A generalized model for social collaboration networks

Chunhui Deng¹, Jing Liu² and Huifang Deng², ³

¹School of Computer Engineering, Guangzhou College, South China University of Technology, Guangzhou, China, 510800
²School of Computer Science and Engineering, South China University of Technology, University Town, Panyu, Guangzhou, China, 510006
³Email: hdengpp@qq.com

Abstract. In this paper, we constructed a three-parameter general social collaboration network model (GCNM) using both similarity-based strategy and copying-based strategy. In this model, a connection can be introduced between two nodes if they are similar in attributes or similar in structure (i.e., have many common friends). This model can match various types of real-world social networks by just choosing different values for three parameters, and therefore, can reproduce observed real network characteristics. Particularly, three widely studied models (similarity-based, copying-based, and pure preferential attachment (PA)) are only the special cases of this new model. Specifically, this new model possesses following new characteristics which existing models have difficulty to fulfill in all: 1) it not only exhibits the high average clustering coefficient, but also the high global clustering coefficient and explicit community structure (great number of triangles); 2) it not only makes the connections between the new node and existing nodes, but also between existing nodes; and 3) it has not only linear, but also non-linear relationship between the linking probability and the degree (the connectivity of nodes). Numerical results, which are in good agreement with real-world datasets from different fields, demonstrated all of these.

1. Introduction

Social network is widely used in describing the social structures determined by relationships between individuals, groups and organizations. The development of models describing the structure, function and evolution of such networks has already aroused many researchers’ interests which make it a very hot and active research area. As a result, several models have been put forward [1-11]. However, most of them have difficulty in generating an overall real-world network’s characteristics. There are several well-known networks such as citation networks [1, 2], co-authorship networks [3-5], social ties and traffic networks [6]. In this paper, we are only concerned with the collaboration networks [3, 4, 7, 8] which consist of a variety of entities such as organizations and people that collaborate with each other to achieve common goals. Scientific collaboration networks are being intensively studied at present where two scientists are considered to be connected if they have co-authored at least one paper. It is believed that such studies can reveal the patterns of collaborative relationship among people, to help deepen understanding of the collaboration between scientists.

Many works [12-14] have been done in the field to understand the important structural properties of social networks, such as the small shortest mean path length, the high average clustering coefficient and net clustering coefficient, the assortative mixing by degree, the huge number of triangles, the clear community structure, and the prominent "fat tails" of the degree distribution. The "scale-free" or
"power-law" characteristic of social network topology has been observed in many real cases [15-16]. In recent researches, it has been found that real-world networks have more dense structures than generally thought [1, 17-19]. However, the classical preferential attachment (PA) mechanism [15], which has been used to randomly generate scale-free networks with small average distance, fails to explain the formation of strong neighborhood clusters. That is, the number of triangles is very likely to be underestimated by a pure PA model and clustering coefficients tend to vanish with continuous addition of new nodes. Therefore, the pure preferential attachment has been extended to produce higher clustering coefficient and more triangles. "Similarity" and "Copying" are the two most common strategies used to improve clustering coefficients and the numbers of triangles in literatures.

Similarity-based models [2, 17, 19-21] have been thoroughly discussed in many literatures. The primary motive is that linking probability between nodes in networks is not only affected by the connectivity of nodes, but also strongly influenced by the similarity between them. It is believed that "nodes with more similarities have a higher chance to get connected even if they are not popular". These similarities are also known as homophile in social sciences, For example, scientists are willing to collaborate with others on similar research topics; people prefer to make friends with those who share many interests in common; and web pages tend to link a new page to the most popular one among pages with similar contents.

