Soc. Netw. Anal. Min. (2013) 3:597–609
DOI 10.1007/s13278-013-0105-0

ORIGINAL ARTICLE

Model for generating artificial social networks having community structures with small-world and scale-free properties

Arnaud Sallaberry · Faraz Zaidi · Guy Melançon

Received: 15 October 2012 / Revised: 30 January 2013 / Accepted: 11 February 2013 / Published online: 6 March 2013

Abstract Recent interest in complex systems and specially social networks has catalyzed the development of numerous models to help understand these networks. A number of models have been proposed recently where they are either variants of the small-world model, the preferential attachment model or both. Three fundamental properties attributed to identify these complex networks are high clustering coefficient, small average path length and the vertex connectivity following power-law distribution. Different models have been presented to generate networks having all these properties. In this study, we focus on social networks and another important characteristic of these networks, which is the presence of community structures. Often misinterpret with the metric called clustering coefficient, we first show that the presence of community structures is indeed different from having high clustering coefficient. We then define a new network generation model which exhibits all the fundamental properties of complex networks along with the presence of community structures.

A. Sallaberry
University of California, San Francisco, USA
e-mail: asallaberry@ucdavis.edu

F. Zaidi (✉)
Karachi Institute of Economics and Technology,
Karachi, Pakistan
e-mail: faraz@pafkiet.edu.pk

G. Melançon
CNRS UMR 5800 LaBRI and INRIA Bordeaux–Sud-Ouest,
Talence Cedex, France
e-mail: melancon@labri.fr

Keywords Social networks · Communities · Small-world networks · Scale-free networks · Network generation models

1 Introduction

Most of the real-world systems can be modeled as graphs where different fields of study use extensively the node-link representation to represent information. Wide use of this representation has been witnessed in social networks (Wasserman and Faust 1994). A social network can be defined as a set of people, or groups of people interacting with each other (Scott 2000; Wasserman and Faust 1994). These interactions can be classified into several types like friendship or business relationship. Although many examples have been studied for social networks, two classic examples that have attracted extensive attention in the computer science community and the social network community at large, are the Actor collaboration network from Internet Movie database (IMDB) and the Science collaboration network.

Social network modeling and analysis allows us to understand the different types of relationships that can either facilitate or impede knowledge creation and transfer in a society on the whole, in an organization in particular, and in individuals, providing an insight into the underlying patterns and the social structures present in these networks (Scott 2011; Cross et al. 2000).

The study of networks in general, and of social networks in particular, was revived by the pioneering work of Watts and Strogatz (1998) on the properties of small-world networks. Equally important was the work from Barabasi and Albert on the growth of networks and the property of scale-free degree distribution (Barabási and Albert 1999).

Although random graphs had been studied extensively in the past (Newman 2003), most of the real-world networks have the properties of small-world and scale-free networks.

A small-world network, as defined by Watts and Strogatz (1998), is a network when compared with a random graph of same node–edge density, has higher clustering coefficient and the typical distance between any two nodes scales as the logarithm of the number of nodes. The two structural properties used to define a small-world network are the average path length and the clustering coefficient. The most popular manifestation of the concept of low average path length is the 'six degrees of separation', uncovered by the social psychologist Stanley Milgram, who concluded that there was a path of acquaintances with a typical length of about 6 between most pairs of people in the USA (Milgram 1967). More precisely, the path length refers to the minimum number of edges traversed to go from node A to node B. The average path length is the average calculated for all pair of nodes in a network. Another important characteristic of these networks is the average clustering coefficient of nodes (Watts and Strogatz 1998), sometimes referred as Transitivity (Newman 2003) to avoid confusion from the concept of community structure (or clusters) (Scott 2000; Wasserman and Faust 1994). The concept is very well known in social networks and can be described as the friend of your friend is likely to be your friend. Mathematically, for a graph G with nodes V and edges E, the clustering coefficient (CC) for a node v is defined as:

$$\mathrm{CC}(v) = \frac{r(N(v))}{|N(v)|(|N(v)| - 1)/2}$$

where $u \in V$ and $(u, v) \in E$. The neighborhood of a node v is defined as the set of nodes in the neighborhood of v denoted by $N(v)$. The number of elements in set $N(v)$ is given by $|N(v)|$. And the notation $r(N(v))$ represents the number of edges (u, w) such that $u, w \in N(v)$ and $u \neq w$. To calculate the clustering coefficient of the entire network, we take the average for all nodes in the network.

A scale-free network is a network in which a few nodes have a very high number of connections (degree) and lots of nodes are connected to a few nodes. Generally, it was believed that the degree distribution in most networks follows a poisson distribution but in reality, real-world networks have a highly skewed degree distribution. These networks have no characteristic scales for the degrees, hence they are called scale-free networks (Päivinen 2007). In other words, the degree distribution of scale-free networks follow a power-law distribution (Barabási and Albert 1999).

Apart from the small-world and scale-free properties, another important characteristic of real-world systems and specially social networks is the presence of community structures. A more generic formalism for the term community is clusters, where sociologists use the term *community* (Coleman 1964) as compared to the statistical and data mining domain where people use the term *clusters* (Tryon 1939) to refer to the same concept. Roughly speaking, we like to define a community as a decomposition of a set of entities into 'Natural Groups'. There is no universally accepted definition of clustering (Everitt et al. 2009), most researchers describe a cluster by considering the internal homogeneity and the external separation as the fundamental criteria for defining a cluster (Gordon 1981; Almeida et al. 2012; Jain et al. 1999). A number of algorithms are present in the literature to study clusters and clustering problem in social networks (Newman 2004; Jia et al. 2011; Gilbert et al. 2011).

Different network generation models have been proposed to generate artificial networks having both the small-world and scale-free properties. These models do not generate graphs with community structures by construction as the probability of connection is based solely on the degree of a node and, in some cases, the immediate neighborhood of this node.

In this study, we present a new network generation model which incorporates the properties of small-world and scale-free networks with the additional advantage of having distinct community structures. We explicitly target social networks and argue that using three fundamental concepts from the social network study, we can generate artificial networks replicating real-world social networks. Clustering or community detection remains an important technique to organize and understand complex systems (Girvan and Newman 2002; Jain et al. 1999; Schaeffer 2007; Xu and Wunsch 2005) having a wide range of applications in various fields. For empirical evaluation of algorithms, metrics and analytical methods, it is important to be able to reproduce networks having community structures with small-world and scale-free properties that are close to real-world networks. The proposed model, by construction, incorporates the presence of community structures as we determine the connectivity of nodes based on a pre-generated clustering. We explain the details in the coming sections.

