Identifying the influential spreaders in multilayer interactions of online social networks

Mohammed Ali Al-garadi [1*], Kasturi Dewi Varathan[1*], Sri Devi Ravana[1], Ejaz Ahmed [2,] Victor Chang[3]

[1] Department of Information System, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, Malaysia.

[2] Centre for Mobile Cloud Computing Research, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[3] International Business School Suzhou, Xi'an Jiaotong Liverpool University, Suzhou, China

Abstract. Online social networks (OSNs) portray a multi-layer of interactions through which users become a friend, information is propagated, ideas are shared, and interaction is constructed within an OSN. Identifying the most influential spreaders in a network is a significant step towards improving the use of existing resources to speed up the spread of information for application such as viral marketing or hindering the spread of information for application like virus blocking and rumor restraint. Users communications facilitated by OSNs could confront the temporal and spatial limitations of traditional communications in an exceptional way, thereby presenting new layers of social interactions, which coincides and collaborates with current interaction layers to redefine the multiplex OSN. In this paper, the effects of different topological network structure on influential spreaders identification are investigated. The results analysis concluded that improving the accuracy of influential spreaders identification in OSNs is not only by improving identification algorithms but also by developing a network topology that represents the information diffusion well. Moreover, in this paper a topological representation for an OSN is proposed which takes into accounts both multilayers interactions as well as overlaying links as weight. The measurement results are found to be more reliable when the identification algorithms are applied to proposed topological representation compared when these algorithms are applied to single layer representations.

Keywords: Online social networks, complex network, multilayer interaction, influential spreaders

[*] Corresponding authors. E-mail: mohammedali@siswa.um.edu.my & kasturi@um.edu.my

1. Introduction

Online social networks (OSNs) have billions of users and they have been a dynamic source for various research disciplines. OSNs' lens provide researchers and scientists with exceptional prospects to understand individuals at scale and to analyze human behavioral patterns, otherwise impossible [1]. The data generated by OSNs users have been utilized in various applications [2-4]. The huge rise of OSNs driven by communication technology revolution has intensely renovated the platform of human interactions. Human communications facilitated by OSNs could confront the temporal and spatial limitations of traditional communications in an exceptional way, thereby presenting a qualitatively new layers of social interactions [5], which coincides and collaborates with current interaction layers to redefine the multiplex social networks [5-7]. These several network layers or communication channels in a multiplex network do not act completely separately nor dependently[5]. Although each layer can provide roles within its purpose, it is the interaction and interplay between these layers that can accomplish the full functionality of the network and might provide an increase in significant and unexpected collective outcomes, which can better explain the diffusion process within the network.

Spreading of information influentially is a pervasive process; it refers to variety of applications [8-13]. Targeting these influential spreaders in information propagation is significant for the development of the approaches for either quickening the speed of propagation such as the application of viral marketing [8-10] or blocking the diffusion of undesirable information, such as rumors and viruses [11-13].Therefore, several algorithms have been proposed to identify the most influential spreaders in OSNs. The output of recent researches in identifying influential spreaders in OSNs has triggered an extensive debate. For example, in an OSN (for example Twitter), OSNs structure contains links that are obviously known by users and links that are implicitly detected by network interaction. These links form multi-interactions layer (social friendship layer, retweet interaction layer, mention interaction layer) as shown Figure1. These processes induce connection diversity and multi-layer interaction networks within a single OSN. In network theory, nodes are commonly assumed to be linked to a single type of static link that describes the relationship between them, although, in numerous circumstances, this hypothesis simplifies the complexity of the network. Ignoring the reality of multiple relationships between users [14, 15], as well as the importance of the nodes with respect to the entire structure [16, 17]. Consequently, this idea induces the wrong identification of the most influential nodes (users) with a network [15]. Similarly, identifying influential spreaders in an OSN by modeling a single layer interaction network and ignoring the other interaction will generate a partial relationship information representation, and consequently, uncertain identification results. Therefore, multiple types of interaction between users should be considered for better understanding the information diffusion process and precise influential spreaders identification.

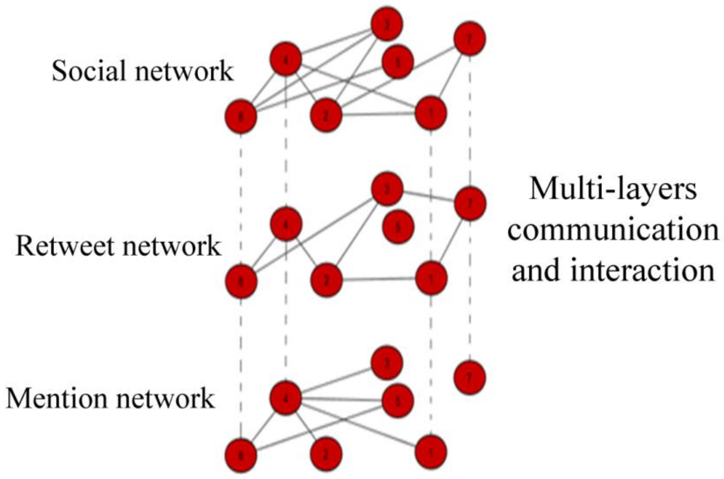

Figure 1 Schematic illustration of the connections in multi interaction layers in an OSN (Twitter network)

In this paper, the topological representation of OSNs of the network structure are argued to have a vital role such that there are possible circumstances under which the top ranked nodes identified by a prominent algorithm applied on single interaction network which poorly correlates with the real dynamic of information diffusion may have a small role in information spreading process. Whereas top ranked nodes identified by a prominent algorithm applied to rich topological network representation, which highly correlates with the real dynamic of information diffusion will have a substantial result that leads to diffusion through a large fraction of the network. Therefore improving the effectiveness of influential spreaders identification is not only depending on the improvement of identification algorithm but also on how the topology of the network is represented. Questions like "how the performance of influential spreaders identification algorithms in an OSN gets influenced by applying them on different topological network representation. "What is best topological network representation to precisely identify the influential spreaders in OSNs context? " are not yet fully investigated and need to be answered. Using real two datasets from Twitter (these two datasets contain large-scale interdependent/interconnected multiplex/multilayer networks; where one-layer represents the social structure and two layers encode different types of user interactions dynamics), the performance of most prominent influential spreaders identification algorithms (degree centrality, PageRank, and k-core algorithms) is investigated. These influential spreaders identification algorithms are applied on different topological network representation. Then the effectiveness of different identification algorithms on different topological network representations are evaluated by comparing ranking list obtained generated by each identification algorithm with ranking list obtained by tracking diffusion links in real spreading dynamics of information [18]. The findings of this paper (presented in result section) are significant in understanding information spread with the real OSNs and on selecting the most efficient algorithms for identifying influential spreaders.

The rest of the paper is organized as follows. Section 2 presents related works. Section 3 discusses proposed method**.** Section 4 presents experimental analysis. In Section 5, the performance evaluations are discussed. Section 6 provides detailed results and discussion. The paper is concluded in section 7.

