CONSTRUCTION OF SOCIAL NETWORK FORMATION MODEL BASED ON CLUSTER-CLUSTER AGGREGATION MODEL

YIFENG DOU AND HAILING XIONG*

College of Computer and Information Science
Southwest University
No. 2, Tiansheng Rd., Beibei District, Chongqing 400715, P. R. China
1204840733@qq.com; *Corresponding author: xionghl@swu.edu.cn

ABSTRACT. *The paper focuses on the formation of social network, putting forward a social networking model based on Cluster-Cluster Aggregation, whose algorithm is combined with social networking features of transitivity and polymerizability. This being the model's evolution rule, MATLAB is adopted to perform 3D simulation, visualizing the forming process and analyzing features like node degree and its distribution. Through the analysis of the total number N of particles, cube side length L and the weight maximum M_W all have influence on the node degrees; meanwhile, the distribution of node degree is akin to Poisson distribution, and the experimental results indicate the model has high clustering coefficient.*
Keywords: Social network, Cluster-Cluster Aggregation, Visualization, Node degree

1. **Introduction.** Social network [1, 2, 3, 4] which is social networking services (SNS), refers to the social network structure established on friendship, kinship, trade, network link, the spread of diseases, ideals, hobbies, etc. Researches show that social network covers all aspects of society ranging from the family relationship to the national diplomacy. Against the background of big data and the Internet+ era, through the analysis of the social networking site data, the important features in the process of social network formation and evolution can be summed up as follows [5]. (1) Property of growth. As the number of new users is increasing, the total number is increasing. (2) Coexistence between preferential connection and random connection. New users will establish connection with those who they are familiar with after being enrolled in the social network, and meanwhile they can randomly connect to some users to make friends. (3) Polymerizability. Users who have the same properties are usually closer and they can form a small cluster. In fact, the entire network consists of many such small clusters which connect to each other. (4) Transitivity. Users establish direct relationship with the friends of theirs. For instance, if user A establishes relations with user B and user C respectively, B and C are directly connected. Social network is different from ordinary basic networks which are based on real relationships, so the study of the structural characteristics of the network to explore how to quickly build a social network is very meaningful to further promote the development of network.

The rest of the paper is organized as follows. Section 2 reviews the research situation about social network formation. Section 3 briefly introduces Cluster-Cluster Aggregation (CCA) model [6] and social network formation model. Also a model based on CCA formation model is put forward, which implements visualization of the formation process. Section 4 presents and analyzes the statistics of node degrees of the social network formation model and their distribution characteristics. And Section 5 draws the conclusion and gives an expectation on the research direction of future social network model construction.

2. **Related Work.** Through comparison between the social network and Barabasi-Albert (BA) scale-free network model [7], Xiong et al. [5] found that social network had different personalized features, such as clustering and random connections of a probability. They proposed a hybrid reconstruction model to better describe the formation characteristics of the social network. Liu [8] proposed a social network construction model based on random walk algorithm. It improves the network connectivity, and can better restore network structure. However, through polymerized character and transitivity mentioned in the introduction, it can well reflect clustering feature of the complex network. The computational method of the clustering feature can be equivalent to the ratio between the number of triangles connected to the node i and the number of triangle elements connected to node i. The number of triangle elements connected to node i represents the maximum of connection probably existing, and the triangles connected to node i stand for the relation among nodes, which means there is probably correlation between node i and the neighbor nodes related to node i. Here, this kind of transitivity is called the network joint relationship, which is the joint relationship that gives the social network high clustering feature. The joint relationship is taken as an evolution rule, based on which CCA model is adopted in order to show the forming process with visualization of 3D simulation.