By changing the weights of degree and similarity, several strategies have been developed to simulate a better network structure. Degree-similarity mixture (DSM) model was proposed by Menczer [19] to study document networks, in which linking probability is determined by the sum of standard PA and content similarity. Degree-similarity product (DSP) model [17] is a modification of the DSM model, which uses the product of node connectivity and content similarity to ensure that linking probability is proportional to both popular and similar. In [20], the individual is associated with an attribute vector, whose value of each component is generated by uniform distribution function within a -dimensional space. Based on similar distances to the new nodes, existing vertices are divided into distinct groups. With a certain probability, the closest group is selected as the target for new links, and edges are established following the typical PA within the group. In [21], each node is randomly placed on a plane with polar coordinates where is a random angular position, and the radial coordinate is used to represent the popularity of nodes (older nodes have more chances to become popular in a growing network model). Thus, the linking probability between nodes and depends on hyperbolic distance: , which is a single-metric representation of the preferential attachment between radial popularity and angular similarity. With theoretical analysis and abundant experiments, all of these similarity-based models claim that they can reproduce a large-scale social network with strong community structure and high clustering coefficient, while the degree distribution continues to be the power law and the linking probability of a -degree node is also a linear function in , which is very similar with standard PA. Despite all this, it is hard for similarity-based models to reproduce actual networks in terms of the number of triangles and net clustering coefficient, let alone to reproduce the collaboration networks. The analysis and discussions in Section 3.1 could support this conclusion.

Copying is another common strategy used in practices for the formation of triads [1,5,18,22]. The idea is based on the fact that we often make new friends introduced by mutual friends in our daily life, also known as meeting friends of friends [23]. Generally speaking, in most copying-based models, the first neighbor of a new node is selected randomly from all existing vertexes or based on standard PA. Then, another connection is set up by a triad formation step (creating one or more connections to the friends of a neighbor) with a certain probability. What's more, the user specified proportion of triad formation for each new node could ultimately affect the clustering coefficient. It's worth mentioning that Holme and Kim (HK) model [18] and forest fire model [24] choose a copied node from neighbor's friends randomly. While in the DAC model (Degree – Aging preferential attachment and Clique neighborhood attachment model) [1], the clique structure is considered: new node links to all the vertexes in the same clique. The authors showed that the number of triangles and the clustering
The clustering coefficient of the networks generated by DAC model remarkably matched that of two real-world datasets, HEP-PH dataset [25], and PNAS dataset [19]. However, the growth process that the DAC model represents does not often occur in the real collaboration network. Moreover, as is shown in Figure 2, the average clustering coefficient and community structure of networks predicted by copying-based models are still well below the actual ones of collaboration network.

Both "copying" and "similarity" mechanisms are very useful in modeling social networks with characteristics such as small-world and scale-free, and more importantly, clustering coefficient does not vanish as the size of network grows. However, many empirical networks appear to be more sophisticated. 1) It is really hard to generate as many triangles as real networks have, for example, there are more than 3 million triangles in Arxiv HEP-PH dataset [25], which has only 12 thousand nodes and less than 20 links for each node. If each additional node is connected to all the friends of a specified neighbor, it would help improve the number of triangles, just as the DAC model shows [1]. However, it is not always easy to explain the behavior that you collaborate with all the neighbors of a new friend immediately after meeting him. 2) New links can only take place between new node and existing nodes in the network. For example, the new node duplicates one or more links from its neighbors as its own, but nothing is happening among the existing nodes, even if two nodes may share many common neighbors; 3) Although it is helpful to obtain a power-law distribution of degree by keeping a linear relationship between linking probability and degree, it is an often observed fact that the degree distribution of real-world networks is much closer to the truncated exponential distribution [6,10] rather than a simple power law distribution [9,11,15,16]. In [6], Zhang et al. found that the distribution of nodes of low degree behaves more like an exponential function. The formation of real-world networks is a complicated process realized by mixing and matching various strategies that we may or may not know. As a result, the combination of similarity-based method and copying-based method is promising to more closely reproduce the real-world networks.

In this paper, we proposed a general social collaboration network model (GCNM), in which a connection can be added between two nodes if they are similar to each other or have many mutual friends. This hybrid model is not just a simple combination of the copying-based and similarity-based strategies, but it also possesses some new advantages as shown in Section 3: 1) the generated networks could reveal the main characteristics that actual networks have, such as the short average path length, the fat-tailed distribution of degree, the high average clustering coefficient and net clustering coefficient, the assortative mixing by degree, and explicit community structure (plenty of triangles); 2) the connections can be established between existing nodes as well as the new node and existing nodes; 3) The relationship between the linking probability and the degree not only can be linear, but also can be non-linear. Hence the degree distribution of such simulated network could be more realistic.