2. Related work

Several researchers have developed many algorithms to identify influential spreaders. Classical centrality measures, such as degree, closeness centrality, betweenness centrality and eigenvector centrality, are direct methods for recognizing the influential spreaders. However, closeness centrality and betweenness centrality have very high computational complexity, hence, it is not suitable to be applied into very large-scale OSNs [18, 19]. This limitation has made impractical for large OSNs. In other hand eigenvector centrality is not inefficient, especially in scale-free networks, due to the weight will be assigned to few number of nodes (hub), whereas the remaining majority the others have considerably small weights, therefore, they will not be ranked accurately [20]. However, the degree distribution for OSNs such as Facebook network [21] is proved to be the scale-free network. As consequence eigenvector, centrality may lead to improper ranking if applied to such networks. Various studies have used PageRank and its extensions to identify the influential spreaders in OSNs [22-27]. Kitsak *et al* found, in contrast to common belief, there are plausible circumstances where the best spreaders do not correspond to the most highly connected or the most central people [28]. The research [28] showed that the most efficient spreaders are those located within the core of the network as identified by the k-shell decomposition analysis. Recently Pei *et al* have conducted a research [18] with large OSNs datasets and reported that the most influential spreaders are placed in the k-core. The k-core algorithm performs better than degree centrality and PageRank. The performance of aforementioned algorithms have been tested on single layer OSN such as followers network on Twitter [22, 29] , retweet network [25, 26] or mention network [18].

However, even though many researchers developed effective algorithms to identify influential spreaders but investigation of how different representations of OSNs network effects the result of these algorithms has poorly understood. These effects specifically in OSNs where multi social interactions play a diverse role in information spreading should be carefully investigated. For example, applying a proposed algorithm on single layers interaction (such as social network, retweet network or mention network) yielded different results[30]. Consequently, claiming any improvement in result may not be due to the effectiveness of proposed algorithm but it could be due to variation of the network representation, which has a higher correlation with information spreading. Moreover, in OSNs, the lack of knowledge about connection strength between the users can lead to networks with heterogeneous relationship strengths (e.g., acquaintances and best friends mixed together) [31] . Therefore, the binary relationship (the relationship that describes only if the connection exists or not without considering the strength of it) will lead to unreliable connection representation, consequently, variable identification results and the effectiveness and efficiency of the algorithms will be varied with different network representations. The connection strength can be better understood by considering different interaction layers rather than considering a single network [15], for example in twitter the followers network layers explain the social network relations between the users while retweet and mention can give more understanding of connection strength between the users [31-34].

3. Methods

In this section, firstly comprehensive methods to represent single and multilayer of complex OSNs which consist of N nodes and M layers are described, each layer presents a different interaction type between the users within the same OSN. Secondly, the most prominent algorithms used to find the most influential spreaders in OSNs are applied to these networks. Thirdly, how the different multilayer network topological representations influence the accuracy of these algorithms is discussed.

3.1 Network topological representation

Consider a multi-layers network involving several types of links between its nodes. When it is plausible to differentiate the nature of the ties, a successful approach to describe the network comprises in embedding the links in diverse layers based on their category. In this section, the topological representations of the network layers are mathematically described.

3.1.1 Definition 1: Single layer network

A complex social network can be represented as single-layer graphs in which the nodes are connected by links. Nodes represent the users and the links represent the relationship among the users throughout the networks. It is assume that a network can be viewed as graph *G= (N, E)*, where *N* denotes nodes (users), and *E* denotes edges or links (relationship). In network theory, it is common to suppose that nodes are linked by a single type of static edge that summarizes their relations, although in a many of situations this assumption generalizes the complexity of the network. However, various types of connections between nodes can be nowadays appropriately examined on the basis of multilayer networks.

3.1.2 Definition 2: *Aggregating into multi-layer network*

Let us consider network consists of N nodes and M layers. It is considered all the links at all layers to be unweighted layers [17, 35-37]. Each layer α $, \alpha = 1, 2, \dots, M$ is associated an adjacency matrix $A^{\alpha} = \{a_{ij}^{\alpha}\}$. Such a network can be represented by the set layer $\boldsymbol{A} = [\mathrm{A}^{[1]}, \mathrm{A}^{[2]}, \mathrm{A}^{[3]} \dots\dots\dots.. \mathrm{A}^{[\mathrm{M}]}]$ whose elements are $\mathrm{N} \times \mathrm{N}$ adjacency matrices of the M layers [17, 37]. The degree of a node i on a given layer α is denoted by $k_i^{\alpha} = \sum_j a_{ij}^{[\alpha]}$. Therefore, the degree of node i in a multiplex network is the vector.

$$\boldsymbol{k_i} = [k_i^{[1]}, k_i^{[2]}, \dots\dots\dots k_i^{[M]}] \qquad i = 1,2,3 \dots\dots\dots\dots N.$$

Vector variables $\boldsymbol{A}$ and $\boldsymbol{k_i}$ are essential to correctly present multiplex networks.

Aggregating all relationships to a single aggregated network can be presented by aggregated adjacency matrices. In aggregated network, the fact that the links belong to different layers are ignored [17]. The aggregated topological adjacency matrix $A = [a_{ij}]$ of a multiplex network is described similar to Ref [17], where

$$a_{ij} = \begin{cases} 1 & if \ \exists \ \alpha \ \epsilon a_{ij}^{[\alpha]} \\ 0 & otherwise \end{cases}$$

This is the adjacency matrix of the unweighted network achieved from the multi-layer network by combining all pairs of nodes i and j, which are connected by a link in at least one layer of the multiplex network, and ignoring the probability of multi-link existence between a pair of nodes and the nature of each link as well. For the degree of node i on the aggregated topological network,

$k_i = \sum_j a_{ij}$

Summing k_i over all elements of the network is obtained

$\sum k_i = 2K$

Where K is links counts (also called the size) of the aggregated topological network. Hence, such aggregated topological network can be studied using the widely used measures defined for single layer networks. An essential characteristic, which is lost in the topological aggregated representation, is that in the multiplex network the same pair of nodes can be connected by a link of different kinds of relations.

3.1.3 Definition 3: Aggregating into multi-layer network considers the overlaying links as weight

To eliminate the limitation of aggregating all links to a single aggregated network such as a network in Figure 2 that same pair of nodes in multiplex network can be linked by links of different relations. Therefore, here the links describing the different relation between users are taken into account.

Similar as presented in [17], the overlapping of links due to multi-link relations between i and j of two layers α and β

$O_{ij}^{\alpha,\beta} = a_{ij}^{[\alpha]} + a_{ij}^{[\beta]}$

This can be represented in all layers as:

$O_{ij} = \sum_\alpha a_{ij}^{[\alpha]}$

From which follows that $0 \leq O_{ij} \leq M, \forall_i j$.

Therefore, the aggregated overlapping adjacency set will be constructed by:

$ovelap\ adjancy\ matrix = \{O_{ij}\}$

In a multiplex network, the essential questions to be investigated are the following: How can one take into account all the interactions between the different multi-layer networks considering that not all of them hold the equal significance? It is important, to state that in order to calculate the centrality of a node, it is essential to take into account how the centrality (importance, influence) of a node is disseminated within the entire network through a different layer that is not necessarily additives [38]. For example, OSNs such as Twitter is characterized by very heterogeneous interactions [38]. Here the aggregated multi-layer network, which considers the overlaying links as weight, has richer structure compared to the purely topological network.

3.2 Influential Spreaders Identification Algorithms in Complex Networks

A number of different measures aimed at identifying influential spreaders were suggested over the years [39]. The most prominent ones include classical centrality measures in complex networks such as degree centrality [40-42], betweenness centrality [43], closeness centrality [44], and eigenvector centrality [45-47] , PageRank [22-24, 48] and it extensions and k-core algorithm [28, 49, 50] . Classical centrality measurements rely on network topology.

