symmetry

MDPI

Article

Online Social Networks (OSN) Evolution Model Based on Homophily and Preferential Attachment

Jebran Khan and Sungchang Lee *

School of Electronics and Information Engineering, Korea Aerospace University, Deogyang-gu, Goyang-si, Gyeonggi-do 412-791, Korea; jebran.khan@kau.kr

* Correspondence: sclee@kau.ac.kr; Tel.: +82-2-300-0127

Received: 11 October 2018; Accepted: 15 November 2018; Published: 19 November 2018

Abstract: In this paper, we propose a new scale-free social networks (SNs) evolution model that is based on homophily combined with preferential attachments. Our model enables the SN researchers to generate SN synthetic data for the evaluation of multi-facet SN models that are dependent on users' attributes and similarities. Homophily is one of the key factors for interactive relationship formation in SN. The synthetic graph generated by our model is scale-invariant and has symmetric relationships. The model is dynamic and sustainable to changes in input parameters, such as number of nodes and nodes' attributes, by conserving its structural properties. Simulation and evaluation of models for large-scale SN applications need large datasets. One way to get SN data is to generate synthetic data by using SN evolution models. Various SN evolution models are proposed to approximate the real-life SN graphs in previous research. These models are based on SN structural properties such as preferential attachment. The data generated by these models is suitable to evaluate SN models that are structure dependent but not suitable to evaluate models which depend on the SN users' attributes and similarities. In our proposed model, users' attributes and similarities are utilized to synthesize SN graphs. We evaluated the resultant synthetic graph by analyzing its structural properties. In addition, we validated our model by comparing its measures with the publicly available real-life SN datasets and previous SN evolution models. Simulation results show our resultant graph to be a close representation of real-life SN graphs with users' attributes.

Keywords: SN model; SN data; SN topology; homophily; interactions; interests

1. Introduction

Online social networks (OSNs) have experienced dramatic growth, due to recent advancements in information and communication technologies. OSNs today have become the most important platforms for social interaction, communication, information processing, social influence and information diffusion, user/social opinion analysis, user behavior and personality investigation, business analytics and other commercial applications and services. Due to their wide range of applications, social network analysis has grabbed the attention of researchers, not only from the computer sciences, but also from social sciences, psychology, mathematics, computational linguistics, artificial intelligence and machine learning.

Understanding of social network (SN) structure, data distribution, and evolution helps in the development of SN models for different applications, such as recommendation systems [1], trust and reputation modeling [2], defense against Sybil attacks [3], community clustering [4,5], trust and interest based modeling [6], disaster detection [7] and other commercial applications. The literature suggests that SN structures along with user attributes, such as user profile information, provide a more in-depth understanding of SN and many useful patterns and trends, extracted from SN data, are used for applications in different domains. Several studies have been conducted to analyze SN.

However, considering the wide range of SN applications, there is room for research to provide more comprehensive and promising models for improving the performance of SN applications and services.

The SN are usually composed of two entities, namely nodes (i.e., users or items) and edges, representing the relationship, between nodes. The relationship can be attributed in terms as one of the interaction types, such as friendship, co-authorship, SN activities between nodes and group membership. However, an extended definition of SN can be explored in the area of synthetic SN, i.e., SNs are sets of people or groups of people with common interests and interactions patterns, usually having similar attributes [8]. In SNs, users tend to connect with similar users based on spatial and demographic similarity or mutual interests despite influential users. In most SN evolution studies, the network topological structure is considered to be the basis of changes in the network. Topologically, due to continuous addition of new nodes and their preferential attachments with high degree nodes, the nodes connectivity follows the property of scale-free power law distribution [9,10].

The SN topology based evolution models produce synthetic networks that can structurally represent real-life SN and can be useful to evaluate topology based models and applications, such as modularity based community clustering [11], influence modeling [12], and information diffusion modeling [13]. However, these evaluation models are not suitable for applications where users' attributes and the relationship strengths are critical, such as recommendation systems [14], trust models [15], and interest-based models [16]. Moreover, such models also have limitations when dealing with new users having low attachment degrees by assigning them low attachment probabilities. However, in real-life SN, new users are recommended to other users with more preference and many similar users are suggested to new users as well, and the probability of establishing relationships among new users is much higher than the existing ones.

Moreover, in SN analysis, access to real-life, complex SN data is critical to evaluating and validating SN models and applications. However, SNs provide APIs for data acquisition, and the process is generally time- and resource-consuming. Moreover, to ensure users' privacy, social networks, such as Facebook, Instagram, and Twitter, provide a limited amount of information. To train and evaluate SN models, several datasets are available for different SN applications. However, these datasets are more application oriented and cannot be used to evaluate multi-facet models and frameworks.

In this paper, in order to deal with the limitations of topological evaluation models, we propose a novel evolution model based on the principal of homophily and preferential attachment by considering user attribute similarity as the basis for the connection formation in SN. Moreover, using the proposed model, personalized activities are generated, which, in combination with the relationship graph and profile information, results in a synthetic dataset. The proposed model is evaluated, by comparison with real SN datasets and models, to authenticate its feasibility as an approximation of real SN.

The main contributions of the paper can be summarized as:

- We propose a novel SN evaluation model based on the principles of homophily combined with preferential attachments.
- We generated a synthetic dataset with the proposed model that can be used as an approximation of data from real-life SN and can be used for the evaluation of the SN models for different application domains.

The rest of the paper is organized as follows. In Section 2, we provide a detailed review of the literature. Section 3 provides a detailed description of the proposed model. In Section 4, we provide detailed description of the conducted experiments, experimental results and detailed analysis of the results. Section 5 draws conclusions of the work.

2. Background and Related Work

The scope of this work is related to social networks evolution models, data distribution and synthetic graph generation. This section briefly presents the related work of social network topology-based evolution, similarity-based evolution, data distribution and synthesis.

2.1. Social Network Topology Based Evolution Modelling

Social Networks' (SNs) structural characteristics and node attributes are important in SN research, in order to evaluate SN models and applications. An extensive study is available on structural characteristics of real-world SN. These structural characteristics include degree distribution, clustering, centralities, modularity, size of network, average path distance, network diameter, network connectivity, ties strength, and link predictions. These structural characteristics are evaluation metrics to evaluate synthetic SN structural graph models. Many studies in SNA analysis are conducted to analyze the changes in the single network. These studies attempt to map the changes as a function of structural characteristics of the network [17–20]. Many formal models explain the structural evolution and give statistical examination of online SN.

In [21], bloggers demographics, friendships and their activity pattern graph have been studied to explain and analyze friendships, and to demonstrate the small-world phenomenon and spreading properties of the target graphs [22–24]. To understand the concepts and characteristics of large and complex graphs, one can refer to [25–31]. Most of these studies are conducted on static graphs. However, real-world SN graphs are evolving. Many papers are available about SN graph evolution. Typically, for analysis, the researchers used to take snapshots of real SN at different points in time and analyze them to make their assumptions about SN evolution over time. This approach was used by [32], to study connection patterns, trends and the advent of bursty communities in blog space. Fetterly et al. [33] and Cho et al. [34] studied the structural properties of the World Wide Web by using different snapshots.

Over the years, many SN evolution models are proposed to generate synthetic graphs, with real-life SN like characteristics and patterns. Exponential random graph models (ERGMs), are statistical models used to analyze SN. The random graph model [35], with arbitrary degree distribution, for a unipartite, i.e., acquaintance, and bipartite, i.e., affiliation, graphs were proposed and evaluated to have close agreement with that of real SN data. Erdos–Rnyi random model [36], an undirected and simplest model, is very unlikely to produce real-life SN, but, still, it can predict a certain number of network properties like small diameter, path length, and giant components. This model is unable to explain degree distribution and clustering [37,38]. The Watts–Strogatz model [39], proposed by Duncan J. Watts and Steven Strogatz in 1998, is a random graph model with small world properties like short average path and high clustering. The major drawback of this model is the generation of unrealistic degree distribution. From studies on SN evolution, it has been concluded that large-scale complex SN are scale-free and obey the power law degree distribution [27,39]. In the past decade, the mechanism responsible for the scale free nature of the SN has been researched thoroughly, and preferential attachment is found to be the most suitable explanation for the scale-free structure of the SN [40]. In preferential attachment, based on the principle of the rich get richer, nodes with a high degree have high link establishment probability. In this mechanism, new nodes are more likely to connect with high degree existing nodes. The degree of the nodes in this model is proportional to the age of the node [39,40]. Chakrabarti et al. [41] proposed R-MAT. This method is based on matrix recursion. R-MAT has many real-life networks like characteristics, i.e., power-law degree distribution, small diameter, and community structures. Using the recursion idea proposed in R-MAT [41] and Kronecker multiplication, Leskovec et al. [25] proposed a new graph model. The idea is to generate self-similar graphs recursively. They proposed a fire-forest graph model to explain the evolution of citation graphs. There are many other structural models for synthetic graph generation such as Girvan–Newman (GN) model [4] , Lancichinetti–Fortunato–Radicchi (LFR)

model [42], Kleinberg model [43], Chung–Lu (CL) model [44,45], Waxman model [46], etc., which can be used as required.

2.2. Homophily Based Social Network Evolution Modeling

Besides SN structural characteristics, there are other factors that can influence the rate and pattern of network evolution. The nodes in SN have their intrinsic properties, referred to as attributes. In SN, users tend to connect with other similar users [47,48]. In social sciences, this tendency is termed as homophily [49,50]. McPherson et al. [49] described the rate of contact between similar people as much higher than between dissimilar people.

Previously, many studies were carried out in the social sciences and SN analysis, in order to understand the importance of homophily in social relationship formation. In [51], Facebook social graph structures were characterized by different structural and homophily metrics and a strong relationship was found between users with similar demography, i.e., age and location. A study [52], on popularity versus similarity in growing networks, evaluated the role of similarity in the evolution of networks and validated the model and its statistical predictions against different real networks from different domains. Gregory A. H. et al. in [53], analyzed the online dating behavior to find the effect of homophily in political views on relationships formation. They concluded that individuals establish relationships with others who share their political identities and participate actively in political engagements. From experiments, more positive reactions were observed to users of the same political ideologies.

In social psychology, the average degree of user relationships varies a little [54–56], and it depends on the user's demographics, i.e., age, gender, profession, language, religion and political orientation, and type of the network. The connectivity degree is higher in young adulthood which in middle adulthood reaches a plateau and then decreases in older ages [54,57]. The relationship between gender and connectivity; females have slightly larger personal networks than male [55]. A converse concept, i.e., the effect of relationships on users' similarities is simulated in [58] for movie recommendations.

Users, while joining SN, input their demographic information, i.e., age, gender, education, occupation, language, religion, spatial information and interest information, i.e., sports, music, movies, books and other hobbies. This information, combined to form a user profile, is then used by SN service providers to suggest other users, similar in demography, interest or belong to same proximities, in order to make new connections. The users' spatial, demographic and interest information is diversely exploited by the developers for improving the recommender systems and information filtering systems. Recommender systems are some of the powerful tools that are used for various purposes in SN applications. In [59], the extension of the recommendation system is discussed to improve the performance, capabilities and extend the range of recommender systems, by understanding users' and items attributes, and incorporating contextual information into the recommender process. The user's profile information can be used for personalized information access by collecting and analysing user's explicit and implicit information [60]. A friend recommendation system for Twitter-like SN was proposed by incorporating heterophily and homophily values with hybrid contents recommendation algorithm, to make best use of SN [61]. Another friend recommendation systems, by using proximity and homophily along with common friends, was proposed [62] to answer the question of why to add this friend? Recently, in [63,64], a personalized multimedia contents retrieval system was proposed to enhance users' Quality of Experience (QoE), by weighing multimedia contents with users' profile weights. The users' profile weights were calculated from users' interactions and interests. Information filtering aim to expose users' only to the relevant information. An overview about issues, research and information filtering systems [65] discussed the importance and issues of user profiling for information filtering.