For the sake of discussion and explanatory purposes, throughout this paper, we are going to discuss four social networks. Two well-studied and well-structured social networks are the Actor collaboration network where nodes represent actors and two actors are connected to each other if they appear in a movie together. The other network is the Science collaboration network where nodes represent scientists and two scientists are connected to each other if they have written an artifact together. Apart from these two networks, we consider two hypothetical cases from everyday life. Consider a person joining a new organization

as an employee and a person joining a sports club as a leisure activity. We will refer them as Actor, Author, Employee and Club networks, respectively, throughout this paper.

The rest of the paper is organized as follows: the next section contains a review of the existing network generation models for small-world and scale-free networks. In Sect. 3, we discuss the metric clustering coefficient and compare it with the presence of community structures as being two separate concepts. We then discuss assortativity, transitivity and preferential attachment in social networks in Sect. 4 and argue that with a little modification to these concepts, we can understand how networks having community structures evolve in the real world. We then present a network generation model in Sect. 5. We introduce three networks in Sect. 6 that are used for experimentation and comparative analysis with the artificial network generation models. In Sect. 7, we show that the existing network generation models not only produce graphs without community structures but also have other limitations. Finally, we conclude giving possible future directions of our research in Sect. 8.

2 Existing network generation models

In this section, we review a number of network generation models proposed in the literature having small-world and scale-free properties. A comparative summary of these models is presented in Table 1.

Holme and Kim (2002) modified the well-known Barabasi and Albert model (Barabási and Albert 1999) to obtain graphs that are small world as well as scale free. The idea is pretty simple and effective. A triad formation step is added after the preferential attachment step where every node, introduced in the network, connects not only to node w but also to a randomly chosen neighbor of w, thus resulting in a triad formation. The idea is similar to another model separately proposed by Dorogovtsev and Mendes (2002) in the same year where every new node added to the network is connected to both ends of a randomly chosen link where one of the nodes of this link is selected through preferential attachment. These models inspired Jian-Guo et al. to introduce another similar model (Liu et al. (2005). The network starts with a triangle and at each time step, a new node is added to the network with two edges. The first edge would choose a node to connect preferentially, and the second edge will choose a node connected to the first node, again based on preferential attachment. This is different from the model of Holme and Kim where the second node is randomly chosen. Wang et al. (2006) proposed a similar model to that of Dorogovtsev et al. where at each time step, a new node with two edges is added to the network and the two edges are connected to the two ends of a randomly chosen existing edge.

Table 1 Comparing and summarizing different artificial network generation models for small-world and scale-free networks existing in the literature

Model	Year	Nodes added per step	Edges added per step	Innovation
Holme and Kim	2002	1	m	Triad formation step, forcing a new node to connect to the neighbors of the first node it links to, in order to have triangles and increase the clustering coefficient
Dorogovtsev and Mendes	2002	1	2	Randomly chose an edge and attach both ends of this edge with the new node where the probability of choosing an edge is based on the degree of the nodes at its ends
Jian-Guo et al.	2005	1	2	Each new node attaches to existing node with preferential attachment and chooses one of its neighbors again based on preferential attachment (and not randomly as compared to Holme and Kim)
Wang et al.	2006	1	2	For each edge, a new node with two edges is added, which is attached to both end nodes of the edge. Produces Fractals rather than a random graph
Fu and Liao	2006	1	m	Once a new node attaches to a node, its neighborhood has a higher probability of connecting to the new node
Klemm and Eguiluz	2002	1	m	Activate and deactivate nodes based on node degree where nodes having low degree have a high probability of getting deactivated
Catanzaro et al.	2004	1	m	Assortativity and allows growth in old nodes by allowing new edges
Guillaume and Latapy	2005	1	m	Bipartite structure identified as a fundamental characteristic for real-world graphs
Bu et al.	2007	1	m	n-partite structure, where nodes do not connect to similar node types
Wang and Rong	2008	n	m	Add m new nodes and any two nodes in the m new nodes link together from each other and they link to existing nodes based on preferential attachment

Fu and Liao (2006) proposed another extension to the Barabasi and Albert model which they called the Relatively Preferential Attachment method. At each time step, the newly introduced node in the network connects to a node w with preferential attachment, the nodes in the immediate neighborhood of w have higher probability of connecting to this new node as compared to other nodes. The only difference in this model with the already proposed models is that the new node can have m edges instead of two edges where the value of m is chosen as an initial parameter which remains constant throughout the execution of algorithm.

Klemm and Eguiluz (2002) also proposed a model, where each node of the network is assigned a state variable. A newly generated node is in the *active* state and keeps attaching links until eventually deactivated. At each time step, a new node is added to the network by attaching a link to each of the z active nodes. The new node is set as *active*. One of the existing nodes is deactivated where the probability of a node being deactivated is inversely proportional to its degree, i.e., lower the degree, higher the probability of deactivation. To reduce the average path length of the entire graph, at every step, for each link of the newly added node, it is decided randomly whether the link connects to the active node or it connects to a random node.

Catanzaro et al. (2004) present a model taking into consideration the assortativity of social networks. Assortativity is the tendency of nodes to preferentially connect to nodes that are similar to them. This similarity in general can consider any attribute, but in case of social networks, it is referred to as the connectivity or the node degree of the nodes. At every step, a new node is added to the network based on preferential attachment and a new edge is added between two existing nodes. These existing nodes are chosen on the basis of their degree, thus forcing links between similar degree nodes. The model is innovative as it allows addition of new links between old nodes.

Another interesting model was proposed by Guillaume and Latapy (2005). They identify bipartite graph structure as a fundamental model of complex networks by giving real world examples. The two disjoint sets of a bipartite graph are called *bottom* and *top*. At each step, a new *top* node is added and its degree d is sampled from a prescribed distribution. For each of the d edges of the new vertex, either a new *bottom* vertex is added or one is picked among the pre-existing ones using preferential attachment. A more generalized model based on similar principles was proposed by Bu et al. where instead of using the bipartite structure, a network can contain t disjoint sets (instead of just two sets, as is the case of the bipartite graph), where the example of sexual web (Lilijeros et al. 2001) was considered as a model. A sexual web is a network where nodes represent men and women having relationships to opposite sex, and similar nodes do not interact with each other. At each time step, a new node and m new edges are added to the network with the sum of the probabilities equal to 1. The preferential attachment rule is followed as the new node links with the existing nodes with a probability proportional to the degree of the nodes.

