Exploiting Online Social Network Structural Properties for Information Spreading

Edward Yellakuor Baagyere, Zhen Qin, Hu Xiong, and Qin Zhiguang

***Abstract*—The ability to influence individuals on online social networks for dissemination of information is crucial for commercial advertising, online marketing, political campaigning and, for the general public. However, there is still a research gap in understanding the underlying structure of these networks, their structural properties and how these properties can be leveraged in other research areas. Though information dissemination is a key objective of most online social networks, several influence models that are proposed in the literature are based on simulations, greedy and heuristic approaches, which sometimes are computationally expensive. Thus, these approaches do not take advantage of the underlying properties of these networks for effective and efficient information dissemination. This is because these network structural properties are not well-studied couples with the computationally expensive algorithms for implementing information diffusion on them. To this end, we propose to address these gaps in three folds. Firstly, the structural properties of several online social networks are studied to have a thorough overview of their underlying structure. Secondly, an efficient information diffusion algorithm is proposed and implemented with a less computational time that scales $O(N)$, where N is the number of nodes. Thirdly, we apply the algorithm to these networks and an influence index is calculated on them in order to study the impact of their structural properties on information diffusion and also as a way to characterize them. The results show that the networks structural properties of online social networks such as the average clustering coefficient, average degree, degree entropy, edge entropy among others, are effective in disseminating information as they correlate well with the influence index.**

***Index Terms*—Network structural properties, influence radius, information dissemination, online social networks.**

I. Introduction

An online social network consists of the interaction between two or more individuals called "nodes". These interactions could be for sharing common interest, knowledge, beliefs or friendship, depending on the kind of online social network.

The need to model these interrelations with the right tool is paramount to understanding and utilizing these complex networks and has been a hotbed of research in recent times. Graph theory tool is used to model these relationships and then further analyze to decode the intricate underlying network properties behind these systems.

Presently, online social networks are the portal through which information is disseminated to individuals. User traffic on these social platforms now rivals that of the traditional web. These networks play an important role in the spread of information and influence and, therefore, are of interest to business analysts, politicians and even the human society as a whole. For instance, a huge amount of money is spent yearly on digital ad advertisement on online social networks. It was estimated that worldwide social network ad spending will reach $16.10 billion in 2014, an increase of 45.3% from 2013 and will push social networks share of the overall digital ad investment to 11.5% [1], and will collectively spend more than $58 million on paid digital advertising in 2015 for the sixth consecutive year. [2] and [3] argued that information and influence can be propagated even more powerfully via the Internet than by traditional channels. This argument is more profound in this Web 2.0 era that has social networking services embedded in the traditional web 1.0. Many advertisers, celebrities and politician leverage on these social platforms to enhance their influence or attract followers. However, characterizing the topological properties of online social networks and leveraging on these properties in order to maximize the full potential of these networks is still an ongoing research.

Information dissemination is a key objective of most online social networks and many research efforts are directed at maximizing the influence of nodes on these networks for information diffusion [4]–[6]. The major limitations of most of these research efforts are that simulation, greedy and heuristic methods are used in studying these phenomena [6]–[8] which are either not data driven, or do not leverage on the underlying network structure or are directed at very few networks. Furthermore, the non-greedy and heuristic methods are computationally expensive coupled with the fact that most of these network structural properties are not well studied and therefore cannot be efficiently harnessed for information diffusion.

Importantly, almost all these methods focus largely on the use of a subset of nodes in the network that has the potential of greater influence on the entire network without considering the impact of the network's structural properties. Also, advertisers, marketing analysts, and society at large are faced with the challenge of knowing which online social network can efficiently disseminate information. To this end, we seek to address each of these problems by first exploiting the structural properties of several online social networks in order to characterize them and using an efficient and less computationally expensive algorithm for measuring nodes

Manuscript received January 6, 2016; revised July 14, 2016. This work was supported in part by the National Science Foundation of China (No.61133016, No.61300191, No.61202445 and No.61370026), the Sichuan Key Technology Support Program (No.2014GZ0106), the National Science Foundation of China - Guangdong Joint Foundation (No.U1401257), and the Fundamental Research Funds for the Central Universities (No.ZYGX2013J003 and No.ZYGX2014J066).

The authors are with the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China (e-mail: ybaagyere@uds.edu.gh, qinzhen@uestc.edu.cn, xionghu.uestc@gmail.com, qinzg@uestc.edu.cn).

influence maximization on these online social networks.

We then proposed which of these networks are efficient and can be leveraged for information dissemination and being able to achieve the following:

1) The structural properties of 12 different kinds of online social networks from several parts of the world are studied using concepts from graph theory in order to have a comprehensive understanding of online social network topological structure.
2) We proposed and implemented an efficient algorithm for information dissemination and defined an *influence-index* for characterizing online social networks for information dissemination.
3) We demonstrated how the structural properties of online social networks can be leveraged for information diffusion using an efficient information diffusion algorithm.
4) We experimentally verified that the influence-index of the networks depends on particular structural properties, and these include the average clustering coefficient, average node degree, the network degree entropy and network normalized edge entropy among others.
5) Consequently, we proposed that to effectively share, or sell ideas on any online social networking site, the information diffusion index of that network should be known.

