Analysis of Clustering Coefficients of Online Social Networks by Duplication Models

Duan-Shin Lee, Cheng-Shang Chang, Wen-Gui Ye and Min-Chien Cheng

Institute of Communications Engineering
National Tsing Hua University
Hsinchu 300, Taiwan, R.O.C.
Email: lds@cs.nthu.edu.tw, cschang@ee.nthu.edu.tw, foliage246@gmail.com, great770203@gmail.com

***Abstract*—In this paper we propose to model the formation of online social networks by a duplication model. In this model vertices are added into the network one at a time. Each vertex is first attached to a randomly selected vertex. Each neighbor of the attached vertex establishes an edge with the new vertex with a probability. A main contribution of this paper is that we derive analytically the clustering coefficient for this model. Numerical studies show that the range of mean degree and the clustering coefficient of the duplication model is quite large. By properly choosing values for the parameters of our model, the mean degree and the clustering coefficient match well with those of popular online social networks.**

keywords: social network, network growth, duplication model

I. Introduction

Social networks have attracted a lot of attention recently mostly because of the explosive popularity of online social networking sites such as Facebook, Flickr, Orkut and etc. These sites offer an integrated environment for users to make friends, chat, distribute and share images/video, play games and etc. In some part of the world, access to social networking sites constitutes nearly ten percent of all access to web sites [12]. Users may form social links with their real-life acquaintances or online contacts who share common interests. Networks formed in this way are called online social networks. Online social networks receive great interests because they offer opportunities for potentially new business models. For instances, products or concepts are conventionally promoted by celebrities such as movie or sports stars. However, potential customers may be more willing to take recommendation from friends. Thus, viral marketing over online social networks can be a power tool for business.

Social networks have been widely studied in the past. One of the most well known studies was due to Milgram [14]. Milgram's result leads to the idea of "six degrees of separation", a common belief that there are only six hops between any two people in the world. This result is referred to as the small world property. Mathematically, the small world property says that the average diameter of a social network is of the logarithmic order of the number of vertices in the network. Other important properties of social networks include [16], [20]

- large clustering coefficients;
- power law degree distributions;
- positive degree-degree correlation;
- existence of community structures.

There has been a great research interest on modeling and understanding the micro operations that lead to the formation of social networks. The objective is to devise a growth model producing random networks that possess the properties listed in the last paragraph. Such a network model helps to understand how a social network is formed and may be useful to the study issues, such as viral marketing, on social networks. Long time ago social network analysts have observed a behavior called "*triadic closure*" [16]. Suppose that individual A has two friends, N and B, who are not friends to each other as shown in Figure 1. Quite likely A may introduce N and B to each other. N and B may become friends as a result. On a network graph, this operation "closes" an "open" triad of vertices by adding of an edge between N and B. Newman [17] has showed empirical evidence of triadic closure in scientific collaboration networks. Network models with operations similar to triadic closure have been proposed in the literature to model the formation of social networks. Kumpula *et al.* [13] proposed a model for social networks with community structures. In their model, at each time each vertex with at least one neighbor performs a random search to a second neighbor. If this second neighbor is not a neighbor, perform a triadic closure to this second neighbor. Kumpula *et al.* [13] called this operation *cyclic closure* or *local attachment*. Each vertex with no neighbor is made to connect to a random vertex with a probability. This operation is called *focal closure* or *global attachment*. Finally, in Kumpula's model vertices and edges can be removed. In addition, edges are weighted. Simulation methods were used to study the model. In Davidsen's model [5] a vertex is randomly selected from the network. If this vertex has more than one neighbors, randomly select two of its neighbors and connect them by triadic closure. This operation is called *transitive linking* by Davidsen *et al.* If the randomly selected has only one neighbor, connect this vertex with a random vertex. Finally, randomly select a vertex and remove all its edges. Then, connect this

vertex to a randomly selected vertex. In addition, Holme et. al. [9] and Szabó et. al. [19] studied network formation problems involving with preferential attachment and operations similar to triadic closure. We refer the reader to [9], [19] for more details.

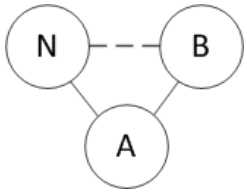

Fig. 1. Illustration of triadic closure. Vertices N and B are not friends, but have a common friend A. A introduces N and B. A new edge, shown in dashed line, closes the triad.