2.3. Social Network Synthetic Graph Generation

Synthetic SN graph and data generation are useful for many purposes, i.e., model evaluation, algorithms benchmarking, interaction, and many more. Many SN datasets and API's are available to get SN data, but these datasets have some issues of missing attributes, missing values, high time and resource cost, and risk of privacy breaches. These limitations can damage the statistical correctness of the SN algorithms and models. To resolve this issue, one way is to use SN graphs and attribute generators, which can provide a common benchmark for researchers to evaluate their models on the same dataset. There are many modeling approaches available to generate SN synthetic graphs; (i) Use structural properties, (ii) principle of homophily and (iii) hybrid approach.

In SN, the users' relationships formation, interests, and activities depend on users' demographics. For synthetic data generation, it is necessary to consider such associations, trends, and dependencies between SN user attributes, interests and activities. It is predicted that aged users have fewer relationships [66–68] and are less frequent to contact [67,69,70].

Many studies carried out to generate synthetic data for SN. In [71], a synthetic SN users' skills data was generated for specific types of SN, like LinkedIn and ResearchGate, and the network growth was simulated based on the skills' endorsement scores. In [72], synthetic data was generated and populating to SN topology. A novel system, synthetic high fidelity social media data generator (SHIELD), was proposed by [73] to generate synthetic interactions graph and text data for social media.

In SN analysis, like social attributes, social interactions are also significant for many applications such as social ties strength, nodes trust, recommender systems, nodes ranking, information spreading, and opinion mining. To the social interactions for research, crawling can be used. SN data crawling is used to collect SN users/items attributes and users' generated content data. However, as the social graph crawling is time- and resource-consuming, there exist few anonymized datasets, shared with the researchers. These datasets are sometimes not able to fully satisfy the statistical confidence in simulation results. Many secure OSNs do not allow crawling due to the risk of user privacy breach. The social interaction graph and data can be synthetically generated. In OSN, higher degree nodes generate more interactions and high degree nodes cover the largest proportion of the total interactions on OSN [9]. In [10], a generic model, for synthetic social interaction graph generation, is presented. It is concluded that, in SN, social interactions have power law distribution.

3. OSN Evalution Model Based on Homophily and Preferential Attachment

In this section, we discuss the challenges associated with synthetic network generation, and also discuss how the proposed model can cope with such challenges.

3.1. Challenges in Synthetic Network Generation

In SN like synthetic network generation, many challenges must be considered while generating graphs and distributing attributes. These challenges are:

- **Attributes distribution:** What is the distribution of attribute values? In SN, the node attributes can have high diversity in values. Therefore, it is necessary to determine what the percentage of different values for each target attribute is. These percentages can be obtained from real-life SN datasets, and SN statistics, e.g., the attribute gender has two possible values—male and female, and their percentage on Facebook is 47% and 53%, respectively [74].
- **Profile data distribution:** What are the trends in the combination of user attributes to form different profiles? This is also an important factor that needs to be considered while generating synthetic SN data and graphs. In SNs, some node attributes can be used to predict the values of other attributes. These attributes are referred as inter-related attributes, e.g., if the age is in the range of 65+, there is a high probability of having interest in news.
- **Communities structure:** What is the community structure? There are many bases to form communities in SN. These bases range from structural parameters to profile similarity.

Selecting the basis for community formation is application dependent, such as with information spreading, where the connectivity can be the basis, while, for recommendation systems, the interest similarity is a better choice.

- **Synthetic network topology:** What is the topology? Many SN topologies are presented in the literature. SN topology can be obtained from real-life data sets. Previously, it was deduced from many studies that generally SNs have scale-free power-law degree distribution and have small world properties.
- **Activities distribution:** What are the activities distribution? From studies in SNA, it has been observed that the social activities obey the power-law distribution. In [9], it was concluded that, in SN, we do not need to follow all users in a group, and, out of all, only a proportion of users generate about 80% of activities. These trends in activities' generation can be extracted from SN datasets, previous research, and surveys.
- **Correlation between attributes distribution:** What is the correlation between these distributions? These distributions are associated with one another. These correlations can be deduced from SN datasets and surveys.

3.2. Preliminaries

The synthetic SN graph is represented as $G(V, E)$, where V is set of nodes, and E is the set of edges between the nodes. In this work, we have represented SN node attributes into three categories, i.e., spatial information, represented by Sp, demographic information, represented by Dm, and interest information, represented by In. Sp is a set of location information of each user in the network, where, for the ith user, the location is represented by Sp_i, and each Sp_i contains two values, i.e., latitude (ϕ) and longitude (φ) values. Dm, represents a set of demographic information for SN users, with its elements for each ith user as Dm_i, the Dm_i contains the age, gender, occupation, religion, language, political view, and initial interests' information of each user i, which they enter while joining SN. The set of items interests is represented by In, where In_i represents the interest information of every ith user in the network. The spatial, demographic and initial interests information combines to form user profile (Pr), and, for every ith node, the profile is represented by Pr_i, where i ranges from $[1, N]$, and N is the number of users in the network. As this is a dynamic/evolving network, the number of nodes can increase by the addition of new nodes over time; therefore, the value of N is incremented by the number of nodes newly added to the network. With the addition of new nodes, the demographic, spatial and initial interest information of new nodes are populated to their corresponding sets. For items' interests, we used Filmtrust, a publicly available dataset, which contains user-item ratings.

The set SI contains the social interactions generated by users on SN. These social interactions are posts, likes, comments, shares and other information, generated by users on SN. The social interactions are further divided into two categories: i.e., social actions and social responses. The social actions, (SI_{ij}), are the actions performed by a user i to user/items j and the social responses (SI_{ji}) are the responses that a user i receives from user j. The social interactions may be in the form of likes, comments, tags, posts, shares, re-shares, tweets, re-tweets, followings, followers, tweets, and re-tweets. The properties, like allowable contents, lengths, and formats, of these interactions may vary due to the business model of different SN. Some interactions may be present in one media but absent in other, depending on the SN applications domain. Therefore, we categorized these interactions into the aforementioned two broad categories. There exist other social media content that the SN users generate, but here we have only considered the social interactions. We have combined users into different groups, and users with similar properties are assigned to the same group, known as communities, based on homophily and demoted by $C_k \in C$, where C, is the set of all communities and C_k, is the kth community in the network. These notations are listed with brief description in Table 1.

Table 1. List of Notations and Description.

Notations	Description
$G(V, E)$	Graph with users/nodes V and edges E
Sp	Set of Spatial information
Sp_i	Spatial information of user i
Dm	Set of demograpic information
Dm_i	Demograpic information of user i
In	Set of Interest information
In_i	Interest information of user i
Pr_i	ith profile
SI	Set of social contents
SI_{ij}	Social interaction from user i to j
SI_{ji}	Social contents to user i from j
$p_{deg(j)}$	Probability of attachment based on degree (preferential attachment)
C_k	kth community in the network
$deg(j)$	Degree of node j

3.3. Proposed SN Evaluation Model

Figure 1 provides a block diagram of the proposed model. The proposed model is composed of seven phases, namely (i) node generation, (ii) data generation, (iii) data combination or profile generation, (iv) data population, (v) similarity based clustering, (vi) network graph generation, and (vii) social activities generation.

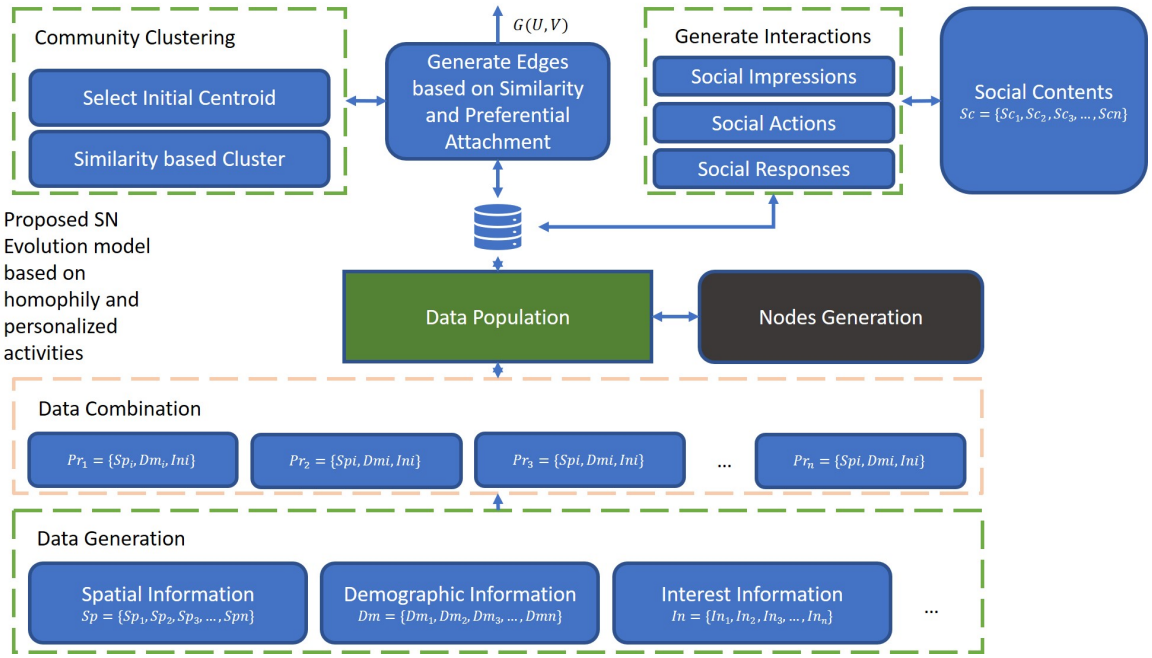

Figure 1. Proposed social network operational/dynamic model based on homophily, preferential attachment and personalized activities generation.

As a first step, we generated N number of nodes in nodes generation phase of Figure 1. Example of labelled generated nodes are shown in Figure 2. The node generation process is followed by the data generation phase. The data generation phase is application dependant, and represent SN nodes' demographic, spatial and interest attributes.

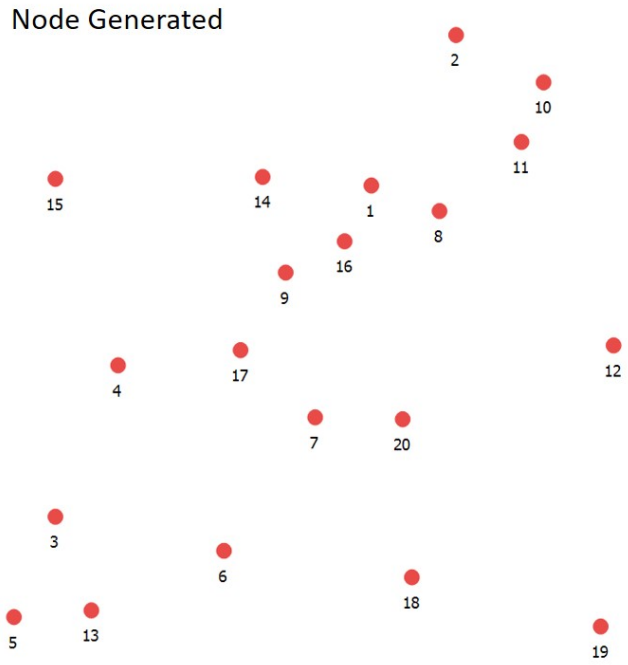

Figure 2. Example of Social Networks nodes generated.

In order to provide an approximation of real-life SN, for the selection of our model's attributes, we rely on the statistics from public sources, such as US government census data [72], and publicly available SN statistics and insights, such as Facebook [74,75]. Table 2 shows the example attributes and their proportion. There exists some attribute with inter-related values. The user's initial interests' information in SN can be associated with users' demography like age, gender, profession, and location information. In addition, interests in SN can be inferred based on user information, such as social interactions, and user's neighbors' interests. In SN user's interests can be predicted from other users, with higher demographic similarity [76]. Similarly, in [77], the effect of various SN user's information on interest similarity is investigated by video based user profiling. We need to generate lookup tables to assign such attributes. For example, nodes in age group 18–25, it is more likely to have profession = student and marital status = single, and interests = Sports Teams. Another option, to get these proportions, is to use publicly available SN datasets from sources like kaggle, SNAP, UCInet, dataworld, etc. We generated random geo-locations in the North-America region as shown in Figure 3.