The rest of this paper is organized as follows: In Section 2.1 and 2.2 we introduce the motivations and the basic framework of our similarity and mutual friend model (GCNM model), where similarity and the number of friends in common are taken into account in forming new connections. In Section 2.3, we give a discussion on how the GCNM model contributes to the formation of structures with high clustering coefficient, sufficient triangles and broad degree distributions. In Section 3, the experimental results are presented to illustrate the performance of GCNM model. The final section concludes the paper.

2. GCNM model

2.1. Motivation for the model

As mentioned above, similarity-based and copying-based algorithms are the mainstream mechanisms for modeling social network with high clustering coefficients. While in a more general view, the copying-based method can be regarded as a kind of structurally similar method: two nodes are similar if they have one or more shared neighbors (similar structural features). Moreover, common neighbors have been shown to have substantial effects on linking prediction in many literatures [12], that is, the greater the number of joint neighbors, the higher the possibility of two nodes to be linked in the future.
Figure 1 shows the relationships between the number of shared coauthors for each pair of individual and the average linking probability three year (2003-2005) and five year (2003-2007) later. The number of shared coauthors is measured by calculating the DBLP dataset [26] from 1991 to 2002. It can be seen from Figure 1 that as the number of shared coauthors rises, the linking probability increases very rapidly. Linking probability also increases with time, which means that two scientists will be more likely to work in partnership with each other as time goes by. Consequently, the proposed network growth model should satisfy two characteristics: the probability that two nodes will be linked in the future is closely related to the number of mutual friends (the degree of similarity in structure) and the time being potential friends with each other (they have at least one neighbor in common).

![Figure 1](image-url)  
**Figure 1.** Probability of future coauthorship with another author as a function of the number of shared coauthors.

Both similarities in attributes/properties and in structure can contribute to the improvement of standard PA model, the key distinction between those two "similarities" is that attribute similarity is globally effective while structural similarity can be seen as local features. Particularly, the structural similarity can be expected to generate more triangles and more compact structures within groups, while attribute similarity is likely to produce higher clustering degree and larger communities with a comparatively loose structure. Therefore, we have good reason to believe that the distinct advantages could be obtained at the same time when integrating these two aspects.

2.2. The model algorithm

Based on observations described earlier, the proposed similarity and mutual friends’ model (GCNM) consists of two main processes: 1) A new node initiates several connections to the existing nodes according to the degree-similarity preferential attachment. The degree of similarity between two nodes can be calculated by comparing the values of their attributes. Here, each node in the network is associated with a vector, representing various properties of it. This process gives rise to the power-law distribution of degree. 2) New links occur between any random pairs of disconnected nodes if they are similar to each other in structure. This model leads to an undirected network and is particularly useful for collaboration networks. Detailed algorithm of the model is described as follows:

1. Following most other grown models, we start with a very small seed network;
2. For each time $t$, a new node is added into the network. First choose $am$ old nodes (here $0 \leq a \leq 1$, $m$ is defined as a positive integer, representing the average links of each node) and then make arcs to them. The probability that a node $i$ is selected is proportional to the product of its degree and similarity with the new node: $\prod_i \propto k_i \cdot \kappa(sim(i))$, where $k_i$ is the degree of node $i$ and $sim(i)$ denotes cosine similarity with the new node. The influence of similarity can be adjusted by parameter $\gamma$:

$$
\kappa(sim(\_)) = \left\{ \frac{2.0}{sim(\_) + 1.0} - 1.0 \right\}^{\gamma}, \quad -1 \leq sim(\_) \leq 1
$$

(2.1)
(3) Pick up \((1 - \alpha)m\) pairs of disconnected nodes in the network (including the new node) and then make a fresh connection to each selected pair. The probability of a specific pair (say, nodes \(i\) and \(j\)) to be selected, depends on the number of neighbors they shared in common:

\[
\prod_i \propto \kappa \left( { \left| {n_i \cap n_j} \right| } \right)^\alpha , \quad \text{where} \quad n_i \text{ and } n_j \text{ are the neighbors of node } i \text{ and } j \text{ respectively, and} \quad \left| {n_i \cap n_j} \right| \text{ is the number of mutual neighbors of node } i \text{ and } j . \]

We can impose incentive or penalty on the high-degree common neighbors by adjusting the parameter \(\lambda\).

(4) Repeat steps (2) and (3) until the size of network grows up to the desired scale.