3. **Construction of Social Network Model Based on CCA.**

3.1. **The mechanism of CCA model.** First of all, a three dimensional array, $CUBE[L][L][L]$, is used to represent the cubic box system, for L is the side-length of the cube. The number of all the single particles (monomers) N can be calculated by the following Equation (1) with the particles concentration C given.

$$N = C \times L^3 \tag{1}$$

Then these N particles are randomly distributed in the cube with non-overlapping. Each particle occupies an element of the three-dimensional array and is labeled with a label using a positive integer, therefore on-lattice. In on-lattice simulation, the cube is divided into equal-size little boxes [9], and side-length of the little box is regarded as a unit length. Initially we can regard each single particle as a special cluster of size 1. When initial work is over, in the cube system, every cluster can be selected randomly to do potential Brownian motion by the diffusion probability P_{move}. Assuming some factors can be ignored, the cluster's diffusion probability P_{move} is only affected by its mass S, which is expressed by the number of particles contained in the cluster. An Equation (2) is simply launched in the following for P_{move}:

$$P_{move} = h \times S^\gamma \tag{2}$$

where γ ($-0.5 < \gamma < 0.5$) and h denote diffusivity exponent and diffusivity coefficient respectively. P_{move} is set to 1 if greater than 1. For the potential movement, a random number X uniformly distributed over the range 0 to 1 is generated and the cluster gets movement only if $P_{move} > X$. In the process of moving, the selected cluster diffuses a unit length with its all particles, it can move to six directions (up, down, left, right, front or back) and for each moving step, the moving direction is chosen randomly among the six directions. If the selected cluster does not collide with another one, the algorithm will go on with choosing another one for next step. If a collision event occurs between two clusters (one consists of i particles, and the other consists of j particles) during movement, they stick together to form a new larger one by the sticking probability P_{ij} [10, 11]:

$$P_{ij} = P1 \times (i \times j)^\sigma \tag{3}$$

where parameter $P1$ is the sticking probability of two single particles, and parameter σ ($-0.5 < \sigma < 0.5$) is the sticking exponent. P_{ij} is set to 1 if greater than 1. A random

number Y is generated over the range 0 to 1, and aggregation is considered effective only when $P_{ij} > Y$ [6].

3.2. **Social network formation specific algorithm based on CCA model.** The social network formation model simulates the network changing from the randomly distributed non-adjacent particles to size N. This embodies the growth of the network. Whether indirect connection is formed between new nodes i and all neighbor nodes of node j according to the introduced external parameters α reflects the joint relationship between network nodes.

An $N \times N$ adjacency matrix $W = (w_{ij})_{N \times N}$ is used to represent the network model, and the elements w_{ij} in the matrix W denote whether there is a connection between nodes i and j. If the value of w_{ij} is zero, it means that there is no connection between two nodes. We know that each individual in the network corresponds to a node; direct connection between nodes means that the two nodes have a direct relation; the weight between two nodes is the correlation strength between individuals. The algorithm is divided into initialization and outward spreading growth, and the details are as follows.

(1) Initialization

Set a cube of fixed length L.

Randomly distribute n particles within the cube (assuming a single particle is a cube whose side length is 1 and density is C), and make sure any two particles are not adjacent.

Each particle can only move in 6 positive directions, up, down, left, right, front and back. And they must move 1 unit each time.

In an $N \times N$ matrix, if the element w_{ij} of W is not -1, it denotes there is some connection between i and j, and w_{ij} shows the weight between the two points.

Set an external parameter $MAX\_WEIGHT(M\_W)$ to control indirect connection between nodes.

(2) Increasing

Choose one of the particles and a direction randomly.

Judge whether the particle moves in the range. If not, return to step.

Move the selected particle/particle group 1 unit in the selected direction.

Judge whether moved particle or every particle in the group is adjacent to new particles. If not, return to step.

Enroll new particles encountered in step in the moving particles or particle group and set the $WEIGHT$ of the connection between the two particles as 1.

Judge whether new enrolled particles have a correlation with the original neighbor of the particles adjacent to the new ones. According to the control of the parameter M_W, repeat steps until the particle group's size is N.