Closeness centrality [43] emphasizes on the extensiveness of influence measurement on whole network. In the succeeding equation, $c_c\,(n_i)$ is the closeness centrality, and $d(n_{i,}\, n_j\,)$ is the distance between two vertices in the network. Calculated as following

$$c_c\,(n_i) = \sum_{i=1}^{N} \frac{1}{d(n_{i,}\, n_j\,)}$$

Betweenness centrality $c_B(n_i)$ [43, 51] is constructed on the number of shortest paths passing over a node. It is assumed that the node with a high betweenness have the significance position of linking different communities. In the succeeding formula, g_{jik} is all geodesics connecting node j and node k which pass through node i; g_{jik} is the geodesic distance between the nodes of j and k.

$$c_B(n_i) = \sum_{j,k\neq i} \frac{g_{jik}}{g_{jk}}$$

However closeness centrality has high computational complexity; hence, it is unsuitable to be applied into significantly large-scale OSNs. Similarly, the best algorithm for betweenness centrality requires a computational time equal to $O(NM)$ for unweighted networks with N nodes and M edges. Also eigenvector centrality is inefficient, particularly in scale-free networks, [20, 21]. Consequently, because betweenness centrality, closeness centrality, and eigenvector centrality are infeasible to calculate for large-scale social networks, therefore in this paper degree centrality, PageRank, k-core are applied to different network presentation. These algorithms are defined as follow

Degree centrality is a direct and widely used topological measure of user influence. Commonly in a network, a high-degree node is assumed to be in authority for the largest spread processes [52, 53]. Users with high connectedness have the opportunity to influence the behavior of others [54].

PageRank is a network-based diffusion algorithm. It is the famous Google algorithm for ranking websites that was initially proposed by Brin *et al*. [48]. PageRank is a global ranking of all web pages, regardless of their contents, based solely on their connected links and locations on the web graph. PageRank scores recursively and two key metrics are considered, namely, incoming links counts and the PageRank value of all incoming links. PageRank was initially used in ranking the pages on the World Wide Web.

PageRank is expressed as follows.

$$P\,R(u) \;=\; (1 \,-\, d)/N \,+\, d \sum_{v \in M(u)} PR(v)/L(v)$$

where *N* is the total number of web pages in the network; *L (v)* is the number of outgoing links from page *v*; *M (u)* refers to the set of web pages pointing to web page *u*; and *d* (with $0 \leq d \leq 1$) is a damping factor that is usually set to 0.85.

K-core ranking is based on the k-shell decomposition of the network. Each node is assigned the k-shell number, k_s, that is, the order of the shell to which it belongs. In k-shell decomposition, all of the nodes with degree $k = 1$ are initially removed, and pruning processes will continue until no node with $k = 1$ exists. Similarly, the pruning processes will be applied to the next k-shells. This process will continue until the k-core of the network is found [49].

4. Experimental analysis

In this section, datasets used are deliberated, and how the different network topological interaction discussed in the previous section are constructed from these datasets, finally discussed how the algorithms performances are evaluated and compared on network topological representations.

4.1 Dataset

Two real online social network datasets are used to investigate and evaluate the performance of influential spreaders algorithms on different network structure in order to compare different topological network representation. These two datasets have been anonymized, in such way the same user ID is used for all networks (Social network, Retweet network and Mention network). More importantly is that these two datasets are large-scale interdependent/interconnected multiplex/multilayer networks; where one-layer represents the social structure and two layers encode different types of user interactions dynamics.

4.1.1 Dataset 1

Dataset 1 contains directed twitter network used in research [7]. This dataset is Higgs dataset, which has been constructed after observing the spreading processes on Twitter before, during and after the declaration of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. The messages posted on Twitter about this discovery between 1st and 7th July 2012 are considered. Social network dataset contain 456626 nodes, 14,855,842 edges. Retweet network contains 256491nodes, 328132 edges and Mention network contains 116408 nodes and 150818 edges.

4.1.2 Dataset 2

This dataset comprises of 121,807,378 tweets generated by 14,599,240 unique users [55]. Then, they constructed an undirected, unweighted social network based on reciprocal following relationships between 595,460 randomly selected users, as bidirectional links that reflect more stable and reliable social connections. Two other types of networks constructed based on Retweets and mentions were built. .

Hence based on these datasets, the social network nodes is used to construct the network and the number of Retweets, Mention corresponding of each user are extracted from Retweet network, Mention network, and they are use to build the weight for this social network.

4.2 Real Network Topological Representation

In this section different networks representation of real network from dataset 1 and 2 are described based on mathematical representation of the network topology introduced in section 3.1

4.2.1 Single layer networks

Using the above mention datasets, following networks are constructed:

Social network (SN): it is assumed that a network can be constructed as graph *G= (N, S),* where *N* denotes nodes (users), and *S* denotes links (following relationships).

$$S_{ij} \;= \begin{cases} 1 & if\ i \leftrightarrow\ j\ \ in\, SN \\ 0 & otherwise \end{cases}$$

Retweet network (RN): it is assumed that a network can be constructed as graph *G= (N, R)*, where *N* denotes nodes (users), and *R* denotes links (Retweet relationships).

$$R_{ij} = \begin{cases} 1 & if\ i \leftrightarrow j\ \ in\ RN \\ 0 & otherwise \end{cases}$$

Mention network (MN): it is assumed that a network can be constructed as graph *G= (N, T)*, where *N* denotes nodes (users), and *T* denotes links (mention relationships).

$$T_{ij} = \begin{cases} 1 & if\ i \leftrightarrow j\ \ in\ MN \\ 0 & otherwise \end{cases}$$

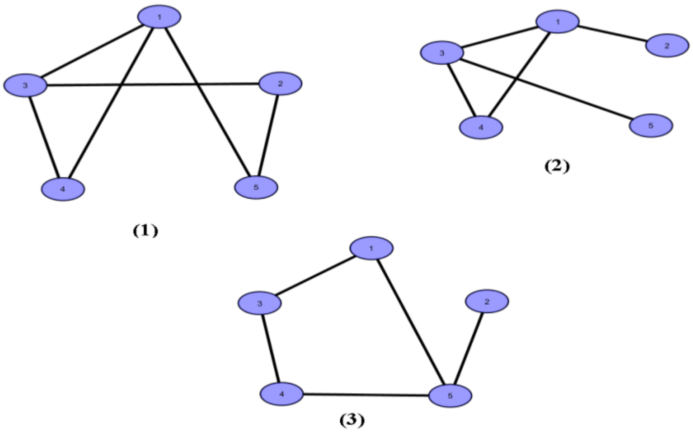

Figure 2 Single layer networks

4.2.2 Aggregated multilayer network

Here the network is constructed by aggregating different relationships (following, retweet, mention) into a single network.

Let us consider a network consists of N nodes and M layers where in our case M = 3 . All the links at all layers to be unweighted layers. each layer α ,$\alpha = 1, 2, \ldots, M$ are associated an adjacency matrix $A^{\alpha} = \{a_{ij}^{\alpha}\}$. Such a network can be represented by the set layer for example in our case it could be represented as $\boldsymbol{A} = [\mathrm{A}^{[\mathrm{SN}]}, \mathrm{A}^{[\mathrm{RT}]}, \mathrm{A}^{[\mathrm{MN}]}]$ whose elements are $\mathrm{N} \times \mathrm{N}$ adjacency matrices of the 3 layers [17, 37]. Aggregating all relationships to a single aggregated network can be presented by aggregated adjacency matrices. In aggregated network the fact that the links belongs to different layers is neglected [17]. The aggregated topological adjacency matrix is defined as $A = [a_{ij}]$ of a multiplex network, where

$$a_{ij} = \begin{cases} 1 & if\ \exists\ \alpha\ \epsilon a_{ij}^{[\alpha]} \\ 0 & otherwise \end{cases}$$

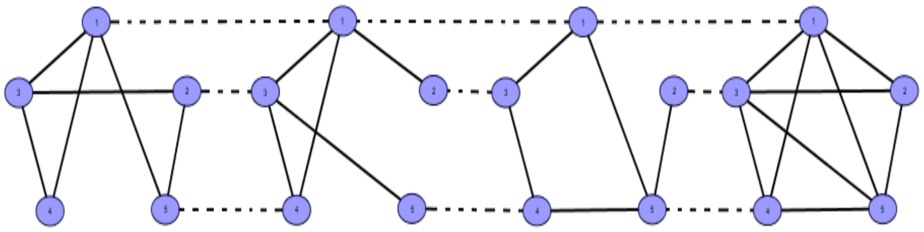

Figure 3 Aggregated multilayer network

4.2.3 Aggregated multilayer network with overlapping links consideration

The overlap of links i and j between three layers SN , RN and MN are presented as

$$O_{ij}^{SN,RN,MN} = a_{ij}^{[SN]} + a_{ij}^{[RN]} + a_{ij}^{[MN]}$$

This can be represented in all layers as:

$O_{ij} = \sum_{\alpha} a_{ij}^{[\alpha]}$

From which follows that $0 \leq O_{ij} \leq M, \forall_i j$.