Table 2. Example attributes, attribute-values, their spatial, demographic and interest proportions.

Attribute	Values
Age	"18–24" (15%), "25–34" (24%) , "35–44" (19%), "45–54" (16%), "55–64" (13%), "65+" (11%) [74]
Gender	male (46%), female (54%) [74]
Location	Random latitudes and longnitudes generated in North-American region; shown in Figure 3
Religion	"Christian" (31.9%), "Hindu" (14.8%), "Jewish" (0.2%), "Muslim" (27.1%), "Sikh" (0.3%), "Traditional Spirituality" (0.1%), "Other Religions" (12.9%), "No religious affiliation" (12.7%) [72]
Language	"English" (64%), "Spanish" (17%), "Portuguese" (15%), "French" (11%), "German" (9.9%), "Indonesian" (7.7%), "Japanese" (6.6%), "Vietnamese" (6.5%), "Arabic" (6.4%), "Hindi" (6.2%) [74]
Marital status	"Single" (31.5%), "Married" (51.4%), "Divorced" (10.5%), "Widowed" (6.6%) [72]
Profession (ISCO-08 structure)	"Manager" (12.2%), "Professional" (17.1%), "Service" (13.9%), "Sales and office" (17.8%), "Student" (23%), "Natural resources construction and maintenance" (7.0%), "Production transportation and material moving" (9.0%) [72]
Political orientation	"Far Left" (9.4%), "Left" (34.7%),"Center Left" (18.1%), "Center" (18.0%), "Center Right" (10.5%), "Right" (8.0%), "Far Right" (1.3%) [72]
Interests	Brands, Celebrities, Sports Teams, Movies, Tv Show, Games, News, Organizations [75]

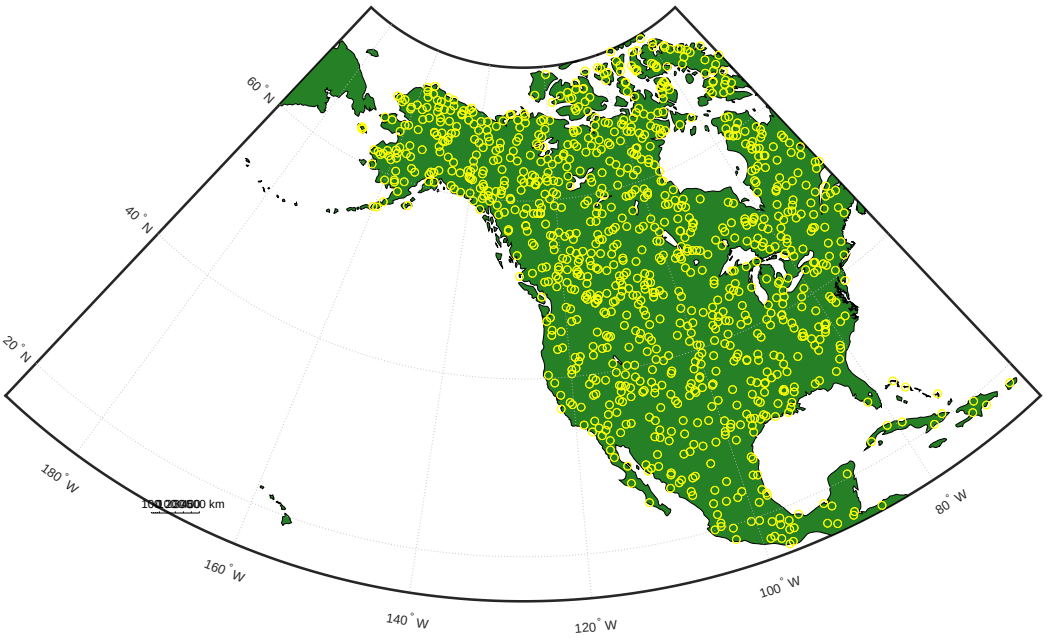

Figure 3. Random geo-location generation in the North-American region; each point on the map consists of latitude and longitude value.

After the nodes and attributes generation, we combined the attributes according to the trends and patterns extracted from real-life datasets and studies to generate profiles. The SNs are global platforms and can have different combinations of these attributes and form a diverse range of profiles. In this work, as an example, we have considered a limited number of nodes, attributes and attribute values. These parameters can be extended based on the application area where this model is used. This model is a generic one and can be adopted according to the application. In Figure 4, the combination of social attributes referred to as users' profile are assigned to each generated node. In the Figure 4 example, values are assigned to each attribute of the users' profiles, e.g., three interest, age range, gender and location. The locations in this example are labelled as L1, L2, L3, L4 and L5. These locations contain latitude and longitude values of the location where the user belongs. In profile generation, we also consider the inter-dependency of attributes. For instance, for users with age in the range of 18–25, it is more likely for them to be single, and, if the gender is male, it is more probable to have interest in sports teams. On the other hand, the probability for having interests in brands is higher for females. For the aged user, it is more likely to have an interest in news.

After the SN nodes generation and profile information population, we calculated similarities between nodes to generate a similarity matrix. The similarity matrix depicts how much a node is similar to all other nodes. The similarity is calculated by using Equation (1) [14]:

$$USim(u,v) = \alpha.SpSim(u,v) + \beta.DmSim(u,v) + \gamma.InSim(u,v) + w.OpSim(u,v), \quad (1)$$

where $USim(u,v)$ is the normalized weighted cumulative similarity between users u and v. This is weighted sum of spatial similarity, demographic similarity, and interest similarities. The α, β, γ, and w are the weights assigned to each similarity based on their importance in SN relation formation.

For computing spatial similarity ($SpSim(u,v)$) between users u, and v, we used cosine similarity [78]. We normalized the spatial similarity by the geo-distance between users, as shown in Equation (2):

$$SpSim(u,v) = \frac{1}{\left[1 + \frac{Dist(u,v)}{1000}\right]} \left(\frac{Sp_u.Sp_v}{\parallel Sp_u \parallel . \parallel Sp_v \parallel} \right), \quad (2)$$

where $Dist(u, v)$ is the geo-distance between the users u and v. Equation (3) finds the geo-distance by using Harvesine formula [79,80]:

$$\begin{gathered} a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_{Sp_u}).\cos(\phi_{Sp_v}).\sin^2\left(\frac{\Delta\varphi}{2}\right); \\ where\ \Delta\phi = \mid \phi_{Sp_u} - \phi_{Sp_v} \mid, \Delta\varphi = \varphi_{Sp_u} - \varphi_{Sp_v} \\ c = 2\arctan\left(\frac{\sqrt{a}}{\sqrt{1-a}}\right) \\ Dist(u, v) = R \times c. \end{gathered} \tag{3}$$

In Equation (2), the Sp_u and Sp_v are the GPS coordinates of the users u and v and consists of pair of latitude (ϕ) and longitude (φ) value. ϕ_{Sp_u} and ϕ_{Sp_v} are the latitudes of the users u and v and φ_{Sp_u} and φ_{Sp_v} are the longitudes of users u and v, respectively.

The demographic similarity $(DmSim(u, v))$ between users u and v is calculated by using cosine similarity [78], as shown in Equation (4):

$$DmSim(u, v) = \frac{\sum_{i=1}^{D} Dm_{i_u}.Dm_{i_v}}{\sqrt{\sum_{i=1}^{D} Dm_{i_u}^2}.\sqrt{\sum_{i=1}^{D} Dm_{i_v}^2}}, \tag{4}$$

where Dm_{i_u} and Dm_{i_v} are the ith demographic attribute of users u and v and i ranges from 1 to D, i.e., total number of demographic attributes in Dm.

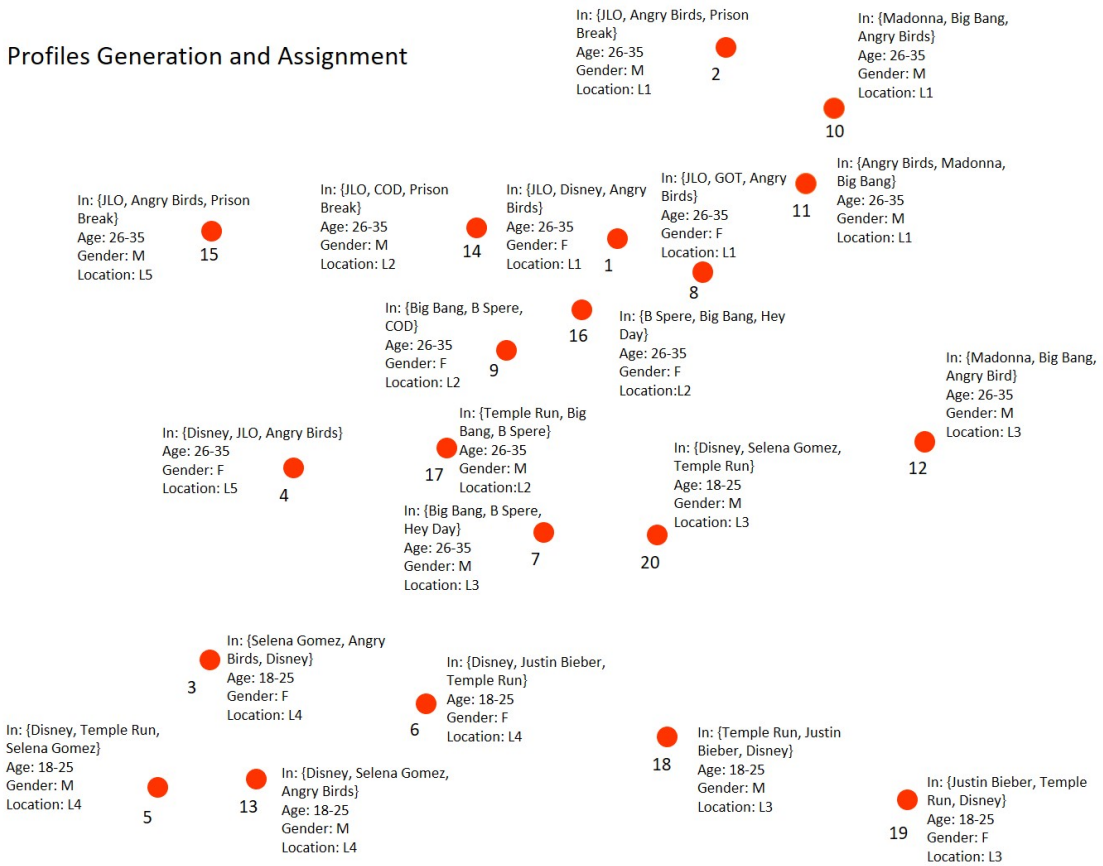

Figure 4. Profiles' generation and assignment; the profiles are generated by combining the SN users' attributes.

The users' interest similarity are the users' interests in different categories such as some users like sports, some users' have interests in books and movies, some may like to follow celebrities, and some

may have interest in politics. Equation (5) computes the interests similarity ($InSim(u,v)$) between users u and v by using a Jaccard similarity formula [81]. In Equation (5), the interest similarity of two users is computed by dividing the number of common interests with the total number of interests that both users have. When the number of users' common interests increases, the interest similarity is increased and vice versa. The users' interest information is binary, e.g., a user is interested in sports or not; therefore, we used Jaccard similarity to measure the interest similarity of users:

$$InSim(u,v) = \frac{[In_u \bigcap In_v]}{[In_u \bigcup In_v]}, \tag{5}$$

where In_u and In_v are the interest information of users u and v.