The paper is further organized into sections as follows:

The related work in the literature is studied in Section II. Section III discusses the background information with regards to online social networks and their topological properties. In Section IV, measurements are made on 12 selected online social networks. The mathematical framework for defining a node influence radius is outlined and applied in Section V. Conclusion and future work are shown in Section VI.

II. Related Work

Ref. [9] used click stream data to identify patterns in social network workloads and social interaction. Their studies show how click data can be used to characterize user behavior in social networks. Moreover, [10] analyzed the topological characteristics of 3 online social networks and established certain online social networking services encourage online activities that cannot be easily copied in a real life. [11] also studied the topological properties of two online Chinese social networks and observed that their topological properties of these networks possessed "small world" and scale-free together with other well known complex network properties. The evolution of structure within large online social networks is investigated by [12] and presented a series of measurements of two large real networks. They established that these networks can be segmented into three regions; *singletons*, *isolated communities*, and a *giant component*. They characterized the evolutions of each of these three regions.

Online social networks with either positive and negative links are studied by [13] and showed that the signs of links in the underlying social networks can be predicted with high accuracy by using generalized models. The findings, they said, shed more light on the theories of balance and status from the perspective of social psychology.

TABLE I: Notation Used in the Paper

Notation	Meaning
G	A Graph or network
V	The set of vertices of graph G
L	The set of edges or links in graph G
n	Number of nodes in graph G
m	Number of edges in graph G
$A_{(i,j)}$	The adjacency matrix of graph G
I	Node Influence Index
SW.Index	Small-World Index of graph G
$\langle k \rangle$	Average nodes degree of graph G
$\langle d \rangle$	Average distance between nodes in graph G
$\langle cc \rangle$	Average nodes clustering coefficient of graph G
γ	Power-law exponent of graph G
$\langle K \rangle$	Average nodes' neighbor degree
σ	Standard deviation of network degrees distribution
H_{NDE}	Node Degree Entropy of graph G
Gini.Value	The gini coefficient of graph G
$AC_{deg}(G)$	Degree Assortativity Coefficient
$CC(G)$	Clustering Coefficient of graph G
$Diam(G)$	The Diameter of graph G
$C_{c(size)}(G)$	Size of Connected Component

Furthermore, [14] developed an unsupervised model to estimate relationship strength from interaction activity and user similarity. They evaluated their model on Facebook and LinkedIn data.

Wang *et al*. [15] proposed influence maximization based on the initial seed user problem by assuming that each user needs a cost to accepting a particular information and outlined several contributions. Subsequently, [16] further proposed a PageRank-like scheme called PRDiscount, for selecting initial seeds based on a heuristic approach for diffusion maximization in social networks. They showed that the PRDiscount has an advantage over the DegreeDiscount and has achieved a comparable level of performance with that of greedy algorithms. Moreover, [4] addressed the problem of influence maximization in complex social networks by using the concepts of Independent Cascade Model (ICD). The proposed an efficient method of obtaining an approximate solution to the influence maximization problem for the case where the propagation probabilities through the network link are small. The effectiveness of their model was tested on real data. A High-PageRank heuristic algorithm for searching for initial seeds for influence maximization within a small portion of the nodes contain high-PageRank nodes is proposed by [8]. Their algorithm is said to scale to reduce search time when compared with classical algorithms and scale favorably without losing influence. Baagyere *et al.* [17] characterized several complex network properties and applied these on epidemic modeling and reported impressive findings.

The major difference between the work of these researchers and our approach is that we focused just characterizing the topological properties of some few

networks, but several online social from different countries and then further demonstrated how their underlying structural properties can be exploited for information dissemination. We proposed a computationally efficient algorithm for studying influence maximization on 12 selected online social networks based on data-driven approach and not on heuristics or simulations. Also, we came out with a baseline on how to select online social networks for the purpose of maximizing influence. Our findings, therefore, complement in a greater extend the ongoing efforts of getting a better way of studying nodes influence maximization.

III. Background

An Online Social network (OSN) is defined as a web-based system where the individual has the following features:

- an individual user is an actor or a node who has privacy setting as either public or semi-public
- where a user can create both explicit or implicit links among themselves or to a content item, and
- a user can transverse these social connections to some extent by looking into the profiles, friends or content items.

This definition is consistent with that used in previous studies [18]. These social networks could be a platform for friends to share their interest, photos, posts, news, and so on with each other [19]. Users and their interrelationships are the principal components of online social networks.

A. A Model of the Social Relationship as a Network

The social relationships between individuals can be formalized as a network using the concept of *graph theory*. Thus an online social network can be represented as $G=(V,L)$, where *V* denotes the social entities and *L* is the interconnections among these entities. This concept is used to modeled 12 online social networks and studied so as to understand their underlying structures and how these can be leveraged in information dissemination.

1) Topological properties of online social networks

At the very surface of online social networks, they look complex to understand their formation and, therefore, their underlying structure. Studying their topological properties help to understand quantitatively the interrelations among them. Methods and concepts developed under a field of mathematics called *graph theory* and combinatorics are employed to study these topological properties.

a) The adjacency matrix

The foremost of these tools is the adjacency matrix and is defined as follows:

Let $G=(V,L)$ is a simple graph.

The adjacency matrix A of G is

$$A(i,j)=\begin{cases}1 \text{ if there is an edge between } i \text{ and } j\\ 0 \text{ otherwise}\end{cases}$$

For instance, if there is a social relationship between i and j then there is a chance of an opinion to be formed or spread among i and j.

