The Structural Properties of Online Social Networks and their Application Areas

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

***Abstract*—Online Social Networks (OSNs) are here to stay as they have exploded in popularity and currently rival the traditional Web in term of user traffic. The extreme popularity of these networks coupled with their rapid growth present a unique opportunity to researchers never before to study, understand, and leverage their properties in other fields of study.**

Several research efforts are in the literature on OSNs, however, the structural and statistical properties of these networks are yet to be leveraged on other areas of research. In this paper, a comprehensive study on OSNs is given in view of the structural properties together with their underlying mathematical framework. Open problems with their associated application areas are also enumerated of which answers to them would have a great impact on social network theories and their related application areas such as commerce, law enforcement, algorithm optimization and product recommendation.

***Index Terms*—Online Social Networks, Social network analysis, Structural Properties, Network Metrics**

I. Introduction

A SOCIAL NETWORK consists of the interaction between two or more individuals or social entities called "nodes". Sociologists consider these interactions as social tides and the nodes as actors.

The interaction between the social entities could be that of sexual relations, common interest, knowledge and beliefs exchange or friendship, etc. and the interaction often become complex in the process of time.

Social networks have various levels of interconnection/interaction, ranging from the level of friends to that of nations and these interconnections can be of importance with regards the flow of information among friends and that of goods and services among nations.

In order to analyze the social network structure that resulted from these interactions, the social system is modelled as a graph that consists of nodes and edges.

An Online Social Network (OSN) is then defined as a web-based system where

(a) an individual user is the actor or node who has privacy setting as either public or semi-public
(b) where a user can create both explicit or implicit links among themselves or to a content item, and
(c) 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 [1]. Online social networking sites are currently the platforms where users share their interest, photos, posts, news, and so on with each other [2]. Social network analysis is therefore the holistic application of network theories on OSNs in order to study the interrelationship between humans and other systems. The analysis of a site or an organization for example, can offer in the form of a picture the interdependency or links among nodes or companies and other social systems. The picture illustrates how members of these social systems or individuals in an organization are actually interconnected as oppose to the traditional "pyramidal organizational" models.

Social network analysis measures can be harnessed in order to gain more insights into the collective capacity of individuals or organizations by minimizing some of the negative effects that come along with these social network analysis at the level of human-computer interaction.

Social networks and their related benefits are in every facet of the human society. The followings are some of the related benefits of social network analysis. Social network analysis can:

- Identify individuals, companies, and components who play central roles in a system.
- Be harnessed to strengthen the efficiency and effectiveness of existing communication channels in an organization.
- Be leveraged for peer support in an organization or company
- Be used for innovation and learning.
- Be used to discern or predict information breakdown, to identify bottlenecks, structural holes and isolated units in an organization or a system.
- Be used to refine business strategy or develop an effective Organogram for an organization

The remaining parts of the paper are organized into sections as follows. Section II gives a comprehensive history on online social networks and the various network models. In Section III, an overview on OSN properties is outlined. The theoretical metrics for OSNs analysis are discussed in Section IV. Section V gives related works on the paper. The paper is concluded in VI with open problems on OSNs and their potential application areas.

II. Social Network Sites and network models

In this section we outline the history behind OSN sites and the various classes of networks in the literature.

Manuscript received September 22, 2015; revised December 19, 2015. 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).

E.Y. Baagyere is a Ph.D candidate in the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu. Email: (ybaagyere@uds.edu.gh).

Z. Qin, H. Xiong and Q. Zhiguang are with the School of Information and Software Engineering, University of Electronic and Technology of China, Chengdu. Emails:(qinzhen@uestc.edu.cn, xionghu.uestc@gmail.com, qinzg@uestc.edu.cn)

TABLE I
NETWORK PROPERTIES FOR THE FOUR DIFFERENT NETWORKS

Network Properties	Reference Network (cit-HepTh)	Erdös and Réyni Network model	Power-Law Network (Scale free Network model)	Small World (Watts-Strogatz Network) model
Number of nodes	27,770	27,770	27,770	27,770
Number of edges	352,807	352,807	352,807	361,010
Average Node Degree	25.41	25.41	25.41	26
Network Diameter	37	5	5	6
Network Radius	0	4	4	5
Average path length	8.460	3.524	3.239	4.452
Av. clustering coeff.	0.1195	0.0009	0.0041	0.5277
Network Density	0.00046	0.00092	0.00092	0.00094
Assortativity coeff.	0.00172	0.00147	-0.00399	0.00065
Modularity coeff.	0.54639	0.08447	0.06054	0.87216

A. Brief History On Social Network Sites

The history behind OSNs sites date back to the 90s but became very popular in the mid 2000 partly due to the advent of the Web 2.0 and the mobility of internet enabled devices. `danah m. b.` and `Ellison N. B.` [1] in their paper gave a comprehensive outline on the history behind OSNs systems or sites and were able to reveal their functionalities, rise and fall. Figure 1 shows the launched dates of some major social network sites between 1997 and 2012. The list enumerated in this history is however not meant to be exhaustive, as new sites are being created by the day. A more complete and up -to - date list of the notable OSN sites can be found in the Wikipedia [2].

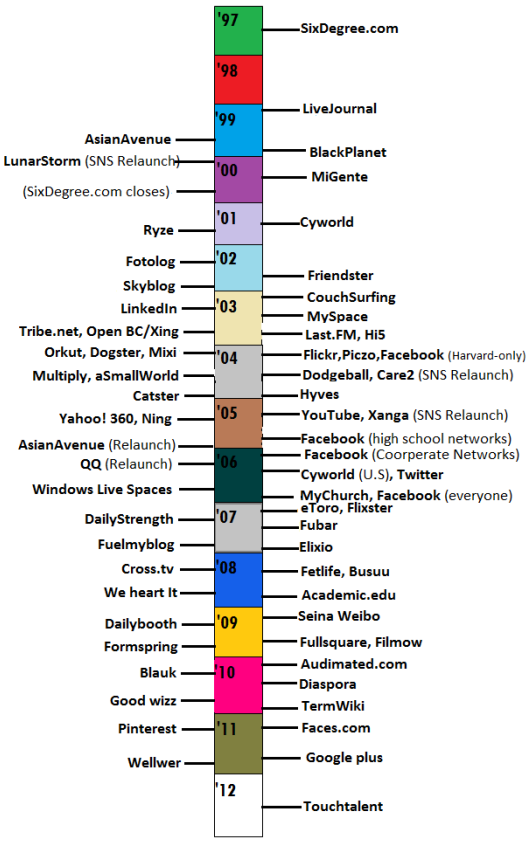

Fig. 1. Some major OSNs and their year of establishment

B. Type of Networks

The major network models in the literature and their theoretical properties are outlined in Table II. In order to confirm some of these theoretical properties, the paper citation network of `Arxiv High Energy Theory category` (cit-HepTh) [3] is used as a reference network to simulate these network models.