3.3. **Visualization realization.** Within the cube of length $L = 80$, 5120 non-adjacent particles are initialized randomly. The total number of communities is 5120. The results are shown in Figure 1(a). A particle i is selected randomly to move 1 unit in the arbitrary direction. If the adjacent particle j is discovered, particle i is added in the community j. The connection between i and j generates $WEIGHT[i][j]$, and $WEIGHT[i][j]$ is 1. Meanwhile it is necessary to know whether there is an indirect correlation between particle j and particle k that is indirectly related to particle i, and the same with particle i. By judging whether $WEIGHT[k][i]$ plus $WEIGHT[i][j]$ is less than M_W, the size of the M_W can be adjusted to control probability of indirect connection between nodes. Polymerization process is shown in Figures 1(b)-1(h). n stands for the number of the rest communities; (b) means the number of the rest communities is reduced to 4096 after polymerization; (c) is the number of the rest communities which is reduced to 3002; (d) is the rest communities reduced to 2160; (e) means the rest communities are reduced to 1200; (f) is when the rest communities are reduced to 320; (g) is when the rest communities

are reduced to 20; (h) is when the rest communities are reduced to one, namely the final condition of the polymerization.

In the process of particle aggregation, the number of communities within the whole cube is decreasing until the number is 1. When the polymerization process is completed, all particles gather in a community and form a whole, with this method the social network model being quickly built.

4. **Discussion.** It is known that the degree distribution of complex networks and network topology structure are closely related. We analyzed in statistics the node degree of the social network formation model and the node degree distribution.

4.1. **The analysis of node degree.** The degree k_i of node i in complex networks is defined as the number of the other nodes that the node is connected to, and the average of all the node degrees is called node average degree of a network, recorded as $\langle k \rangle$. Node degree reflects the status of a node in the whole community. In the social network, the bigger node degree a person has, the greater influence the person has. In this model, the weight of two directly connected nodes defaults to 1, and the weight of two non-adjacent nodes is the sum of weights of indirectly connected nodes. Therefore, the smaller weights the connection has, the more likely it is to provide more nodes with indirect connection. The greater the mutual influence between the nodes is, the smaller their weight becomes. Through the analysis the total number N of particles, cube side length L and the weight maximum M_W all have influence on the node degrees.

Figure 2(a) shows when M_W is 10 and L is 30, the total number N of particles influences node degrees. When the side length L and the maximum M_W are fixed, the greater the number N of particles is, the more compact the final community becomes. As the indirect connection among nodes increases, the average node degree becomes bigger. So in social network, with the increase of the total number of users, the more mutual communications users have, the more complex the network becomes.

Figure 2(b) shows the influence of side length L on the node degree. When other conditions stay the same, the longer the length L is, the smaller the average node degree is. It is known that when the side length becomes longer, the density of particles in the space gets relatively small, making the probability to form an object with complex structure less and the relation between nodes less closer. In social network, it is always easy to make friends with those who people are close to. In the circle of friends, there is relatively complicated interpersonal relationship, so it is usually hard to make friends with those who they are unfamiliar with.

Figure 2(c) shows the impact of the maximum M_W on node degree when L is 20 and N is 500. With the increase of M_W, the relationship between nodes is more complicated. If $DEGREE$ is regarded as a person's influence, the greater the influence is, the smaller the value of $DEGREE$ is. In other words, they are in inverse proportion. Therefore, M_W is the biggest influence scope of a person. From the chart above, it can be seen that when the influence scope of a person is large enough, he is likely to affect all the other people in his circle. However, in real life, as the law of 150 people shows us, the number of people each person may meet in our life is limited, so in social network, not all users will be affected by someone.