Consequently, the aggregated overlapping adjacency will be constructed by:

$ovelap\ adjancy\ matrix = \{O_{ij}\}$

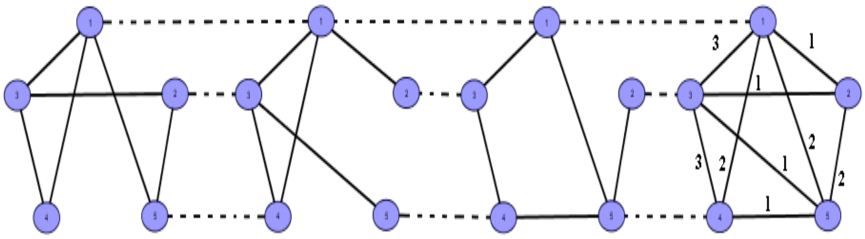

Figure 4 Aggregated multilayer network with overlapping links consideration

4.3 Evaluation and Comparison of Algorithms Performance on Network Topological Representation

In this subsection, the evaluation model and effectiveness of influential spreaders algorithm applied to different topological network representations is presented.

4.3.1 Evaluation models

Information spread can be modeled in probabilistic frameworks [56]. Several research works have Intensive implemented of classical disease models like susceptible-infectious-recovered (SIR) model and susceptible-infectious- susceptible (SIS) to model in information diffusion and information spread [57]. Unfortunately, these models [58] are developed based on basic belief of human behavior which might not be representative and illustrative of real dynamics information diffusion [18, 57]. Therefore studies have reported that the measurement which is based on artificial models are not suitable in practice [59, 60]. Moreover, the spread of diseases and spread of information are found to be different [59, 61]. Based on these observations this study is validated using real dynamics of information diffusion in real-world social network similarly to study [18]. In order to construct real dynamics of information diffusion for the datasets used in this study, the retweet network is used related to all the users in the social network. The retweet network is best representative network which can explain how the information is diffused in Twitter [7]. In Retweet network, if user i retweets a tweet of user j, the information propagates from j to i, thus creating diffusion link from j to i. In this way, diffusion graph of the networks are built. Overall spreading efficiency of each user is calculated and ranking list of the users is generated.

4.3.2 Effectiveness of spreaders identification algorithms

The spreading efficiency of an individual origin i through the number of the users in the region of influence is calculated and denoted as M_i. In order to evaluate which algorithm are more accurate to measure the spreading capability of nodes, degree, PageRank, and k-core are compared by calculating the imprecision function ϵ_k ϵ_{PR} ϵ_{k_s} for degree, a a PageRank , and k-core respectively, proposed in ref [28]. Where imprecision function of ϵ_k is calculated as following:

$$\epsilon_k(p) = 1 - \frac{M_k\,(p)}{M_{eff}(p)}$$

Similarly, imprecision function of ϵ_{PR} ϵ_{k_s} are calculated

$$\epsilon_{PR}(p) = 1 - \frac{M_{PR}\,(p)}{M_{eff}(p)}$$

$$\epsilon_{k_s}(p) = 1 - \frac{M_{k_s}\,(p)}{M_{eff}(p)}$$

Where p is the fraction of network size $N(p \in [1,0])$, $M_{(k)(pr)(k_s)}(p)$ is the average spreading efficiencies of pN nodes with highest (degree, PageRank, and k-core, values and $M_{eff}(p)$ is average spreading efficiencies of pN nodes with largest spreading efficiency. The smaller imprecision function (ϵ) value, the more accurate the algorithm is to identify the most influential spreaders. A value for (ϵ) close to 0 denotes a very efficient process, since the nodes that are chosen are practically those that contribute most to information diffusion. imprecision function of (1%, 5%, 10%, 15%, and 20%) top influential spreaders in network identified by degree, PageRank, and k-core are compared as shown in figure 5 and 7 for dataset one and two respectively.

Even though imprecision function can quantify the spreading efficiency well, it is unclear which algorithm can better locate individual influential spreaders. Therefore, recognition rate $r(f)$ proposed in [18] is used to verify the performance of each algorithms in recognizing influential spreaders. Is recognition rate $r(f)$:

$$r(f) = \frac{|I_f \cap P_f|}{|I_f\,|}$$

Where I_f and P_f ranking lists in the top f fraction obtained by tracking diffusion links in real spreading dynamics (node influence) and obtained by algorithms (degree, PageRank, and K-core) respectively the higher recognition rate indicate that algorithm identify the influential spreaders more accurate . Top network fraction (1%, 5%, 10%, 15%, and 20%) recognition rate of degree, PageRank, and k-core is compared as shown in Figure 6 and 8 for dataset one and two respectively.

5. Results and discussions

First the imprecision function and recognition rate for a degree, PageRank and k-core applied to on different topological representation of network (single layer :social network , retweet network , and mention network ; aggregated multilayer network and; aggregated multilayer network with overlapping links as weight) are calculated. The imprecision functions and recognition rate of ten real topological representation networks extracted from two datasets are shown in fig. 5, 7 and fig 6, 8 respectively. Contrary to common belief, there is no any algorithm, which always performs well in all topological representation of the different networks. How the dataset is extracted as well as, how the network is represented is an important factor for determining the ranking accuracy. However, in dataset 1, the k-core has performed well in all networks as shown in figure 5 and 6. in dataset 2 as shown in figure 7 and 8 the degree performed well in three networks (retweet network and weighted aggregated network) while k-core perform well in one network (social network) and all algorithms have approximately similar accuracy in mentioned network. With respect to network representation, both retweet network and proposed weighted aggregated network has given comparably a well representation of information diffusion. In dataset 1, the k-core applied to retweet network achieved best-ranking result (lowest imprecision and highest recognition rates) compared other algorithms. In dataset 2, the degree provided best result lowest imprecision and highest recognition rates) compared other algorithms. Initially, this indicates the retweet is considered as a good topological network representation as it provides the ranking algorithms, informative network data, which result in lowest imprecision function and highest recognition rates. But deep analysis of the results showed that even though k-core and degree algorithms perform well in retweet network in dataset 1 and dataset 2 respectively the remaining ranking algorithms (degree and PageRank in dataset 1 and PageRank and k-core in dataset 2) failed to perform well in retweet networks. This is due to the topology perturbations of the retweet network, which affect the ranking values provided by the ranking algorithms. This effect has been observed by the diverse imprecision functions values and recognition rates of the ranking algorithms applied to retweet network [62]. In contrary to the proposed weighted aggregated topological network representation, this has provided informative network data, which result in comparable low imprecision function and high recognition rate for all algorithms in both datasets. This indicates the weighted aggregated topological network representation is more reliable to represent diffusion process as ranking algorithm are not much varied when applied to this network representation compared to when ranking algorithms applied to retweet network.