The network consists of 27,770 papers (nodes) and 352,807 citations (edges). It is a directed network of the form of paper i citing paper j. The results of the simulations are shown in Table I confirming these theoretical properties.

One of the important characteristics of networks is their degree distribution. Figure 2 shows the degree distribution of the four networks namely the `cit-HepTh`, the `Erdös and Réyni` Network model, `power-law/scale free` network model, and the `Watts-Strogatz small world` network model as shown in Table I. The true or reference network stands for the paper citation network of `Arxiv High Energy Theory category`, ER stands for the `Erdös and Réyni random network model`, W-S stands for the `Watts-Strogatz small world network model`. From the Figure, the difference between these networks are highlighted. The scale free characteristic of the True network and that of the modeled scale free network

TABLE II
NETWORK MODELS AND THEIR THEORETICAL PROPERTIES

Network Type	Degree Distribution (p_k)	Average path length (l) or Diameter (d)	Clustering Coefficient (C)	Main Characteristics
Random Network (The Erdös and Réyni model [4])	$p_k = e^{-c}\frac{c^k}{k!}$ for large N. Average Degree: $\langle k\rangle = p(N-1) \cong pN$ or $\langle k\rangle = \frac{2m}{N}$ (for undirected network).	$l_{rand} \sim \frac{lnN}{\langle k\rangle}$, $d = \frac{logn}{logk}$	$C = p = \frac{\langle k\rangle}{N} \ll 1$	Have very small clustering coefficient as compare to that of regular lattice. Have short geodesic distances. For large value of N, the degree distribution is poisson with mean degree value of $p(\langle k\rangle)$
Power-Law Network (Barabási and Albert model[6])	$p_k \propto k^{-\alpha}$ or $p_k = \frac{2m}{k^3}$	$l = \frac{lnN}{lnNlnN}$	$C \sim N^{-0.75}$ $C(k) = k^{-1}$	These networks differentiate continuously by the removal or addition of nodes. These networks always have a single connected component. New nodes attached preferentially to well connected nodes Examples are Internet topologies [7], the Web [8], [9], social networks [10], neural networks [11].
Scale free Network	$p_k \propto k^{-\gamma}$, $2 < \gamma < 3$	$d \sim lnNlnN$	Clustering is coefficient larger than random networks	Degree distribution is power law. The clustering coefficient decreases as the node degree increases. This distribution also follow power law. Have a small average path length as compare to a highly ordered network as a lattice graph for $2 < \gamma < 3$
Small-World Network (Watts and Strogatz model)	$p_k = e^{\langle -k\rangle p}\frac{(\langle k\rangle p)^k}{(k-\langle k\rangle)!}$ where $\langle k\rangle$ is node average degree and p is the probability of an edge connecting to a node. Thus p_k is similar to that of the ER random graph model	Have small average distance, Diameter scales logarithmically with the size of network, thus $d \sim lnN$	Have large clustering coefficient	This is a homogenous network where all nodes have approximately the same number of edges. Similar to the ER random graph model. The name "Small-world" is due to the small diameter [5]

are very pronounced. These networks have few nodes with extremely high nodes degree connectivity as opposed to the ER and W-S network models whose degree distribution are Poisson in nature, clustered around the mean value.

Another key feature of networks is their path length distribution. "Small world" networks are known to have smaller path lengths and high clustering coefficient when compared to random networks and are therefore very effective in information transmission. In Figure 3 the path length distributions of the four networks are compared with each other. It can be seen that the "small world" network, the scale free network and the Erdös and Réyni network models have a small path length for comparatively the same number of nodes and edges with that of the Cit-HepTh networks thereby justifying their theoretical properties in the Table II.

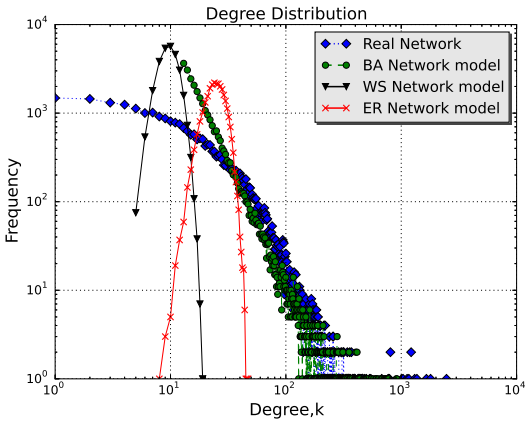

Fig. 2. Nodes Degree Distribution for the four networks

III. Overview of Online Social Network Properties

Social networks are arguably the first classes of networks ever studied rigorously with their long history dating back

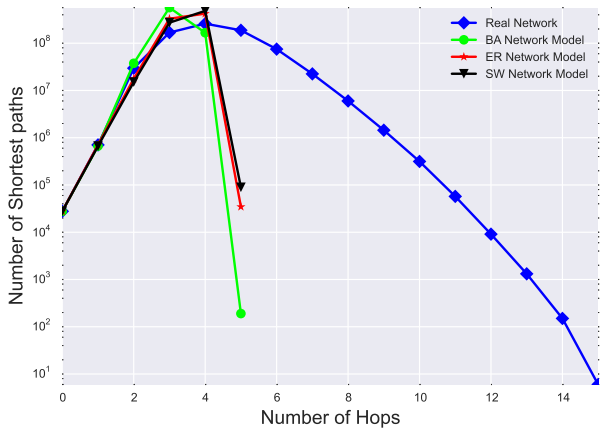

Fig. 3. Network Paths Length Distribution

to at least the late 1930's. Sociologists have arguably been the first to have used the notion and concept of network in studying human behavior. Barnes, J. A., among others [17] is credited with coining the name social networks. Social network theory study how people are connected to one another and the way these connections affect their belief and behavior as individuals, group, or organization.