In this model, a new node and \(m\) connections are added into network at each time step. Networks with various kinds of characteristics can be generated by changing the values of three main parameters \((\alpha, \gamma, \lambda)\). For example, we can change the value of \(\alpha\) to adjust the weight of "similarity" (step (2)) and "shared neighbor" (step (3)) in the process of generating network. Especially, the GCNM model can immediately reduce to the similarity-based one if \(\alpha = 1\), or the copying-based one if \(\alpha = 0\), or even the pure PA model if \(\alpha = 1, \gamma = 0\). This means all these widely studied three types of models are only the special cases of our new model. It is interesting to note that two distinct nodes have no chance to be connected together in step (3) if they do not have any neighbor in common. In the next section, we will take a closer look at the evolution properties of GCNM model.

2.3. Temporal characteristics of degree distribution

Here, we will focus on the evolution of degree distribution \(N_k(t)\) (the number of nodes with \(k\) links at time \(t\)) of GCNM model. To determine degree distribution, the rate equation \([27-29]\) is a common analytical approach, in which attachment kernel \(A_k\) is often used to represent the weight that \(k\)-degree node obtains a new link. The rate equation for \(N_k(t)\) is

\[
\frac{dN_k}{dt} = \frac{A_k}{A} N_{k-1} N_k - A_k N_k + \sigma_{s, a} \quad \text{(2.2)}
\]

This is a typical expression to show the gain and loss in \(N_k\) at each time step. Here, denominator \(A = \sum_{k \geq a} A_k \cdot N_k\) is normalization factor, and \(A_k / A\) is the rate of increase in \(N_k\), while loss in \(N_{k-1}\). The last term on the right comes from the new node with \(a\) links. There is a consensus that the power-law distribution is attributed to the linear attachment kernels, \(A_k \propto k\), as \(k \to \infty\). However, the proposed GCNM model is for a more sophisticated case, where \(A_k\) is influenced by two essential processes, "similarity" and "common neighbors". We denote attachment kernel in step (2) and step (3) as \(A_{s, a}\) and \(A_{c, a}\), respectively.

1) At each time step, the increasing rate of \(N_k\) in step (2) is related to the sum of the product of degree and similarity for all nodes with \((k-1)\) degree. From the derivation of equation (2.3), attachment kernel in step (2) is still linearly proportional to the degree. Therefore, we can get a conclusion that degree distribution does not seem to be greatly affected by similarity, which is consistent with early studies \([8, 12-15]\).

\[
A_{s, a} \cdot N_{k-1} = \sum_{j \geq a, d_j = k-1} s_j \cdot (k-1) = \overline{s} \cdot (k-1) \cdot N_{k-1} \sim (k-1) \cdot N_{k-1} \quad \text{(2.3)}
\]

Where \(d_j\) is the degree of node \(j\), \(s_j\) denotes the similarity between node \(j\) and the new node, and \(\overline{s}\) denotes the average similarity for all the nodes with degree of \(k-1\). All nodes in network can be divided into separate sets by degree. Average similarity of each set is the same due to the randomly selected values of attributes for each node. And this is the reason why equation (2.3) is valid and relevant.

2) In step (3) of GCNM model, attachment kernel is not directly related to degree, but we can derive the relationship based on two assumptions:
Assumption 1. The positive degree-degree correlations of our model networks can be observed, that is, neighbors of a high-degree vertex are also likely to be high-degree vertices. This assumption is supported by the growth process: those nodes which appear earlier have higher probability to increase connections as new nodes entering the network. In addition, old nodes also have more chances linking to each other by step 3. So the average links of a k-degree node's neighbors can be approximated as \( r \cdot k \), where \( r \) is the correlation coefficient of assortative mixing by degree.