Wang and Rong (2008) proposed a slightly different model, which is still a modified form of the preferential attachment model. Instead of adding one node at a time, the model proposes to add n nodes at each time step which are connected in a ring formation. Any two nodes in the n new nodes are connected to the existing network where these connections are determined through preferential attachment.

Generation models for clustered graphs exist in the literature such as the work of Condon and Karp (1999) and Virtanen (2003), where the idea is to generate graphs that are already clustered as opposed to random graph models of Rapoport (1957) and Erdos and Renyi (1960). A recent work by Zaidi (2012) addresses the issue of generating clustered small-world networks which are not scale free and the clusters are randomly connected to each other. These generation models as such do not produce graphs with small-world and scale-free properties, which are fundamental to most real-world networks. Thus, the study and comparison of these other models remain out of the scope of the paper.

Comparing the different network generation models (See Table 1), first five models are quite similar to each other, as they try to force the triad formation step, one way or the other. Another common aspect in the first five models is that, in every step, only one node and two edges are added to the network. The only other taxonomical grouping possible is the last two models where the bipartite and n-partite structures are used as the fundamental property of real-world networks. The model of Wang and Rong is slightly different as it allows the addition of m new nodes at every time step. The ideas of Klemm and Eguiluz and Catanzaro et al. are quite original, and provide another way to look at the evolution and structure of complex networks.

3 Clustering coefficient and community structures

Largely due to the terminology used to define the metric clustering coefficient, often it is misunderstood that a network having high clustering coefficient suggests the presence of clusters or community structures in a network. Clustering coefficient, by definition, determines the local cohesiveness of a set of nodes, i.e., it focuses on the immediate neighborhood of nodes but fails to capture the presence of communities on the whole as argued by different researchers (Brandes and Erlebach 2005; Girvan and

Newman 2002; Zaidi and Melanço 2010). High clustering coefficient only indicates the presence of a large number of triads, i.e., three nodes connected to each other through three edges. This property is often present in social networks where it refers to the phenomena that if you know two people, there is a high probability that the two people know each other as well. This metric also measures the presence of cliques as they are a composition of triads but it cannot be used to identify the presence of densely or sparsely connected nodes in a network as we explain below using examples.

Figure 1 is an example graph that depicts the differences between clustering coefficient and community structure. Figure 1a clearly has four communities with high connectivity between nodes of the same community, and Fig. 1b has several nodes sharing common neighbors, but visually no distinct groups. Both these graphs have the same number of nodes and edges where the clustering coefficient for graph a is 0.70 and b is 0.69. No information about the presence of four communities can be deduced from the clustering coefficient of graph a.

Another interesting example is shown in Fig. 2 a and b as clusters obtained from some graph. Both contain the same number of nodes and edges. Thus, the density (ratio of number of edges and number of nodes) of the two clusters is exactly the same. Cluster (a) is constructed using Quads instead of triads, where a quad is a set of four nodes connected through four edges in a ring. We then compare this cluster with a cluster constructed with triads. Both these clusters are shown in in Fig. 2a and b. The clustering coefficient of cluster (a) is 0.0 representing the absence of triads and as compared to cluster (b) with a value of 0.69. This is no surprise as clustering coefficient, by definition, measures the quantity of triads in a graph. Another important metric used to classify a network as small world is the average path length. Calculating the average path lengths of the two clusters, (a) has a lower value with 2.3 as

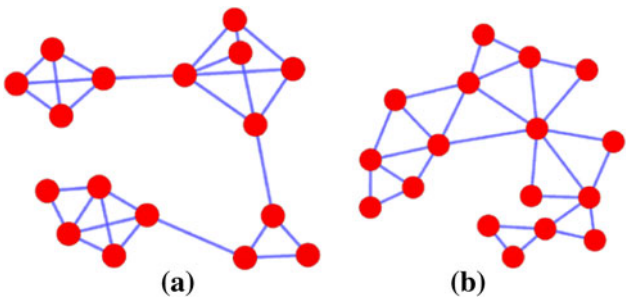

Fig. 1 Two graphs with the same number of nodes and edges. **a** Four groups of nodes well connected within and sparsely connected with other groups. **b** Nodes sharing neighbors in the form of triads. Clustering coefficient for graph **a** is 0.70 and **b** is 0.69. High values for clustering coefficient do not necessarily imply the presence of community structures in a network as shown in graph (**b**)

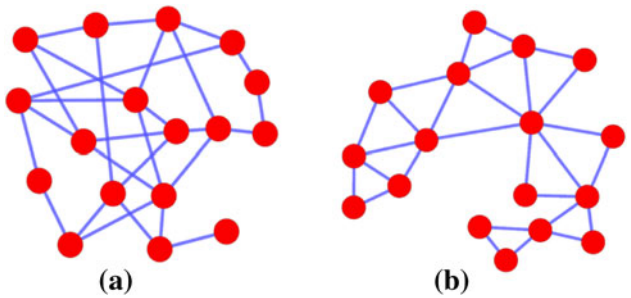

Fig. 2 Consider two clusters of some graph with same number of nodes and edges and thus having the same density in terms of number of nodes and number of edges. **a** Nodes well connected to each other forming quads, **b** nodes sharing neighbors to form triads. Clustering coefficient for cluster **a** is 0.0 and **b** is 0.69 representing the absence of triads in cluster (**a**). The average path length of **a** is 2.3 and of **b** is 2.6 showing that cluster (**a**) is more compact and on average, the nodes are much closer to each other than cluster (**b**). This example shows that a cluster can exist even with a low value of clustering coefficient as shown in **a**

compared to (b) with a value of 2.6 showing that cluster (a) is more compact and on average, the nodes lie closer to each other as compared to cluster (b) and thus is a better cluster even though its clustering coefficient is 0.

From the above two examples, we can conclude that a graph having high clustering coefficient does not necessarily suggests the presence of distinct group of nodes tightly connected to each other and loosely connected within themselves. Moreover a cluster can be a good cluster even if its nodes have a low average clustering coefficient.

4 Assortativity and triads in social networks

In this section, we present three fundamental concepts associated with the theory of social networks. First, we briefly introduce these concepts and then argue that combining these concepts, we can produce a network generation model with small-world and scale-free properties having distinct clusters.