As online social networks are gaining popularity, couple with the availability of computer resources, computer scientists, mathematicians, among others are studying these properties in a large scale in order to verify various sociological hypothesis. Due to this, mathematical frameworks are formulated to enlighten our understanding on the structural properties of these networks, and how their structural properties can be used in several areas of applications.

In this section, we outline the various research efforts that have been done on OSNs categarized under:

(a) `information spreading`
(b) `user attribute prediction` and
(c) `community detection`.

A. On Information Spreading

Social networks can be portrayed as a structure that has some inherent properties that can be leveraged for the dissemination of information. The concept of information dissemination or cascading in Social Networks has thus received good research efforts in recent times. The first significant empirical study on social networking is the hypothesis raised by Stanley Milgram[18] that two people on the surface of the earth are separated by at most 6 degrees of separation. This hypothesis termed as the "small - world problem", was tested in the context of MSN messenger data and it was shown that on average, there is 6.6 degree of separation between two Americans [19]. This study supports the notion that the world is "small" when you think of it as how few hops of friends is it to take you to almost every other person within it.

Social networks have an essential role to play when it comes to information flow. One of such role is to acts a conduit or bridge between two or more nodes that are connected locally or globally. The fundamental roles social networks play in information flow and how these roles shape the evolution of the network have been studied by Granovetter [20]. For example, it is shown by Granovetter that social network structure can be grouped into "weak" and "strong" ties which have consequences to the flow of information within the network.

Meeyoung *et al.* conducted a research on Twitter in 2009 [21] and argued that Twitter social network comprises of three types of information spreaders and these information spreaders play different distinct roles in the spreading of information. And their relative importance in each role they play is also depended on the kind of information that is available.

A large-scale traces of information dissemination in the Flickr social network is done by [22]. The research showed how widely and fast information cascade within social network, and the importance of word-of-mouth exchanges between users.

The factors that influence business brand pages dissemination on Facebook are studied by Chua A. Y. K. *et al.* [23]. They discovered three factors that are related to users' attention toward businesses brands post disemination. These include *incentives, vividness and interactivity.*

Kwak H. *et al.* [24] investigated on the topological properties of Twitter and its influence as a medium of information flow. Their analysis showed that the topological properties of the Twitter network in term of its follower-following does not follow a power-law distribution (which is a deviation from the nature of most social networks), has a short effective diameter, and low reciprocity.

The production, flow, and consumption of information in Twitter in the context of the microblogging service, is studied by [25]. They analysis the data by exploiting the "Lists" feature within Twitter and thereby were able to distinguish two kinds of users; elite users and other formal organizations, termed ordinary users. By this classification, they found that about 50% of URLs are generated by just 20K elite users, and that there is a significant level of homophile within categories.

B. On User Attribute Prediction

User attributes are common features of OSNs and are used in prediction link formation, cliques formation, and user behavior, among others. For example, link prediction has relevance in determining important future connections that are likely to take place within social network. These links can provide a blueprint for predicting potential future interconnections that might emerge or the probable missing relations in the social network. Link prediction base on user attributes is also useful in an adversary and terrorist networks analysis as the full linkages of these networks are not always known aprior.

Homophily among users is studied by [26] using LiveJournal data. Their investigation showed that common interest is a key factor in determining the likelihood of two users being friends and vice versa two users have a high chance of sharing common interest if they are friends.

Mislove A. *et al.* [27] investigated on how user attributes in combination with their social network graph properties can be used to predict the attributes of another user in the network. Using fine-grain data from Facebook, it was

found out that users with common attributes form a dense community and are likely to be friends. With only a small attributes percentage as 20%, the research showed that certain user information can easily be inferred with high accuracy.

Student/Faculty relationships is investigated by [28] using Facebook dataset in order to understand how contacts on Facebook were influencing student perceptions of Faculty. Their result showed that contacts on Facebook have no impact on students ratings of Professors.

C. On Community Detection

The main challenge associated with community detection is strongly related to the challenge of finding users or nodes with similar structural attributes. Networks or sub networks that have structurally related or similar attributes are often known as as communities.

Community detection in social networks to discover social groupings, and their pyramidal organization by using only the data embedded in the social network structure has received some research efforts in recent times.

One of such research efforts consists of searching for community structure within the working groups of a government agency due to Weiss and Jacobson [30]. They observed a matrix of working relationships between members of the agency. They further observed that by removing members working with people of different groups that acts as boundary spanners (links) between them can lead to the formation of various communities.

Collaborative filtration for communities is also investigated by [31] using Orkut data set based on two sets of algorithms; the `association rule mining` (ARM), which finds association between users and their communities and the `latent Dirichlet allocation` (LDA), that finds the co-occurrence of users and their communities.

Hansen *et al.* [32] worked on personal and organization Email data and was able to identify some interesting metrics within the email social network. For example, the research revealed individuals that are important and play unique roles in both the personal and organization levels. Also subgroups that show collaborative activities among individuals and departments were exposed in analyzing the email data.

The karate club study by the anthropologist Zackary, is another notable case of community detection and has been used as a benchmark in evaluating community detection algorithms in social networks [33]. Also, the influence of cultural identity on the development of the network community is investigated by [34] using LiveJournal user data.

Term clustering is a community structure discovering technique used in textual data analysis and it helps in understanding the similarity between words within the text. Yang, J. *et al.* [35] proposed words clustering method using the relative contribution of each word. Their method achieved comparative gains when compared with other text clustering techniques.

[46] proposed a Gauss chaotic map particle swarm optimization method for studying clustering. They experimented their clustering method on six data sets and compared the results to 8 other clustering methods and indicated that their method significantly perform better than other clustering methods.

IV. THEORETICAL METRICS OF ONLINE SOCIAL NETWORKS

In this section, we examine the mathematical framework together with some concepts from graph theory on which metrics that are used to measure the structural properties of social networks are formulated. These frameworks and concepts form the basis on which Online Social Networks are studied.

A. Some Concepts from Graph Theory

Online social network can be viewed as a graph $G = (V, E)$, where the the set of vertices V corresponds to number of users and the set of edges E corresponds to the number of social relations among the users. This social relationship can either be a directed one, meaning each relationship is sourced at one node or user and terminated at another node or user, or an undirected one, in which relationship between two notes or users has no source or destination, etc.