The user-items opinion similarity is computed from the Filmtrust dataset by using Equation (6). In SN, the users adopt items based on the rating and opinions of other users with similar interests, and such users are more prone to forming a connection. The user-item opinion similarity is computed from ratings which the SN users give to different items such as movies, food and other products:

$$OpSim(u,v) = 1 - \frac{\sum_{i \in (ISet_u \bigcap ISet_v)} \cdot \mid R_{u,i} - R_{v,i} \mid}{R_{max} \mid ISet_u \bigcap ISet_v \mid}, \tag{6}$$

where $ISet_u$ and $ISet_v$ are the sets of items rated by users u and v, respectively. $R_{u,i}$ and $R_{v,i}$ are the ratings given by users u and v to item i and R_{max} is the maximum rating.

After computation of similarity matrix, K seed nodes are selected, based on the similarity measures, to generate K clusters. The initial nodes are selected in such a way that the seed nodes selected must be similar to more nodes in the network. In addition, all the seed nodes must be dissimilar from each other. An example of seed nodes is shown in Figure 5. The nodes are then clustered based on the similarity measures in the previous phase. Figure 6 shows nodes assignment to their respective kth clusters C_k. Each cluster contains a set of similar users. The initial seed nodes allow the nodes to lie in their respective communities based on their similarities. The initial seed nodes and initial links generation are important to initialize the network based on similarity, rather than random initialization. When the number of seeds increased, the network convergence time is slightly decreased.

To initialize the graph, links are established between the seed nodes and other member nodes of their corresponding clusters. After the network initial nodes, more links are generated between the nodes other than the seed nodes based on their similarity and preferential attachment. The similarity is the primary parameter for the connection establishment with the degree of the nodes as the secondary parameters, i.e., nodes with higher similarity and degree are more likely to form connection. Figure 7 shows an example of initial links establishment and Figure 8, shows an example of links other than within the community. We limit the number of links in the synthetic graph based on the minimum degree that each node must achieve and average degree of the network, and each node is assigned a random number of links according to the aforementioned criteria, shown in Equation (7). In simulations, the minimum degree and average degree, whichever occurs first, was set as the model convergence criteria. The simulation results are shown in Table 4 in Section 4:

$$P_{attach} = w_s.USim + w_{deg}.P_{deg(j)}. \tag{7}$$

In Equation (7), the P_{attach} is the probability of a node to attach to other nodes, $USim$ is the user similarity, computed by using Equation (1), and the $P_{deg(j)}$ is the degree based attachment probability. The w_s and w_{deg} are the weights of the similarity and degree-based attachments, respectively. The computational formula for degree based attachment [10], is shown in Equation (8):

$$P_{deg(j)} = \frac{deg(j)+1}{\sum_{k=1}^{N} deg(k)}, \tag{8}$$

where $deg(j)$ is the degree of the target node and the $deg(k)$ is the degree of all other nodes, where k ranges from 1 to N, and N is the total number of nodes in the network.

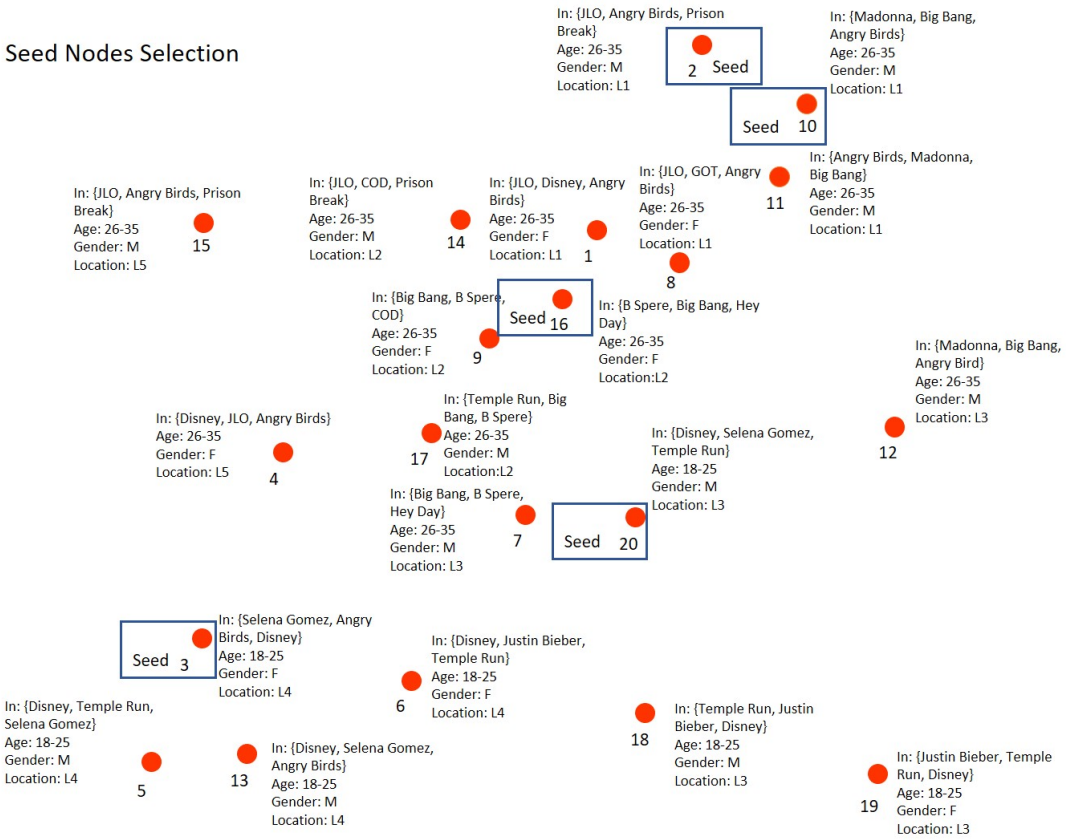

Figure 5. Seed nodes selection for clustering. The selected seeds are similar to most of the generated nodes and each seed node attributes are different from other selected seed nodes.

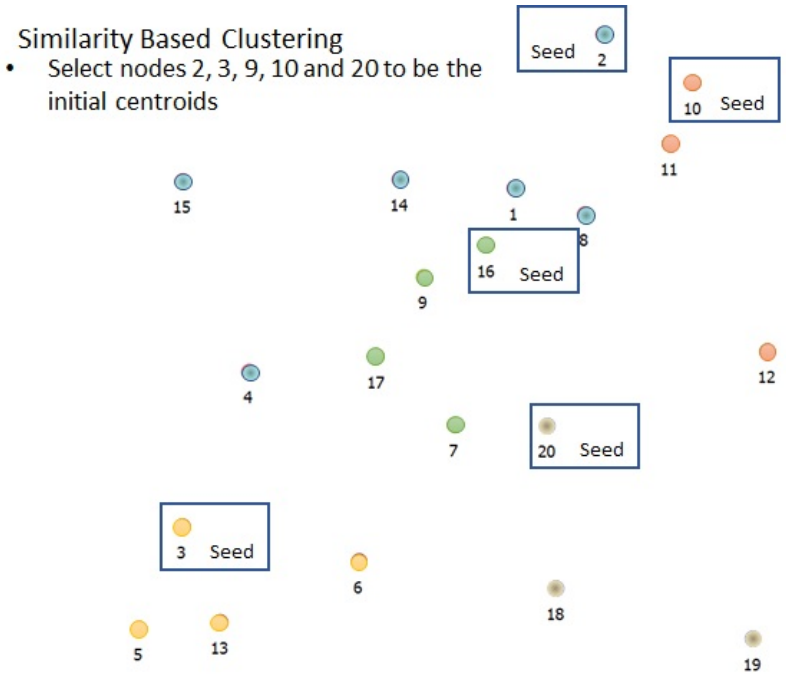

Figure 6. Similarity based clustering. Clusters are represented by different colors.

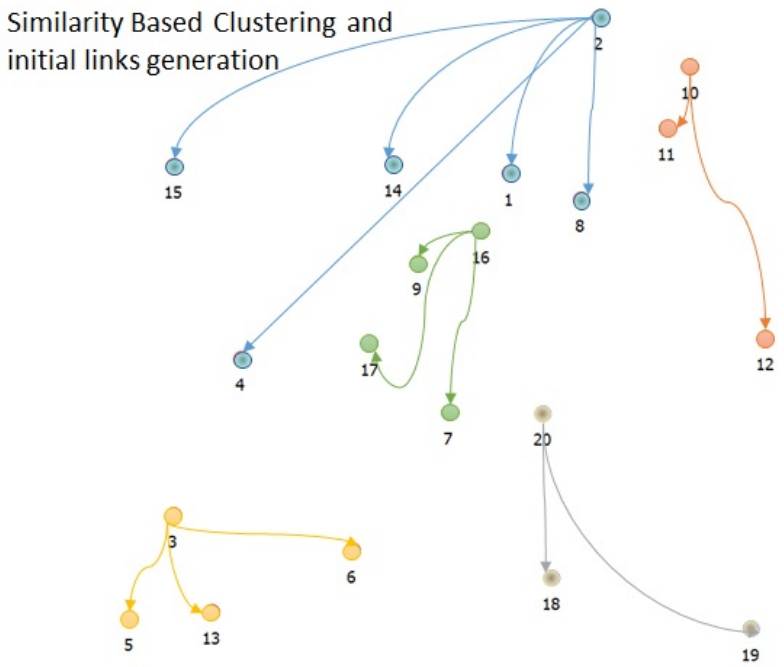

Figure 7. Initial links generation within the clusters. Nodes in each cluster are connected with their respective cluster centers.

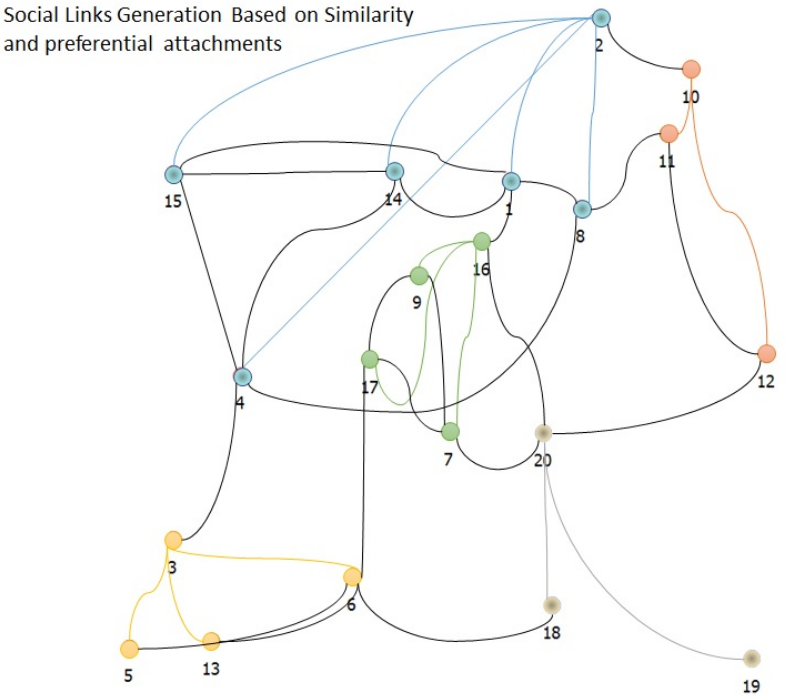

Figure 8. More links generation, outside and within communities, based on similarity and preferential attachments.

We considered the average degree to be 100 for a network of 1000 nodes with minimum degree set to 10. The number of links depends on these initial conditions. When we increase the minimum degree, each node must achieve the number of links in the network will increase, and vice versa. Similar is the case of the average degree of the network. We generated random connectivity degree $deg(j)$ for each node and selected the top $deg(j)$ most similar nodes for each jth node. The proposed model is dynamic,

and, in our implementation, 10 new nodes are added to the network at each iteration, where each node is connected to the most similar node in their respective community, as their initial connection.

The connectivity of SN users is a projection of profile similarity of the users with other existing users and the newly added users, and the human characteristic, in order to manage/keep limited friends at a time.