The theory of *Assortativity* or *Assortative Mixing* refers to the principle that in a network, similar nodes tend to attach to each nodes. This similarity can be based on one or more than one attributes. An important application of this theory is the assortative mixing in social networks where nodes attach to other nodes having similar degree. This differs from biological and technological networks that exhibit disassortative mixing (Newman 2002, 2003). Disassortative mixing refers to the phenomena where dissimilar nodes tend to connect to each other. A good example is the Sexual Web (Lilijeros et al. 2001) where nodes representing men or women connect to nodes with opposite sex.

We move on to another important concept in social networks, the formation of *triads* introduced by Simmel and Wolff (1950) as a fundamental structure for social networks. In fact, the smallest and most elementary social unit, a *dyad* is a social group composed of two members while a *triad* is a social group composed of three members. *Groups* of larger size are also possible but since a variety of relationships can form in them, they are less stable (Simmel and Wolff 1950) and often less studied in sociology.

Finally, the principle of *Preferential Attachment* introduced by Barabasi and Albert (1999) has an ingredient for growing network model with power-law degree distribution. Some times, referred to as the 'Rich gets richer', the idea is that in real-world networks, nodes having high degree have a high probability of attracting more connections as compared to nodes with low connectivity. In terms of social networks, this means that a famous person is likely to become more famous as compared to a person who is not well known in the social community. The idea is the direct implication of the human trait of extraversion–introversion (Jung 1921). Extroverts, who are open to meeting new people and developing new relationships are expected to have high degree of connectivity in a social network as compared to Introverts, who tend to be more reserved, less outgoing, and less sociable.

Extending the principle of assortativity, we argue that in theory, since nodes tend to connect to similar nodes, it is not always the case that the similarity is based on node degree. We consider examples from two of the most studied social networks in computer science, the actor network and the authorship network (Watts and Strogatz 1998; Barabási and Albert 1999; Freeman 2004; Wasserman and Faust 1994). If the degree was the only criteria for associating to other nodes, in case of an actor network, actors beginning their career will never get to play a role in a movie with well-known actors. This is contrary to the reality as often, new actors are given supporting roles along with well-known actors and thus the similarity is based on some other criteria. Similarly, for the authorship network, if degree was the only criteria, an experienced professor will never take a doctorate student who has only a few publications under his supervision which is generally not the case.

Our deduction from the above two examples is that the nodes tend to associate to other nodes having similarity based on the context and not the node degree. For example, an actor starting his career as a comedian has a higher probability to act in a comedy film and thus act with a well-known actor in the domain of comedy films. In this case, the similarity is based on the domain of the two actors. Similarly, for the case of authorship network, a student having done a *masters* in a particular domain such as computer networks, has a high probability of collaborating with a well-know researcher of the same domain, probably as a doctorate student. Again, the connectivity preference is due to the domain or subject of research. Generalizing from this concept, for other social networks in real world, consider the example of a person joining a new organization as employee. He has a high probability of interaction with his fellow employees, people working on the same project or sharing the same office. Another example is that of a person joining a sports club. He has a high probability of interaction with people sharing the same sports activity like Tennis.

Returning to the formation of *triads* in social networks, our perspective is that usually when a person enters a new social network, it is not just the triads that are formed but *groups* of larger size, or cliques are formed. From the actor example, it is quite clear that a new actor will probably act with a well-known actor, but the social interaction will take place within the entire cast of the movie. This interaction will be represented with a clique where all the nodes representing the actors will be connected to each other. The authorship network is no different as people co-authoring an artifact will form a clique. Similarly, in real world, usually *groups* of larger size are formed. Continuing with the two examples, a new employee will interact with not just only one or two more people, but with different colleagues in the same organization which work together on the same project or with whom he shares an office, and for a person joining a sports club, he will interact with people sharing similar activities instead of just one or two others.

Addressing the principle of *Preferential Attachment*, we argue that for every node in a *group* (or *clique*), few nodes have a higher number of connectivity with other nodes. For example, in a group representing the actors playing in the same movie, the famous actors will have many connections with others as they would have played a role in many movies. Similarly, in the authorship network, an experienced researcher would have published an artifact with many other researchers and thus would have a high number of connections.

Combining the above principles together, we claim that People in social networks are most likely to interact with similar people where similarity is based on the context and the domain. People form groups of larger size and not just triads, there are few people who have a very high degree of connectivity as compared to others. Based on these ideas, we present a new network generation model in the next section.

5 Proposed network generation model with communities

The basic idea of the proposed algorithm comprises three major steps. Instead of adding one node at a time, we add

cliques of various sizes. This results in the network having high clustering coefficient. Next, we associate a possible connectivity attribute drawn from a degree distribution following power law. This insures that the degree distribution of the final network follows a scale-free property. Finally, to obtain community structures where some nodes are densely connected within and sparsely connected to other nodes, we generate a cluster tree which represents the possible communities for this network. Based on the connectivity attribute, and the distances in the cluster tree, nodes within the cliques are merged together creating highly dense groups of nodes well connected within and sparsely connected to nodes distant in the cluster tree.

The proposed algorithm comprises several steps where each of these steps is explained in detail below. The following mathematical notations are used throughout the explanation: $G(V, E)$ represents an undirected multigraph where V is a set of n nodes and E is a set of e edges. The graph G is initially empty and the nodes and edges are added as the algorithm progresses. $\mathcal{C}$ represents a set of cliques such that $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ are different cliques each comprising one or several nodes. $\mathcal{T}$ represents a tree where its leaves are equal to k (the number cliques in set $\mathcal{C}$).

5.1 Step 1: Clique generation

In contrast to existing network generation models, instead of adding one node or triads at a time, to generate the network, we start by adding cliques of variable sizes to G. The algorithm takes as parameter, the number of cliques to be generated (k), the minimum (minSize) and the maximum size (maxSize) of the cliques to be generated. A random number is generated between these two limits and for each random number, a clique C_i is added to the graph G such that nodes and edges of the clique become member of V and E, respectively. As a result, G contains nodes that are well connected to each other within a clique, and nodes from different cliques are not connected to each other. G becomes a graph comprising $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ as shown in Fig. 3.