Thus, the matrix A contains in it the connection relationships within the online social networks under study.

b) Node degree

Another important and common way to characterize online social network is by the node degree. This is the number of social relationships an individual has, and it is denoted as k in this paper.

The node degree can be written in terms of the adjacency matrix A as:

$$k_i=\sum_{i=1}^{N}A_{ij}$$

The average node degree is therefore defined as:

$$\langle k\rangle \overset{def}{=}\frac{1}{N}\sum_{i=1}^{N}d_i$$

The node degree distribution, p_k, denotes the probability that a randomly selected node has a *k* number of links (degree).

It captures key interesting concepts such hubs and it plays a key role in calculating many network properties.

For instance, the average degree can also be obtained in terms of the degree distribution as:

$$\langle k\rangle \overset{def}{=}\sum_{k=0}^{\infty}kp_k$$

Networks with a significant number of hubs are known to have an effect on information flow and robustness of the network. The node degree is, therefore, important in characterizing networks. In particular, the average node degree can be seen as the average contacts per individual within a given online social network.

c) Node degree distribution

The node degree is commonly used to generate a network's degree distribution $Pr(k)$.

The $Pr(k)$ is the probability that a randomly selected node has a *k* degree. The degree distribution describes how the links in the network are distributed.

d) Power-law exponent

The power-law exponent value has an association with the network node degree distribution. The number of nodes with degree k is proportional to $k^{-\gamma}$, where γ is the power-law exponent.

2) Degree entropy

Let $d(k)$ be the degree of a node *k* in network $G=(V,L)$. The degree entropy is defined as:

$$H_{NDE}=\sum_{n=1}^{kmax}-\frac{d(k)}{n}\ln\frac{d(k)}{n}$$

a) Lorenz curve

An alternative method of showing the node degree distribution a network is by the use of the Lorenz curve [20].

The Lorenz curve is a tool used in microeconomics to visualize how their resources are distributed among

individuals within a given society. In an undirected network $G=(V, L)$, where each edge is attached to at least two nodes, the Lorenz curve can be used to visualize the distributions of edges among nodes in a given network and this distribution, unlike the power law degree distribution, is independent of the underlying network degree nature. The *Gini coefficient* that shows the distribution pattern has a value of 0 for fair distribution of edges among nodes and a value of 1 for single node dominance. Another value of interest as used in [20] is the *balanced inequality ratio* which is denoted as $Q(X\%, Y\%)$. It shows that X% nodes with the smallest degree accounts for Y% of edges within the network under study or could be viewed as Y% of rich nodes own X% of edges.

b) Graph clustering coefficient

Graph clustering can be seen in two perspectives; local and global clustering.

As introduced by Watts and Strogatz [21], the local clustering coefficient of a graph G is defined as follow:

Assuming that graph G is simple, connected and undirected. Let vertex $v \in V(G)$ with neighbor set $N(v)$. Let $n_v = |N(v)|$ and $m_v = |L(G[N(v)])|$ be the number of edges in the subgraph induced by $N(v)$. The local clustering coefficient of a given vertex v with degree $\delta(v)$ in G $cc(v)$ is then defined as [22]:

$$cc(v) \overset{def}{=} \begin{cases} \dfrac{m_{(v)}}{\binom{n_v}{2}} = \dfrac{2m_v}{n_v(n_v-1)} & \text{if } \delta(v) > 1 \\ \text{undefine otherwise} \end{cases}$$

The clustering coefficient of the entire graph denoted as $CC(G)$ is the average over all the $cc(v)$ in graph G. This can also be expressed mathematically as follow:

Consider a simple, connected and undirected graph G. Let V^* denotes the set of vertices $\{v \in V(G) \mid \delta(v) > 1\}$. The clustering coefficient $CC(G)$ of graph G is therefore defined as[22]:

$$CC(G) \overset{def}{=} \frac{1}{V^*} \sum_{v \in V^*} cc(v)$$

The local clustering of a graph tells to what extent the friends of an individual, v are friends among themselves or to what extent are friends on a social network who are adjacent to v are also adjacent to each other. The importance of clustering in network science with regards to information flow is discussed in Section V.

B. Small World Properties

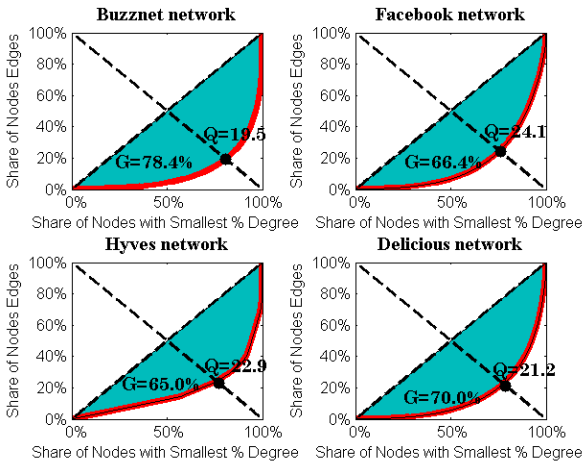
The distance between nodes within a network has relevance to understanding the small world effect. The basic idea of the "small world" network originated from the American sociologist Stanley Milgram from his seminal study on social networks [23] and was later modeled by [21].