Nearly all online social networking sites offer a section called "people you may know". This section contains a list of users that the system recommends. By clicking on a user in the list, this user becomes a friend. Since this list of users contains mostly friend's friends, effectively in online social networks friend's friends are more likely to be friends of the new user. As illustrated also in Figure 1, suppose that N is a new user who just signs up to an online social networking site. Suppose that A is an initial friend of N and that B is one of users in the "people you may know" section that the site recommends to N. N may become a friend of B by establishing a new edge to B. We thus propose the following growth model for online social networks. Specifically, vertices are added into the network one at a time. At each time, a new vertex is added into the network. The new vertex randomly selects and attaches to one existing vertex in the network. Then, each neighbor of the randomly selected existing vertex establishes an edge with the new vertex with probability a. In fact, this network growth model that we just proposed is also called a duplication model in the study of biological networks [4], [3], [6], [8]. Duplication models have many variations. Chung *et al.* rigorously analyzed the scale-free degree distributions of duplication models [4], [3]. They also established bounds on the maximum degrees. A duplication model with edge rewiring was studied by Solè *et al.* [18]. Through computer simulation, Solè analyzed the clustering coefficient and the average path length. Bhan *et al.* [1] considered three duplication models. Through simulations, Bhan *et al.* studied the clustering coefficients, average path lengths and exponents of scale-free degree distributions of the three models. Ispolatov *et al.* [11], [10] studied the average degree and the average number of cliques in a duplication-divergence model. Through computer simulations Zhao *et al.* studied the Pearson degree correlation coefficient for several duplication-divergence models [22]. Boccaletti *et al.* [2] considered a model similar to the duplication model. In Boccaletti's model a new vertex is added to the network at each time. A new vertex randomly selects a vertex from the network and the neighbors of the selected vertex. The new vertex establishes m edges randomly to the selected vertex and its neighbors. Rate equations were derived for the degree distribution and the conditional degree-degree probability of this model.

To our knowledge, the clustering coefficient of Chung's duplication model has not been analytically studied. In this paper we propose to model the formation of online social networks by Chung's duplication model described as follows.

Duplication model:

(i) Initially, a clique of m_0 vertices is given. $t \leftarrow 1$.
(ii) A new vertex N is introduced.
(iii) A vertex, denoted by A, is randomly chosen from the existing network and an edge between vertex N and vertex A is added.
(iv) For every neighbor of vertex A, an edge between vertex N and that neighbor of vertex A is added with probability a. This is independent of everything else.
(v) $t \leftarrow t + 1$. Repeat (ii).

A main contribution of this paper is to derive a closed form expression of the clustering coefficient as a function of time for duplication model described above. Specifically,

$$C = \frac{2\tau(t)}{s(t) - k(t)}, \tag{1}$$

where $k(t)$ and $s(t)$ are the first and the second moments of the degree of a randomly selected vertex in the network at time t. In (1), $\tau(t)$ is the expected number of triangles that a randomly selected vertex has at time t. Later in this paper we shall derive differential equations for $k(t)$, $s(t)$, and $\tau(t)$. Our numerical experience with this model indicates that the range of mean degree and clustering coefficient of this model are quite large. By properly choosing values for the parameters, one can match the mean degrees and the clustering coefficients of some observed online social networks.

The rest of the paper is organized as follows. In Section II we review the definition of clustering coefficients. In Section III, Section IV, and Section V, we derive the first and the second moments of degree and the expected number of triangles of a randomly selected vertex. These quantities are needed in order to compute the clustering coefficient. Numerical and simulation results are presented in Section VI. The conclusions of the paper are presented in Section VII.