5.2 Step 2: Scale-free degree distribution

To have the degree distribution of G follow a scale-free behavior, we generate a separate scale-free graph G' with the same number of nodes as in G, using (Barabási and Albert 1999). Next, we assign the degree of a node in G' as an attribute of a node in G chosen randomly and call this attribute *sf Deg*. Nodes once processed are not reconsidered for another assignment. The Pseudo Code for the process is given in Algorithm 1. Every time the procedure *getNode* is called, it picks a unique and randomly selected node from a given graph until all the nodes have been selected. The

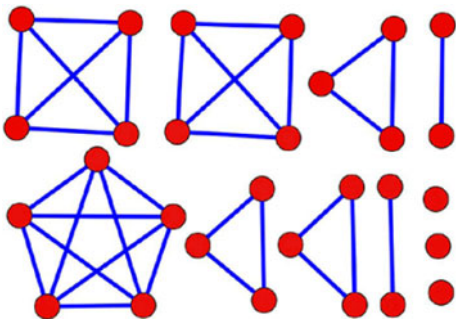

Fig. 3 *Step 1*: A graph G containing only cliques of different sizes. Parameters used for this example are minSize $= 1$, maxSize $= 5$ and $k = 11$

assignment of values from nodes of G' to G is random but since there are few nodes with very high node degree, there is a high probability that they are divided among the cliques sparsely. Thus, we end up with one or two nodes in a clique with a high node degree as shown in Fig. 4. This step assures that the final graph G has a scale-free degree distribution.

Algorithm 1: Scale Free Degree Association to nodes of G

```
Input: Graph G and G'
Output: G(V, E)
begin
    node n, n';
    attribute sfDeg(G);
    n ← getNode(G);
    n' ← getNode(G');
    foreach node ∈ G do
        if deg_G(n) < deg_G'(n') then
            sfDeg(n) ← deg_G'(n');
        else
            sfDeg(n) ← deg_G(n);
end
```

5.3 Step 3: Merger of nodes

The next step is the merger of nodes from different cliques to form a single connected network. The merger is a simple step where two nodes to be merged are replaced by a single node and all the edges connected to the two merged nodes are connected to this new node. Figure 5 shows how two nodes from two different cliques are merged forming a connected network of two cliques. The exact details of how to select two nodes and how many nodes are selected for merger are explained in the following sub-steps.

5.3.1 Step 3.1: Calculation of number of mergers

As a first step, we need to calculate for each node, how many merges will it perform with other nodes. This calculation is based on the attribute *sfDeg*. The idea is pretty simple, the more a node is merged with others, the more

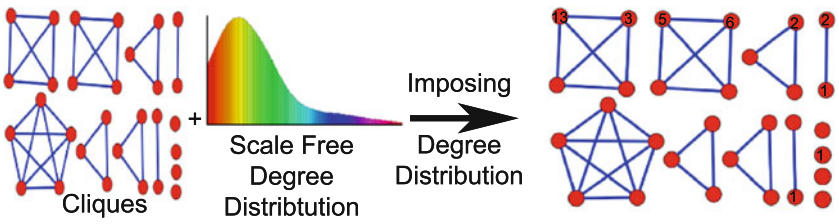

Fig. 4 *Step 2*: A scale-free degree distribution is imposed as attribute of nodes in graph G containing cliques

higher its degree will be in the final network. This phenomenon is shown in Fig. 5 where the two merged nodes result in a single high degree node. The number of merges for each node $n \in G$ is calculated using *sf Deg* as follows:

$$Node\_Merges(n) = \left\lfloor \frac{sfDeg(n)}{Avg\_Node\_Degree(G)} \right\rfloor. \quad (1)$$

We use the following equation to calculate the total number of merges for a clique $C_j \in \mathcal{C}$:

$$Clique\_Merges(C_j) = \sum_{\forall n \in C_j} Node\_Merges(n). \quad (2)$$

5.3.2 Step 3.2: Generation of cluster tree

As the main objective is to have distinct clusters in the graph, we generate a random tree $\mathcal{T}$ with the number of leaves exactly equal to the number of cliques generated in step 1. Each clique $C_j \in \mathcal{C}$ is assigned to a leaf of the cluster tree $\mathcal{T}$ as shown in Fig. 6. The tree can have varying depths to generate a hierarchical clustering where one such tree is shown in Fig. 6.

5.3.3 Step 3.3: Merging nodes of cliques to form clusters

For every clique $C_j \in \mathcal{C}$ in the cluster tree $\mathcal{T}$,, we calculate a vector of probabilities P_{j_i}, where j represents the clique for which this vector is being calculated and i represents the clique with which the probability of connecting j is calculated. This probability is inversely proportional to the distance between two cliques in $\mathcal{T}$ and is spread equally over the branches of $\mathcal{T}$ as shown in Fig. 7. The vector P_{j_i} thus obtained represents the probability of two cliques having their nodes merged.

For example, to calculate the probability of connection of the encircled node with other nodes in $\mathcal{T}$, the probability is uniformly divided among the three branches (1/3 in this example) for each branch leading outward from the encircled node. One of these branches leads to the root of the tree which is again uniformly divided among two of its children as shown in Fig. 7.

Using probability vector P_{j_i} and Clique_Merges(C_j) from Eq. 2, we calculate the exact number of pairwise merges using the equation below:

$$Pairwise\_Merges(C_j, C_i) = \lfloor Clique\_Merges(C_j) * P_{j_i} \rfloor \quad \forall\, j, i \in \mathcal{C} \quad (3)$$

Pairwise_Merges(C_j, C_i) is a directed vector representing the exact number (as integer) of merges between each pair of clique (C_j, C_i). Based on these integer values, nodes from different cliques are merged to form connections between cliques which result in a fully connected network with the desired properties. This calculation is depicted in Fig. 8 where we show the probability vector for C_0, P_{0_i} and its corresponding Pairwise_Merges(C_0,C_b) $\forall b \in \mathcal{C}$.