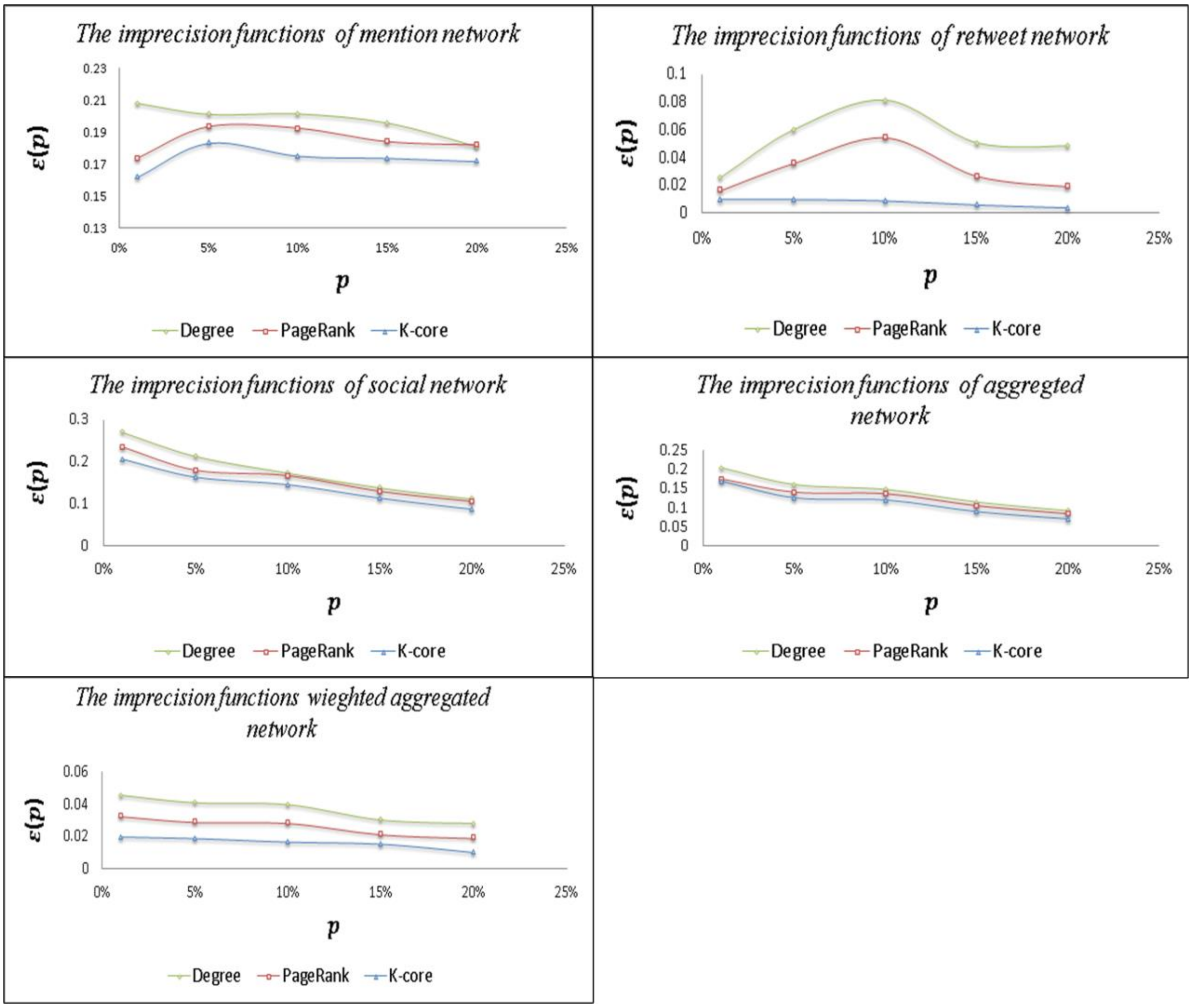

Figure 5 Imprecision function of identification algorithms applied to different topological network representations of dataset 1

.

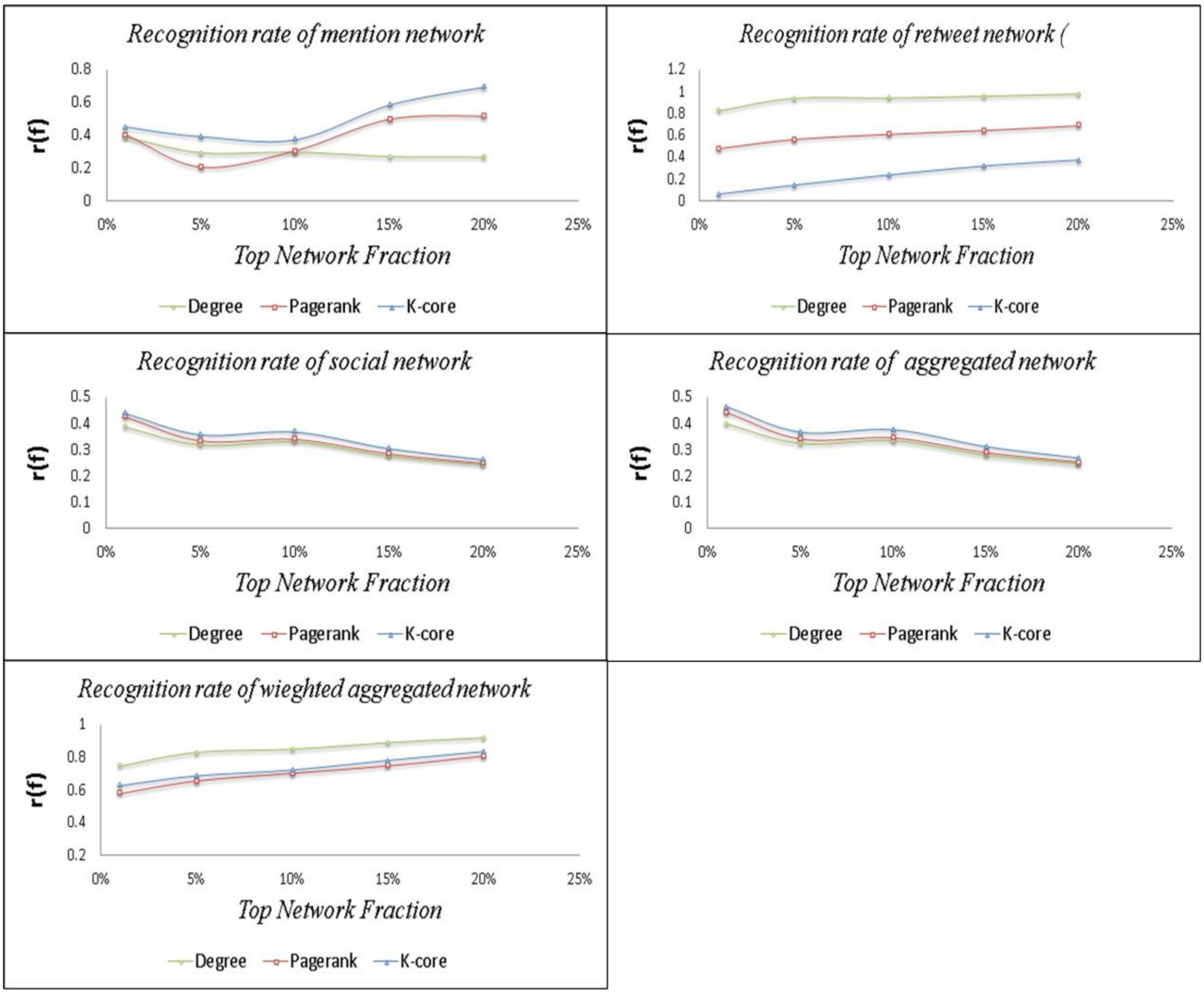

Figure 6 Recognition rate of identification algorithms applied to different topological network representations of dataset 1.