A user or node degree is the number of social links that are incident on the user or node within the social network. The degree of a node or user is categorized into incoming links, and outgoing links, technically called in-degree and out-degree respectively.

These concepts from graph theory are used modeled user-user relationship into a network with behavior, that exhibits certain properties. The mathematical definitions or formulations that quantify or define these properties are termed as `network metrics`.

In the following, the metrics that are commonly used to measure some of these network properties are reviewed.

B. Network Metrics

Theories and algorithms for calculating important structural properties of social networks have been developed by mathematicians, computer scientists, social scientist, and physicists. These structural properties give a quantitative measure of these network properties that help in doing a systematic study of the social network.

The positions of nodes and their relative importance, how information flow over these nodes, and the nature of the social network over time can be analyzed by using these theories and algorithms. Some of the metrics that are used to quantify these structural properties are global in scope while others are local in scope. The global network metrics take into consideration the entire network. In the following, the various metrics used in network analysis are outlined, narrowed down to only the frequently used ones.

1) Aggregate Network Metrics: Aggregate network metric is a global network metric that is used to describe the entire network system. One of such metrics is the network density. The network density is used to describe how well connected a given network is. This social network metric can be viewed as the ratio of the number of social relationships observed in the network to the total number of possible relationships that could exist among nodes. It is therefore the proportion of ties within the network. For networks that are simple and undirected, the density is expressed mathematically as:

$$D = \frac{2|E|}{|V|(|V| - 1)} \tag{1}$$

Whereas directed simple graphs density is defined as the proportion of arcs present in the digraph. It is expressed as the number of arcs, E divided by all possible number of arcs within the digraph.

$$D = \frac{|E|}{|V|(|V|-1)} \tag{2}$$

The density of a graph quantitatively captures important network properties such as cohesion, solidarity, and network membership. Its value ranges from a minimum of 0, if no arcs are present, to a maximum of 1, if all arcs are present thereby forming a mutual relationship. In Table I the densities of the four four networks are calculated with the cit-HepTh network having the least density value of 0.00046.

2) Node-Specific Networks Metrics: Node specific network metrics are associated with the relative positions of individual nodes within the network. The centrality measure, which describes how an individual node is in the "center" of a network, is the key metric when dealing with node-specific network metrics. It measures the extend to which nodes have central influence or importance within the network.

The `Centrality` and `Prestige` of the *Florentine Families* network is used to explain some of the centrality measures discussed in this Section. As seen in this network, the *Medici family* had a position of centrality in the social community through marriage which was harnessed for communication and other deals. This network is show in Figure 4.

(i) **Degree Centrality:** Consistent to previous work, the degree centrality of a node is calculated as the total number of neighbors that are linked to that node. The degree centrality in some context can be regarded as a popularity index, but in an unrefined way, without considering who one is connected to who. The degree centrality measure is categorized into two types for directed networks; `in-degree` and `out-degree`. The in-degree measures the number of links that points toward or incident at a given node in a network. However, the out-degree metric counts the number of outgoing links from a given node.
The degree centrality $C_D(v)$ of vertex v within a simple graph $G = (V, E)$ that has n vertices is expressed as [1]:

$$C_D(v) = \frac{deg(v)}{n-1} \tag{3}$$

For example, from Figure 4, the most central node in terms of node's degree within the Medici family marriage network is*Medici* followed by *Guadagni*. This means that Medici and Guadagni are well connected and famous individuals that have a high level of influence in their immediate neighborhood. Because information of any kind must pass through them to the rest of the social community.

(ii) **Betweenness Centralities:** Paths are crucial in the study and analysis of social networks in particular. How far apart are two people (nodes) within a network is a everyday question in network analysis. Betweenness centrality B_C of an edge e within a network is defined as the ratio of the total number of shortest paths between all pairs of vertices in the graph that cross e to the total number of possible shortest paths between the two nodes that include e. This definition is consistent that proposed by Girvan and Newman [36]. The betweenness centrality for an edge can therefore be expressed mathematically as

$$B_C = \sum_{u \in V, v \in V} \frac{\phi_e(u,v)}{\phi(u,v)} \tag{4}$$

The node with highest betweenness centrality within the `Florentine families` network is *Medici* followed by *Guadagni* again. The betweenness centrality of an edge within a network can be viewed as a metric that quantifies the importance of an edge in the network, because edges with a higher betweenness centrality fall on more shortest paths, and are therefore more important for the structure of the network. Thus *Medici* and *Guadagni* are very critical individuals in the transmission of information within the network. For their removal can disrupt the flow of information.
There are variants of betweenness centrality of an edge in literature. Some of the them are:

- **Closeness Centrality:** This centrality measure is quite different from the other network metrics. It captures the average distance between a given node and every other node in the network. For examples, if nodes are assumed to be channels through which information are relayed, or influence can be exerted from, then a low closeness centrality of a node shows how fast it will take information to flow directly from that node to all other nodes that are connected to it. High closeness centrality scores are associated with nodes at the outer perimeter of the network. The high scores indicate how long it may take information to flow to the core of the network. Closeness centrality can then be seen as the mean shortest path between a given node v and all other nodes that are within it reach.
  Mathematically, Closeness Centrality C_C is expressed as:

$$C_C = \sum_{t \in V \setminus v} \frac{d_G(v,t)}{n-1} \tag{5}$$

  Where n ≥2 is the size of the network's 'connected component' V of which t is reachable from v [37]. Using the Florentine families network shown in Figure 4 as an example, the two nodes with the smaller closeness centrality is *Rodlfi* and *Medici* and therefore could aid in the quick spread of information within the marriage network.
- **Eigenvector Centrality:** This measure quantifies the relative importance of a node in a network. High eigenvector centrality score of a node shows how important that node is in respect to its contribution to the score of a particular node under consideration. Thus if a node with low degree has high eigenvector centrality, then it means that the low degree node were themselves connected to other "well connected " nodes in the network.
- **PageRank Centrality:** PageRank, the trade name given by the Google web Search Cooperation is used in social network to measure the importance of a node. The PageRank centrality a node derives from its neighbors is proportional to their centrality