Assumption 2. The local clustering coefficient \( C_k \) is a monotonously decreasing function of nodes' degree \( k \). In other words, high-degree vertices often exhibit relatively weak clustering than those low-degree vertices. This assumption satisfies most of the facts. Simulated results presented in the next section back this assumption. On average, each neighbor of a \( k \)-degree node introduces \( (1 - C_k) \cdot r \cdot k \) potential friends because another \( C_k \cdot r \cdot k \) nodes are already friends of the \( k \)-degree node. In other words, each neighbor could be calculated \( (1 - C_k) \cdot r \cdot k \) times when we count the number of common neighbors between a \( k \)-degree node and all potential friends. Therefore, the total number of shared neighbors can be approximated as \( (1 - C_k) \cdot r \cdot k^2 \). The attachment kernel in step (3) could be expressed as:

\[
A_{k,c} = \sum_{i \neq j, i < n} \left| n_i \cap n_j \right| \sim \left[ (1 - C_k) \cdot r \cdot k \right]^2 - \Delta
\]  

(2.4)

As you can see, under the assumed conditions, the equation (2.4) is a very good approximation to the relationship between attachment kernel \( A_{k,c} \) and degree. The result is greatly affected by parameter \( \lambda \). Especially when \( \lambda \to 1 \Rightarrow \Delta = 0 \), \( A_{k,c} \) is exactly equal to the number of common friends between a \( k \)-degree node and all potential nodes. Moreover, the power-exponent is greater than 1 if \( \lambda \geq 1/2 \), which means that attachment kernel depends on degree in a super linear way.

Let us put it simple, the final attachment kernel is a combination of steps (2) and (3), \( A_k = A_{k,s} + A_{k,c} \), and the power-exponent of \( A_k \) is not necessarily smaller than 1. As Krapivsky and Redner have pointed out [27], the "winner takes all" phenomenon can easily be discovered by homogeneous superliner attachment kernels. That is, there exists a single dominant "gel" node, which is connected to a great (but finite) number of nodes in the network while most nodes have very few links. Similar behaviors also occur in our model, and the only difference is that there may be more than one local "gel" nodes that can be found because the whole network is divided by similarity. Although a closed-form solution to the degree distribution is hard to derive here, we have conducted a thorough numerical analysis on the effect of two different mechanisms on the formation of distinct features. Moreover, both the real-world datasets and simulation results showed in the next section will support the above analytical seasoning.

3. Model simulation results

In this section, we first analyze the characteristics of the GCNM model under various adjustable parameters, and then make the comparison between the results obtained from simulated tests and four real-world collaboration networks. It will show that our model can reproduce the characteristics of network as observed in these real networks.

3.1. Effects of various model parameters

The effects of three main parameters on simulation results are discussed. They are the weight of "similarity" and "shared neighbor" \( \alpha \), similarities in attributes \( \gamma \), and similarities in structure \( \lambda \) are respectively. The number of nodes for each simulation network is set as 200,000, and the average links for each node is 10. Each test was run 100 times and the mean (average) value is taken as final simulation results. In Figure 2(a, c, e), we fix the value of parameter \( \alpha \) as 0.6. And in Figure 2(b, d, f), the value of parameter \( \gamma \) is set to be 1.6.

Let us now examine the change of average clustering coefficient, net clustering coefficient and number of triangles with different values of three main parameters. When the weight of "similarity" and "shared neighbor" is fixed ( \( \alpha = 0.6 \) ), we can easily find that average clustering coefficient is
especially sensitive to the similarity in attributes. With the parameter $\gamma$ increasing its value from 0.5 to 3.5, the value of average clustering coefficient changes from 0.1 to 0.7. On the other hand, the net clustering coefficient and number of triangles are greatly affected by changes in the values of parameter $\lambda$. We all know that the net clustering coefficient represents the characteristics of a global clustering. So, we believe that similarity in structure can be helpful to form bigger communities.

$$\alpha = 0.6$$

$$\gamma = 1.6$$

![Graphs showing changes in clustering coefficients and number of triangles for different parameter values.](image)

Figure 2. The effects of three main parameters in GCNM model. Each test was executed 100 times and the mean is taken as final simulation results.
From Figure 2(b, d, f), we can see that the weight of "similarity" and "shared neighbor" $\alpha$ can also has great effect on network features. If there are more connections which are established by finding friends with similarity in structure (the value of $\alpha$ becomes smaller), then we are led to a higher net clustering coefficient and more number of triangles. Contrarily, if there are more links which are generated by step (2) in GCNM, the average clustering coefficient will increase gradually. In Figure 2, we can find that different characteristics responses to these parameters ($\alpha, \gamma, \lambda$) will be different. Crucially, however, all network features will be influenced to some extent if the value of any parameter is changed. So, the GCNM model can match various types of social networks in the real world by just setting different values for the parameter set.