II. Clustering Coefficient

Clustering coefficients are designed as a measure of network transitivity, which is a very important property of social networks. Recall that there are two definitions of clustering coefficients [16]. A network-wide definition of clustering coefficients for a random network is

$$C = \frac{\text{(expected number of triangles)} \times 3}{\text{(expected number of connected triples)}}. \tag{2}$$

Let $\tau(t)$ be the expected number of triangles that a randomly selected vertex has at time t. Let $n(t)$ be the number of vertices at time t. Since each triangle has three vertices, it follows that the expected number of triangles of the entire network is

$$n(t)\tau(t)/3. \tag{3}$$

Consider a randomly chosen vertex and let its degree be X. The expected total number of connected triples of the network is

$$n(t)\mathsf{E}\left[\begin{pmatrix} X \\ 2 \end{pmatrix}\right] = n(t)\frac{s(t)-k(t)}{2}, \tag{4}$$

where $s(t)$ is defined as $\mathsf{E}[X^2]$ and $k(t) = \mathsf{E}[X]$. Substituting (3) and (4) into (2), we obtain (1). From (1), one needs to evaluate $k(t)$, $\tau(t)$ and $s(t)$ in order to compute the clustering coefficient C. We shall analyze these quantities in the following sections.

III. Expected Degree

Let $k(t)$ denote the expected degree at time t. We shall derive a differential equation for $k(t)$. We will first derive a difference equation by equating the total expected number of edges in the network right before and after time t. We then approximate the difference equation by a differential equation.

Since there are $m_0 + t$ vertices in the network at time t, clearly the total expected degree of the entire network is $(m_0 + t)k(t)$ at time t. Since each edge has two ends, the expected number of edges at time t is $(m_0 + t)k(t)/2$. The new vertex that arrives at time t randomly selects and attaches to an existing vertex. This produces a new edge. Additional $ak(t)$ edges are generated on average due to the TA operation. Thus, on average $1 + ak(t)$ new edges are generated. Thus, we have

$$\frac{(m_0+t+1)k(t+1)}{2} = \frac{(m_0+t)k(t)}{2} + 1 + ak(t). \tag{5}$$

Eq. (5) is a difference equation. We propose to approximate this difference equation by a differential equation. To achieve this, we rewrite (5) as

$$k(t+1) - k(t) = \frac{2 + (2a-1)k(t)}{m_0+t+1}. \tag{6}$$

We approximate the left hand side of (6) by derivative $k'(t)$ and obtain

$$k'(t) = \frac{2 + (2a-1)k(t)}{m_0+t+1} \tag{7}$$

with initial condition $k(0) = m_0 - 1$. If $a \neq 1/2$, (7) is a separable differential equation whose solution is

$$k(t) = c_1(m_0+t+1)^{2a-1} - \frac{2}{2a-1}, \tag{8}$$

where c_1 is a constant determined by the initial condition $k(0) = m_0 - 1$ and

$$c_1 = (m_0 - 1 + \frac{2}{2a-1})(m_0+1)^{1-2a}. \tag{9}$$

Solution (8) is not valid for $a = 1/2$. This special case can also be solved easily. We omit the details.

IV. Expected Number of Triangles

In this section we shall first derive a difference equation for $\tau(t)$ by equating the expected total number of triangles in the network right before and after time t. We then approximate the difference equation by a differential equation.

In the current model, the total number of vertices at time t is $n(t) = m_0 + t$. Since each triangle has three vertices, the expected total number of triangles at time t is $(m_0+t)\tau(t)/3$. Thus, we reach the following identity

$$\frac{(m_0+t+1)\tau(t+1)}{3} = \frac{(m_0+t)\tau(t)}{3} + ak(t) + a^2\tau(t). \tag{10}$$

We now explain the last two terms in the right hand side of (10). Suppose that at time t a new vertex N is attached to vertex A as shown in Figure 2. With probability a, an edge between vertex N and a neighbor of A is established. These two new edges introduce two new triangles NAA_1 and NAA_2. In addition, with probability a^2 triangle NA_1A_2 is formed. Thus, $k(t)$ new triangles could be introduced. Each is introduced with probability a independently. In addition, $\tau(t)$ new triangles could also be introduced. Each is introduced with probability a^2 independently. We approximate $\tau(t+1) - \tau(t)$ by $\tau'(t)$. Eq. (10) can be approximated by the following differential equation

$$\tau'(t) = \frac{3ak(t) + (3a^2-1)\tau(t)}{m_0+t+1} \tag{11}$$

with initial condition

$$\tau(0) = (m_0-1)(m_0-2)/2. \tag{12}$$

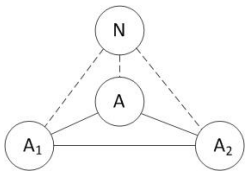

Fig. 2. A new vertex, denoted by vertex N, is attached to vertex A. Two new triangles NAA_1 and NAA_2 are formed, each with probability a. Triangle NA_1A_2 is formed with probability a^2.