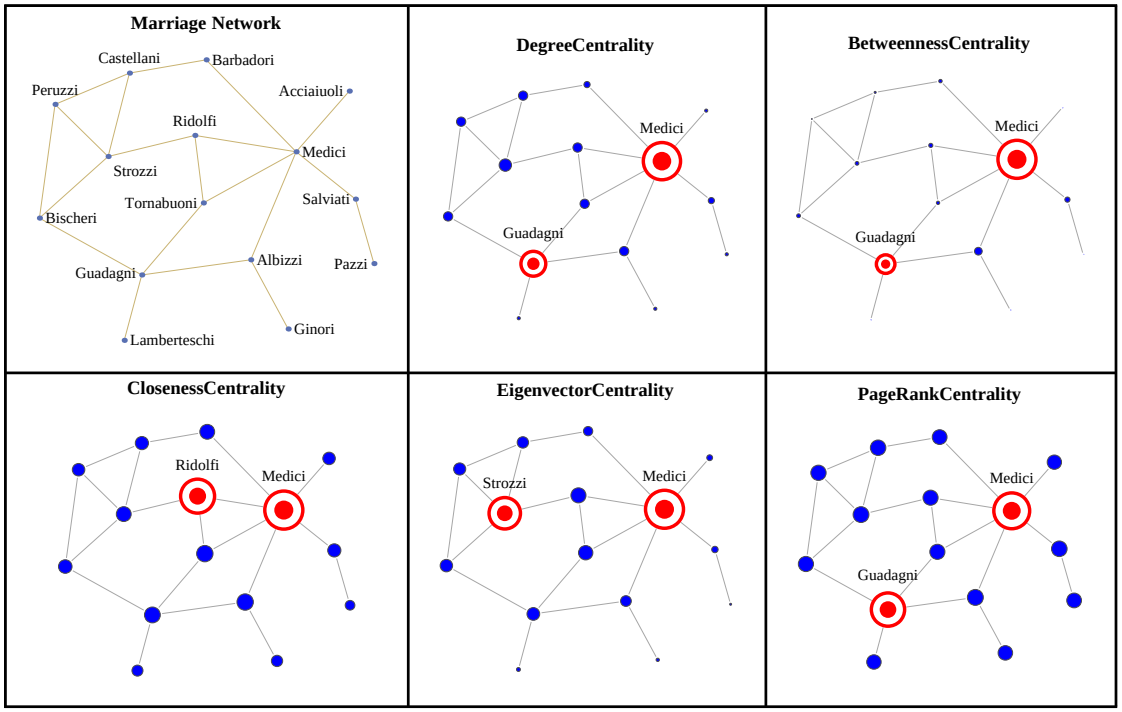

Fig. 4. Various Centralities measures of the Florentine Families Network drawn with Mathematica 9

divided by their out-degree. This is expressed mathematically as[50]:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta \tag{6}$$

This can in turn be expressed in matrix form as:

$$x = \alpha A D^{-1} x + 1 \implies x = (I - \alpha A)^{-1}.1 \tag{7}$$

where D is the diagonal matrix with elements $D_{ii} = max(k_i^{out}, 1)$.

From Figure 4, *Medici* and *Guadagni* are the most important families in the network as highlighted in the figure.

(iii) **Clustering coefficient:** A common phenomenon in social networks is the formation of cliques or communities. This inherent tendency to cluster is quantified by the clustering coefficient [19]. Thus, if a node v's neighbors have n directed connections between them, then the clustering coefficient of node v is expressed mathematically as

$$C(v) = \frac{n}{d_v(d_v - 1)} \tag{8}$$

The clustering coefficient of the entire network is then the average of all individuals $C(v)$'s. $C(v)$ is generally ≤ 1, and if $C(v) = 1$ then every node in the network connect to every other node. The average clustering coefficients of the four networks are shown in Table I. The SmallWorld network has the highest clustering coefficient compared with the other networks and this is correlates to their properties outlined in Table I. Figure 5 further shows the graphical representation of the average clustering coefficient of the four networks plotted against their various nodes degree. From the Figure, the clustering coefficient of *cit-HepTh* and

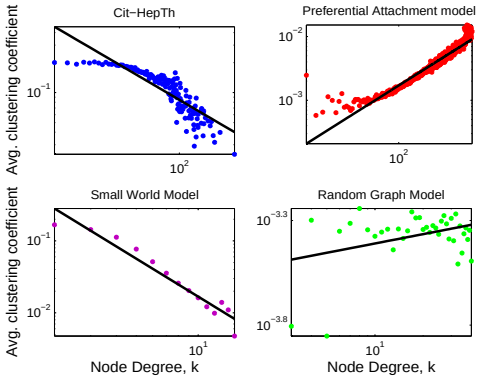

Fig. 5. A Plot of Clustering Coefficient against Node Degree

small world networks are higher for nodes of low degree. This means that there is high clustering among low-degree nodes in these networks. However, the clustering coefficient of the *preferential attachment model* of the same cit-HepTh network tends to have high clustering for nodes of high degree. For that of the *small world*, the clustering is relatviley the same for all nodes. These emperical results throw more light on differences in the internal wiring of these networks that in turn supports the mathematical framework that govern these networks.

(iv) **Modularity:** The modularity measure, first proposed by Newman [39] is an objective function that seeks to maximize the strength of division of a network into communities when studying communities in networks. It is therefore the fraction of edges of the network that fall within the communities minus the expected such fraction if the edges were assumed to be distributed randomly without regards for communities.

Thus *Modularity* can be expressed mathematically as

$$Q = k - w_i^2 \tag{9}$$

where $k = \sum_i e_{ii}$ is the fraction of edges in the network within the same community and $w_i = \sum_i e_{ij}$ is the fraction of edges that are assumed to be distributed randomly.

The value of Q lies in the range $(-1, 1)$ with a value 0 being a network without any community structure as compare to a random graph. Real-world networks with good community structure has a modularity value of 0.3 or high [39]. For, from Table I, the most modular network is the W-S network model. This network has a modularity value of 0.87216 showing the strong community structure within it.