### 3.2. Comparison with real-world networks

In this section, four real-world datasets from different fields are used to illustrate the proposed model.

- **Arxiv HEP-PH and Arxiv ASTRO-PH** [25] are two collaboration networks from the e-print arXiv, which cover the papers from High Energy Physics and Astro Physics during the period from January 1993 to April 2003 (124 months). In the two datasets, collaboration relations were established between two authors if they have published at least one paper together.

- **The DBLP** [26] computer science bibliography provides a comprehensive list of research papers in computer science. Here, we focus on symbiosis among authors.

- **Actor network data** [30] represent the collaboration graph of movie actors, in which two actors are connected if they have acted in the same movie.

#### Table 1. The basic statistical information of all these real-world data sets and the corresponding calculation results. Each test was executed at least 100 times and the mean is taken as final simulation results.

|                  | Actor | GCNM-Actor | Astro-Ph | GCNM-Astro-Ph | Hep-Ph | GCNM-Hep-Ph | Dblp | GCNM-Dblp |
|------------------|-------|------------|----------|---------------|--------|-------------|------|-----------|
| Nodes            | 382219| 382219     | 18772    | 18772         | 12008  | 12008       | 1199464| 1199464   |
| Average degree   | 78.69 | 78.69      | 21.1     | 21.1          | 19.7   | 19.7        | 7.88  | 7.88      |
| Average clustering coefficient | 0.78  | 0.728      | 0.631    | 0.645         | 0.611  | 0.616       | 0.635  | 0.640     |
| Net clustering coefficient | 0.166 | 0.173      | 0.318    | 0.336         | 0.660  | 0.664       | 0.180  | 0.188     |
| Number of triangles | 3.47e+8 | 3.86e+8   | 1.35e+6  | 1.42e+6       | 3.36e+6 | 3.58e+6     | 1.05e+7 | 1.16e+7   |

#### Table 2. The summary of parameter values for each simulation process.

|                  | GCNM-Actor | GCNM-ASTRO-PH | GCNM-HEP | GCNM-DBLP |
|------------------|------------|---------------|----------|-----------|
| $\alpha$         | 0.74       | 0.76          | 0.61     | 0.70      |
| $\gamma$         | 3.23       | 3.5           | 1.9      | 3.3       |
| $\lambda$        | 1.2        | 1.57          | 2.2      | 1.13      |

The basic statistical information of all these real-world data sets and the corresponding calculation results using the GCNM model are presented in Table 1. Table 2 is the summary of parameter values for each simulation process. By comparing the essential features of each data set, we could find that every network has its unique characteristics, e.g. The DBLP is the largest social network, has more than one million nodes, and ACTOR has the greatest number of triangles. The size of HEP-PH is smaller than ASTRO-PH, but the net clustering coefficient and the number of triangles formed by HEP-PH are much greater than the ones of ASTRO-PH. For this reason, we can infer that the network structure of HEP-PH is more closely clustered than the structure of ASTRO-PH. One possible
A possible explanation for this finding is that researchers who work in High-Energy physics need stronger cooperative relationship than those who are dedicated themselves to ASTRO Physics. In spite of the differences, they share some characteristics in common: 1) It is generally known that the clustering of the network produced by PA model tends to vanish as the size grows. However, the clustering coefficient of four real-world networks far outweighs zero. Moreover, they all have small average path length and huge number of triangles. All of these characteristics show a distinct community structure in networks, and are well in line with the general definition of social network. From Table 1, we can observe another fascinating phenomenon that GCNM model proposed in this paper has a strong capability to reproduce the networks with the same characteristics as real-world networks from different areas. Therefore, we could say that GCNM model could well explain the growth processes of various kinds of real-world networks.

In order to make a more detailed comparison of the topological properties between simulated networks (black curves) and four empirically studied ones (red curves), four different groups of curves for degree distribution(a), degree-clustering coefficient(b), degree-degree correlation(c), and degree-common neighbor(d) are displayed in Figure 3. First of all, let us point out that the four different groups of curves for the simulated networks (black) match pretty well the corresponding curves for four real-world networks (red). This consistency indicates the effectiveness of GCNM model. Now, let us give descriptions and analysis of the characteristics of each curve graph as follows.