The average path length or diameter of these networks is known to depend logarithmically on the network size. Hence, "small" means $\langle d \rangle$ is proportional to $\log N$ rather than N or some powers of N. The clustering coefficient of these networks is small when compared to a random network.

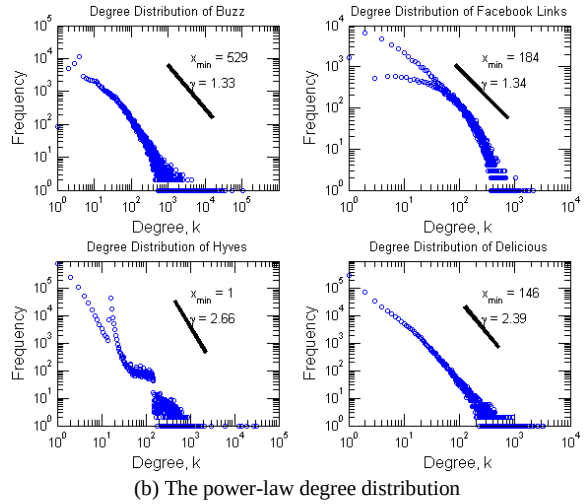
IV. Online Social Network Data

The data used in this paper consist of the friendship relationships among users and though some are directed networks, all are treated as undirected and, therefore, their total degrees are the sum of both in- and out-degrees, thus ignoring their edge directions. For it has been established that online social networks in- and out-degree distributions are similar [19].

These networks include the following: *Durban, Flickr, Flixster, Hyves, Livemocha, Buzznet, Delicious and Digg networks*. These online social networks are obtained from [24] and they span different social network domains and are used in different countries for sharing information. Wiki-Vote and Epinions social networks, *Facebook friendship and Facebook wall post networks* are obtained from [25]. The main reasons for the usage of these network data are that they differ widely in their sizes, countries of origin andtheir purposes and thus can provide a wide sample space for good network analysis without or with little biases. Also, these networks primary objective is for the spreading of opinions, ideas, habits and products and are of a major interest to sociologists, marketers and network scientists.

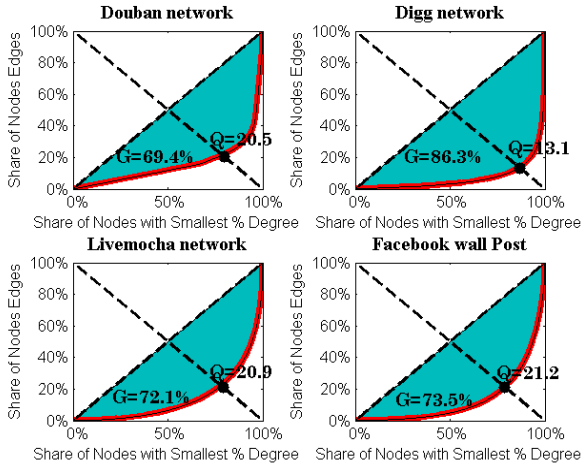


(a) The Lorenz curves of edges distribution

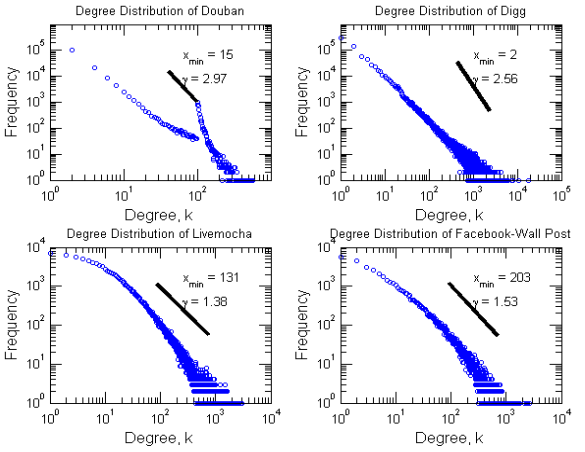


(b) The power-law degree distribution

Fig. 1. Lorenz curve and power law distribution of 4 social networks.



(a) Lorenz curves of edges distribution



(b) Power-law degree distribution

Fig. 2. Lorenz curve and power law distribution of 4 social networks.

A. High-Level Statistical Analysis of Online Social Network Data

The network statistics of the 12 online social networks are shown in Table II. The degree distribution of all the online social networks studied in this paper is characterized by the power-law exponents estimated using the maximum likelihood fitting algorithm by [26] and are shown in Figs. 1(b) and 2(b). Most of them are within the scale-free networks region of $1 \le \gamma \le 3$.

1) Small-world index

The small-world index is a measure that quantifies howgiven network is defined in terms of its clustering coefficient and average distance as compared to that of a random network with the same number of nodes and links. Higher values show that the given network has a high clustering coefficient and high average distance which is a key characteristic of small-world networks. Almost all the networks understudied showed significant small-world properties, some in the order of 4, signifying that online social networks are small-world networks. These statistics are shown in Table III.

What this means is that the information within these networks is easily circulated among the users. This thus further motivates the need to exploit these networks for information dissemination. Another property of the small-world networks is that their average distances depend logarithmically on the network size and these online social networks confirm this property as the $\log\langle k\rangle / \log N$ ratio is $\approx \langle d\rangle$, shown in Table IV.