4.2. **The analysis of node degree distribution.** The degree distribution of nodes is denoted as $P(k)$, which is a distribution function. Degree distribution represents the probability as k of a randomly selected node degree. Figure 3 shows the node degree distribution curve when the L is 60, N is 2000 and M_W is 10. When the node degree N is distributed near the average node degree $\langle k \rangle$, nodes are distributed very closely within this value, which means most of node degree values are close to the average node degree $\langle k \rangle$. It is completely contrary to the situation away from the average node degree
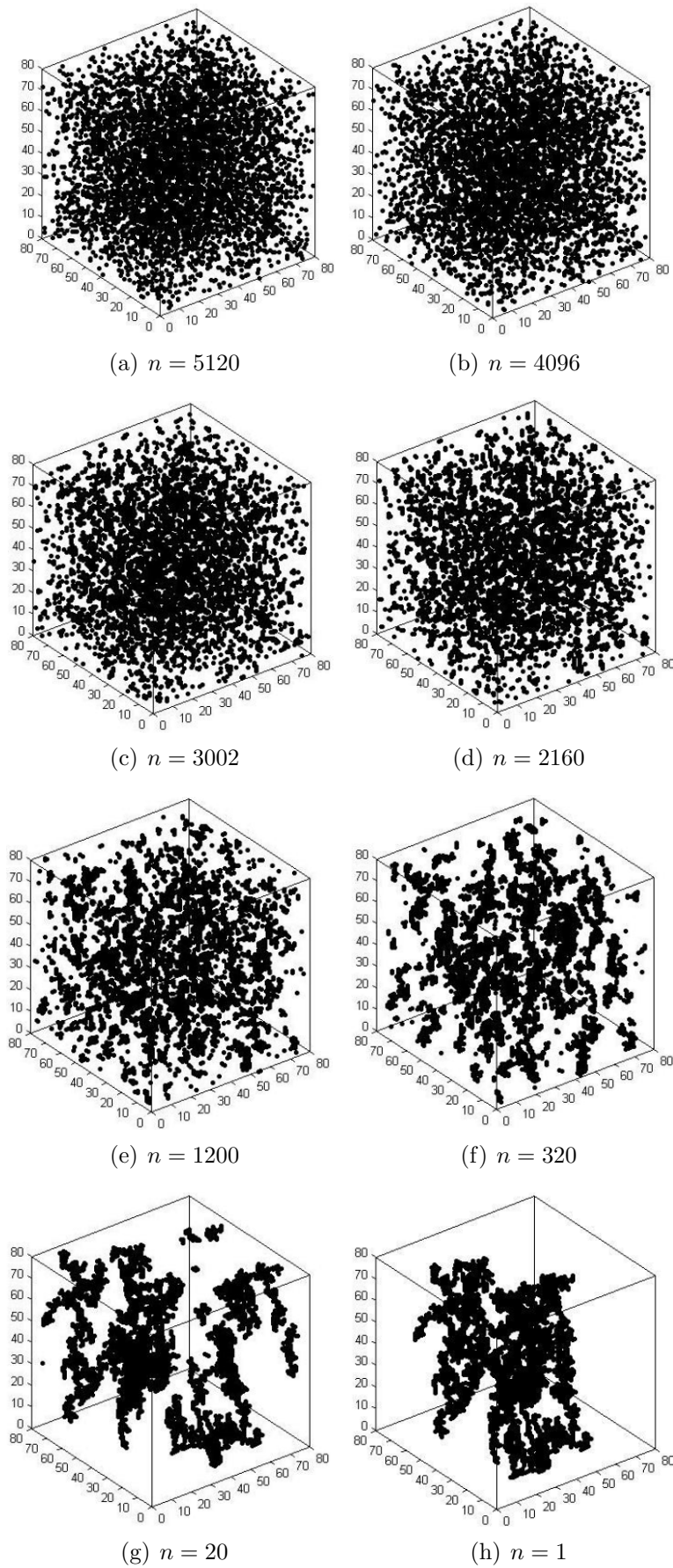
(a) $n = 5120$