In order to find out the reason for the poor performances of ranking algorithms under different network representation, the topological characteristics of the studied real networks are explored. In dataset 1, the results are quite consistent: k-core performs better than degree and PageRank. This result shows that k-core catches the common properties of the diffusion process, which let the k-core powerful influential spreaders identification algorithm across different network representation. In dataset 2 the degree certainly performs well in three network representation which indicates that the reciprocal properties of the diffusion process can be captured by local structural of the users and degree can provide efficient ranking results compared to the other two algorithms. It is important to note here, the dataset 2 were constructed in such way, the following relationships between randomly selected users presented as bidirectional links, which can reflect stronger social connections. However, this has led to biasing the diffusion process to bidirectional links, which is not always true. In a Twitter network, most of the users pairs with any link between them are connected in the one-way direction[30] and OSNs have circumstances where information spreads between two users even if they are not connected by a social link. Hence, the dataset 2 construction has affected the topological representation of the most networks in this dataset. The top spreaders were limited to those who have a bidirectional relationship.

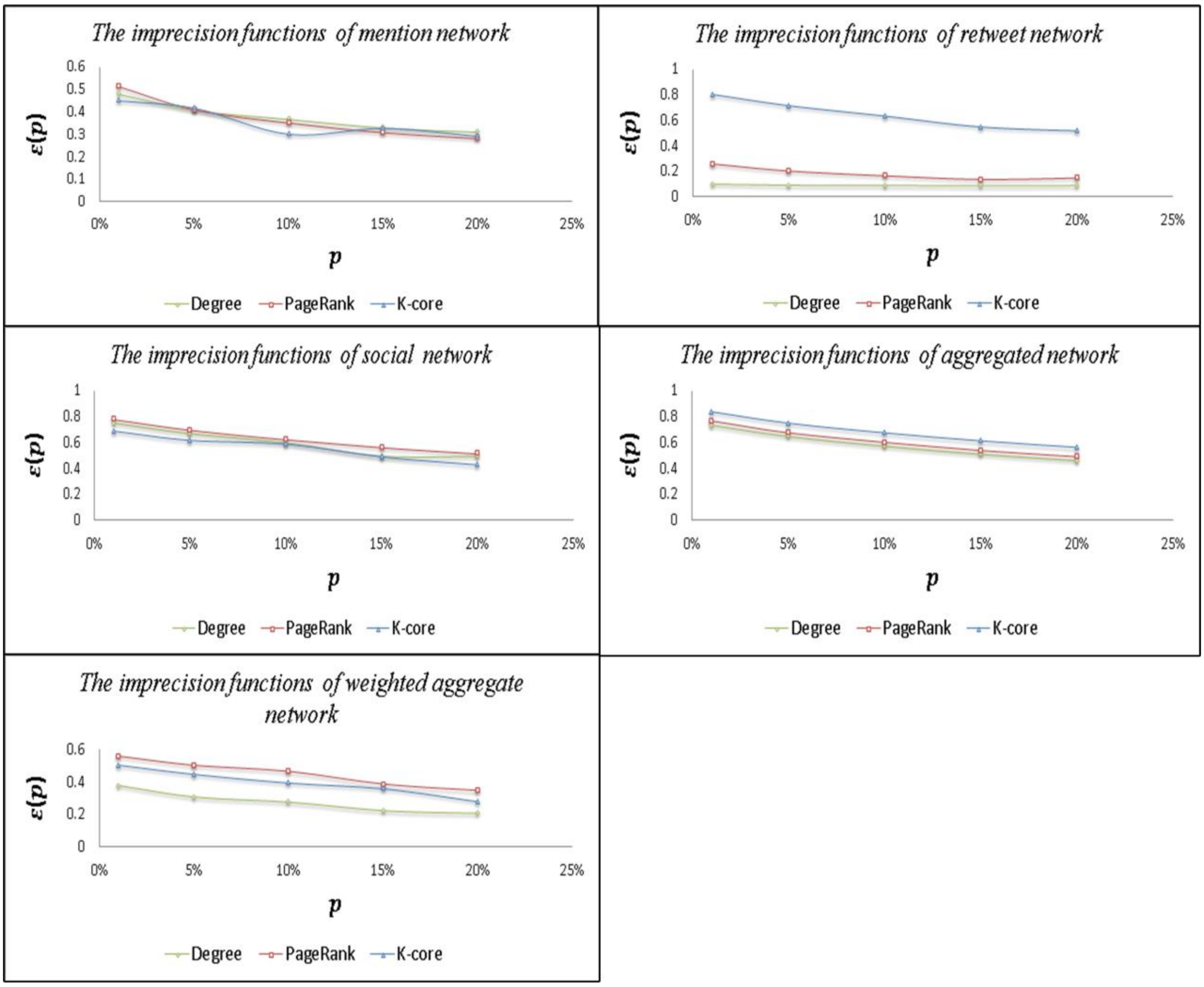

Figure 7 Imprecision function of identification algorithms applied to different topological network representations of dataset 2.