(v) **Radius and Diameter:** The *radius* and *diameter* of a graph represents how far away nodes are from each other in the network. The *eccentricity* $\epsilon(v)$ of a node v is the maximal geodesic distance between v and any other nodes within the network. The smallest eccentricity value of graph is its *radius* and the maximum value is its *diameter*. Thus, the *radius r* and the *diameter d* of a graph are expressed respectively as:

$$r = \min_{v \in V} \epsilon(v) \tag{10}$$

and

$$d = \max_{v \in V} \epsilon(v) \tag{11}$$

The Table I shows the various radius and diameter values for all the four networks with cit-HepTh and Small-World networks having the highest diameter and radius respectively.

(vi) **Average Path Length:** The average path length of a network l_G is defined as the distance between a pair of adjacent nodes divided by the total distance between all possible pairs of node within the network.

In terms of symbols, the average path length of graph l_G is defined as

$$l_G = \frac{1}{n(n-1)} . \sum_{i,j} d(v_i, v_j) \tag{12}$$

where $d(v_i, v_j) = 1$ if there is a path from node v_i to node node v_j and 0 otherwise. For a online social network for example, the average path length could be the average number of friends or hops existing in the shortest chain linking two people or entities together. The average path length of the four networks are displayed in Table I. Small average path lengths in online social network can lead to the rapid cascade or diffusion of information within the network.

For a complete outline on OSN metrics with their mathematical framework, the reader is referred to the research works by S. Wasserman [37] and [41].

V. Related Works

Social networks are very relevant in many application areas and several research efforts are directed toward knowing and understanding the structure and dynamics of these networks. In the following, the related research efforts on applying the structural properties of online social networks are reviewed.

Fraud detection in large scale auction online network is investigated by [43] using an algorithm called *NetProbe*. The NetProbe is claimed to have the ability to locate ubiquitous fraudsters and even other potential fraudsters. The algorithm was tested using a live data from eBay. However, their work is limited to that of bipartite cores. A general architecture for fraud detection is highly appropriate due to the complexities and dynamics of these networks.

Effective use of social network tools to mine criminal data has great importance in crime investigation. For example, link analysis can be used to fight organized crime [44], [45]. In [45], criminal network is studied by using shortest - path algorithms, priority-first-search (PFS) and two - tree PFS to identify association paths, or geodesics between two or more entities within the network. The method was 70% efficient as compared with other methods in the literature. As a complement to this work, degree centralities and community structure within online social networks can be another effective way in studying and fighting criminal networks.

Text data mining is another key method of extraction relationship or correlation of text data. [46] used text data mining method to extract relevant information from in-patient nursing records by using a KeyGraph tool.

Link prediction is studied by [47] using the structural properties of online social networks. The approach used is based on the average distance between nodes in the network and the weights associated with the links. However, the application area of their finding is lacking. Also the finding can be extended to do what-if-analysis, which could be vital for extrapolation, provisioning and optimal algorithm design.

The influence that social interactions have on politics is studied by [48]. The paper argued that these social platforms could be the right place for politicians to sell their political views and thereby become politically relevant to the masses. But another way of measuring the influence of online social network on politics could be by the use of network structural and statistical properties. This method is more general and can easily be modeled for different network categories.

Kahanda I. and Neville J. [49] developed a supervised learning approach to predict link strength in online social network using transactional information. This prediction methodology was used to do utility comparison based on attributes, topological and transactional features of public data from the Purdue Facebook social network and was be able to predict strong relationships among users. But how these findings can be leveraged in other areas of research are not given. Also the approach is limited in scope and cannot be generalized to take into consideration the dynamics or evolutions associated with online social networks.

VI. Open Questions and their application areas

Due to the widespread of network connectivity, billions of people have changed their lives within just the past few years, by creatively using online social networks. Through the usage of online social networks families and friends are brought closer together, neighbors and colleagues are reached out to, market products and services are investigated into and so on.

In the following subsections, we outline some future works of which answers to them have a great potential application areas in product views, law enforcement, real - time citizen journalism, trusted systems, human behavior analysis, political campaign, among others. The questions are grouped into various headings for easy understanding of the framework of this research.

A. The Structure of Online Social Networks

OSNs are complex interconnections of nodes and these networks grow as new nodes and links are added to the them. Research has shown that various social network sites have their unique network layout and structural properties, however, with some commonalities that cut across them. Therefore, the following questions could be of interest to the research community:

- It would be interesting to investigate the structural and statistical properties of online social networks and then model these properties to capture the dynamics of the online social networks. This could help predict users behavior within the network.
- The microscopic mechanism underlying the formation of online social network is less understood by the social network community. Why and how do large social networks come about with such structure is a good research direction as this will help the research community to understand the microscopic forces that shape the formation of social networks in general.
- Social networks are said to be resilient against random attack. However, a quantitative measure of this resilient is yet an open question. For example, the percentage of nodes that needs to be removed to affect the social network connectivity and other metrics of the network is still an on going research. A detail quantitative measure of this needs to done in order to know wether this measure varies or remains constant for some types of networks.

Answers to the above questions would be a great contribution in the area of law enforcement and disease control. For example, with regards to law enforcement, tracking and tracing the activities of criminals and other deviants in society can easily be achieved by knowing their network structure.

B. Information Spreading

Spreading or cascading of information on online social networks is receiving a good research efforts in recent times. However, the nature of these cascades and how they can be leveraged in other areas of human endeavors are still open questions to investigate. The topology of these networks has great influence on the overall behavior of information or epidemic spreading in the network.

The Twitter social network is a unique kind of network as compare to the other online social networks in that it is made up of the traditional mass media and the social media (word - of - mouth propagation). According to [42] the flow of information in Twitter is carried out through three channels. They include: (a) the media[1] (b) the grassroots[2] and (c) Evangelists[3]

- It will be a good research effort to measure quantitatively the influence levels of the mass media, grass roots and evangelists using their network properties and then study the growth pattern of these information diffusion channels over a time frame. This influence measure could be a reference point in classifying users into any of these three categories. It can also help in viral marketing, product reviews by the use of hash tags to monitor the popularity of a new product and how users are commenting on the product. Also, spammer detection and other decision making process within the social communities of users can be tracked and analyzed.
- Also the Twitter blogosphere is classified into three levels of communications based on a data driven approach. However, it would be interesting to classify these Twitter users base on size of the hubs, the degree distribution of hubs, growth pattern of hubs over a time frame, and hub behaviors and attributes and then analyze their structural and statistical properties.
  It would be of interest to investigate the behavior of clusters or communities to flow of information or sentiments within online social network. It would be of interest to know whether clusters or communities inhibit or facilitate the flow of information or sentiments. Also the nature and duration of information cascade and how these cascades are predicted in a given time frame is still an on going research.