In general, (11) is a first-order linear differential equation that can be solved by the technique of integrating factors. Specifically,

$$\tau(t) = c(m_0+t+1)^{3a^2-1} + (m_0+t+1)^{3a^2-1}\int \frac{3ak(t)}{(m_0+t+1)^{3a^2}}\,dt, \tag{13}$$

where c is a constant to be determined by the initial condition (12). If $a \neq 1/2$ or 2/3, substitute (8) into (13) and integrate. We obtain

$$\tau(t) = \frac{3ac_1(m_0+t+1)^{2a-1}}{2a-3a^2} - \frac{6a}{(2a-1)(1-3a^2)} + c_4(m_0+t+1)^{3a^2-1}, \tag{14}$$

and

$$c_4 = \Big(\frac{(m_0-1)(m_0-2)}{2} - \frac{3ac_1(m_0+1)^{2a-1}}{2a-3a^2} + \frac{6a}{(2a-1)(1-3a^2)}\Big)(m_0+1)^{1-3a^2}. \quad (15)$$

Solution (14) is not valid for $a = 1/2, 1/\sqrt{3}, 2/3$. For these special values of a, the corresponding differential equation can also be solved easily. We omit the details.

V. Second Moment of Degrees

In this section we shall first derive a difference equation for $s(t)$. We then approximate the difference equation by a differential equation. Let X be the the degree of A. Let Y_i be the degree of the i^{th} neighbor of A, where $i = 1, 2, \ldots, X$. In addition, let $U_1, U_2, \ldots,$ be a sequence of independent and identically distributed Bernoulli random variables with $\Pr(U_i = 1) = a$. Let

$$\phi(t) \stackrel{\text{def}}{=} \mathsf{E}\left[\sum_{i=1}^{X} Y_i\right]. \quad (16)$$

Recall that $s(t)$ is the second moment of the degree of a random selected vertex in the network at time t. Thus, $(m_0 + t)s(t)$ is the total second moment of degrees of the network at time t. We need to examine all the vertices whose degrees have changed from time t to $t+1$. Suppose that the degrees of a vertex at time t and $t+1$ are k and $k + \Delta k$, respectively. In the current model, $\Delta k \geq 0$. This vertex contributes

$$\mathsf{E}[2k \times \Delta k + (\Delta k)^2] \quad (17)$$

to the difference

$$(m_0 + t + 1)s(t+1) - (m_0 + t)s(t). \quad (18)$$

- Consider vertex A. Its degree at time t is X and the change is one. In view of (17), A's contribution to the quantity in (18) is

$$\mathsf{E}[2X + 1] = 2k(t) + 1. \quad (19)$$

- Consider vertex N. Obviously, its contribution to (18) is

$$\mathsf{E}\left[\left(1 + \sum_{i=1}^{X} U_i\right)^2\right]. \quad (20)$$

- Consider the neighbors of A. Consider the i^{th} neighbor of A. Its degree at time t is Y_i and the degree change is one. Thus, its contribution to (18) is $\mathsf{E}[2Y_i + 1]$. Hence, the total contribution from all neighbors of A is

$$\mathsf{E}\left[\sum_{i=1}^{X}(2Y_i + 1)U_i\right]. \quad (21)$$

We now simplify (20) and (21). We first work on (20). We have

$$\mathsf{E}\left[\left(1 + \sum_{i=1}^{X} U_i\right)^2\right] = \mathsf{E}\left[1 + 2\sum_{i=1}^{X} U_i + \left(\sum_{i=1}^{X} U_i\right)^2\right]$$
$$= 1 + 2\mathsf{E}\left[\sum_{i=1}^{X} U_i\right] + \mathsf{E}\left[\sum_{i=1}^{X}\sum_{j=1}^{X} U_i U_j\right] \quad (22)$$

From the Wald's equation [7],

$$\mathsf{E}\left[\sum_{j=1}^{X} U_i\right] = a\mathsf{E}[X] = ak(t). \quad (23)$$