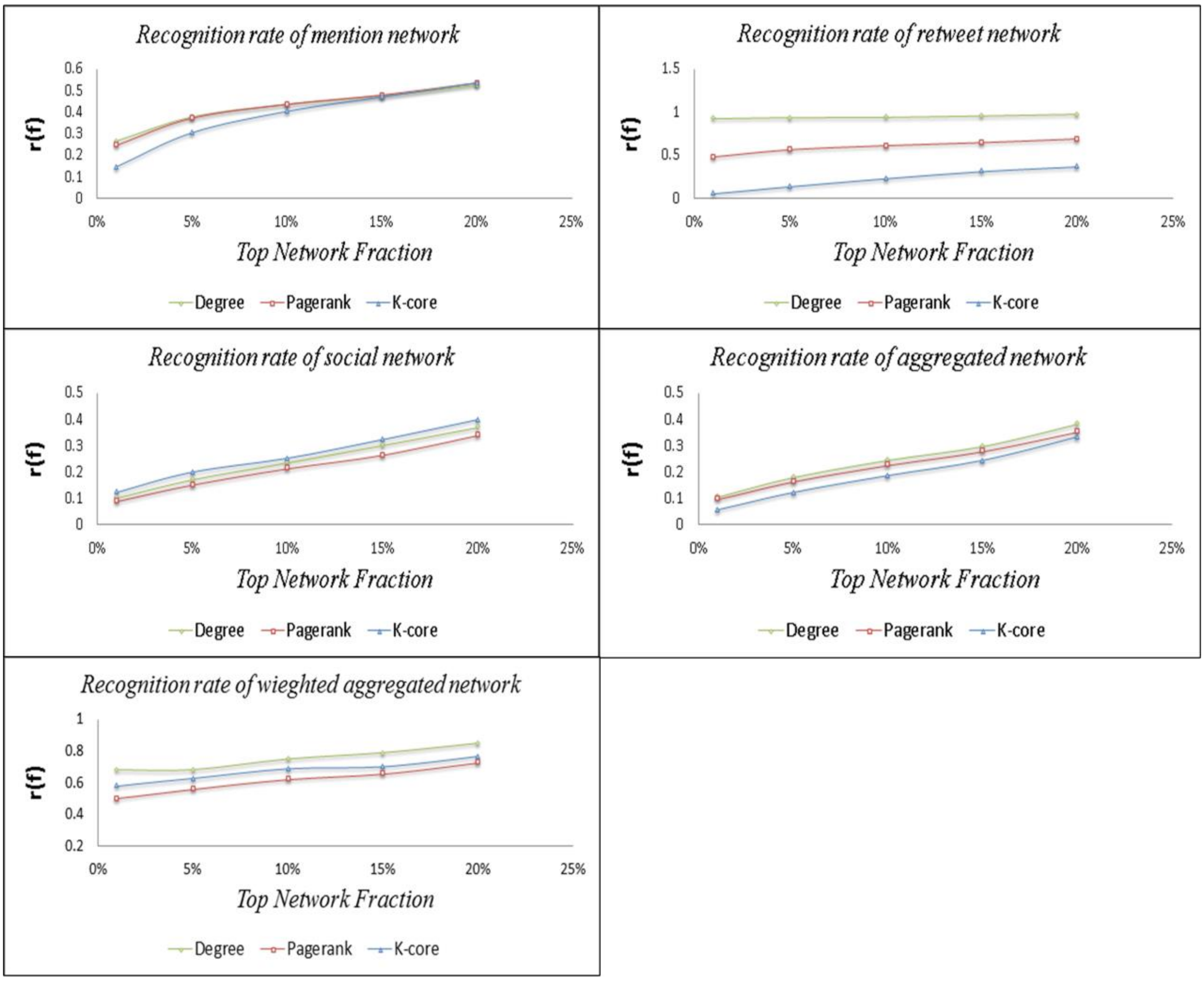

Figure 8 Recognition rate of identification algorithms applied to different topological network representations of dataset 2.

From the result Figures 5, 6,7, and 8, it is observed that the PageRank has failed to detect influential spreaders in most of the studied network, as both datasets represent incomplete network data of OSNs and the measurements given by PageRank are responsive to perturbations in network topology, rendering it unreliable for incomplete or noisy networks [62]. However, the complete OSN structure is unavailable due to the inherent limitations of OSNs caused by API restrictions and user privacy. Consequently, the PageRank algorithm is an unreliable measurement for OSNs. Moreover, the finding of this paper reconfirmed that the accomplishment of PageRank in web network, while it failed in OSNs, was due to the unintentional result of the scale-free nature of the web graph [62]. If the web graph was an exponential network, the ranking generated by PageRank would have been unreliable given the incompleteness of the web graph [62].

Generally in complex networks, the most connected nodes are usually considered to authoritative for the largest information dissemination and are viewed as the most influential nodes [52]. An inadequacy of this method is that hubs may form tightly-knit groups called “rich-clubs” [63]. Approaches based on degree measures will highly rank these rich-club hubs [20]. However, reasonable situations exist in which the influential spreaders do not correspond to the most highly connected users [28]. In this study, the success of degree in identifying the influential spreaders in the three-network representation of dataset 2 indicates that the properties of the diffusion process can be apprehended by local structural of the users. This can be inferred as the reciprocal properties of this dataset has produced highly connected users in the network and tracking the information in propagation network is limited to reciprocal links which lead to a highly correlation between the outcomes of degree method and real dynamic of information diffusion. However, in dataset 1 and social network of dataset 2, the degree method does not perform well. The failure of degree method due to the local features of nodes (number of links) are not always represented the spreading efficiency of nodes in the network. The position of users within the network as well as the spreading efficiency of their connected users plays a major role in the diffusion of information within in OSNs. These factors cannot be captured by the degree method, which simply represent the local connection features of the users.

The k-core measures the spreading efficiency of the users more effectively than other algorithms in all network of dataset 1 as well as in a social network of dataset 2 but it fails to identify the influential spreaders in a retweet, and aggregated network of dataset 2. This can be explained as influential spreaders in these networks were identified more accurate by a direct number of connections. The k-core defines the most influential nodes as those that are located within the core of the network, and they can be successfully identified by the k-core decomposition method [28]. The limitations related to the k-shell decomposition such as considering only the links between the remaining nodes and entirely ignoring the links connected to the removed nodes has led to failure k-core in theses network. This can be explained as most influential spreaders in this network were connected to many users that have low k_s values and were removed in beginning stage. Therefore, k-core was not able to detect these users. These influential spreaders were detected by direct number of their links regardless to their position in the network or to whom they are connected. this finding leads need more investigation on roles of low-degree users in information diffusion specifically those who have significant broker role in the network [20].

The conclusion based on our preliminary analysis exposed that the improvement, of influential spreaders identification accuracy is not only based on the improvement of ranking algorithms but also developing a network topology that represents the information diffusion well. In addition, it should consider the multi-layers interaction between users for better understanding the social influence and spreading processes. Therefore, the network multiplexity needs to noticeably be considered to understand and predict spreading dynamics accurately in OSNs. Our result has shown there is not a single influential spreaders identification algorithm, which always performs well in any topological networks. It is required to understand how the network dataset is extracted and how the users within the network are connected and interacted in order to identify the best possible algorithms.

6. Conclusion

The huge rise of OSNs has intensely renovated the platform of human interactions. Several network layers or communication channels in such multiplex network do not act completely separately nor dependently [5]. Although each layer can provide roles within its purpose, it is the interaction and interplay between these layers that can accomplish the full functionality of the network and might provide an increase in nontrivial and unexpected collective outcomes, which can better explain the diffusion process within the network.

This study has concluded that based on our preliminary analysis, improving the accuracy of influential spreaders identification is not only based on the improvement of identification algorithms but also on developing a network topology that represents the information diffusion as well. In addition, multi-layers interaction between users and spreading processes need to be looked into more carefully. Therefore, the network multiplexity needs noticeably be considered to understand and predict the spreading dynamics accurately in OSNs. Our result has shown that there is not a single influential spreader identification algorithm which always performs well in any topological networks. It is required to understand how the network dataset is extracted and how the users within the network are interacting in order to identify the best possible algorithms. The results obtained have shown that topological representation of the OSNs which takes into account both multilayers interactions as well as the weight of the interaction has given results that are more reliable.

However, the future of OSNs platform will represents multilayers of network that allows not only the users within the same online network to be connected but also people and smart devices within the community will be connected in multi and different layers [64-66]. Consequently this may introduce networks with different interaction creating intersecting researches fields as future directions of different multilayer OSNs and their role in developing smart cities [67].

7. Acknowledgements

We would like to take this opportunity to thank University of Malaya Research Grant (RP028D-14AET) for funding this research.

8. References

[1] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and Tracking Political Abuse in Social Media," in *ICWSM*, 2011.

[2] M. A. Al-garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, "Using online social networks to track a pandemic: A systematic review," *Journal of Biomedical Informatics,* 2016.

[3] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior,* 2016.

[4] V. Chang, "A Cybernetics Social Cloud," *Journal of Systems and Software,* 2015.

[5] B. Min, S.-H. Gwak, N. Lee, and K.-I. Goh, "Layer-switching cost and optimality in information spreading on multiplex networks," *Scientific reports,* vol. 6, 2016.

[6] J. Borge-Holthoefer, A. Rivero, I. García, E. Cauhé, A. Ferrer, D. Ferrer*, et al.*, "Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study," *PloS one,* vol. 6, p. e23883, 2011.

[7] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi, "The anatomy of a scientific rumor," *Scientific reports,* vol. 3, 2013.

[8] D. J. Watts, J. Peretti, and M. Frumin, *Viral marketing for the real world*: Harvard Business School Pub., 2007.