Figure 8 also shows the probability of C_0 divided uniformly among C_1, C_2 and the rest of the cliques in the cluster tree. The close neighbors of C_0 in the tree C_1, C_2 have a very high probability of 0.33 each of merging with C_0. The merger of nodes with close neighbors results in lots of connections being built between the cliques nearer to each other in the tree and thus represents clusters in the

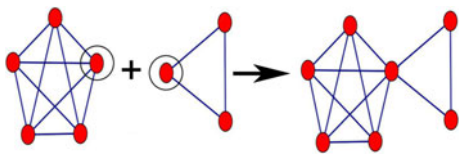

Fig. 5 Merging two nodes from two different cliques. Two nodes from different cliques are removed and a new node is added which takes all the connections of the two removed nodes

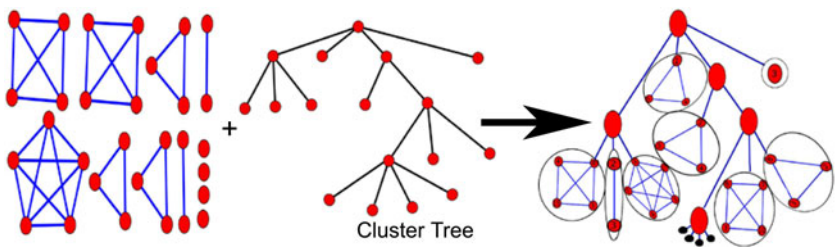

Fig. 6 *Step 3*: A cluster tree $\mathcal{T}$ is generated and cliques are assigned to its leaves to decide how the nodes (in the cliques) will be connected to each other forming clusters

final graph. Algorithm 2 contains the pseudo code for the merger of two nodes.

Algorithm 2: Merger of Nodes in the Cliques to Form Clusters in G

Input: Graph G, Pairwise_Merges, Node_Merges, $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$
Output: $G(V, E)$
begin
  foreach $C_a \in \mathcal{C}$ **do**
    foreach $C_b \in \mathcal{C}$ **do**
      while *Pairwise_Merges*$(C_a, C_b) > 0$ **do**
        Call Merge(C_a, C_b);
        Pairwise_Merges(C_a, C_b)=Pairwise_Merges(C_a, C_b)-1 ;
end

Procedure Merge(C_a, C_b);
begin
  node n_u, n_v;
  $n_u \leftarrow Select\_One\_Node_{rand}(C_a)$;
  Node_Merges(n_u)=Node_Merges(n_u)-1 ;
  $n_v \leftarrow Select\_One\_Node_{prob}(C_b)$;
  Merging_Nodes(n_u, n_v) ;
end

Algorithm 2 uses a procedure *merge* where two functions are used to select nodes from a clique named *Select_One_Node*$_{rand}$ (C_a) and *Select_One_Node*$_{prob}$ (C_b). The implementation of these functions is very simple. The function *Select_One_Node*$_{rand}$ (C_a) chooses a node n randomly such that $n \in C_a$ and $C_a \in \mathcal{C}$ and *Node_Merges*$(n) > 0$. Note that the equality in Eq. (2) is always preserved during the execution of algorithm. The function *Select_One_Node*$_{prob}$ (C_b) uses the *sfDeg*(n) to calculate a probability which is proportional to the node degree of the node. Thus, nodes having high connectivity have a high probability of being selected as compared to nodes with low connectivity.

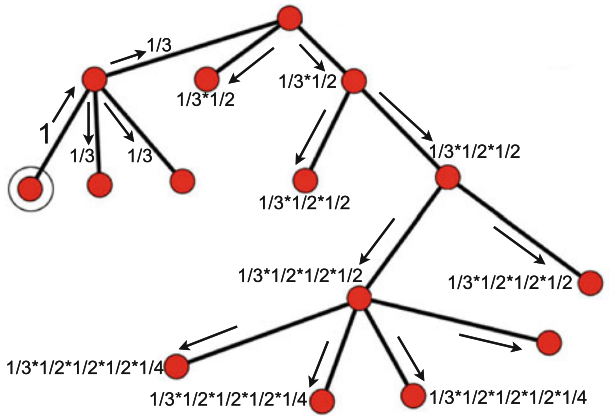

Fig. 7 Calculation of probabilities of merger of the *left most* leaf containing a clique (*encircled*) with other leaf nodes based on distances in the cluster tree $\mathcal{T}$ where the probabilities represent the likeliness of a node in the *encircled* leaf (containing a clique) to be merged with other nodes in the cliques

5.4 Further explanations and possible variants to the proposed model

In this section, we provide explanations of the different steps of the proposed model and relate these explanations to real-world social networks. This helps understand how characteristics of real-world networks are incorporated in the proposed model. We also discuss possible variations in the different steps that can change the behavior of the network generated. These variations demonstrate the robustness and flexibility of the proposed model as it can be used to generate networks with varying properties.

The first possible variation to the model is in the very first step explained in Sect. 5.1 where we add cliques of different sizes. The size of the cliques can be forced to be exactly 3, in which case we would have forced the presence of only triads just as the other network generation models presented in Sect. 2. Due to the presence of cliques (or triads), the average clustering coefficient of the entire graph increases as compared to a random graph which is a fundamental property to identify a small-world network.

The assignment of values in Sect. 5.2 is easy to comprehend once considered in the context of real world. This assignment represents that, in certain social groups, there are people who have relatively high connectivity with others. Continuing with our two example social networks, a famous actor who plays in many films will have a high number of connections with other actors and similarly, a senior professor will have a high number of connections with other researchers. This value is used in step 3 of the model to determine how different cliques of step 2 are merged together to form a single connected network.

A variation to this step can be the assignment of a normal degree distribution or a uniform degree distribution. The choice results in what the final degree distribution would be for the generated network. This flexibility is quite useful as the model can be used to generate networks with any kind of degree distribution.

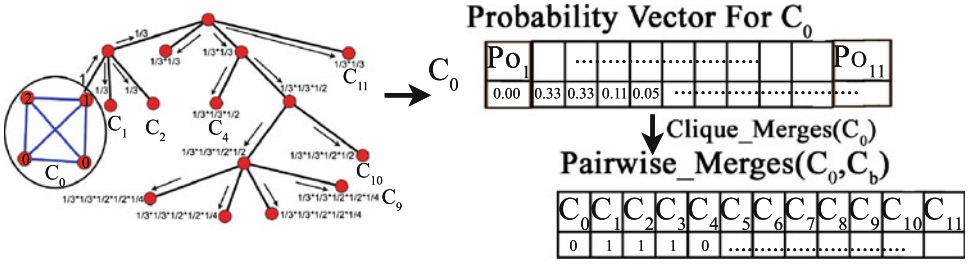

Fig. 8 Calculation of Pairwise_Merges(C_0,C_b) using distances in $\mathcal{T}$ and probability vector of C_0 to determine the number of merges between C_0 and C_b where $C_b \in \mathcal{C}$ and $C_0 \neq C_b$

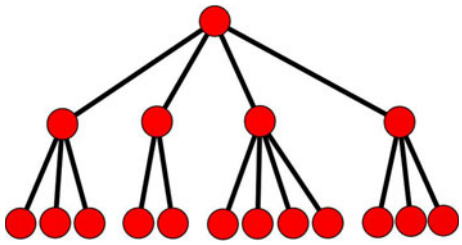

Fig. 9 A cluster tree $\mathcal{T}$ to generate flat or partitional clustering with the leaf nodes containing the cliques at the *bottom level*, the clusters at the *second level* and the root regrouping the clusters at the *top level*

In Sect. 5.3.2, we discuss how a cluster tree is used to generate a network with hierarchical clusters. A possible variation is the generation of a flat or partitional clustering. We can generate only a tree with depth 2 where we have all the cliques at the bottom level, merging with other cliques at level 1 to form clusters and the root represents the regrouping of all the clusters as shown in Fig. 9, which contains 4 clusters and 12 cliques.