After the network generation, we generated user activities. According to [9], user activities also obey the power law distribution, i.e., there are only a small proportion of users who produce a significant proportion of activities over SN. In [10], it was observed that almost 80% of user activities, within an SN group, are produced by 20% of the users in the group. We produced different types of user activities/interactions with a proportion from [9,10,82].

The time complexity of the proposed systems depends on the number of nodes N, number of initial seeds K and number of users' attributes. The numerical results are shown in Section 4, Table 5.

4. Evaluation and Simulation Results

In this section, we evaluated our model by generating a synthetic SN graph, based on homophily and preferential attachment. From simulations, we observed that the synthetic SN graph obey the SN principles of "Birds of feather flocks together", and the "Rich get richer". We validated our results by comparing the closeness of the structural and similarity properties of our synthetic network with that of real-life SN, obtained in different studies. We found that the properties, i.e., degree distribution, clustering coefficients, modularity, spatial similarity, demographic similarity, interest similarity, network diameter, shortest path, of our synthetic network are similar to that of the real-life SN.

We have generated a test network of 1000 nodes with a random distribution of SN nodal attributes as discussed in Section 3. For each node we have generated, random latitudes and longitudes in the US region, demographic attributes, and interest information. The set of demographic attributes considered in the simulation of our model consists of age, gender, religion, language, marital status, profession political orientation and a set of initial interests. These attributes are generally specified by the users at the time of joining an SN. The distribution of demographic information is shown in Table 2. We generated a network of 46,780 edges by connecting the nodes based on similarity and preferential attachment. For graph representation and statistical analysis, we exported our synthetic network to Gephi. Figure 9 shows modularity based clustered synthetic SN graph, generated in Gephi.

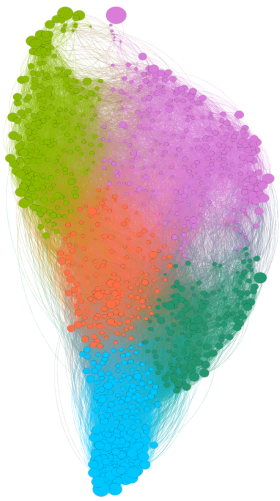

Figure 9. Synthetic network graph.

Previously, it has been observed that real-life SN are scale-free and they obey the degree distribution obey power law distribution. From the Figures 10 and 11, it can be observed that the

degree distribution of our synthetic network graph follows the power law distribution, which is similar to that of the real-life SN. As most of the synthetic graph nodes are low degree nodes, the average degree of our undirected synthetic graph is 94.56, and this value corresponds to 50% of the CDF, as can be seen in Figure 10. This phenomenon is due to the people psychology to make friendship mostly with similar people only and there is a limited number of friends they make. Such pattern can be observed both in online and offline friendships, as discussed in Section 3. Figure 12 shows the distribution and relationship between age and connectivity. It can be seen in Figure 12 that most of the friends for a user belong to the same age group, i.e., young age users are more likely to connect with other young age users and the elder users are more probable to connect with other old age users. This age-based connectivity is due to their common interests. In [51], a study on the Facebook social graph, it was concluded that users on SN make friends with other users of same age group, but this pattern is more prominent in young individuals, and the age range for neighbors of young users is smaller compared to that of the older ones. We observed similar properties in our synthetic social graph. The connectivity of users with other age group users in the synthetic SN graph is due to attribute similarity other than age i.e., gender, religion, language, profession and initial interests.

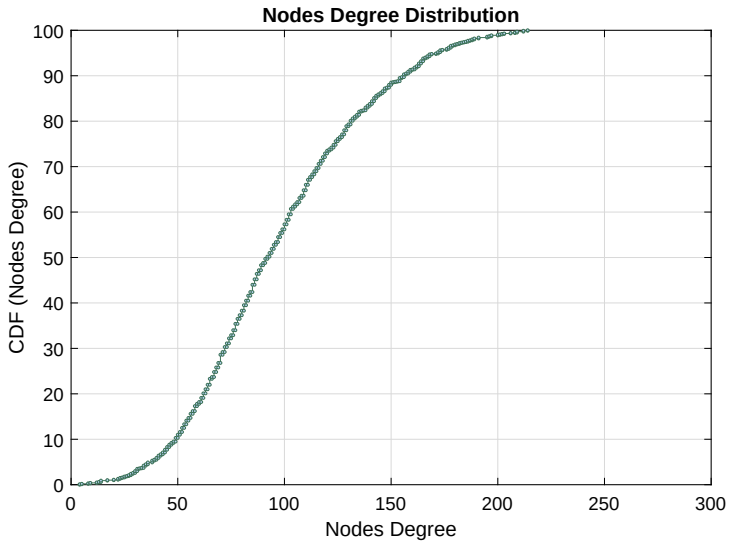

Figure 10. Degree distribution; CDF of nodes degree distribution in the synthetic graph.

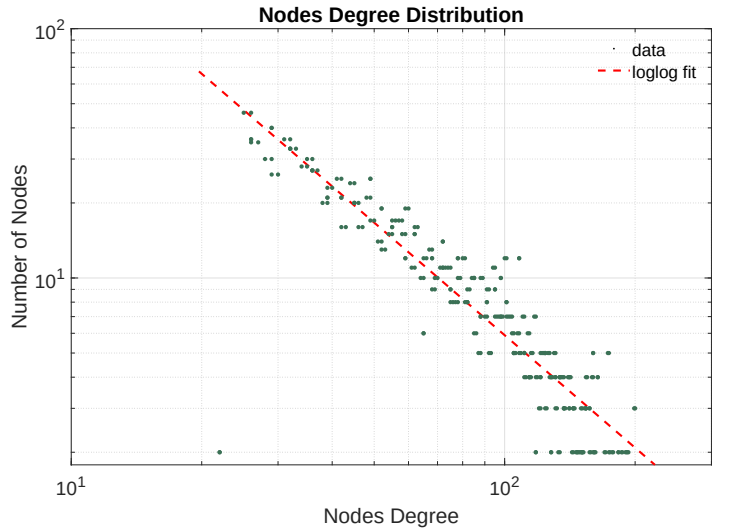

Figure 11. Degree distribution; nodes degree distribution representation in log–log graphs. The synthetic network obeys the power-law degree distribution.

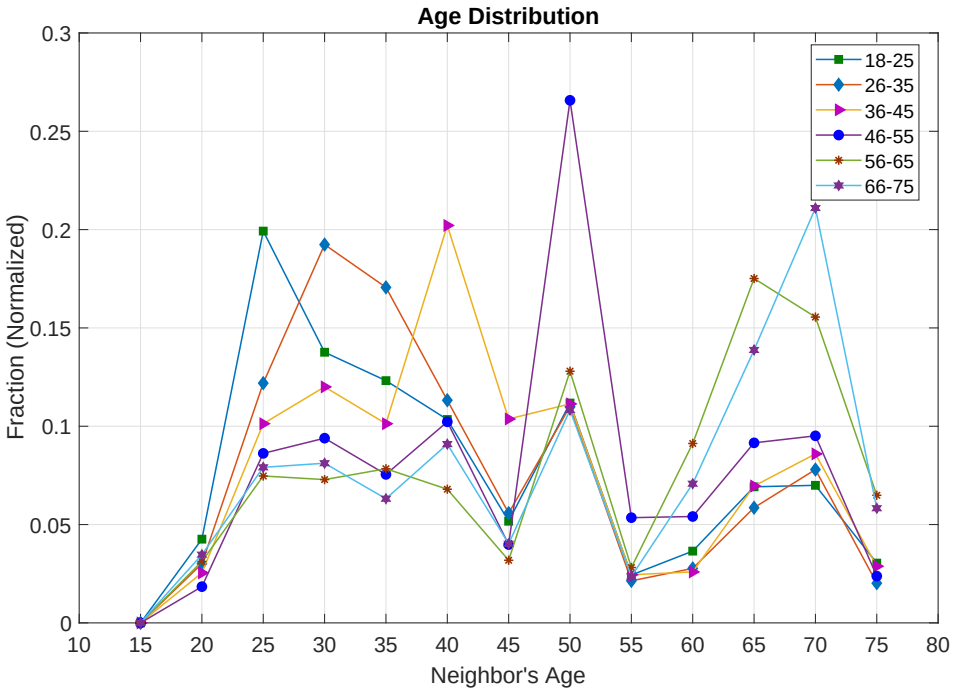

Figure 12. Age and relationships distribution; similar age users are more likely to connect with each other.

In [51], it is concluded that, in SNs, users establish links with other users that belong to the same locality. Figure 13 shows the geodesic distance between connected users. By comparing the geodesic distances between connected nodes in our synthetic SN graph, it is found to have a similar distribution. In Figure 13, it can be observed that most of the users are connected with other users, which belongs to the nearby geo-location. In [51], it was observed that, on Facebook, a user is more likely to establish a connection with a user from the same country and, from simulation, it was concluded that about 84% of the total edges are within the country. Therefore, the SN users can be divided into clusters of locations for analysis.

Figure 14 shows the relationship of demographic similarity with the links' establishment between nodes in our synthetic SN. It can be observed that, like real-life SNs, a high proportion of links are established between nodes that are demographically more similar. The links' establishment in our model depends on multiple attribute and demographic similarity is one of the contributing factors and is not totally dependent on demographic similarity. In addition, the demographic attributes were distributed randomly, and it is more likely for most of the nodes to have a different combination of demographic attributes and only a small fraction of users' have demographic similarity in the range 0.8–1. Figure 15 shows the interest similarity of connected users in our synthetic SN graph. It is observed that the connected users have high interest similarity. This property, also referred to as "birds of a feather flock together", is observed in real-life SNs, i.e., Facebook, Twitter, and Instagram. This pattern is the base for many marketing and business models on SN. In the real-life SNs, new friends are recommended to the users based on their common or shared interests. However, the fraction of links for very small interests similarity are comparable because there are some items that both connecting users' have not rated/experienced. In addition, the connection formation in our model depends on multiple attributes and, as other attributes, the opinion similarity also contributes to the cumulative similarity of all attributes, and there exist users' having very high demographics, and spatial similarity but low opinion similarity. Such users are more likely to connect due to their demographic and spatial similarities, and vice versa.

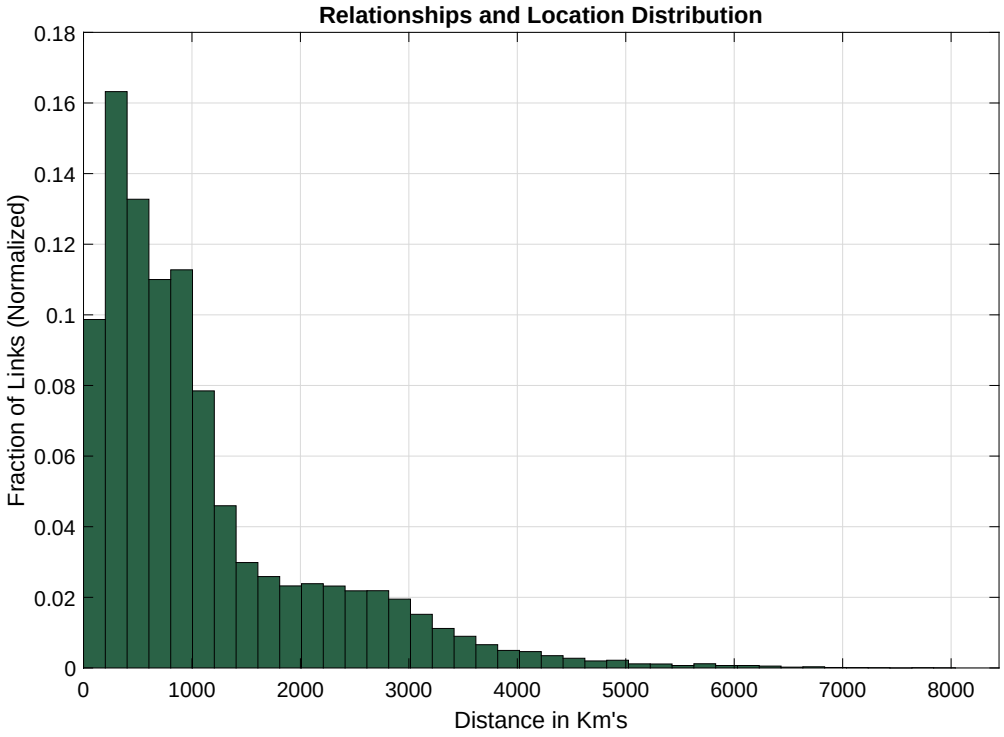

Figure 13. Location distribution and relationships; users with less geo-distance are more probable to connect.