[9] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 61-70.

[10] M. R. Subramani and B. Rajagopalan, "Knowledge-sharing and influence in online social networks via viral marketing," *Communications of the ACM,* vol. 46, pp. 300-307, 2003.

[11] K. Li, Y. Tan, W. Zhang, and W. Wei, "The research of e–mail virus spread based on complex network," *International Journal of Computing Science and Mathematics,* vol. 6, pp. 188-200, 2015.

[12] L. Zhao, Q. Wang, J. Cheng, Y. Chen, J. Wang, and W. Huang, "Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal," *Physica A: Statistical Mechanics and its Applications,* vol. 390, pp. 2619-2625, 2011.

[13] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 2013, pp. 1103-1108.

[14] S. Gomez, A. Diaz-Guilera, J. Gomez-Gardeñes, C. J. Perez-Vicente, Y. Moreno, and A. Arenas, "Diffusion dynamics on multiplex networks," *Physical review letters,* vol. 110, p. 028701, 2013.

[15] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, "Ranking in interconnected multilayer networks reveals versatile nodes," *Nature communications,* vol. 6, 2015.

[16] A. Halu, R. J. Mondragón, P. Panzarasa, and G. Bianconi, "Multiplex pagerank," 2013.

[17] F. Battiston, V. Nicosia, and V. Latora, "Structural measures for multiplex networks," *Physical Review E,* vol. 89, p. 032804, 2014.

[18] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," *Scientific reports,* vol. 4, 2014.

[19] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica a: Statistical mechanics and its applications,* vol. 391, pp. 1777-1787, 2012.

[20] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature,* 2015.

[21] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Extraction and analysis of facebook friendship relations," in *Computational Social Networks*, ed: Springer, 2012, pp. 291-324.

[22] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261-270.

[23] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PloS one,* vol. 6, p. e21202, 2011.

[24] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted LeaderRank," *Physica A: Statistical Mechanics and its Applications,* vol. 404, pp. 47-55, 2014.

[25] L. B. Jabeur, L. Tamine, and M. Boughanem, "Active microbloggers: identifying influencers, leaders and discussers in microblogging networks," in *String Processing and Information Retrieval*, 2012, pp. 111-117.

[26] Z.-y. Ding, Y. Jia, B. Zhou, Y. Han, L. He, and J.-f. Zhang, "Measuring the spreadability of users in microblogs," *Journal of Zhejiang University SCIENCE C,* vol. 14, pp. 701-710, 2013.

[27] A. Silva, S. Guimarães, W. Meira Jr, and M. Zaki, "ProfileRank: finding relevant content and influential users based on information diffusion," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, 2013, p. 2.

[28] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley*, et al.*, "Identification of influential spreaders in complex networks," *Nature Physics,* vol. 6, pp. 888-893, 2010.

[29] W. Chen, S. Cheng, X. He, and F. Jiang, "Influencerank: An efficient social influence measurement for millions of users in microblog," in *Cloud and Green Computing (CGC), 2012 Second International Conference on*, 2012, pp. 563-570.

[30] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 591-600.

[31] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 981-990.

[32] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Machine learning and knowledge discovery in databases*, ed: Springer, 2011, pp. 18-33.

[33] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the SIGCHI Conference*

on Human Factors in Computing Systems, 2009, pp. 211-220.

[34] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 519-528.

[35] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, "Structural reducibility of multilayer networks," *Nature communications,* vol. 6, 2015.

[36] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter*, et al.*, "Mathematical formulation of multilayer networks," *Physical Review X,* vol. 3, p. 041022, 2013.

[37] V. Nicosia, G. Bianconi, V. Latora, and M. Barthelemy, "Growing multiplex networks," *Physical review letters,* vol. 111, p. 058701, 2013.

[38] L. Solá, M. Romance, R. Criado, J. Flores, A. G. del Amo, and S. Boccaletti, "Eigenvector centrality of nodes in multiplex networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science,* vol. 23, p. 033131, 2013.

[39] S. Pei and H. A. Makse, "Spreading dynamics in complex networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2013, p. P12002, 2013.

[40] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007, pp. 29-42.

[41] J. Jiang, C. Wilson, X. Wang, W. Sha, P. Huang, Y. Dai*, et al.*, "Understanding latent interactions in online social networks," *ACM Transactions on the Web (TWEB),* vol. 7, p. 18, 2013.

[42] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science,* vol. 286, pp. 509-512, 1999.

[43] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry,* pp. 35-41, 1977.

[44] K. Faust, "Centrality in affiliation networks," *Social networks,* vol. 19, pp. 157-191, 1997.

[45] H. He, "Eigenvectors and reconstruction," *the electronic journal of combinatorics,* vol. 14, p. N14, 2007.

[46] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*: John Wiley & Sons, 2012.

[47] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social networks,* vol. 28, pp. 466-484, 2006.

[48] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer networks,* vol. 56, pp. 3825-3833, 2012.

[49] V. Batagelj and M. Zaversnik, "An O (m) algorithm for cores decomposition of networks," *arXiv preprint cs/0310049,* 2003.

[50] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "K-core organization of complex networks," *Physical review letters,* vol. 96, p. 040601, 2006.

[51] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," *Journal of the American Society for Information Science and Technology,* vol. 60, pp. 2107-2118, 2009.

[52] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *nature,* vol. 406, pp. 378-382, 2000.

[53] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters,* vol. 86, p. 3200, 2001.

[54] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics,* vol. 74, p. 47, 2002.

[55] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Scientific reports,* vol. 3, 2013.

[56] D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, X. Lan, and S. Suri, "Sequential Influence Models in Social Networks," *ICWSM,* vol. 10, p. 26, 2010.

[57] S. Pei, L. Muchnik, S. Tang, Z. Zheng, and H. A. Makse, "Exploring the complex pattern of information spreading in online blog communities," 2015.

[58] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM review,* vol. 42, pp. 599-653, 2000.

[59] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Threshold-limited spreading in social networks with multiple initiators," *Scientific reports,* vol. 3, 2013.

[60] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters,* vol. 12, pp. 211-223, 2001.

[61] D. Centola and M. Macy, "Complex contagions and the weakness of long ties1," *American Journal of Sociology,* vol. 113, pp. 702-734, 2007.

[62] G. Ghoshal and A.-L. Barabási, "Ranking stability and super-stable nodes in complex networks," *Nature communications,* vol. 2, p. 394, 2011.

[63] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks," *Nature physics,* vol. 2, pp. 110-115, 2006.

[64] E. Ahmed, A. Gani, M. Sookhak, S. H. Ab Hamid, and F. Xia, "Application optimization in mobile cloud computing: Motivation, taxonomies, and open challenges," *Journal of Network and Computer Applications,* vol. 52, pp. 52-68, 2015.

[65] E. Ahmed, A. Gani, M. K. Khan, R. Buyya, and S. U. Khan, "Seamless application execution in mobile cloud computing: Motivation, taxonomy, and open challenges," *Journal of Network and Computer Applications,* vol. 52, pp. 154-172, 2015.

[66] U. Shaukat, E. Ahmed, Z. Anwar, and F. Xia, "Cloudlet deployment in local wireless networks: Motivation, architectures, applications, and open challenges," *Journal of Network and Computer Applications,* vol. 62, pp. 18-40, 2016.

[67] F. Aisopos, A. Litke, M. Kardara, K. Tserpes, P. M. Campo, and T. Varvarigou, "Social Network services for innovative Smart Cities: the RADICAL platform approach," *Journal of Smart Cities,* vol. 2, 2016.