(b) $n = 4096$

(c) $n = 3002$

(d) $n = 2160$

(e) $n = 1200$

(f) $n = 320$

(g) $n = 20$

(h) $n = 1$

FIGURE 1. Non-adjacent N particles from initialization to polymerization
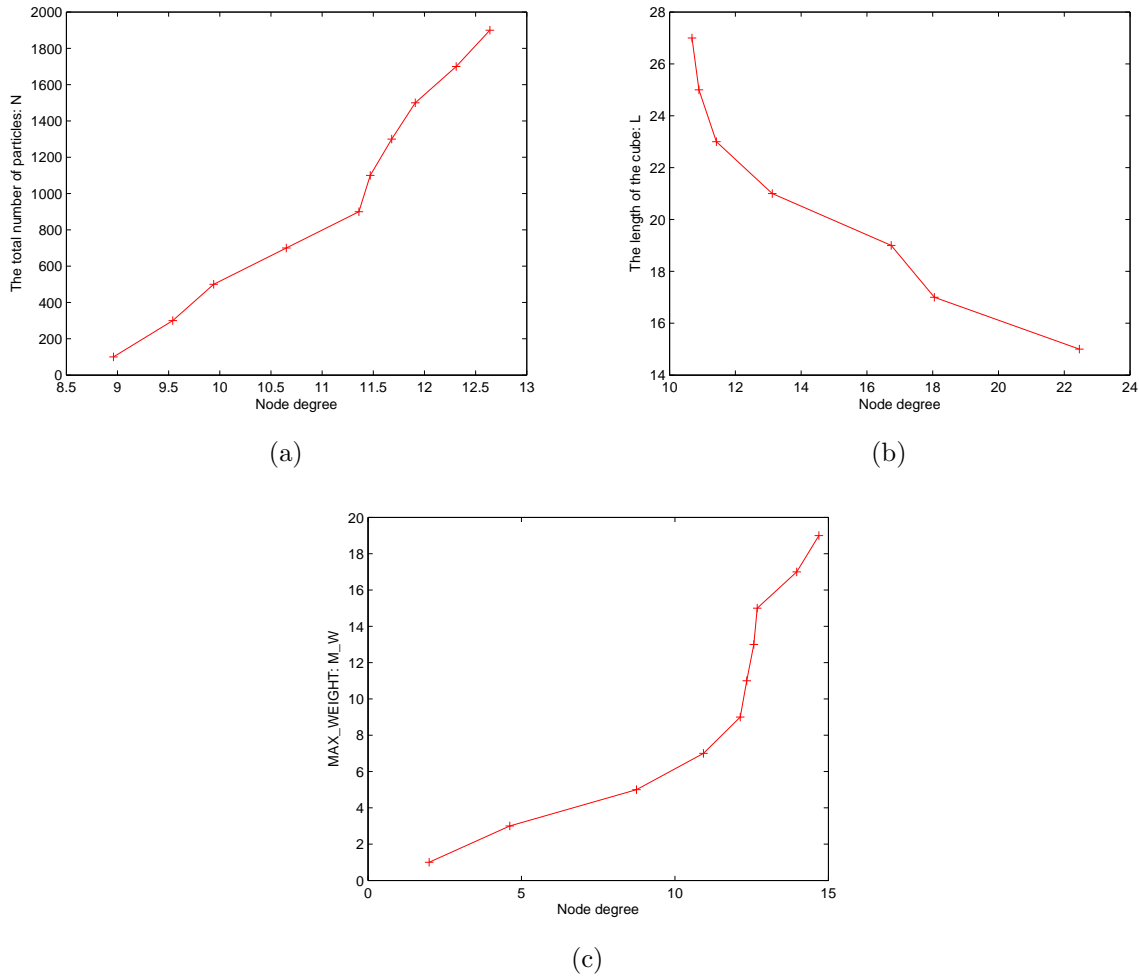
(a)

(b)

(c)