In addition,

$$\mathsf{E}\left[\sum_{i=1}^{X}\sum_{j=1}^{X} U_i U_j\right] = \mathsf{E}\left[\sum_{i=1}^{X} U_i^2\right] + \mathsf{E}\left[\sum_{i=1}^{X}\sum_{j\neq i} U_i U_j\right]. \quad (24)$$

The first term in the right hand side of (24) can be simplified by Wald's equation, which gives

$$\mathsf{E}\left[\sum_{i=1}^{X} U_i^2\right] = ak(t). \quad (25)$$

The last term in (24) can be simplified by conditioning on X. Specifically,

$$\begin{aligned}\mathsf{E}\left[\sum_{i=1}^{X}\sum_{j\neq i} U_i U_j\right] &= \mathsf{E}\left[\mathsf{E}\left[\sum_{i=1}^{X}\sum_{j\neq i} U_i U_j \,\Big|\, X\right]\right] \\ &= \mathsf{E}[a^2 X(X-1)] \\ &= a^2(s(t) - k(t)) \end{aligned} \quad (26)$$

Substituting (23), (24), (25) and (26) into (22), we obtain

$$\mathsf{E}\left[\left(1 + \sum_{i=1}^{X} U_i\right)^2\right] = 1 + 3ak(t) + a^2(s(t) - k(t)). \quad (27)$$

Next we consider (21). To analyze (21), we first obtain

$$\begin{aligned}\mathsf{E}\left[\sum_{i=1}^{X} Y_i U_i\right] &= \mathsf{E}\left[\mathsf{E}\left[\sum_{i=1}^{X} Y_i U_i\right]\right] \\ &= \mathsf{E}\left[\sum_{i=1}^{X} \mathsf{E}[Y_i U_i \mid X]\right] \\ &= \mathsf{E}\left[\sum_{i=1}^{X} \mathsf{E}[Y_i|X]\mathsf{E}[U_i|X]\right] \\ &= \mathsf{E}\left[\sum_{i=1}^{X} a\mathsf{E}[Y_i|X]\right] \\ &= a\mathsf{E}\left[\mathsf{E}\left[\sum_{i=1}^{X} Y_i|X\right]\right] \\ &= a\mathsf{E}\left[\sum_{i=1}^{X} Y_i\right] \end{aligned} \quad (28)$$

The above identity implies that

$$\mathsf{E}\left[\sum_{i=1}^{X}(2Y_i+1)U_i\right] = 2a\phi(t) + ak(t). \tag{29}$$

Considering (19), (27) and (29), we have

$$\begin{aligned}(m_0+t+1)s(t+1) - (m_0+t)s(t) = \\ (2k(t)+1) + (1+2ak(t)+ak(t)+a^2(s(t)-k(t))) \\ +(2a\phi(t)+ak(t)).\end{aligned}$$

Approximating $s(t+1)-s(t)$ by $s'(t)$, the last equation yields the following differential equation

$$\begin{aligned}s'(t) &= \frac{1}{m_0+t+1}\Big((-1+a^2)s(t) + 2a\phi(t) + 2 \\ &\quad +(2-a^2+4a)k(t)\Big).\end{aligned} \tag{30}$$

To analyze the differential equation in (30), we prove the following proposition.

Proposition 1. *If initially the network starts from a clique with m_0 vertices, then*

$$s(t) \stackrel{\text{def}}{=} \mathsf{E}[X^2] = \mathsf{E}\left[\sum_{i=1}^{X} Y_i\right] \stackrel{\text{def}}{=} \phi(t). \tag{31}$$

Note that if each neighbor of A has the same number of neighbors as A does, (31) clearly holds. Thus, the above identity implies that the duplication model behaves as if it were a regular network in which all vertices have the same degree. The proof of Proposition 1 can be found in the appendix.