6 Real-world social networks

For the analytical study of the network generation models, we compare the networks generated by existing models with real-world networks using a number of metrics (see Sect. 7). We consider three social networks, two of which are author networks and the third one is an actor network.

The author network is a network where nodes represent scientists and an edge between them represents a collaboration in terms of co-authoring a scientific artifact like a book or an article. The two datasets are the Network Science dataset and the Geometry dataset. The Network Science data was compiled by Newman (2006) from the bibliographies of two review articles on networks, Newman (2003) and Boccaletti et al. (2006), with a few additional references added by hand. The network contains a single connected component with 379 nodes and 914 edges.

The other author network is the authors collaboration network in computational geometry. It was produced from the BibTeX bibliography obtained from the Computational Geometry Database geombib,[1] version February 2002. Problems with different names referring to the same person are manually fixed and the data base is made available by Vladimir Batagelj and Andrej Mrvar from the Pajek datasets website.[2] Only the biggest connected component was considered for experimentation where the reduced simple network contains 3,621 vertices and 9,461 edges.

The Actor network is a network where nodes represent actors and two actors are connected to each other if they have acted in a movie together. The dataset we use here is a subset taken from the IMDB[3] database of movies made until the year 1999 and used by other researchers such as Auber et al. [3] and Archambault et al. [2]. This network contains 7,640 nodes and 277,029 edges.

The choice of selecting these models is based on two criteria. First, we wanted to use graphs that are publicly available and have been studied by other researchers. Moreover, networks having varying density and size so as to see the behavior of the different models in terms of scalability and flexibility could be evaluated.

7 Results and discussion

We calculate a number of statistics using various network generation models and compare them with the real-world networks of equal sizes. The results are shown in Tables 2, 3 and 4. In some cases, the models are not parameterized and thus the node–edge density could not be controlled. We tried to generate models of similar size in terms of number of nodes, and where possible, similar number of edges. An important observation about these networks is that since all of them use the preferential attachment to produce the scale-free property, the degree distribution for all the models follows a power law. To the best of our knowledge, there is no metric which tries to identify the presence of communities in a network by analyzing the graph on the whole in a global perspective; thus, the presence of community structure in the proposed model is only justified by construction. Using the cluster tree, the way the nodes connect to each other can be controlled and thus any network that is produced has densely connected nodes which are sparsely connected to other nodes.

Lets have a look at some individual results for the various models in comparison to the real-world networks. For example, graphs generated using the model of Guillaume and Latapy, the node–edge density in every case is very high and could not be controlled. The model of Fu and Liao, in all the three examples, have a very low clustering coefficient as compared to the respective real-world network and thus could not really be classified as generating similar networks to the real-world networks used as examples in our study. Looking at the clustering coefficient of the model by Wang and Rong in Table 3, it is quite clear that the model fails to generate a high clustering coefficient for a similar size network. In the model of Holme and Kim (Table 4) where the node–edge density of the network is comparatively high to other two networks but the network has a large size, the clustering coefficient drops considerably. The model of Klemm and Eguiluz scales well in

[1] http://www.math.utah.edu/∼beebe/bibliographies.html.

[2] http://vlado.fmf.uni-lj.si/pub/networks/data/.

[3] http://www.imdb.com/.

terms of clustering coefficient, but in case of low node–edge density (see Table 2), the average path length is considerably high to be a small-world network. Also, from Table 4, the average path length in case of a number of models is 1.99, which is a direct implication of a node having a very high degree. As a result, most of the nodes are connected to this high degree node and thus have almost a distance which reduces the average path length of the entire network.

From the above examples, one obvious problem that can be inferred is that these models have problems with scalability, as the node–edge density is varied for a network, the models are not able to reproduce comparative values with real-world networks for various statistics. On the other hand, the proposed model in this paper has the ability to control the size of cliques as the starting point, which helps us to gauge the density and at the same time, and generate small-world and scale-free networks. The values are quite close to the ones expected and thus the proposed model is quite flexible.

8 Conclusion and future research directions

In this paper, we have studied the concepts of assortativity, triads and preferential attachment as the building blocks for

Table 2 Comparing different models with the collaboration network of scientists from the network science data

Model	Nodes	Edges	Average path length	Clustering coefficient	Maximum node degree
Network science	379	914	6.04	0.74	34
Zaidi et al.	364	935	4.7	0.65	34
Holme and Kim	379	757	4.86	0.77	42
Fu and Liao	379	744	4.03	0.01	31
Klemm and Eguiluz	379	755	9.08	0.5	33
Catanzaro et al.	379	898	2.42	0.58	197
Guillaume and Latapy	379	5,315	2.30	0.54	109
Bu et al.	379	755	3.05	0.37	80
Wang and Rong	379	943	4.32	0.37	14

Table 3 Comparing different models with the collaboration network of scientists from the computational geometry data

Model	Nodes	Edges	Average path length	Clustering coefficient	Maximum node degree
Geometry	3,621	9,461	5.31	0.53	102
Zaidi et al.	3,567	9,433	5.4	0.66	127
Holme and Kim	3,621	7,241	7.3	0.79	90
Fu and Liao	3,621	10,662	4.22	0.005	101
Klemm and Eguiluz	3,621	10,857	2.27	0.72	197
Catanzaro et al.	3,621	8,896	2.47	0.48	1,720
Guillaume and Latapy	3,621	528,499	*	*	1,275
Bu et al.	3,621	10,856	3.13	0.24	607
Wang and Rong	3,621	10,828	4.6	0.10	30