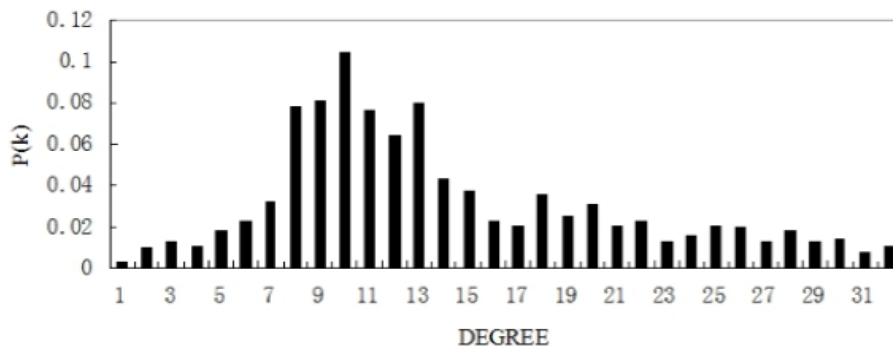FIGURE 2. The analysis of influence factors of node degree

FIGURE 3. The degree distribution when L is 60, N is 2000, M_W is 10

$\langle k \rangle$. Clustering coefficient reflects the relevance of network nodes. Clustering coefficient is usually used to describe affinity or disaffinity between neighbor nodes. This result is akin to Poisson distribution, and the experimental results indicate the model has high clustering coefficient.

5. **Conclusions.** In this paper, we mainly study the formation process of the social network and propose the social network formation model based on CCA model, realizing the visualization of the formation process through 3D simulation with MATLAB and carrying out the statistical analysis of node degree and its distribution characteristics. In the network model built in this paper, the connection number of a single particle is

limited. The future study may focus on a sphere instead of the present form with more random moving directions, making the process more diverse. Meanwhile, we will compare with other methods in the previous works to illustrate the advantages of the model in future studies.

REFERENCES

[1] D. M. Boyd and N. B. Ellison, Social network sites: Definition, history, and scholarship, *Journal of Computer-Mediated Communication*, vol.13, no.3, pp.210-230, 2007.

[2] J. H. Fowler and N. A. Christakis, The spread of obesity in a large social network over 32 years, *New England Journal of Medicine*, vol.357, no.4, pp.370-379, 2007.

[3] J. E. Garrison, Clinical construction of action social networks, *International Journal of Family Therapy*, vol.3, no.4, pp.258-267, 1981.

[4] P. S. Chinowsky, J. Diekmann and J. O'Brien, Project organizations as social networks, *Journal of Construction Engineering & Management*, vol.136, no.4, pp.452-458, 2014.

[5] X. Xiong, W. Cao, X. Zhou et al., Research on the feature model of the formation and evolution of social networks, *Sichuan Daxue Xuebao*, vol.44, pp.140-144, 2012.

[6] C. Li and H. Xiong, 3D simulation of the cluster-cluster aggregation model, *Computer Physics Communications*, vol.185, no.12, pp.3424-3429, 2014.

[7] A. L. Barabasi and R. Albert, Emergence of scaling in random networks, *Science*, vol.286, no.5439, pp.509-512, 1999.

[8] H. Liu, Random walk algorithm based on social network construction, *Journal of Hubei Adult Education Institute*, vol.2014, no.1, pp.11-12, 2014.

[9] H. Xiong, H. Li, W. Chen et al., Data structure for on-lattice cluster-cluster aggregation model performance optimization, *Computer Physics Communications*, vol.185, no.3, pp.836-840, 2014.

[10] P. Meakin, A historical introduction to computer models for fractal aggregates, *Journal of Sol-Gel Science and Technology*, vol.15, no.2, pp.97-117, 1999.

[11] P. Sandkühler, M. Lattuada, W. Hua et al., Further insights into the universality of colloidal aggregation, *Advances in Colloid & Interface Science*, vol.113, nos.2-3, pp.65-83, 2005.