Substituting (31) into (30), we obtain

$$\begin{aligned}s'(t) &= \frac{1}{m_0+t+1}\Big((-1+a^2+2a)s(t) + 2 \\ &\quad +(2-a^2+4a)k(t)\Big).\end{aligned} \tag{32}$$

The solution of the above differential equation with $k(t)$ in (8) is

$$\begin{aligned}s(t) &= \frac{a^2-4a-2}{a^2}c_1(m_0+t+1)^{2a-1} \\ &\quad +\frac{2a^2-4a-6}{(2a-1)(-a^2-2a+1)} \\ &\quad +c_6(m_0+t+1)^{a^2+2a-1},\end{aligned} \tag{33}$$

where c_6 is a constant determined by the initial condition $s(0) = (m_0-1)^2$.

VI. Numerical and Simulation Results

In this section we present numerical and simulation results. First, we simulate the duplication model one hundred times and calculate the mean degree and the expected number of triangles per vertex. Ninety-five percent confidence intervals were collected based on the repeated simulation of one hundred times. We choose $m_0 = 4$. The mean degree as a function of time is shown in Figure 3 for $a = 0.3$ and $a = 0.6$ respectively. Note that 95% confidence intervals are also shown in these figures. From these figures we see that the solution of the differential equation agree very well with simulation results. For other values of a, simulation results and the solutions from equations also agree very well.

t	simulation	confidence interval	analysis
500	0.3093	[0.3053, 0.3134]	0.3082
1000	0.2967	[0.2930, 0.3004]	0.2958
1500	0.2900	[0.2866, 0.2935]	0.2897
2000	0.2864	[0.2831, 0.2897]	0.2858
2500	0.2835	[0.2802, 0.2867]	0.2830
3000	0.2814	[0.2782, 0.2845]	0.2808

TABLE I
CLUSTERING COEFFICIENTS OF THE DUPLICATION MODEL OBTAINED BY SIMULATION AND ANALYSIS. $a = 0.3$

t	simulation	confidence interval	analysis
500	0.2755	[0.2724, 0.2787]	0.2736
1000	0.2333	[0.2304, 0.2361]	0.2317
1500	0.2113	[0.2085, 0.2140]	0.2098
2000	0.1965	[0.1938, 0.1992]	0.1953
2500	0.1859	[0.1832, 0.1885]	0.1846
3000	0.1773	[0.1748, 0.1799]	0.1763

TABLE II
CLUSTERING COEFFICIENTS OF THE DUPLICATION MODEL OBTAINED BY SIMULATION AND ANALYSIS. $a = 0.6$

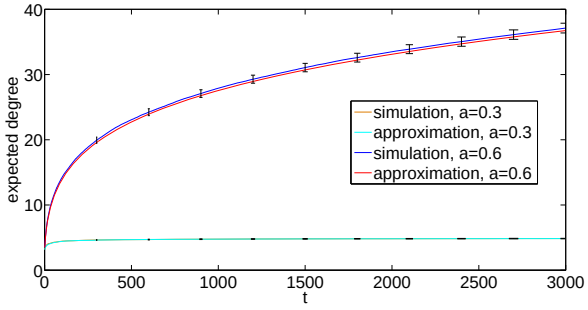

Fig. 3. Expected degree of the duplication model as a function of time with $a = 0.3$ and 0.6.

We simulate the clustering coefficient and compare the simulation result with the result obtained from (1). The result with $a = 0.3$ is shown in Table I. As one can see, the analytical result is very accurate and falls within corresponding confidence intervals at all t. Changing a to 0.6, the result is shown in Table II. Again we see that the simulation and the analytical results are very close.

Finally we acquire topologies of popular online social networks such as Flickr, Orkut and Livejournal from [15]. We also acquire a Facebook network from [21]. We compute their mean degrees (denoted by $\tilde{k}$) and global clustering coefficients (denoted by $\tilde{C}$) using software package igraph. The results along with their names and sizes are shown in the first four rows in Table III. We remark that the data sets of Orkut and Livejournal are too large for igraph. In order to use igraph, we have pruned their sizes down to what are shown in the table. The Facebook network acquired from [21] is denoted by Facebook I in the table. We have also developed

sites	Facebook I	Facebook II	Flickr	Orkut	Livejournal
size	63697	4884	1800K	400K	500K
$\tilde{k}$	29.23	32.15	29.74	113.26	55.27
$\tilde{C}$	0.148	0.390	0.112	0.107	0.073
m_0	350	200	1700	1400	800
a	0.246	0.177	0.196	0.273	0.289
$\langle k \rangle$	28.543	28.033	27.834	111.369	57.497
C	0.148	0.390	0.112	0.107	0.073