The research findings from these problems under information spreading could reveal some interesting phenomena such as the resilient properties of online social networks, and how long information take to reach other nodes within the network. These findings would be of great importance to law enforcement agencies, real-time citizen journalism, recommendation systems, disease control efforts, and algorithm optimization.

C. Community Detection Base On User Attributes

Community formation is a common phenomenon in OSNs and detecting these communities accurately is not trivial, however, community detection has a lot of importance in social network analysis. For example, identifying these communities using user attributes have a lot of application areas.

- Therefore, it is a good research effort to investigate on the distribution of the `Grassroots` and `Evangelists` on Twitter using data base on gender, age and geographical location, and then use the distribution to predict users attributes and behavior within the Twitter social network.
- Another research direction worth investigating is how to group users base on gender and age and by this bring in information control policy into social media to prevent

[1]These are the traditional media such as BBC, CNN, e.t.c. who can reach a large audience but do not actively follow others.

[2]These are ordinary users, who are not followed actively by large number of users.

[3]These consist of opinion leaders, politicians, celebrities and local businesses, who are socially connected and actively follow others.

some information not to propagate to certain kinds of users, thereby protecting some users from "social epidemics". This could be seen as segregation of social network for information diffusion.
- Also, categorizing the followers pattern of the Twitter social network base on content and personality is still an open question. For example, recommendation systems can be built to suggest objects such as music, movies, activities, *et cetera*, base on the interest of other individuals with overlapping characteristics.

The potential applications of these open questions under community detection using user attributes are numerous. Some of them are listed below:

- Identifying clusters or communities within OSNs have a good application in product reviews as customers within the same cluster or community are likely to have the same view about particular goods or services offered by an online retail service. These reviews can be of great importance in enhancing the social outlook of the retail service.

  Also systems which show advertisements can use an individual who is most likely to be interested and receptive to the advertisements as a baseline to introduce the same advertisement to members of that individual community.
- Clustering users on online social network base on their interest, age and geographic locations can have application in human behavior analysis, shared interest analysis, law enforcement and political campaigning efforts as these communities can easily be tracked within the entire social network.
- Classifying vertices according to their structural positions on online social networks can aid in the development of optimize algorithms for online product search.
- Nodes that are few distance apart in a social network have a high chance to trust one another as compare to nodes that are distributed randomly and are several hops apart. Therefore, understanding the community structure of online social networks will offer a good ground in building trusted networks that are secure and robust. These trusted networks can be leveraged in online marketing, online banking, online auctions, and so on. Also new friends, services or contents can be suggested to individuals by using their interest, demographics or experiences.

D. Node Degree Analysis

Linear preferential attachment or the *Rich-get-Richer* phenomenon is hypothesized to be a common phenomenon within social networks, where nodes with high degree tend to increase their connectivity faster than those with low degree. Also these nodes with high degrees have a good possibility of being connected to other high degree nodes. This preferential attachment could be both an advantage and a disadvantage. For example, they can be targeted by "epidemics"or be a good channel for information flow.

The following are important issues that merit investigation:

- Wether preferential attachment has a relationship with age, level of education or social class and gender. This will help determine the degrees of correlations between these nodes. Data from Facebook, Twitter, and LinkIn networking sites could help analyze the levels of correlations between these preferential attachments.
- It would be interesting to investigate whether user behavior could be predicted based on the structural properties of nodes within the social network.
- Some researchers have revealed that a small percentage of the nodes in online social network hold majority of the social graph, thus the social network system is controlled and held together by a few minority nodes. It will be interesting to study the social network structure of these few minorities users with large degree distributions in order to understand their network properties and their growth pattern over a time frame. The finding may help in the optimization of search algorithms; understand the distribution of information and also building security measures for these top level nodes/users. Answers to these questions can be useful in marketing applications. For example, the interest of users and other attributes encoded on the social network structure can be used to predict other nodes that have high propensity to be interested in the same product or service offered by an online sales platform like *www.amazon.com*

References

[1] danah m. boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," Journal of Computer-Mediated Communication, 2007; Vol. 13, pp. 210-230.

[2] Wikipedia. "List of Social Networking Sites", URL: http://en.wikipedia.org/wiki/List of social networking websites, [Access Date = 14th August, 2013]

[3] arxiv hep-th (kdd cup) network datast-KONECT, April 2014.

[4] Erdös, P., and A. Rényi. "On random graphs I." Publ. Math. Debrecen 1959; Vol.6, pp. 290-297, 1959

[5] Kochen, Manfred, ed. *The small world.* Norwood, NJ: Ablex, 1989.

[6] Barabási, Albert-Lászól and Réka Albert. "Emergence of scaling in random networks." Science; Vol. 286, No. 5439, pp. 509-512, 1999.

[7] Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos. "On power-law relationships of the internet topology." In ACM SIGCOMM Computer Communication Review; Vol. 29, no. 4, pp. 251-262. ACM, 1999.

[8] Kumar, Ravi, *et al*. "Trawling the Web for emerging cyber-communities." Computer networks; Vol. 31, No.11,pp. 1481-1493,1999.

[9] Barabási, Albert-Lászól and Réka Albert. "Emergence of scaling in random networks." Science; Vol. 286, No. 5439, pp. 509-512,1999.

[10] Adamic, Lada, Orkut Buyukkokten, and Eytan Adar. "A social network caught in the web." First Monday 2003; Vol. 8, No.6.

[11] Braitenberg, Valentino, and Almut Schz. Anatomy of the cortex: Statistics and geometry. Springer-Verlag Publishing, 1991.

[12] Amaral, Luis A. Nunes, Antonio Scala, Marc Barthélémy, and H. Eugene Stanley."Classes of small-world networks." Proceedings of the National Academy of Sciences 2000; Vol. 97, no. 21, pp. 11149-11152.

[13] Broder, Andrei, et al. "Graph structure in the web." Computer networks 2000; Vol.33, No.1, pp. 309-320.

[14] Albert Réka, Hawoong Jeong, and Albert-László Barabási. "Internet: Diameter of the world-wide web." Nature 1999; Vol. 401, No. 6749, pp. 130-131.