After the scale-free theory of social network was proposed a decade ago, degree distribution became a basic measure of social network. As showed in Figure 3 (a)(1-4), fat-tailed distribution of degree can be easily observed, which is consistent with the previous recognition. Another thing to note is that exponential cut-off function fits the degree distribution better than power-law distribution when degree is small (e.g. k<50).

In Figure 3 (b) (5-8), degree-clustering coefficient curves represent the phenomenon that clustering coefficient decreases as the degree increases. What surprises us is that the Figure 3 (b)(5-8) show that “star” nodes (the nodes with relatively or very high degrees) are also likely to have relatively or very strong local clustering effects as indicated by respective peak values in the figure. This means that friends of an influential person (who has very many cooperators) may as well link with each other. More or less, each neighbor of a "star" node is probably another "star" node, which can be evidenced in the degree-degree correlation curves shown in Figure 3(c) (9-12). 3) As it can be seen from Figure 3(c) (9-12) that there exists a clear assortative mixing in a variety of types of social networks. Particularly, high degree nodes are more preferred to connect with other "star" nodes.

Now let us analyze the relationship between degree and the number of potential friends, which is little discussed in previous literatures. The number of probable neighbors equals the value of $A_{k,2}$ when parameter $r$ is set to 1. As we can be seen from the curve set for the degree-common neighbor in Figure 3(d)(13-16), that the number of potential neighbors grows with the degree in a super linear way, which backs the conclusion made in section 2.3. However, if the degree is big enough (e.g. exceeds 200 in Figure 3(d)(15)), the increasing rate of the number of possible friends is expected to have little change or decrease. At the beginning, the more neighbors you have, the more probable friends you will make in the future. But, when the node gradually becomes a "star", most of its neighbors also become friends with each other and the degree of potential new friends is going down. Therefore, the number of potential new friends introduced by neighbors is at a slower rate of increase or even decrease. This provides a certain explanation for why we cannot find a super "star" node in the real-world, who connects to all the rest nodes in each network. In other words, in a world of finite resources, any single individual is unlikely to possess all the resources.
Figure 3. The detailed comparison of the topological properties between simulated networks (black) and the real-world ones (red). Group (a) (1-4) show the curves for degree distribution curve, group (b) (5-8) show the curves for degree-clustering coefficient, group (c) (9-12) show the curves for degree-degree correlation, and group (d) (13-16) show the curves for degree-common neighbour.

4. Conclusions
In this paper, we proposed a general collaboration network model (GCNM model), in which a connection can be made between two nodes if they are similar in attributes or similar in structure. Theoretical analysis and experimental results indicate that structural similarity can lead to generate more triangles and more compact structures within groups, while attribute similarity is likely to produce higher clustering degree and larger communities with a relatively loose structure. The GCNM model is able to generate a simulation network with an enormous number of triangles and very high
global clustering coefficient. And the comparison with the results obtained from simulated tests and four real-world collaboration networks demonstrates that our model can reproduce the characteristics of network as observed in these real collaboration networks.

The combination of similarity-based method and copying-based method is promising to obtain the advantages of both strategies. Through adjusting parameters, the weight of "similarity" and "shared neighbor" in the formation of simulation networks can be changed, and the attachment kernel of GCNM model could depend linearly or super linearly on degree. So, it is not surprised that the GCNM model can well match varying types of collaboration network in real-world by simply choosing different values of three main parameters. Particularly, the widely studied three types of models: the similarity-based model, the copying-based model, and the pure PA model are only the special cases of our new model. In the future, we will further test our model on other types of social networks, such as citation networks and travel route networks.

Acknowledgments
This work was supported in part by Department of Education of Guangdong Province under Special Innovation Program (Natural Science), grant number 2015KTSCX183 and in part by South China University of Technology under “Development Fund” with fund number x2js-F8150310.