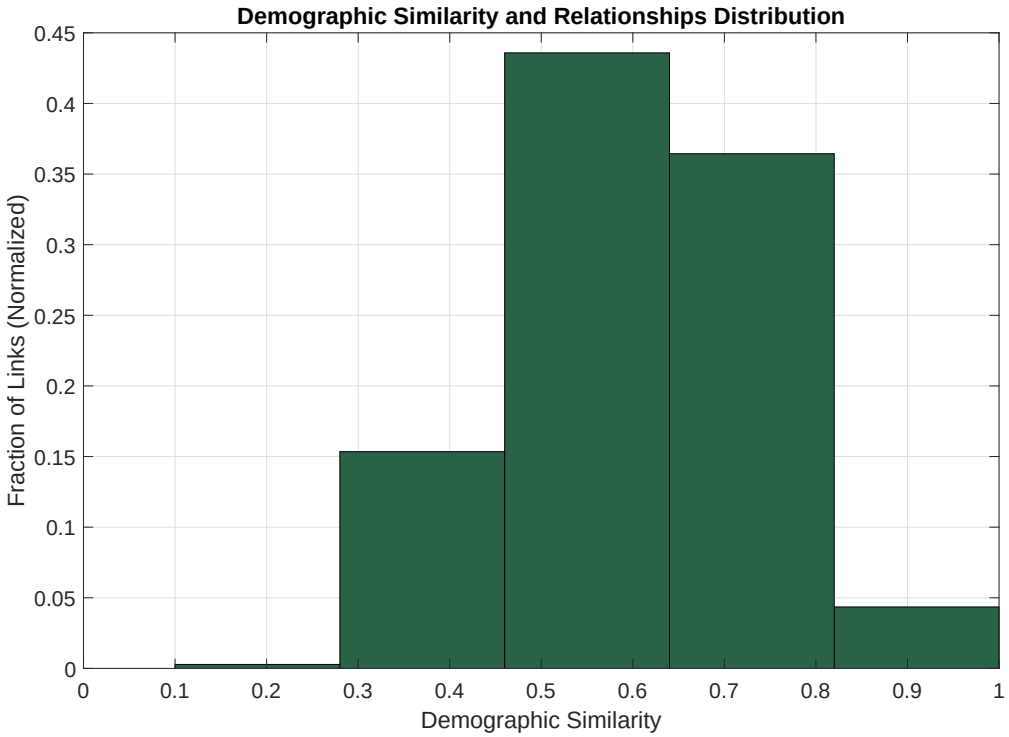

Figure 14. Demographic similarity and relationship distribution; users with high demographic similarity are more probable to connect with each other.

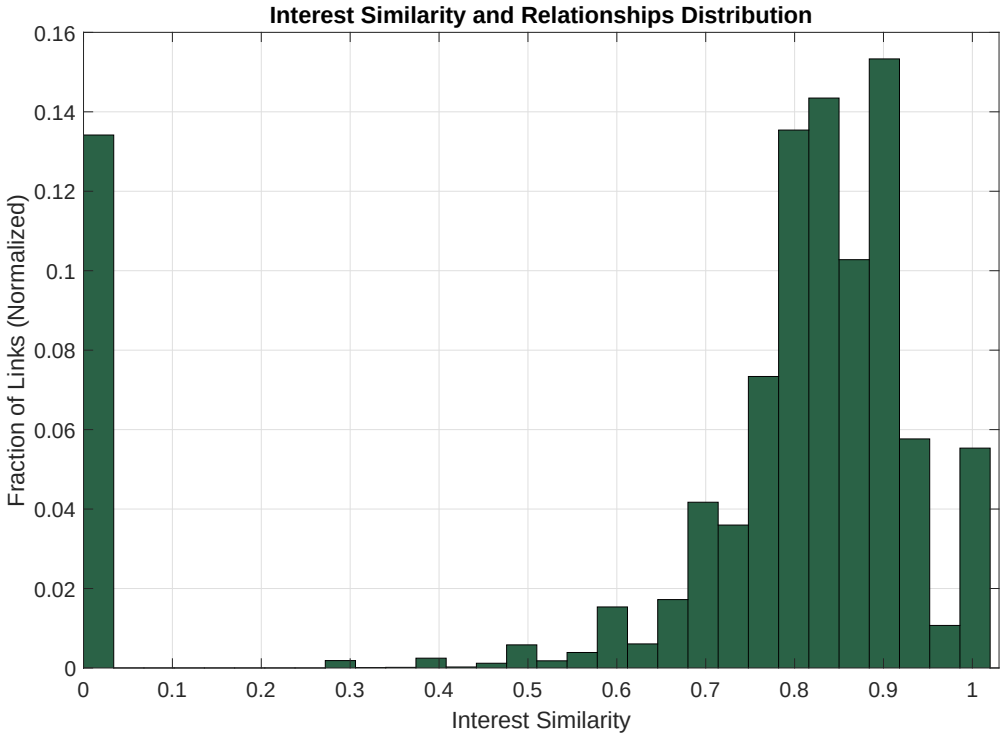

Figure 15. Opinion similarity and relationships distribution; users with the same opinion have a high tendency to connect.

Figure 16 shows a CDF of clustering coefficients against a fraction of users, for our graph in comparison with the BA model. From Figure 16, we can see that the clustering coefficient is improved by our model because, in our model, the connections are established between the most similar users that are more likely to lie in the same community. From Figure 16, we can observe that most of the users in our synthetic graph have relatively high clustering coefficients and the range of clustering coefficient for more than 90% users is from $[0.4, 0.7]$.

In Figure 17, the distribution of social interaction is shown. It is observed that, like real-life SN, the interactions generated by our model obeys the power-law distribution, i.e., and only a small proportion of users are highly active and produce most of the SN interactions. Such users are the most influential users in SN. The distribution of our synthetic social interactions was close to the real-life SN interactions, as observed by [10].

In Table 3, the synthetic SN is compared with the real-life SN datasets to validate their structural closeness based on standard graph measures, i.e., average degree, average path length, average clustering coefficient, modularity, graph density, and graph diameter. We used SN datasets publicly available on Stanford Large Network Dataset Collection [83] and Arizona State University Social Computing Data Repository [84]. The datasets we used are Livejournal [85], Facebook [86], Twitter [84], Friendster [85], Amazon [85], Douban [84], Digg [84], Karate Club [87], Les Mesrible [88], Netsciences [11], Enron email [89,90], CollegeMsg [91], contact network [92],and random graph generated in Gephi. These datasets are accessible and widely used. From Table 3, we can see that, based on the comparison measures, the performance of our model is comparable to that of the real-life SN datasets.

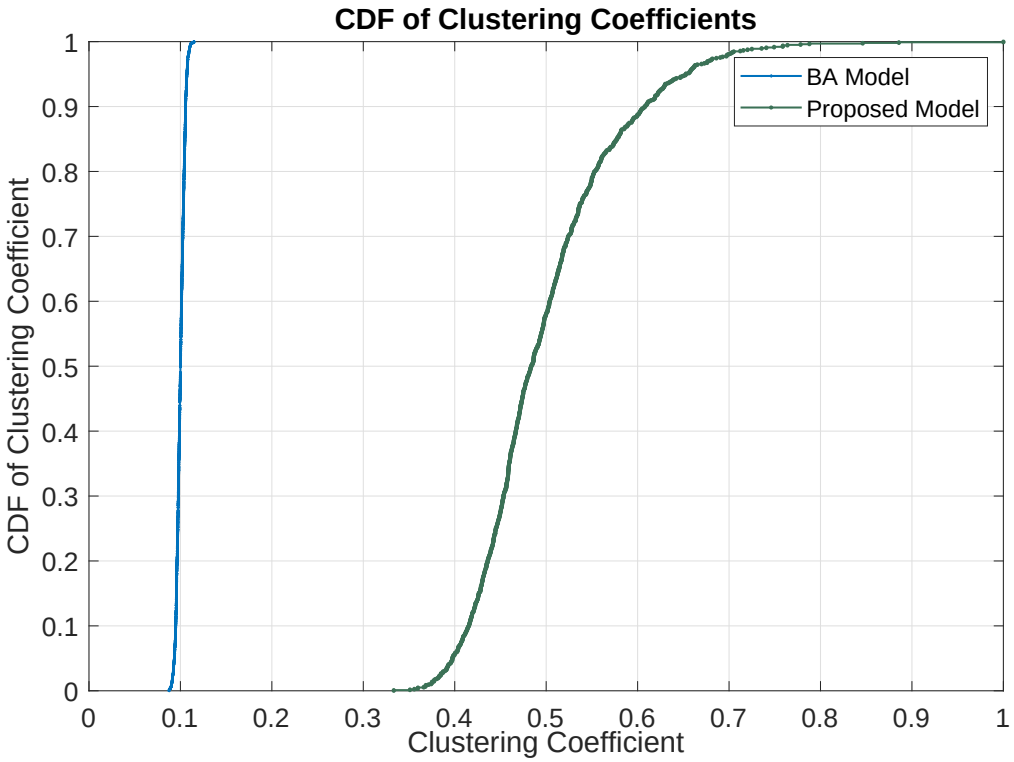

Figure 16. Clustering coefficient distribution; comparison of clustering co-efficient of the proposed model with the BA-model (preferential attachment based approach).

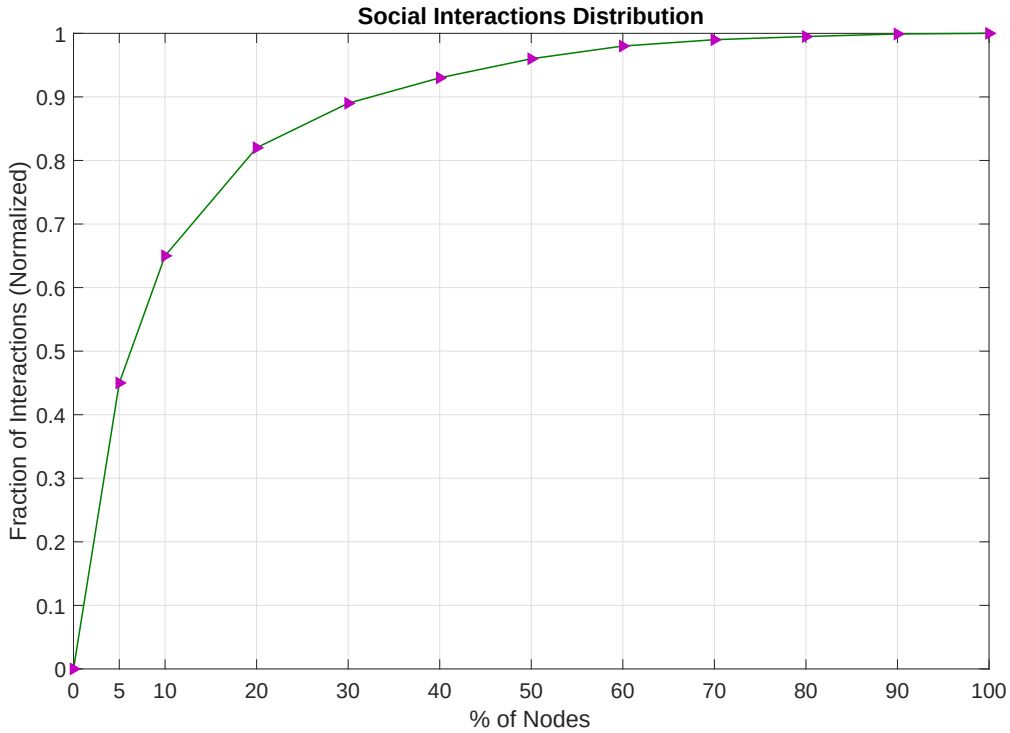

Figure 17. Social interactions distribution; most of the interactions are generated by only a small fraction of users. Such users are referred to as the most active users.