[15] Newman, Mark EJ. "The structure of scientific collaboration networks." Proceedings of the National Academy of Sciences 2001; Vol. 98, No. 2, pp. 404-409.

[16] Adamic, Lada, Orkut Buyukkokten, and Eytan Adar. "A social network caught in the web." First Monday 2003; Vol. 8, No. 6.

[17] Barnes, John Arundel. *Class and committees in a Norwegian island parish*. Plenum, 1954.

[18] Milgram, Stanley. "The small world problem." Psychology today 1967; Vol. 2, No. 160-67.

[19] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world'networks." nature 1998; Vol. 393, No. 6684, pp. 440-442.

[20] Granovetter, Mark S. "The strength of weak ties." American journal of sociology: 1973; pp. 1360-1380.

[21] Cha, Meeyoung, et al. "The world of connections and information flow in twitter." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 2012; Vol. 42, No. 4, pp. 991-998.
[22] Cha, Meeyoung, Alan Mislove, and Krishna P. Gummadi. "A measurement-driven analysis of information propagation in the flickr social network." Proceedings of the 18th international conference on World wide web. ACM, 2009.
[23] Chua, Alton YK and Banerjee, Snehasish. "How Businesses Draw Attention on Facebook through Incentives, Vividness and Interactivity.", IAENG International Journal of Computer Science, vol. 42, no. 3, pp. 275-281, 2015.
[24] Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?" Proceedings of the 19th international conference on World wide web. ACM, 2010.
[25] Wu, Shaomei, et al. "Who says what to whom on twitter." Proceedings of the 20th international conference on World wide web. ACM, 2011.
[26] Lauw, Hady, et al. "Homophily in the digital world: A LiveJournal case study. " Internet Computing, IEEE 2010; Vol. 14, No. 2, pp. 15-23.
[27] Mislove, Alan, *et al.* "You are who you know: inferring user profiles in online social networks." Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.
[28] Hewitt, Anne, and Andrea Forte. "Crossing boundaries: Identify management and student/faculty relationships on the Facebook. " Poster presentes at CSCW, Banff, Alberta: 2006; pp. 1-2.
[29] Nguyen, Khanh, and Duc A. Tran. "An analysis of activities in Facebook." Consumer Communications and Networking Conference (CCNC), 2011 IEEE.
[30] Weiss, Robert S., and Eugene Jacobson. "A method for the analysis of the structure of complex organizations." American Sociological Review 1955; Vol. 20, No. 6, pp. 661-668.
[31] Chen, Wen-Yen, et al. "Collaborative filtering for orkut communities: discovery of user latent behavior." Proceedings of the 18th international conference on World wide web. ACM, 2009.
[32] Hansen, Derek, Ben Shneiderman, and Marc A. Smith. "Analyzing social media networks with NodeXL" :Insights from a connected world. Morgan Kaufmann:2010; pp. 119- 124.
[33] Zachary, Wayne W. "An information flow model for conflict and fission in small groups." Journal of anthropological research: 1977; p. 452-473.
[34] Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos. "On power-law relationships of the internet topology." ACM SIGCOMM Computer Communication Review 1999; Vol. 29. No. 4. ACM.
[35] Yang, J.M., Liu, Z.Y. and Qu, Z.Y., "Clustering of Words Based on Relative Contribution for Text Categorization". IAENG International Journal of Computer Science, vol. 40, no. 3, pp.207-219, 2013.
[36] Newman, Mark E J, and Michelle Girvan. "Finding and evaluating community structure in networks." Physical review E 2004; Vol. 69, No. 2, pp. 026113.
[37] Wasserman, Stanley."Social network analysis: Methods and applications."Vol. 8. Cambridge university press, 1994.
[38] Mislove, Alan, et al. "Measurement and analysis of online social networks."Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007.
[39] Newman, Mark E J. "Fast algorithm for detecting community structure in networks." Physical review E 2004; Vol. 69, No. 6, pp. 066133.
[40] Ghosh, Saptarshi, et al. "Cognos: crowdsourcing search for topic experts in microblogs." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.
[41] Wikipedia.URL: http://en.wikipedia.org/wiki/ Centrality#Degree centrality, [Access date: 14th August, 2013]
[42] Cha, Meeyoung, et al. "Measuring User Influence in Twitter: The Million Follower Fallacy." ICWSM 2010; No. 10, pp. 10-17.
[43] Pandit, Shashank, et al. "Netprobe: a fast and scalable system for fraud detection in online auction networks." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.
[44] Xu, Jennifer, and Hsinchun Chen. "Criminal network analysis and visualization." Communications of the ACM, Vol. 48, No. 6, pp. 100-107,2005.
[45] Xu, Jennifer J., and Hsinchun Chen. "Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks." Decision Support Systems; Vol. 38, No. 3, pp. 473-487, 2004.
[46] Kushima, M., Araki, K., Suzuki, M., Araki, S. and Nikama, T., "Text Data Mining of In-patient Nursing Records Within Electronic Medical Records Using KeyGraph". IAENG International Journal of Computer Science, vol. 38, no.3, pp.215-224, 2011
[47] Murata, Tsuyoshi, and Sakiko Moriyasu. "Link prediction based on structural properties of online social networks." New Generation Computing; Vol. 26, No. 3, pp. 245-257, 2008.
[48] McClurg, Scott D. "Social networks and political participation: The role of social interaction in explaining political participation." Political Research Quarterly 2004; Vol. 56, No. 4, pp. 449-464.
[49] Kahanda, Indika, and Jennifer Neville. "Using Transactional Information to Predict Link Strength in Online Social Networks." ICWSM. 2009.
[50] Newman, Mark. Networks: an introduction. Oxford University Press, 2010.
[51] Chuang, L.Y., Lin, Y.D. and Yang, C.H., "Data clustering using chaotic particle swarm optimization". IAENG International Journal of Computer Science, vol. 39, no. 2, pp.208-213, 2012

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. Mr. Baagyere is with the Faculty of Mathematical Science, UDS, where he teaches in the Department of Computer Science. He is currently a Ph.D candidate at the University of Electronic Science and Technology of China (UESTC) studying for a degree in Computer Science and Technology. His research interests include Social Networks, 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. He 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 Dean of the School of Information and Software Enigineering at University of Electronic Science and Technology of China (UESTC), where he is also 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.