* Could not be calculated due to large size of networks

Table 4 Comparing different models with the Actor network from the IMDB dataset

Model	Nodes	Edges	Average path length	Clustering coefficient	Maximum node degree
Actor	7,640	277,029	2.94	0.87	1,271
Zaidi et al.	7,413	244,905	3.1	0.98	352
Holme and Kim	7,640	274,865	2.35	0.09	2,303
Fu and Liao	7,640	29,972	4.00	0.004	163
Klemm and Eguiluz	7,640	274,374	1.99	0.97	7,627
Catanzaro et al.	7,640	28,127	1.99	0.78	7,639
Guillaume and Latapy	7,640	2,378,281	*	*	2,614
Bu et al.	7,640	274,935	1.99	0.83	12,151
Wang and Rong	7,640	273,355	3.28	0.94	83

* Could not be calculated due to large size of networks

the structure of social networks. We use these concepts to present a model to generate artificial social networks. We evaluated a number of network generation models that successfully generated small-world and scale-free networks but fail to capture another important characteristic of real-world network, i.e., the presence of community structures. We compared the existing and the proposed network model with real-world social networks using a number of statistics. Results show that the proposed model indeed generates networks that have community structures and are topologically similar to real-world networks as compared to the other existing models that generate small-world and scale-free networks. Moreover, we identified another problem for the existing models, the scalability in terms of node–edge density, where it is difficult to maintain the high clustering coefficient and low average path length as networks of varying sizes are produced.

In this paper, we have focused on social networks and effectively presented a model to generate networks having small-world and scale-free behavior with communities. We intend to extend our study to other types of networks such as biological and technological networks to propose network generation models for these types of networks as well, incorporating several real-world networks.

References

Almeida H, Neto G, Meira W Jr, Zaki MJ (2012) Towards a better quality metric for graph cluster evaluation. J Inf Data Manag 3(3):378–393

Archambault D, Munzner T, Auber D (2007) grouse: feature-based, steerable graph hierarchy exploration. In: EuroVis, pp 67–74

Auber D, Chiricota Y, Jourdan F, Melancon G (2003) Multiscale visualization of small world networks. In: INFOVIS '03: Proceedings of the IEEE Symposium on Information Visualization, pp 75–81

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Boccaletti S et al (2006) Complex networks: structure and dynamics. Phys Reports 424:175–308

Brandes U, Erlebach T (2005) Network analysis: methodological foundations. Lecture Notes in Computer Science. Springer, Berlin

Catanzaro M, Caldarelli G, Pietronero L (2004) Assortative model for social networks. Phys Rev E (Statist Nonlinear Soft Matter Phys) 70(3):1–4

Coleman JS (1964) An introduction to mathematical sociology. Collier-Macmillan, London

Condon A, Karp RM (1999) Algorithms for graph partitioning on the planted partition model. Random Struct Algorithms 18(2):116–140

Cross R, Parker A, Borgatti SP (2000) A bird's-eye view: using social network analysis to improve knowledge creation and sharing. Knowl Dir 2(1):48–61

Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51:1079–1187

Erdos P, Renyi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61

Everitt BS, Landau S, Leese M (2009) Cluster analysis, 4th edn. Wiley, New York

Freeman LC (2004) The development of social network analysis: a study in the sociology of science. Empirical Press, Vancouver

Fu P, Liao K (2006) An evolving scale-free network with large clustering coefficient. In: ICARCV, IEEE, pp 1–4

Gilbert F, Simonetto P, Zaidi F, Jourdan F, Bourqui R (2011) Communities and hierarchical structures in dynamic social networks: analysis and visualization. Soc Netw Anal Min 1:83–95. doi:10.1007/s13278-010-0002-8

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99:8271–8276

Gordon AD (1981) Classification: methods for the exploratory analysis of multivariate data. Chapman & Hall Ltd., London

Guillaume J-L, Latapy M (2005) Bipartite graphs as models of complex networks. In: Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN), Lecture Notes in Computer Science, vol 3405. Springer, pp 127–139

Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Phys Rev E 65:026107

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Jia Y, Garland M, Hart JC (2011) Social network clustering and visualization using hierarchical edge bundles. Comput Gr Forum 30(8):2314–2327

Jung CJ (1921) Psychologischen typen. Volume Translation Baynes HG, 1923. Rascher, Zurich

Klemm K, Eguiluz VM (2002) Growing scale-free networks with small world behavior. Physical Review E 65:057102

Lilijeros F, Edling C, Amaral L, Stanley E, åberg Y (2001) The web of human sexual contacts. Nature 411:907–908

Liu J-G, Dang Y-Z, tuo Wang Z (2005) Multistage random growing small-world networks with power-law degree distribution. Chin Phys Lett 23(3):746 (Comment: 3 figures, 4 pages)

Milgram S (1967) The small world problem. Psychol Today 1:61–67

Newman M (2003) Mixing patterns in networks. Phys Rev E 67:026126

Newman M (2004) Detecting community structure in networks. Eur Phys J B-Condens Matt 38(2):321–330

Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167

Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E (Statist Nonlinear Soft Matter Phys) 74(3):036104

Päivinen N (2007) A quest for the hidden knowledge. PhD thesis, University of Kuopio, Kuopio

Rapoport A (1957) Contribution to the theory of random and biased nets. Bull Math Biophys 19:257–277

Schaeffer SE (2007) Graph clustering. Comput Sci Rev 1(1):27–64

Scott J (2011) Social network analysis: developments, advances, and prospects. Soc Netw Anal Min 1:21–26

Scott JP (2000) Social network analysis: a handbook. SAGE Publications, New York

Simmel G, Wolff KH (1950) The sociology of Georg Simmel/ translated and edited with an introduction by Wolff KH. Free Press, Glencoe

Tryon RC (1939) Cluster analysis. Edwards Brothers, Ann Arbor

Virtanen S (2003) Properties of nonuniform random graph models. Research Report A77, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo

Wang J, Rong L (2008) Evolving small-world networks based on the modified ba model. Inf Technol Int Conf Comput Sci 0:143–146

Wang L, Du F, Dai HP, Sun YX (2006) Random pseudofractal scale-free networks with small-world effect. Eur Phys J B—Condens Matter Complex Syst 53:361–366

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678

Zaidi F (2012) Small world networks and clustered small world networks with random connectivity. Soc Netw Anal Min, pp 1–13. doi:10.1007/s13278-012-0052-1

Zaidi F, Melançon G (2010) Identifying the presence of communities in complex networks through topological decomposition and component densities. In: EGC 2010, Extraction et Gestion de Connaissance, vol E-19, RNTI, pp 163–174