Table 3. Comparison of proposed social network graph with real-life social networks data, and existing SN evolution models based on various metrics and attributes.

	Network	# Nodes	# Edges	Average Degree	Average Path Length	Average Clustering Coefficient	Modularity	Graph Density	Graph Diameter	User Attributes	User-Items Ratings	User Interactions
Datasets	Livejournal [85]	3,997,962	34,681,189	8.67	6.5	0.28	0.15	0...	17	No	No	No
	Facebook [86]	4,039	88,234	43.69	3.69	0.617	0.83	0.011	8	Yes	No	No
	Twitter [84]	472,753	1,048,575	2.21	5.991	0.012	0.606	0...	16	No	No	No
	Friendster [85]	65,608,366	1,806,067,135	27.52	5.8	0.16	0.24	0...	32	No	No	No
	Amazone [85]	334,863	925,872	2.73	15	0.39	0.06	0...	44	No	No	No
	Douban [84]	154,908	327,162	4.22	57.81	0.048	0.57	0...	9	No	No	No
	Digg [84]	256,092	1,019,033	7.96	138.47	0.138	0.574	0...	22	No	No	No
	Karate club [87]	34	78	2.29	2.40	0.58	0.415	0.139	5	No	No	No
	Lesmeserible [88]	77	254	6.59	2.64	0.74	0.57	0.087	5	No	No	No
	Netsciences [11]	1,589	2,742	3.45	5.82	0.878	0.955	0.002	17	No	No	Yes
	Enron Email [89,90]	36,692	183,831	5.01	3.99	0.49	0.34	0...	11	No	No	Yes
	CollegeMsg [91]	1,899	20,296	21.37	3.05	0.138	0.356	0.011	8	No	No	Yes
	Contact Network [92]	236	5,899	49.99	1.86	0.50	0.37	0.21	3	No	No	No
Previous Models	Random Graph [35]	1,000	47,791	47.79	1.90	0.10	0.08	0.096	3	No	No	No
	ER Model [36]	1,000	49,903	49.90	1.9	0.10	0.083	0.010	3	No	No	No
	SW Model [38]	1,000	74,999	74.99	1.85	0.15	0.063	0.015	2	No	No	No
	RMAT [41]	1,000	50,397	50.397	2.16	0.119	0.175	0.051	4	No	No	No
	BA Model [40]	1,000	45,875	45.87	1.92	0.127	0.132	0.092	3	No	No	No
Propsed Model	PM1000	1,000	46,780	46.78	2.49	0.49	0.53	0.094	6	Yes	Yes	Yes
	PM2000	2,000	98,725	49.36	2.32	0.51	0.55	0.098	6	Yes	Yes	Yes
	PM5000	5,000	249,950	49.99	2.29	0.53	0.56	0.099	5	Yes	Yes	Yes
	PM10000	10,000	485,539	48.55	2.39	0.50	0.53	0.096	6	Yes	Yes	Yes

As discussed in Section 3, the convergence of our model depends on the minimum degree that each node needs to achieve and the network average degree. Table 4 shows the simulation results of our model with different initial conditions. From Table 4, it can be observed that the number of links in the network is increased as the value of minimum degree and average degree increased. Our model converges when one of the two initial limiting conditions is achieved. For the simulation with no limit on initial conditions, a time-out was set to converge the model.

Table 4. Number of links' dependency on initial conditions, i.e., minimum node degree and network average degree for network of 1000 nodes.

Average Degree	Min. Degree	# Edges
100	No Limit	49,919
	10	46,780
	20	49,590
	50	55,870
150	No Limit	83,810
	10	64,238
	20	69,569
	50	71,330
200	No Limit	89,711
	10	61,958
	20	63,290
	50	83,891
No Limit	No Limit	101,158
	10	49,191
	20	64,429
	50	83,442

As discussed in Section 3, the time complexity of our proposed model depends on the number of nodes N, number of similarity parameters, and slightly on the number of initial seeds K. Table 5 shows results of simulation time for our model. We have simulated our model in two types of settings, i.e., pre-computed similarity matrix and with similarities computation. We have found from simulation that the similarity matrix computation is very time-consuming and it hugely increases the time complexity of our model. The model simulation time is also raised as the number of nodes is increased. We have simulated our model for thousands of nodes. The simulation was done in MATLAB 2018 (The MathWorks, Inc., Natick, MA, USA). The system we used for simulation was LG (LG Electronics Nanjing Displays Co. Ltd., Nanjing, China), Core i5, 2.3 GHz processor, 8 GHz memory, and Windows 10 Pro 64-bit (Microsoft, Redmond, WA, USA). For space complexity, on the described system setting, when we increase the number of nodes to more than 20,000, a memory error is shown by MATLAB and also the model needs several days to simulate for 20,000 nodes; therefore, we have shown the results of our model for up to 10,000 nodes. Therefore, for current settings and settings, the memory issue can occur.

Table 5. Time complexity of the proposed model with different numbers of nodes, and number of seeds.

	# Nodes	K	Time in sec.
Pre-Computed Similarity Matrix Provided to the Model	1000	10	2.45
		100	2.07
	2000	10	22.48
		100	18.97
	5000	10	297.56
		100	292.34
	10,000	10	2693.58
		100	2335.60
With Similarity Computation	1000	10	103.96
		100	98.24
	2000	10	1589.87
		100	1509.53
	5000	10	18,190.87
		100	18,011.72
	10,000	10	234,481.23
		100	234,392.65

For implementation, we have generated random geo-locations in the North-American region, and user interests' information using MATLAB built-in functions and commands. We have generated seven demographic attributes with different percentages and distribution, as discussed in Section 3 and shown in Table 2, by using MATLAB. The main program consists of nodes profile generation, similarities computation, similarity based clustering and synthetic graph generation.

As a result of the simulations, a synthetic dataset is produced that can be used by researchers to evaluate their social network models. Our model is scalable and general, and researchers can generate N number of nodes with different attributes depending on their target applications.

5. Conclusions

In this paper, we proposed an SN evolution model based on the property of homophily and preferential attachments. The proposed model is simulated, to generate a synthetic SN graph, evaluated, based on SN structural and similarity properties, and validated by using many graph measures and being compared with real-life SN datasets.

From simulations, it is observed that the network graph, synthesized by our model, bears the real-life OSN properties. Like real-life SN, our synthetic graph has a scale-free nature and obeys the power-law degree distribution. The synthetic SN graph comprises the small-world effect, also referred to as six degrees of separation. The average distance between the nodes was found to be 2.49 and the graph diameter was 6, which is comparable to that of the real-life SN. The resultant graph was found to be fully connected, having short average path length, comparable clustering coefficient, average degree, modularity, and graph diameter. The networks with these properties are categorized to be small world networks [38].

We also observed the relationship between SN structure and user attributes. These attributes are demographic attributes, geodesic distance, and user interest. In real-life SNs, like Facebook, it is observed that the users are more likely to establish a connection with other users of similar demographic attributes and belong to the same locality [51]. From simulation, similar patterns have been observed in our synthetic graphs.

Moreover, as a result of this work, a generic SN synthetic dataset is obtained, which contains SN nodal attributes, such as demographic attributes, spatial information and user interests. The resultant SN graph obeys the SN structural properties, which exhibits its closeness to real-life SN graphs.

The proposed model is scalable, generic, and is based on the phenomenon and psychology behind the real-life SN links formation. This model can be used, by the researcher in the area of SNA, for SN synthetic datasets to evaluate their models and applications in many SN domains. The researchers,

while using this model for graph generation, can set attributes, and attribute values distribution and weights according to their target applications.

Author Contributions: The authors contributed equally to this work overall.

Funding: This research was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1A2B1010817).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rose, D.E.; Bornstein, J.J.; Tiene, K.; Ponceleón, D.B. System for Ranking the Relevance of Information Objects Accessed by Computer Users. U.S. Patent Application 10/388,362, 26 October 2010.
2. Mui, L. Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2002.
3. Yu, H.; Kaminsky, M.; Gibbons, P.B.; Flaxman, A. Sybilguard: Defending against sybil attacks via social networks. *ACM SIGCOMM Comput. Commun. Rev.* **2006**, *36*, 267–278. [CrossRef]
4. Girvan, M.; Newman, M.E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821–7826. [CrossRef] [PubMed]
5. Lancichinetti, A.; Fortunato, S. Community detection algorithms: A comparative analysis. *Phys. Rev. E* **2009**, *80*, 056117. [CrossRef] [PubMed]
6. Ullah, F.; Lee, S. Community clustering based on trust modeling weighted by user interests in online social networks. *Chaos Solitons Fractals* **2017**, *103*, 194–204. [CrossRef]
7. Ahmad, K.; Pogorelov, K.; Riegler, M.; Conci, N.; Halvorsen, P. Social media and satellites. *Multimed. Tools Appl.* **2018**, 1–39. [CrossRef]
8. Radicchi, F.; Castellano, C.; Cecconi, F.; Loreto, V.; Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 2658–2663. [CrossRef] [PubMed]
9. Wilson, C.; Sala, A.; Puttaswamy, K.P.; Zhao, B.Y. Beyond social graphs: User interactions in online social networks and their implications. *ACM Trans. Web (TWEB)* **2012**, *6*, 17. [CrossRef]
10. Durr, M.; Protschky, V.; Linnhoff-Popien, C. Modeling social network interaction graphs. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), Istanbul, Turkey, 26–29 August 2012; IEEE Computer Society: Washington, DC, USA, 2012; pp. 660–667.
11. Newman, M.E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef] [PubMed]
12. Kempe, D.; Kleinberg, J.; Tardos, É. Maximizing the spread of influence through a social network. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–27 August 2006; ACM: New York, NY, USA, 2003; pp. 137–146.
13. Guille, A.; Hacid, H.; Favre, C.; Zighed, D.A. Information diffusion in online social networks: A survey. *ACM Sigmod Rec.* **2013**, *42*, 17–28. [CrossRef]
14. Ullah, F.; Lee, S. Social content recommendation based on spatial-temporal aware diffusion modeling in social networks. *Symmetry* **2016**, *8*, 89. [CrossRef]
15. Al Qundus, J.; Paschke, A. Investigating the Effect of Attributes on User Trust in Social Media. In Proceedings of the International Conference on Database and Expert Systems Applications, Regensburg, Germany, 3–6 September 2018; Springer: Berlin, Germany, 2018; pp. 278–288.
16. Bai, Y.; Deng, G.; Zhang, L.; Wang, Y. A Measuring Method for User Similarity based on Interest Topic. *Int. J. Perform. Eng.* **2018**, *14*, 691–698. [CrossRef]
17. Sampson, S. A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships. Ph.D. Thesis, Cornell University, Ithaca, NY, USA, 1968.
18. Burt, M.R. Cultural myths and supports for rape. *J. Personal. Soc. Psychol.* **1980**, *38*, 217. [CrossRef]
19. Johnson, J.C. Social networks and innovation adoption: A look at Burt's use of structural equivalence. *Soc. Netw.* **1986**, *8*, 343–364. [CrossRef]
20. Johnsen, E.C. Structure and process: Agreement models for friendship formation. *Soc. Netw.* **1986**, *8*, 257–306. [CrossRef]