TABLE II: HIGH LEVEL STATISTICS OF 12 ONLINE SOCIAL NETWORKS

Network	n	m	$\langle k\rangle$	AC_{deg}	C_{csize}	σ	*Gini. value*
FB Wall	46952	876993	37.36	0.465	0.94	86.88	73.5%
FB Links	60856	690071	34.45	0.172	0.99	57.86	66.4%
Delicious	536408	1385843	2.63	-0.068	1.00	22.77	70.0%
Buzznet	101163	4284534	25.11	-0.075	1.00	563.98	78.4%
Flixster	2523386	9197337	4.01	-0.259	1.00	44.31	80.4%
Hyves	1402673	2777419	3.13	-0.023	1.00	45.30	65.0%
Digg	771231	7261524	31.92	-0.070	1.00	114.22	86.3%
Flickr	80513	5899882	31.92	0.072	1.00	294.33	71.0%
Douban	154908	654188	8.45	-0.180	1.00	23.49	69.4%
Livemocha	104103	2196188	20.25	-0.147	1.00	109.80	72.1%
Wiki-votes	7115	100762	29.15	-0.068	0.99	60.39	75.2%
Soc-Epinions	75879	508837	13.41	-0.011	1.00	52.67	81.4%

2) Gini coefficient

The *Gini coefficient* values that measure how the edges in the online social networks are distributed together with their *balance common ratios* are shown by the Lorenz curves in Fig. 1(a) and Fig. 2(a) for 8 online social networks. All the networks have comparatively very high *Gini coefficient* values revealing the unfair distribution of the edges between the nodes in these networks. It is further revealed that in all the networks, very few nodes own most of the edges. For instance, in the *Durban network*, about 21.2% of rich nodes own almost 78.8% of edges as shown in Fig. 2(a). The unfair distribution of edges among the nodes is further emphasized by the high standard deviation σ values shown in Table II that characterized by these networks.

The following statistics are also confirmed on the online social networks studied:

1) All the networks are connected as shown by their connected component size C_{csize} \$75% of the networks have negative degree assortativity, thus, there is a high tendency of low degree nodes connecting to high degree nodes and vice versa.
2) The networks have a high average neighbor degree coupled with high average node's degree.
3) The networks are very clustered taking into consideration their average clustering coefficients.
4) Each of the networks has a substantial number of nodes and links.

These effects of these properties on information

dissemination are examined in Section V.

V. Node Influence Radius

Influence maximization of nodes within networks has been of research interest in recent times due to its importance in viral marketing, political campaigning, and social policing, among others. The ability of an individual to influence another individual within a geographical location depends on certain key factors which generally are not easily obtainable. As such, researchers normally resort to simulation methods to study the influence of individuals within a given radius. We propose a mathematical framework of influence maximization using 12 online social network data.

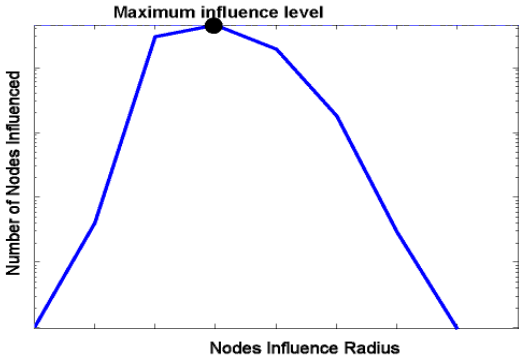


Fig. 3 The general influence maximization curve.

A. The Mathematical Framework of Influence Maximization

The influence a given node $k \in G(V, L)$ has on its neighbors is the set of nodes within a given radius R from node k. Thus, for a given node k, its influence radius is at most R around it. The distance from node k to any other node $B \in V$ is therefore the least number of edges in a path from node k to node $B \in V$. This is expressed as:

$$\omega(R,k) \overset{def}{=} \{B \in V : Influence\_dist(k,B) \leq R\}$$

Literature in marketing research strongly supports the fact that product sales cycles and product adoption by users follow an *S-curve* pattern. An *S-curve* pattern demonstrates that at the beginning, new product sales increase at a rapid rate, and then come to halt and begin to decline with time. The influence ability of nodes within a network also follows the *S-curve* pattern. Initially, nodes within the immediate neighborhood are influenced, which in turn influence those within a particular neighborhood and then finally declines with time due to the reduced number of nodes to be influenced. Fig. 3 demonstrates this scenario in which the influence of nodes in a network G increases with the nodes radius and halt after reaching its maximum influence level and then decreases sharply.

B. Network Influence-Index

Using Algorithm 1, the $network\ influence-index\ (I)$ proposed in this paper is a metric that measures the number of nodes that can be influenced within a given social network normalized by its volume (total number of edges). It is defined as the ratio of the log of the maximum number of nodes influenced to log of the total number of nodes in the network, expressed in the units of $nats$.