TABLE III
MEAN DEGREE AND CLUSTERING COEFFICIENT MEASURED FROM SOCIAL NETWORKING SITES

an explorer program to perform a breadth-first search of the facebook social network formed by users residing in Taiwan. This network is denoted by Facebook II in the table. Since the size of networks in our model is $m_0 + t$ at time t, parameter t is determined by the sizes of the online social networks. We determine the value of m_0 and a in (8), (14), and (33) by trying to match with the measured mean degree and the clustering coefficient of online social networks. Values of m_0 and a are shown in the fifth row and the sixth row of the table. The corresponding mean degree and the clustering coefficient are shown in the last two rows of the table. One can see that the mean degree and the clustering coefficient of the duplication model can match quite well with those of online social networks. In future work, one may want to introduce vertex removal or edge removal operations to the duplication model or the triadic closure model so that the models resemble more closely the real-world networks.

VII. CONCLUSIONS

In this paper we have presented a network formation model based on triadic attachment and triadic closure for social networks. We derive the mean degree and the clustering coefficient for this model. We show that the parameters of this model can be chosen such that the mean degree and the clustering coefficient of the model match well with those of the popular online social networks.

APPENDIX

In this appendix, we prove Proposition 1. To this end, we shall call the first neighbors and the second neighbors of a vertex the *near neighbors* of that vertex. Thus, $\sum_{i=1}^{X} Y_i$ is the number of near neighbors of A, if we count multiple times for vertices that are both first neighbors and second neighbors of A. Suppose that there are n vertices in the network and their degrees are $X_1, X_2, \ldots, X_n$. The degrees of the neighbors of vertex i are $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,X_i}$. The expected total number of near neighbors of the network is $\mathsf{E}[\sum_{i=1}^{n} \sum_{j=1}^{X_i} Y_{i,j}]$. The expected total number of near neighbors of the network is also equal to $n \cdot \phi(t)$, where $\phi(t)$ is the expected number of near neighbors of a randomly selected vertex at time t. Similarly, the expected total squares of degrees of the network is $\mathsf{E}[\sum_{i=1}^{n} X_i^2]$, which is equal to $n \cdot s(t)$.

We prove Proposition 1 by induction. It is very easy to verify that $s(0) = \phi(0) = (m_0 - 1)^2$. Now suppose that $s(t) = \phi(t)$ for some t. To analyze $\phi(t+1)$, again we focus on the difference

$$(m_0 + t + 1)\phi(t+1) - (m_0 + t)\phi(t) \tag{34}$$

which represents the change in the expected total number of near neighbors of the network from time t to $t+1$. For each vertex, say V, whose degree is changed from time t to $t+1$, the number of near neighbors of the neighbors of V are changed. By analyzing the change in the number of near neighbors V's neighbors, we analyze the change in the total number of near neighbors of the network. An example is illustrated in Figure 4.

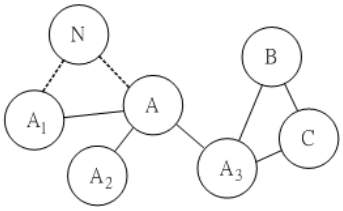

Fig. 4. Two new edges, shown as dashed lines, are introduced between (N, A) and (N, A_1). The degree of A is changed from 3 to 4. As a result, the numbers of near neighbors of A's neighbors are all changed. Specifically, the numbers of near neighbors of vertices A_2 and A_3 are increased by one. The number of near neighbors of A_1 is increased by 3.

- Consider vertex A. The degree of A increases from X to $X+1$. Each one of A's original neighbors increases its number of near neighbors by one. A's new neighbor, vertex N, increases its number of near neighbors by $X+1$ since the new edge between A and N was not present before time t. Thus, the total increase of the number of near neighbors of the network due to vertex A is

$$\mathsf{E}[X \cdot 1 + (X+1)] = \mathsf{E}[2X+1]. \tag{35}$$

- Consider vertex N. The degree of N changes from zero to $1 + \sum_{i=1}^{X} U_i$ as it is introduced into the network. Since the edges of N are all new and were not present before time t, each one of N's new neighbors changes its number of near neighbors by $1 + \sum_{i=1}^{X} U_i$. Thus, the total change is