21. Kumar, R.; Novak, J.; Raghavan, P.; Tomkins, A. Structure and evolution of blogspace. *Commun. ACM* **2004**, *47*, 35–39. [CrossRef]
22. Liben-Nowell, D.; Novak, J.; Kumar, R.; Raghavan, P.; Tomkins, A. Geographic routing in social networks. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 11623–11628. [CrossRef] [PubMed]
23. Dodds, P.S.; Muhamad, R.; Watts, D.J. An experimental study of search in global social networks. *Science* **2003**, *301*, 827–829. [CrossRef] [PubMed]
24. Adamic, L.; Adar, E. How to search a social network. *Soc. Netw.* **2005**, *27*, 187–203. [CrossRef]
25. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994; Volume 8.
26. Strogatz, S.H. Exploring complex networks. *Nature* **2001**, *410*, 268–276. [CrossRef] [PubMed]
27. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47. [CrossRef]
28. Newman, M.E. The structure and function of complex networks. *SIAM Rev.* **2003**, *45*, 167–256. [CrossRef]
29. Dorogovtsev, S.N.; Mendes, J.F. Evolution of networks. *Adv. Phys.* **2002**, *51*, 1079–1187. [CrossRef]
30. Dorogovtsev, S.N.; Mendes, J.F. *Evolution of Networks: From Biological Nets to the Internet and WWW*; OUP Oxford: Oxford, UK, 2013.
31. Kleinberg, J. Complex networks and decentralized search algorithms. In Proceedings of the International Congress of Mathematicians (ICM), Madrid, Spain, 22–30 August 2006; Volume 3, pp. 1019–1044.
32. Kumar, R.; Novak, J.; Raghavan, P.; Tomkins, A. On the bursty evolution of blogspace. *World Wide Web* **2005**, *8*, 159–178. [CrossRef]
33. Fetterly, D.; Manasse, M.; Najork, M.; Wiener, J.L. A large-scale study of the evolution of Web pages. *Softw. Pract. Exp.* **2004**, *34*, 213–237. [CrossRef]
34. Ntoulas, A.; Cho, J.; Olston, C. What's new on the web? The evolution of the web from a search engine perspective. In Proceedings of the 13th International Conference on World Wide Web, New York, NY, USA, 17–20 May 2004; ACM: New York, NY, USA, 2004; pp. 1–12.
35. Newman, M.E.; Watts, D.J.; Strogatz, S.H. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 2566–2572. [CrossRef] [PubMed]
36. Erdds, P.; Rényi, A. On random graphs I. *Publ. Math. Debr.* **1959**, *6*, 290–297.
37. Bollobás, B.; Fulton, W.; Katok, A.; Kirwan, F.; Sarnak, P. *Cambridge Studies in Advanced Mathematics*; Random Graphs; Cambridge University Press: New York, NY, USA, 2001; Volume 73.
38. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world'networks. *Nature* **1998**, *393*, 440–442. [CrossRef] [PubMed]
39. Barabási, A.L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509–512. [PubMed]
40. Barabási, A.L. Scale-free networks: A decade and beyond. *Science* **2009**, *325*, 412–413. [CrossRef] [PubMed]
41. Chakrabarti, D.; Zhan, Y.; Faloutsos, C. R-MAT: A recursive model for graph mining. In Proceedings of the 2004 SIAM International Conference on Data Mining, Lake Buena Vista, FL, USA, 22–24 April 2004; pp. 442–446.
42. Lancichinetti, A.; Fortunato, S.; Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **2008**, *78*, 046110. [CrossRef] [PubMed]
43. Kleinberg, J.M. Authoritative sources in a hyperlinked environment. *J. ACM* **1999**, *46*, 604–632. [CrossRef]
44. Chung, F.; Lu, L. Connected components in random graphs with given expected degree sequences. *Ann. Comb.* **2002**, *6*, 125–145. [CrossRef]
45. Chung, F.; Lu, L. The average distance in a random graph with given expected degrees. *Internet Math.* **2004**, *1*, 91–113. [CrossRef]
46. Waxman, B.M. Routing of multipoint connections. *IEEE J. Sel. Areas Commun.* **1988**, *6*, 1617–1622. [CrossRef]
47. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B-Condens. Matter Complex Syst.* **1998**, *4*, 131–134. [CrossRef]
48. Menczer, F. Evolution of document networks. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5261–5265. [CrossRef] [PubMed]
49. McPherson, M.; Smith-Lovin, L.; Cook, J.M. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **2001**, *27*, 415–444. [CrossRef]
50. Şimşek, Ö.; Jensen, D. Navigating networks by using homophily and degree. *Proc. Natl. Acad. Sci. USA* **2008**. [CrossRef]

51. Ugander, J.; Karrer, B.; Backstrom, L.; Marlow, C. The anatomy of the facebook social graph. *arXiv* **2011**, arXiv:1111.4503
52. Papadopoulos, F.; Kitsak, M.; Serrano, M.Á.; Boguná, M.; Krioukov, D. Popularity versus similarity in growing networks. *Nature* **2012**, *489*, 537–540. [CrossRef] [PubMed]
53. Huber, G.A.; Malhotra, N. Political homophily in social relationships: Evidence from online dating behavior. *J. Politics* **2017**, *79*, 269–283. [CrossRef]
54. Neyer, F.J.; Lang, F.R. Blood is thicker than water: Kinship orientation across adulthood. *J. Personal. Soc. Psychol.* **2003**, *84*, 310. [CrossRef]
55. Doherty, N.A.; Feeney, J.A. The composition of attachment networks throughout the adult years. *Pers. Relationsh.* **2004**, *11*, 469–488. [CrossRef]
56. Gerich, J.; Lehner, R. Collection of ego-centered network data with computer-assisted interviews. *Methodology* **2006**, *2*, 7–15. [CrossRef]
57. Van Tilburg, T. Losing and gaining in old age: Changes in personal network size and social support in a four-year longitudinal study. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **1998**, *53*, S313–S323. [CrossRef]
58. Said, A.; De Luca, E.W.; Albayrak, S. How social relationships affect user similarities. In Proceedings of the 2010 International Conference on Intelligent User Interfaces Workshop on Social Recommender Systems, Hong Kong, China, 7–10 February 2010.
59. Hanani, U.; Shapira, B.; Shoval, P. Information filtering: Overview of issues, research and systems. *User Model. User-Adapt. Interact.* **2001**, *11*, 203–259. [CrossRef]
60. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl.-Based Syst.* **2013**, *46*, 109–132. [CrossRef]
61. Xie, J.; Li, X. Make best use of social networks via more valuable friend recommendations. In Proceedings of the 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), Yichang, China, 21–23 April 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1112–1115.
62. Chin, A.; Xu, B.; Wang, H. Who should I add as a friend? A study of friend recommendations using proximity and homophily. In Proceedings of the 4th International Workshop on Modeling Social Media, Prague, Czech Republic, 23 September 2013; ACM: New York, NY, USA, 2013; p. 7.
63. Pouli, V.; Kafetzoglou, S.; Tsiropoulou, E.E.; Dimitriou, A.; Papavassiliou, S. Personalized multimedia content retrieval through relevance feedback techniques for enhanced user experience. In Proceedings of the 2015 13th International Conference on Telecommunications (ConTEL), Graz, Austria, 13–15 July 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–8.
64. Stai, E.; Kafetzoglou, S.; Tsiropoulou, E.E.; Papavassiliou, S. A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content. *Multimed. Tools Appl.* **2018**, *77*, 283–326.
65. Gauch, S.; Speretta, M.; Chandramouli, A.; Micarelli, A. User profiles for personalized information access. In *The Adaptive Web*; Springer: Berlin, Germany, 2007; pp. 54–89.
66. Ajrouch, K.J.; Blandon, A.Y.; Antonucci, T.C. Social networks among men and women: The effects of age and socioeconomic status. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **2005**, *60*, S311–S317. [CrossRef]
67. Quinn, D.; Chen, L.; Mulvenna, M. Does age make a difference in the behaviour of online social network users? In Proceedings of the Internet of Things (iThings/CPSCom), 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing, Dalian, China, 19–22 October 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 266–272.
68. Cornwell, B.; Laumann, E.O.; Schumm, L.P. The social connectedness of older adults: A national profile. *Am. Sociol. Rev.* **2008**, *73*, 185–203. [CrossRef] [PubMed]
69. Cornwell, B. Age trends in daily social contact patterns. *Res. Aging* **2011**, *33*, 598–631. [CrossRef]
70. Marcum, C.S. Age differences in daily social activities. *Res. Aging* **2013**, *35*, 612–640. [CrossRef] [PubMed]
71. Pérez-Rosés, H.; Sebé, F. Synthetic generation of social network data with endorsements. *J. Simul.* **2015**, *9*, 279–286. [CrossRef]
72. Nettleton, D.F. A synthetic data generator for online social network graphs. *Soc. Netw. Anal. Min.* **2016**, *6*, 44. [CrossRef]
73. Sagduyu, Y.E.; Grushin, A.; Shi, Y. Synthetic Social Media Data Generation. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 605–620. [CrossRef]
74. Facebook Analytics. Available online: https://www.facebook.com/analytics/1701892993437661/?-section=people_demographics/ (accessed on 31 July 2018).

75. Fan Page List. 2015. Available online: http://www.fanpagelist.com/category/top_users/ (accessed on 6 August 2018).
76. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, 42, 30–37. [CrossRef]
77. Han, X.; Wang, L.; Park, S.; Cuevas, A.; Crespi, N. Alike people, alike interests? A large-scale study on interest similarity in social networks. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Beijing, China, 17–20 August 2014; IEEE Press: Piscataway, NJ, USA, 2014, pp. 491–496.
78. Dillon, M. *Introduction to modern information retrieval: G. Salton and M. McGill*; McGraw-Hill: New York, NY, USA, 1983; pp. 491–496, ISBN 0-07-054484-0.
79. Sinnott, R.W. Virtues of the Haversine. *Sky Telesc.* **1984**, *68*, 159.
80. Palmer, M.C. Calculation of distance traveled by fishing vessels using GPS positional data: A theoretical evaluation of the sources of error. *Fish. Res.* **2008**, *89*, 57–64. [CrossRef]
81. Jaccard, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull. Soc. Vaud. Sci. Nat.* **1901**, *37*, 241–272.
82. Huberman, B.A.; Romero, D.M.; Wu, F. Social networks that matter: Twitter under the microscope. *arXiv* **2008**, arXiv:0812.1045.
83. Leskovec, J.; Krevl, A. {*SNAP Datasets*}:{*Stanford*} *Large Network Dataset Collection*; Stanford University: Stanford, CA, USA, USA, 2014. Available online: http://snap.stanford.edu/data (accessed on 19 November 2018)
84. Zafarani, R.; Liu, H. Social Computing Data Repository at ASU ["http://socialcomputing.asu.edu/"]. Tempe, AZ: Arizona State University, School of Computing. *Inform. Decis. Syst. Eng.* **2009**.
85. Yang, J.; Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.* **2015**, *42*, 181–213. [CrossRef]
86. Leskovec, J.; Mcauley, J.J. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems (NIPS) Foundation, Inc.: La Jolla, CA, USA, 2012; pp. 539–547.
87. Zachary, W.W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **1977**, *33*, 452–473. [CrossRef]
88. Knuth, D.E. *The Stanford GraphBase: A Platform for Combinatorial Computing*; ACM Press: New York, NY, USA, 1993.
89. Klimt, B.; Yang, Y. Introducing the Enron Corpus. In Proceedings of the 2004 CEAS, Mountain View, CA, USA, 30–31 July 2004.
90. Leskovec, J.; Lang, K.J.; Dasgupta, A.; Mahoney, M.W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Math.* **2009**, *6*, 29–123. [CrossRef]
91. Panzarasa, P.; Opsahl, T.; Carley, K.M. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 911–932. [CrossRef]
92. Stehlé, J.; Voirin, N.; Barrat, A.; Cattuto, C.; Isella, L.; Pinton, J.F.; Quaggiotto, M.; Van den Broeck, W.; Régis, C.; Lina, B.; et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE* **2011**, *6*, e23176. [CrossRef] [PubMed]