This is mathematically expressed as:

$$I = \frac{\log(max(Number\_of\_nodes\_influenced))}{\log m}$$

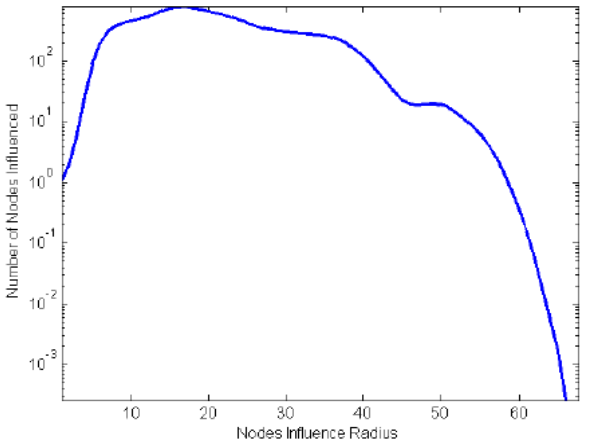
The $network\ influence-index\ I$ is then applied on the 12 online social networks and the result are shown in Table III together with other network measures. The maximum number of nodes influenced within each online social network using Algorithm 1 expressed in the form of an $Influence-index$ are shown in Table III. For a case study two of the influence maximization curves are shown in Fig. 4(a) and Fig. 4(b). The curves show that different online social networks have different radii at which to achieve maximum influence. For example, the *Hyve social network* though having more number of nodes and edges than that of the *douban social network* (see Table II) the *douban social network* is able to influence more nodes at a relatively short radius than that of the *hyve social network*. Thus, there are some inherent topological properties of social networks that can be exploited for influence maximazation and these are carfully examined in the next Section.

TABLE III: Camporison of 12 Online Social Network Properties with That of the Random Graph

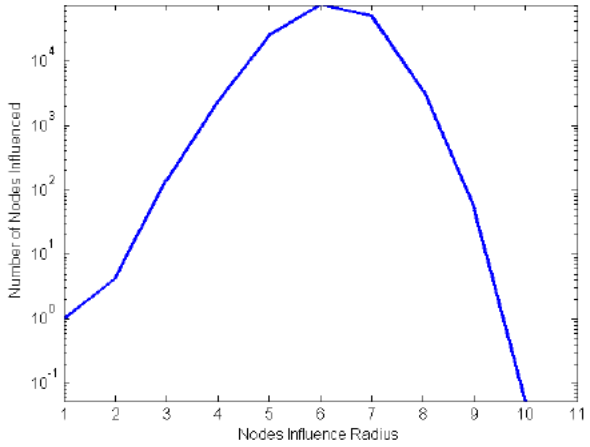
Network	$\langle d \rangle = A$	$diam$	$\langle cc \rangle = B$	$\langle d \rangle_{rnd} = C$	$\langle cc \rangle_{rnd} = D$	$E = \frac{B}{D}$	$F = \frac{A}{C}$	$SWIndex = \frac{E}{F}$	$\frac{logN}{\log\langle k \rangle}$
FB Wall	5.60	18	0.2154	3.30	$7.1E-4$	303.38	1.70	178.46	2.97
FB Links	4.32	15	0.2052	3.13	$7.7E-4$	266.49	1.38	193.11	3.53
Delicious	5.44	20	0.0011	9.80	$9.2E-6$	163.04	0.56	291.14	8.03
Buzznet	2.34	8	0.0259	2.93	$8.4E-4$	185.60	0.80	232.00	2.60
Flixster	4.87	16	0.0005	5.81	$3.8E-6$	447.37	0.84	532.58	7.42
Hyves	6.91	24	1.7×10^{-4}	11.74	$4.7E-6$	35.96	0.59	61.45	10.28
Digg	4.50	22	0.0005	6.24	$3.6E-5$	1202.78	0.72	1670.53	4.62
Flickr	2.90	6	0.0069	2.76	$1.8E-3$	17.67	1.05	16.83	2.26
Douban	5.17	9	0.0161	5.84	$4.5E-5$	357.78	0.89	398.46	5.60
Livemocha	3.21	6	0.0005	3.47	$2.4E-5$	195.02	0.93	383.71	8.02
Wiki-votes	3.25	7	0.0609	2.92	$4.0E-3$	15.22	1.11	13.71	2.63
Soc-Epinions	4.75	8	0.0898	5.00	$2.0E-4$	449.00	0.95	472.63	4.33

Algorithm 1 Compute Node Influence Radius

```
procedure N_InfluenceRadius(N ∈ G(V, L))
    Input: G(V, L)
    Output: Number of Nodes, K Influenced and the Radius, R
    for a given N ∈ G(V, L) do
        Get the number of nodes K ∈ G(V, L) such that N can be at a distance R
    end for
    Let Q be a queue
    for each k_i ∈ K do
        Q.enqueue(k_i)
        Label k_i as been influenced;
    end for
    while size(Q)≠ 0 and ShellSize, (d) < R do
        k_i ← Q.dequeue()
        ∀ neighbors Ω of k_i ∈ K
        if Ω is not labeled influenced then
            Q.enqueue(Ω)
            Label Ω as influenced
        end if
        Return K and Influence radius R
    end while
end procedure
```



(a) Hyves node Influence maximization graph



(b) Douban Nodes Influence Maximization

Fig. 4. Influence maximization of the hyves and douban social networks.

VI. Further Analysis

The impact of the underlying network structure of online social networks for information dissemination is studied using a correlation matrix shown in Table IV. The matrix takes into consideration very important and universal properties of complex network together with the information maximization abilities of these networks expressed as the network influence index. The plot help us to see the relationship between the *influence maximization index I* and the other networks structural properties and this is shown in Fig. 5.