$$\mathsf{E}\left[\left(1 + \sum_{i=1}^{X} U_i\right)^2\right]. \tag{36}$$

- Consider a first neighbor of A, say the i^{th} first neighbor of A. Call this neighbor vertex B. If vertex B is connected with vertex N, then the degree of B changes from Y_i to $Y_i + 1$. The change of the expected number of near neighbors due to the original neighbors of B is U_i. Vertex may have a new neighbor if the triadic attachment to N is successful. The change of the expected number of near neighbors due to vertex N is $(Y_i + 1)U_i$. Thus, the expected total change due to B is

$$\mathsf{E}[Y_i U_i + (Y_i + 1)U_i \cdot 1].$$

and the total expected change of the network is

$$\mathsf{E}\left[\sum_{i=1}^{X}(2Y_i+1)U_i\right]. \tag{37}$$

Note that the increments of $(m_0+t)\phi(t)$ stated in (35), (36) and (37) as t is incremented by one are identical to those of $(m_0+t)s(t)$ stated in (19), (20) and (21). This result in conjunction with the induction hypothesis implies (31) in Proposition 1.

REFERENCES

[1] A. Bhan, D. J. Galas, and T. G. Deway. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.

[2] S. Boccaletti, D.-U. Hwang, and V. Latora. Growing hierarchical scale-free networks by means of nonhierarchical processes. *International Journal of Bifurcation and Chaos*, 17(7):2447–2452, 2007.

[3] F. Chung and L. Lu. Complex graphs and networks. In *Regional Conference Series in Mathematics*, number 107. American Mathematical Society, 2004.

[4] F. Chung, L. Lu, and T. G. Dewey. Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687, 2003.

[5] J. Davidsen, H. Ebel, and S. Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88(12), March 2002.

[6] R. Friedman and A. Hughes. Gene duplications and the structure of eukaryotic genomes. *Genome Res.*, 11:373–381, 2001.

[7] S. Ghahramani. *Fundamentals of Probability with Stochastic Processes*. Pearson Prentice Hall, 3 edition, 2005.

[8] Z. Gu, A. Carvalcanti, F.-C. Chen, P. Bouman, and W.-H. Li. Extent of gene duplication in the genomes of drosophila, nematode, and yeast. *Mol. Biol. Evol.*, 19:256–262, 2002.

[9] P. Holme and B. J. Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(026107), 2002.

[10] I. Ispolatov, P. L. Krapivsky, I. Mazo, and A. Yuryev. Cliques and duplication-divergence network growth. *New Journal of Physics*, june 2005.

[11] I. Ispolatov, P. L. Krapivsky, and A. Yuryev. Duplication-divergence model of protein interaction network. *Phys. Rev. E*, 71:061911, Jun 2005.

[12] B. Kirman, S. Lawson, and C. Linehan. Gaming on and off the social graph: The social structure of facebook games. 2009.

[13] J.M. Kumpula, J.-K. Onnela, J. Saramáki, K. Kaski, and J. Kertész. Emergence of communities in weighted networks. *Physical Review Letters*, PRL 99(228701), 2007.

[14] S. Milgram. The small world problem. *Psychol. Today*, 2:60–67, 1967.

[15] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. 2007.

[16] M. Newman. *Networks: An Introduction*. Oxford, 2010.

[17] M.E.J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(025102), 2001.

[18] R. V. Solé, R. Pastor-Satorras, E. Smith, and T. B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, July 2002.

[19] G. Szabó, M. Alava, and J. Kertész. Structural transitions in scale-free networks. *Physical Review E*, 67(056102), 2003.

[20] R. Toivonen, J.-P. Onnena, J. Saramáki, K. Hyvónen, and K. Kaski. A model for social networks. *Physica A*, 371:851–860, 2006.

[21] B. Viswanath, A. Mislove, M. Cha, and K.P. Gummadi. On the evolution of user interaction in facebook. 2009.

[22] D. Zhao, Z.-R. Liu, and J.-Z. Wang. Duplication: a mechanism producing disassortative mixing networks in biology. *Chin. Phys. Lett.*, 24(10):2766–2768, 2007.