References

[1] Ren F X, Shen H W and Cheng X Q 2012 "Modeling the clustering in citation networks" Physica A: Statistical Mechanics and its Applications 391(12) pp. 3533-3539
[2] Wu Z X and Holme P 2009 "Modeling scientific-citation patterns and other triangle-rich acyclic networks" Physical review E 80(3) 037101
[3] Camarinha-Matos L M and Afsarmanesh H 2005 "Collaborative networks: a new scientific discipline" Journal of Intelligent Manufacturing 16(4-5) pp. 439-452
[4] Newman M E 2001 "The structure of scientific collaboration networks" Proceedings of the National Academy of Sciences vol 98(2) pp. 404-409
[5] Moriano P and Finke J 2013 "On the formation of structure in growing networks" Journal of Statistical Mechanics: Theory and Experiment 2013(06) P06010
[6] Zhang P P et al. 2006 "Model and empirical study on some collaboration networks" Physica A: Statistical Mechanics and its applications 360(2) pp. 599-616
[7] Newman M E 2004 "Coauthorship networks and patterns of scientific collaboration" Proceedings of the National Academy of Sciences of the United States of America vol 101(Suppl 1) pp. 5200-5205
[8] Newman M E 2001 "Scientific collaboration networks. I. Network construction and fundamental results" Physical review E 64(1) 016131
[9] Becatti C, Crimaldi I and Saracco F 2019 "Collaboration and followership: A stochastic model for activities in social networks" PloS One 14(10) e0223768
[10] Sharma A and Bhavani S D 2019 "A network formation model for Collaboration networks" Proceedings of 15th International Conference of Distributed Computing and Internet Technology (Bhubaneswar, India, January, 2019), Eds G. Fahrnberger et al. (Eds.): ICDCIT 2019, LNCS 11319 pp. 279–294 https://doi.org/10.1007/978-3-030-05366-6_24
[11] Crimaldi I, Del Vicario M, Morrison G et al. 2017 "Modelling networks with a growing feature-structure" Interdiscip. Inf. Sci. 23(2) pp. 127-144
[12] Lv L and Zhou T 2011 "Link prediction in complex networks: A survey" Physica A: Statistical Mechanics and its Applications 390(6) pp. 1150-1170
[13] Newman M 2009 Networks: an introduction (Oxford University Press)
[14] Toivonen R et al. 2009 "A comparative study of social network models: Network evolution models and nodal attribute models" Social Networks 31(4) pp. 240-254
[15] Barabási A L and Albert R 1999 "Emergence of scaling in random networks" Science 286(5439) pp. 509-512, https://doi.org/10.1126/science.286.5439.509
[16] Watts D J and Strogatz S H 1998 "Collective dynamics of ‘small-world’networks" Nature 393(6684) pp. 440-442
[17] Cheng X Q et al. 2009 "Triangular clustering in document networks" New Journal of Physics 11(3) 033019
[18] Holme P and Kim B J 2002 "Growing scale-free networks with tunable clustering" Physical Review E 65(2) 026107
[19] Menczer F 2004 "Evolution of document networks" Proceedings of the National Academy of Sciences of the United States of America vol 101(Suppl 1) pp. 5261-5265
[20] Li Y et al. 2009 "Linking via social similarity: The emergence of community structure in scale-free network" in Web Society, SWS'09. 1st IEEE Symposium on. IEEE.
[21] Papadopoulos F et al. 2012 "Popularity versus similarity in growing networks" Nature 489(7417) pp. 537-540
[22] Toivonen R et al. 2006 "A model for social networks" Physica A: Statistical Mechanics and its Applications 371(2) pp. 851-860
[23] Jackson M O and Rogers B W 2007 "Meeting strangers and friends of friends: How random are social networks?" The American economic review pp. 890-915
[24] Leskovec J, Kleinberg J and Faloutsos C 2005 "Graphs over time: densification laws, shrinking diameters and possible explanations" Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining
[25] SNAP: Stanford Network Analysis Platform. http://snap.stanford.edu
[26] DBLP Data Set, http://dblp.uni-trier.de/xml
[27] Krapivsky P and Redner S 2003 "Rate equation approach for growing networks" Statistical mechanics of complex networks (Springer) pp. 3-22
[28] Barabási A L, Albert R and Jeong H 1999 "Mean-field theory for scale-free random networks" Physica A: Statistical Mechanics and its Applications 272(1) pp. 173-187
[29] Bickel P J, Chen A and Levina E 2011 "The method of moments and degree distributions for network models" The Annals of Statistics 39(5) pp. 2280-2301
[30] UCDP Actor Dataset 2.1, 2012 Uppsala Conflict Data Program, www.ucdp.uu.se