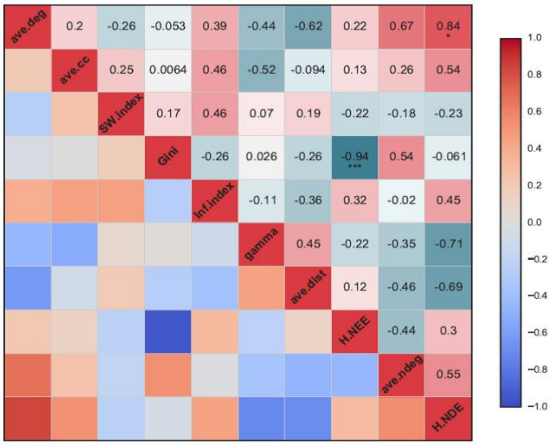


Fig. 5. The correlation plot of social network measures.

The correlation coefficient values of each variable are indicated within the plot. A value of $+1.0$ means strongly positively correlated, 0 no correlation and -1.0 strongly negatively correlated. Taking the *influence index I* as our reference parameter, it is seen that, it correlates fairly strongly with the network average degree ($ave.deg$), the network average clustering coefficient ($ave.cc$), the normalized edge entropy ($H.NEE$) and the node degree entropy ($H.NDE$). This informs that to be able to disseminate information effectively, it is very important to consider the underlying structure of the network by taking into consideration the average number of friends a particular network has coupled with that friend's position within the network. Networks that have their nodes within the core of the other nodes are considered to be a good conduit for information spreading than networks that have a high average node degree but the individual nodes are not situated within the center of other nodes. Furthermore, the normalized edge entropy and node degree entropy correlate positively with the influence-index. This supports the fact that the manner in which edges are distributed among nodes is crucial and has an impact on information dissemination.

The small-world index strongly correlates with the Gini coefficient value, confirming the fact that small-world networks are scale free networks characterize by few nodes with higher degrees.

Therefore, taking into consideration these 12 online social networks, the five most effective networks, based on their structural properties, for information dissemination are as follows:

1) Facebook friendship network
2) Facebook wall-post network
3) Wiki-Vote network
4) Buzznet network
5) Soc-Epinions network

The reason is that, these networks have a high average node degree coupled with high clustering coefficients, the attributes can be confirmed in Table III.

TABLE IV. MATRIX FOR THE CORRELATION PLOT

Network	$\langle k \rangle$	$\langle cc \rangle$	$SWIndex$	$Gini.Value$	I	γ	$\langle d \rangle$	H_{NDE}	$\langle K \rangle$	H_{NEE}
FB Wall	37.36	0.2154	178.46	73.5%	0.8366	1.42	5.60	0.8955	17.05	0.8955
FB Links	34.45	0.2052	193.11	66.4%	0.9176	1.40	4.32	0.9251	55.86	0.9314
Delicious	2.63	0.0011	291.14	70.0%	0.603	2.46	5.44	0.9037	160.38	0.8984
Buzznet	25.11	0.0259	232.00	78.4%	0.7576	2.00	2.34	0.8121	7380.52	0.8537
Flixster	4.01	0.0005	532.58	80.4%	0.6285	3.62	4.87	0.8452	343.95	0.8512
Hyves	3.13	1.7×10^{-4}	61.45	65.0%	0.3859	2.52	6.91	0.8377	2550.50	0.9060
Digg	31.92	0.0005	1670.53	86.3%	0.6330	1.49	4.50	0.8594	2107.0	0.8287
Flickr	31.92	0.0069	16.83	71.0%	0.6244	1.49	2.90	0.8594	750.08	0.9109
Douban	8.45	0.0161	398.46	69.4%	0.9385	2.47	5.17	0.8897	49.73	0.8897
Livemocha	20.25	0.0005	383.71	72.1%	0.6209	1.58	3.21	0.8566	217.22	0.9003
Wiki-votes	29.15	0.0609	13.72	75.2%	0.7447	1.53	3.25	0.8722	140.73	0.8722
Soc-Epinions	13.41	0.0898	472.63	81.4%	0.8432	1.89	4.75	0.8454	159.85	0.8454

VII. CONCLUSION AND FUTURE WORK

Information dissemination is a key reason for the creation of most online social networks. And though these networks in recent years have become the portal of information dissemination, rivaling the traditional web in term of traffic, their underlying topological structures are not well studied collectively in order to leverage their structural properties for information dissemination. The information dissemination models proposed in the literature are either based on simulation and not data driven or are computationally expensive.

This paper addressed these bottlenecks by studying several online social networks to collectively understand their underlying structures and thereby leveraged their properties for information dissemination. Additionally, a cost efficient algorithm which scaled $\boldsymbol{O}(N)$, where N is the number of nodes in the network, is proposed and implemented in order to study information dissemination on these networks. The results showed that online social networks have a unique underlying structure and can be effectively leveraged on information dissemination, and to this end, we proposed an $influence\ index$, a quantitative way of measuring the ability of a given network to disseminate information. This value is seen to correlate with the network average degree, average clustering degree, the node degree entropy and the network edge normalized entropy. Thus behind these complex online social networks, there is a well defined topology that encodes the social interactions among individuals and therefore can be leveraged in a cost effective way for information sharing among these individuals.

