extraction method from twitter messages represented as graphs, Appl. Math. Comput. 240 (2014) 308–325.
[20] M. Litvak, M. Last, A. Kandel, DegExt: a language-independent keyphrase extractor, J. Amb. Intell. Humanized Comput. 4 (3) (2013) 377–387.
[21] C. Akimushkin, D.R. Amancio, O.N. Oliveira Jr, Text authorship identified using the dynamics of word co-occurrence networks, PLoS One 12 (1) (2017) e0170527.
[22] V.Q. Marinho, G. Hirst, D.R. Amancio, Authorship attribution via network motifs identification, in: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), IEEE, 2016, pp. 355–360.
[23] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.U. Hwang, Complex networks: structure and dynamics, Phys. Rep. 424 (4) (2006) 175–308.
[24] W. Liang, Y. Wang, Y. Shi, G. Chen, Co-occurrence network analysis of chinese and english poems, Physica A 420 (2015) 315–323.
[25] W. Liang, G. Chen, Spectral analysis of chinese language: co-occurrence networks from four literary genres, Physica A 450 (2016) 49–56.
[26] Y. Gao, W. Liang, Y. Shi, Q. Huang, Comparison of directed and weighted co-occurrence networks of six languages, Physica A 393 (2014) 579–589.
[27] I. Türker, E. Şehirli, E. Demiral, Uncovering the differences in linguistic network dynamics of book and social media texts, SpringerPlus 5 (1) (2016) 1–18.
[28] S. Wuchty, E. Almaas, Evolutionary cores of domain co-occurrence networks, BMC Evol. Biol. 5 (1) (2005) 24.
[29] Y. Matsuo, Y. Ohsawa, M. Ishizuka, Keyworld: extracting keywords from document s small world, in: nternational Conference on Discovery Science, Springer Berlin Heidelberg, 2001, pp. 271–281.
[30] Y. Ohsawa, N.E. Benson, M. Yachida, KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor, in: IEEE International Forum on Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings, IEEE, 1998, pp. 12–18.
[31] G.K. Palshikar, Keyword extraction from a single document using centrality measures, in: International Conference on Pattern Recognition and Machine, 2007, December.
[32] E. Batziou, I. Gialampoukidis, S. Vrochidis, I. Antoniou, I. Kompatsiaris, Unsupervised keyword extraction using the gow model and centrality scores, in: International Conference on Internet Science, Springer, Cham, 2017, pp. 344–351.
[33] P. Meladianos, A.J.P. Tixier, G. Nikolentzos, M. Vazirgiannis, Real-time keyword extraction from conversations, in: EACL 2017, 2017, p. 462.
[34] J.I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani, K-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases, 2005, arXiv preprint cs/0511007.
[35] S.K. Biswas, M. Bordoloi, J. Shreya, A graph based keyword extraction model using collective node weight, Expert Syst. Appl. 97 (2018) 51–59.
[36] G. Erkan, D.R. Radev, LexRank: graph-based lexical centrality as salience in text summarization, J. Artificial Intelligence Res. 22 (2004) 457–479.
[37] R.A.M.O.N. Ferrer i Cancho, A. Capocci, G. Caldarelli, Spectral methods cluster words of the same class in a syntactic dependency network, Int. J. Bifurcation Chaos 17 (07) (2007) 2453–2463.
[38] M. Litvak, M. Last, Graph-based keyword extraction for single-document summarization, in: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, Association for Computational Linguistics, 2008, pp. 17–24.
[39] A. Bellaachia, M. Al-Dhelaan, Ne-rank: a novel graph-based keyphrase extraction in twitter, in: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, IEEE Computer Society, 2012, pp. 372–379.
[40] E. Baralis, L. Cagliero, N. Mahoto, A. Fiori, GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization, Inf. Sci. 249 (2013) 96–109.
[41] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, S. Ghosh, Extracting situational information from microblogs during disaster events: a classification-summarization approach, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, 2015, pp. 583–592.
[42] M. Shams, A. Baraani-Dastjerdi, Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction, Expert Syst. Appl. 80 (2017) 136–146.
[43] W.X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.P. Lim, X. Li, Topical keyphrase extraction from twitter, in: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, Association for Computational Linguistics, 2011, pp. 379–388.
[44] F. Boudin, A comparison of centrality measures for graph-based keyphrase extraction, in: International Joint Conference on Natural Language Processing, IJCNLP, 2013, October, pp. 834-838.
[45] S. Segarra, M. Eisen, A. Ribeiro, Authorship attribution using function words adjacency networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 5563–5567.
[46] D.R. Amancio, Authorship recognition via fluctuation analysis of network topology and word intermittency, J. Statist. Mech.: Theory Exp. 2015 (3) (2015) P03005.
[47] R.S. Roy, S. Agarwal, N. Ganguly, M. Choudhury, Syntactic complexity of web search queries through the lenses of language models, networks and users, Inf. Process. Manage. 52 (5) (2016) 923–948.
[48] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, Proc. R. Soc. Lond. B: Biol. Sci. 268 (1485) (2001) 2603–2606.
[49] J. Wu, Z. Xuan, D. Pan, Enhancing text representation for classification tasks with semantic graph structures, Int. J. Innov. Comput. Inf. Control (ICIC) 7 (5) (2011).
[50] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440.
[51] P. Erdos, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1) (1960) 17–60.
[52] E. Ravasz, A.L. Barabási, Hierarchical organization in complex networks, Phys. Rev. E 67 (2) (2003) 026112.
[53] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[54] R. Noldus, P. Van Mieghem, Assortativity in complex networks, J. Complex Netw. 3 (4) (2015) 507–542.
[55] M.E. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (20) (2002) 208701.
[56] A. Mukherjee, M. Choudhury, R. Kannan, Discovering global patterns in linguistic networks through spectral analysis: a case study of the consonant inventories, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2009, pp. 585–593.
[57] A. Angel, N. Sarkas, N. Koudas, D. Srivastava, Dense subgraph maintenance under streaming edge weight updates for real-time story identification, Proc. VLDB Endow. 5 (6) (2012) 574–585.
[58] J.I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani, Large scale networks fingerprinting and visualization using the k-core decomposition, in: Advances in Neural Information Processing Systems, 2006, pp. 41–50.
[59] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004.
[60] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Netw. ISDN Syst. 30 (1-7) (1998) 107–117.
[61] A. Bougouin, F. Boudin, B. Daille, Topicrank: Graph-based topic ranking for keyphrase extraction, in: International Joint Conference on Natural Language Processing, IJCNLP, 2013, October, pp. 543-551.
[62] J. Alstott, E. Bullmore, D. Plenz, powerlaw: a Python package for analysis of heavy-tailed distributions, PLoS One 9 (1) (2014) e85777.