As future work, we look forward to applying the spectral properties of online social network data using statistical physics and applying these properties in studying dynamical systems. We intend to formulate a complex network indicator for predicting classes and domains of complex networks base on their spectra and topological properties.

REFERENCES

[1] Worldwide Social Network Ad Spending. [Online]. Available: http://www.emarketer.com/Corporate/Coverage#/

[2] H. Ward, "Book reviews — media virus! Hidden agendas in popular culture by douglas rushkoff," *Ed. Publ.*, vol. 128, no. 5, p. 25, 1995.

[3] J. Scott, "Social network analysis: Developments, advances, and prospects," *Soc. Netw. Anal. Min.*, vol. 1, no. 1, pp. 21–26, 2011.

[4] X. W. Zhu, W. Wu, Y. Bi, L. Wu, and Y. Jiang, "Better approximate algorithms for influence maximization," *J. Comb. Optim.*, vol. 30, pp. 97–108, 2015.

[5] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," *Time*, vol. 67, no. 1, p. 199, 2009.

[6] Q. Naik, S. Anil, and Yu, "Evolutionary influence maximization in viral marketing," *Recommendation and Search in Social Networks*, 2015, pp. 217–247.

[7] Q. Zhao, H. Lu, Z. G. B, and X. Ma, "A K-shell decomposition based algorithm for influence maximization," *Eng. Web Big Data Era Lect. Notes Comput. Sci.*, vol. 9114, pp. 269–283, 2015.

[8] D. Z.-L. Luo, W.-D. Cai, Y.-J. Li. "A pagerank-based heuristic algorithm for influence maximization in the social network," *Recent Progress in Data Engineering and Internet Technology*, 2012, pp. 485–490.

[9] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proc. the 9th ACM SIGCOMM Conference on Internet Measurement Conference, IMC '09*, 2009, p. 49.

[10] Y. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," pp. 835–844, 2007.

[11] F. Fu, L. Liu, and L. Wang, "Empirical analysis of online social networks in the age of Web 2.0," *Phys. A Stat. Mech. its Appl.*, vol. 387, no. 2–3, pp. 675–684, 2008.

[12] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," *Link Min. Model. Algorithms, Appl.*, pp. 337–357, 2010.

[13] J. Leskovec, D. Huttenlocher, and J. M. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. the 19th International Conference on World Wide Web*, 2010, pp. 641–650.

[14] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proc. the 19th International Conference on World Wide Web*, 2010, pp. 981–990.

[15] Y. Wang, W. J. Huang, L. Zong, T. J. Wang, and D. Q. Yang, "Influence maximization with limit cost in social network," *Sci. China Inf. Sci.*, vol. 56, no. 7, pp. 1–14, 2013.

[16] J. Wang *et al.*, "PRDiscount: A heuristic scheme of initial seeds selection for diffusion maximization in social networks," *Intelligent Computing Theory*, 2014, pp. 149–161.

[17] E. Y. Baagyere, Z. Qin, H. Xiong, and Z. Qin, "Characterization of complex networks for epidemics modeling," *Wirel. Pers. Commun.*, vol. 83, pp. 2835–2858, 2015.

[18] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *J. Comput. Commun.*, vol. 13, no. 1, pp. 210–230, 2007.

[19] A. E. Mislove, *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*, ProQuest, 2009.

[20] J. Kunegis and J. Preusse, "Fairness on the web: Alternatives to the power law," in *Proc. 3rd Annu. ACM Web Sci. Conf.*, 2012, pp. 175–184.

[21] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[22] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. Ninth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '03*, 2003, p. 137.

[23] S. Milgram, *Psychology Today*, May 1967, pp. 61-67.

[24] H. Zafarani and R. Liu, "Social computing data repository at ASU," *Sch. Comput. Informatics Decis. Syst. Eng. Arizona State Univ.*, 2009.

[25] A. Leskovec and J. Krevl, "SNAP datasets: Stanford large network dataset collection," 2014.
[26] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, p. 661, 2009.

Edward Yellakuor Baagyere received his BSc. degree (Hons) in computer science from the University for Development Studies (UDS), Tamale, Ghana in 2006, and MPhil. degree in computer engineering from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana in 2011. He is with the Faculty of Mathematical Science, UDS, where he teaches in the Department of Computer Science. Mr. Baagyere is currently a Ph.D candidate at the University of Electronic Science and Technology of China (UESTC). His research interests include social networks, information security, computer arithmetic, cryptography, residue number system and its applications.

Zhen Qin received the B.Sc. degree in communication engineering from UESTC in 2005, the M.Sc. degree in electronic engineering from Queen Mary University of London in 2007, and the M.Sc. and Ph.D. degrees in communication and information system from UESTC, in 2008 and 2012, respectively. Dr. Qin is currently a lecturer with the School of Information and Software Engineering UESTC. His current research interests include network measurement, wireless sensor networks, and mobile social networks.

Hu Xiong is an associate professor in the School of Information and Softwarering Engineering, UESTC. He received his Ph.D. degree from UESTC in 2009. His research interests include information security and cryptography.

Zhiguang Qin is the dean of the School of Information and Software Enigineering at University of Electronic Science and Technology of China (UESTC), where he is also a director of the Key Laboratory of New Computer Application Technology and Director of UESTC-IBM Technology Center. His research interests include computer networking, information security, cryptography, information management, intelligent traffic, electronic commerce, distribution, and middleware.